VOL 3 - PROBABILITY AND STATISTICS

BEFORE MACHINE LEARNING

A STORY ON FUNDAMENTAL MATHEMATICS FOR A.I.

JORGE BRASIL

My wish is that you find this read enjoyable and get proficient in the intricacies of probability and statistics. If you want to complement your reading, you'll find a link below to download a document with the code used on this book.

Link for code!

This book is the last volume of a three-volume series called Before Machine Learning. The first two volumes, which cover linear algebra and calculus, are available on Amazon or my website.

www.mldepot.co.uk

Subscribe for updates!

Subscribe here!

Contents

Chapter 1

Why Probability and Statistics?

Machine learning is fundamentally about vast, complex, and often messy data. To transform this raw data into actionable insights, we must navigate uncertainty, discern patterns, and make informed predictions. This is where probability and statistics become crucial. They provide the mathematical foundation for making sense of data, allowing us to draw meaningful, reliable conclusions.

Linear algebra equips us with the tools to represent data in high-dimensional spaces, manipulate matrices and vectors, and understand the geometric properties of data. It provides the language and framework for expressing and solving problems involving large datasets and complex relationships, albeit within a realm where everything is linear. Calculus, on the other hand, introduces us to the dynamics of change and curves, expanding our capabilities beyond linear structures. This allows us to optimize convex cost functions with gradients, a key element in the training of machine learning models.

Building on this foundation, probability and statistics introduce the essential concepts of uncertainty and inference. In machine learning, we rarely have access to complete or perfect data. Instead, we work with samples representing larger populations, and our task is to make educated guesses about the underlying patterns. Real-world data is inherently uncertain, and probability provides a framework for modeling this uncertainty, enabling us to quantify the likelihood of different outcomes and make predictions that account for variability.

Statistics offers methods to infer the properties of populations based on sample data. Techniques like histograms, measures of central tendency, and measures of spread help us understand the data characteristics and gather fundamental knowledge to create our desired sets of features, leveraging all the linear algebra and calculus we've learned. Moreover, the value of probability and statistics extends beyond the data exploration phase. Methods like hypothesis testing and Bayesian inference are fundamental in conducting and analyzing experiments, which, in a commercial context, are not only invaluable but necessary—for instance, in A/B tests. Additionally, this field enriches our modeling capabilities with its models, such as the Naive Bayes model, a simple but yet powerful framework.

Further deepening our toolkit, Markov Chain Monte Carlo (MCMC) methods enable us to perform complex Bayesian inference by sampling from a probability distribution to construct a Markov chain—fancy names for essential concepts that we will explore into depth throughout the book. So please join me on this journey into the world of probability and statistics. Along the way, we will explore the Great Diamond Heist of Antwerpen, where Notarbartolo and his crew stole 100 million euros worth of diamonds.

Chapter 2

The Event: The Gem of a Plan.

If you are new to these areas, welcome, and let us introduce you to our style of writing and learning. We always start with the simplest element of the topic and build from there. In linear algebra, we started from the vector; with calculus, we did it from a point; now, with probability and statistics, we will kick things off with an event. So, let's get on the floor and do a set of 10 push-ups. No, no, that is not the sort of event I'm talking about. An 'event' in probability is a collection of one or more outcomes from an experiment, each contributing to the probability of the event occurring. For example, consider rolling a fair die. If we define our event as rolling an even number, the possible outcomes that constitute this event are the numbers 2, 4, and 6. Other classical examples include flipping a coin (where an event might be getting heads) or drawing colored balls from a bag (where an event might be drawing a red ball). While these examples are commonly used, I propose we explore a different event.

In February 2003, a group called the 'School of Turin', led by Leonardo Nortarbartolo, robbed the diamond center in Antwerp, the Belgian city. Despite the center's sophisticated security system, including seismic sensors, infrared detectors, and a vault door lock with 100 million possible combinations, the thieves managed to bypass these defenses. Over the course of a couple of days, they cracked the safe, looting over 100 safety deposit boxes filled with diamonds, gold, and other precious jewels.

Leonardo Notarbartolo initiated the heist's planning by securing an office in the Diamond Center, a strategic move to collect intelli-

gence. After a period of meticulous reconnaissance, he would return to Turin, armed with a detailed list compiled from his observations, ready to discuss the next steps with his team.

We are just starting the book, and I already have good news for you: we are going to 'rob' (please don't do it) that same diamond center, but this time, probabilities and statistics are what we will use to plan our heist!

The planning done by Nortarbartolo will serve us as inspiration, but as security measures got tighter after he and his crew robbed the diamond center, we will use statistics not only to plan our heist but also to recruit the other members of our gang. I propose we start with the getaway part of our plan. It so happens that I went to Belgium and did some data collection.

On my reconnaissance mission, I went around the streets surrounding the diamond center. I gathered detailed information about every camera on each road, namely their location and the security camera brand, as represented in table 2.1:

Camera number	Street name	Camera Brand	Lat	Long
1	Rijfstraat	Brand A	51.218	4.404
2	Rijfstraat	Brand B	51.218	4.405
3	Rijfstraat	Brand C	51.218	4.406
4	Rijfstraat	Brand D	51.218	4.407
5	Rijfstraat	Brand E	51.218	4.408
6	Hoveniersstraat	Brand A	51.219	4.409
7	Hoveniersstraat	Brand D	51.219	4.410
7	Hoveniersstraat	Brand C	51.219	4.411
9	Hoveniersstraat	Brand B	51.219	4.412
10	Hoveniersstraat	Brand E	51.219	4.413
11	Schupstraat	Brand B	51.220	4.414
12	Schupstraat	Brand E	51.220	4.415
13	Schupstraat	Brand A	51.220	4.416
14	Pelikaanstraat	Brand D	51.221	4.417
15	Pelikaanstraat	Brand D	51.221	4.418
16	Pelikaanstraat	Brand D	51.221	4.419
17	Appelmansstraat	Brand C	51.222	4.420
18	Appelmansstraat	Brand B	51.222	4.421
19	Appelmansstraat	Brand A	51.222	4.422
20	Appelmansstraat	Brand D	51.222	4.423
21	Vestingstraat	Brand D	51.222	4.423
22	Vestingstraat	Brand D	51.222	4.423
23	Herentalsestraat	Brand D	51.222	4.423
24	Herentalsestraat	Brand D	51.222	4.423
25	Herentalsestraat	Brand D	51.222	4.423

26	Lange Herentalsestraat	Brand D	51.182	4.443
27	Lange Herentalsestraat	Brand D	51.237	4.471
28	Lange Herentalsestraat	Brand D	51.181	4.385
29	Lange Herentalsestraat	Brand D	51.241	4.396
30	Lange Herentalsestraat	Brand D	51.186	4.452
31	Lange Herentalsestraat	Brand D	51.267	4.423
32	De Keyserlei	Brand D	51.175	4.392
33	De Keyserlei	Brand D	51.182	4.456
34	De Keyserlei	Brand D	51.189	4.376
35	De Keyserlei	Brand D	51.232	4.401

Table 2.1: The gemstone grid for a getaway route.

In our table 2.1, we have five columns: the camera n[a], representing a given camera, the street name, describing the street where we collected information, the camera brand, and finally, two location measures, the latitude, and the longitude. The heist unfolded primarily around three key streets: Rijfstraat, Hoveniersstraat, and Schupstraat, which are in the Antwerp Diamond District. We've also included Pelikaanstraat and Appelmansstraat in our dataset to supplement our analysis. These streets are near the diamond center, making them relevant to our study. It's important to note that, while the facts about the heist are accurately presented, some of the data used in this book, particularly the values in our datasets, are hypothetical and created solely for educational purposes. Yes, we will also plan a heist using probabilities and statistics (a fake heist, of course, but let's get into character). So, if we were presented with a table such as the one represented by 2.1, two things would be obvious: the data is raw, which means extracting insights isn't straightforward yet. The second observable fact is that we have two types of entries: text and numbers. In mathematics, we don't always call them text and numbers though. Don't blame me; I would have been perfectly fine with that nomenclature:

- **Numerical variables**: After spending 10 minutes thinking of a good way to define these guys, I came up with this — they are numbers. Examples include the number of cameras in a street, the price of a diamond, the number of years of jail sentence, and latitude or longitude.

- **Categorical variables**: This is the fancy term for variables in text format, such as the 'Street' on table 2.1. Although seemingly inoffensive, they can present us with a big headache,

just like that guy you thought was a dork, then it turned out he was a jiu-jitsu black belt.

Like good thieves we are... mathematicians! I meant mathematicians! A valuable piece of information would be to count the number of cameras one could find on a given street in the Diamond District:

Street Name	Number of Cameras
Rijfstraat	5
Hoveniersstraat	5
Schupstraat	3
Pelikaanstraat	3
Appelmansstraat	4
Vestingstraat	2
Herentalsestraat	3
Lange Herentalsestraat	4
De Keyserlei	4

Table 2.2: Eyes of the diamond district: street surveillance count.

To transition from table 2.1, which lists each camera on the streets, to table 2.2, we count the number of cameras for each distinct street. Since each row in table 2.1 represents an individual camera, this process consolidates the data into a format like the one seen in table 2.2, where the total number of cameras per street is displayed. This simplification is useful, but, as the number of streets increases, imagine scaling up 500 or 1,000 times; clearly the complexity of analyzing the data also rises. The table format becomes less manageable, making it challenging to grasp the distribution of cameras quickly.

To address this, we can introduce a statistical measure that provides an estimate of the average number of cameras per street, offering a more manageable overview and facilitating easier analysis, especially when dealing with large datasets.

We can start by summing all values and dividing by the total number of values, this method ensures that every data point contributes equally to the final result. It provides a single value that represents the entire dataset, taking into account every individual measurement:

$$\frac{5+5+3+3+4+2+3+4+4}{9} = 3.6 \tag{2.1}$$

Equation 2.1 represents the average number of cameras per street in the Diamond District; it comes to 3.6. However, considering a fraction of a camera isn't practical in the context of our task, it's more sensible to round up and assume an average of 4 cameras per street for this specific situation. This approximation suggests that, while some roads may have more than 4 cameras, others might have fewer, but the overall density hovers around 4 cameras per street. Our decision to round up is driven by the need for a cautious approach in planning a getaway strategy; let's play it safe, as we want to avoid jail time. However, if the context was different and required a less conservative estimate, rounding down to 3 cameras per street would also be viable. Ultimately, the approach depends on the specific requirements of the context. We can generalize equation 2.1 for any use case with numerical data as such:

$$\text{Arithmetic Mean} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2.2}$$

Where n represents the number of values in our sample and x_i are the observations. In our particular case, x_i would depict the number of cameras on each street and n the number of streets so the expansion of 2.2 for our example is precisely 2.1. There other types of means, such as the geometric and harmonic but in this book, we will focus on the arithmetic.

2.1 Middle Ground Madness: Exploring Measures of Central Tendency.

2.1.1 In The Mean Time, What's the Typical Outcome?

The 'mean', or 'arithmetic average', is a pivotal concept in statistics. It is part of a group of metrics called 'measures of central tendency'. We have several of those measures, and the mean captures

the typical or central value within a dataset, enabling it to consolidate a vast array of numbers into a singular figure. This single figure elucidates patterns and behaviors, simplifying complex data. For instance, in strategizing a getaway, the mean might indicate an average encounter of 4 cameras per street, highlighting the challenges ahead. However, the mean's reliance on summing all dataset values before dividing by their count renders it vulnerable to outliers, values significantly deviating from the norm. For example, say we have one street monitored by 40 cameras:

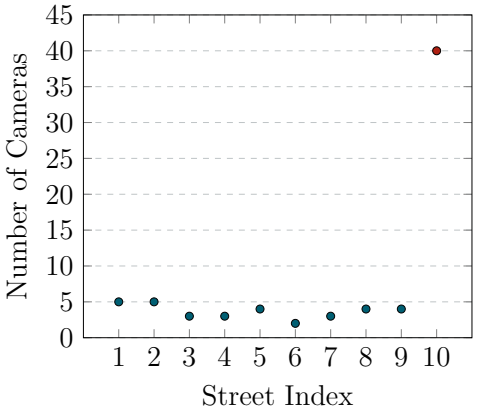


Figure 2.1: Why are you so far mister red?

In 2.1, the y-axis represents the number of cameras on a given street, and the x-axis identifies each street. The streets, marked by blue dots, generally have between two and five cameras. However, there is an exception, as indicated by the red dot, which shows a street with forty cameras. This number significantly deviates from the range observed in other streets. In statistical terms, such data points are referred to as 'outliers' because their characteristics are markedly different from all other observations. In fields like machine learning, addressing outliers is crucial, as they can skew the interpretation of data and adversely affect model performance. While this volume will not go deeply into quantifying the threshold for identifying outliers, the statistical insights gained here will be invaluable for future studies of such anomalies. Such extreme cases, like a street anomalously monitored by 40 cameras, can distort the mean, misleadingly inflating perceived security measures:

$$\frac{5+5+3+3+4+2+3+4+4+40}{10} = 7.3$$

This sensitivity means that the mean can sometimes give a misleading picture of the data, especially when the dataset is skewed by unusually high or low values.

Well, we are trying to leave the scene as smoothly and quickly as possible, so perhaps we should explore another measure of central tendency to see if we can gather any more intelligence on how the distribution of the number of cameras per street behaves. What if we calculate the middle point of the number of cameras per street and then look to what's below and what's above it? Because we won't be doing sums and divisions, this seems like a more robust method to deal with outliers. OK, so if we are after a middle point, the first thing that will help us is to arrange the data; because we need to find the middle point, either ascending or descending order will work:

$$2, 3, 3, 3, 4, 4, 4, 5, 5 \tag{2.3}$$

Now, we can determine the central value by counting the number of entries and selecting the one that splits 2.3 into two equal halves. If the dataset has an odd number of entries, it's straightforward to identify a position that has an equal number of entries on both sides. However, when there are an even number of observations, we need to employ a different approach to determine this central value:

- If the number of data points n is odd, the middle is the value at the $P = \frac{n+1}{2}$ position in the sorted data. This ensures we pick the exact middle value.
- If n is even, the middle is calculated as the average of the two middle values. These are found at the $P = \frac{n}{2}$ and $P = \left(\frac{n}{2} + 1\right)$ positions in the sorted array. This method averages the two central values to determine the midpoint of the dataset.

We have 9 entries so an odd number and because of that the position of the median value comes as:

$$\frac{9+1}{2} = 5$$

If we go back to the array of numbers represented in 2.3 and extract the value that is in position 5 we end up with 4.

2.1.2 Mediating With The Median: The Art of Central Balance.

This new measure of central tendency is called the 'median'; it represents the middle value in a sorted dataset. Unlike the mean, which can be skewed by extremely high or low values, the median provides a more resistant measure of central tendency, especially useful in understanding the 'typical' value in skewed distributions. This makes the median a reliable indicator in many real-world scenarios, offering a clearer picture of the data's central tendency without the influence of outliers.

Considering our dataset, where the median number of cameras per street is 4—which is coincidentally equal to the mean in this case—it's important to note that the median and mean are not always identical. They are separate measures of central tendency, each with distinct characteristics. For instance consider the same set we used as an example of an outlier:

$$2, 3, 3, 3, 4, 4, 4, 5, 5, 40 \tag{2.4}$$

As we saw previously, the mean is sensitive to that 40, as it factors in all values in the dataset equally, leading to a higher average (7.3) that might not accurately represent the most common number of cameras per street. What do you think it will happen to the median ? Well let's check!. Because we now have 10 entries, we must average the values from two positions:

$$P_1 = \frac{10}{2} = 5 \quad P_2 = \frac{10}{2} + 1 = 6$$

Both positions 5 and 6 corresponded to a value of 4, therefore:

$$\text{Median} = 4$$

The median, is less affected by outliers since it merely depends on the middle value(s) after sorting the dataset. Therefore, the median that is 4, provides a more reliable measure of central tendency in this case, as it gives us a more accurate number of cameras if we pick a random street:

Mean	Median
7.3	4

Table 2.3: The original M&M: Mean vs Median.

In e-commerce analytics, understanding user behaviour is crucial. Typically, a large portion of users might visit a website infrequently, while a smaller, more engaged group might visit highly frequently. When attempting to gauge the expected number of visits from an average user, simply calculating the mean can be misleading. This is because, even though high-visit users are less common, their activity disproportionately influences the mean, skewing it upwards. In such cases, the median provides a more accurate reflection of typical user behavior. It effectively neutralizes the impact of outliers by focusing on the middle value of the dataset, offering a clearer view of what a 'normal' number of visits looks like. Therefore, employing the median as a measure of central tendency is particularly advantageous in e-commerce data analysis, ensuring that our insights remain relevant and representative of the majority of users.

The median splits our dataset into two equal-sized parts, effectively 'chopping' the data in half. This division suggests the possibility of further segmentation, perhaps into four parts instead of two. By doing so, each 'chunk' would represent 25% of the data, allowing us to possibility of identifying streets with even fewer cameras. We determined the mid position using the following method:

$$P = \frac{n+1}{2} \tag{2.5}$$

Okay, since we now want to split the data into four chunks instead of just two, the first modification we should make is to the denominator:

$$P = \frac{n+1}{4}$$

Let's see what comes from that:

$$P = \frac{9+1}{4} = 2.5 \tag{2.6}$$

Unfortunately, direct access to position 2.5 isn't possible, but we can access the positions close to it, namely positions 2 and 3.

Considering our objective is to determine the number of cameras that have 25% of the data positioned to its left, a practical approach involves checking the values at positions 2 and 3 of the sample of numbers (refer to sample 2.3) and calculating their average. In this instance, both positions have the same value: 3 cameras.

$$Q_1 = \frac{3+3}{2} = 3 \tag{2.7}$$

2.1.3 The Sharp Knife of Stats: Slicing With the Quantiles.

The Q_1 in equation 2.7 represents a measure we call 'quantiles.' Quantiles are values that divide a dataset into equal-sized subsets; because we are dividing our into four parts, these quantiles are specifically known as 'quartiles' and we will have 3 of them. For instance, Q_1 indicates that 25% of the streets have 3 or fewer cameras. The second quartile, Q_2, represents a number of cameras that have 50% of data to the left (sound familiar?), and, finally, Q_3, a value where most of the data is to its left, or, more precisely 75%.

If Q_1 gives us 25% of the data that sits to the left, to get to 50% and therefore Q_2, we can move 25% to the right of Q_1. This can be achieved by multiplying equation 2.6 that gave us the position P for Q_1 by 2:

$$P = \frac{(9+1)\cdot 2}{4} = 5$$

Moving 50% across the dataset is equivalent to reaching the median, and indeed, we land at position 5, the median of the array 2.3, with a value of 4.

$$Q_2 = 4$$

There we have, the second quartile, which is the same as the median. Now to move along this caterpillar of pretty Q's and compute Q_3, one just has to move 25% from Q_2 to the right. Well that is the same as multiplying 2.6 by 3:

$$P = \frac{(9+1)\cdot 3}{4} = 7.5$$

As we have established we can't operate with decimal positions but, we can once again average the elements of position 7 and 8. So, if we go to 2.3 and get the elements of position 7 and 8, we have 4 and 5:

$$Q_3 = \frac{4+5}{2} = 4.5$$

Our third quartile Q_3 has a value of 4.5. This means that 75% of the streets in our dataset have 4.5 or less cameras. Obviously, we don't have half a camera on a street, so we can round it up to 5 to be more conservative. In other situations we could also round it down; it depends on the context. The critical distinction between positions and actual quartiles values has to be clear. Quartile values represent specific measurements within our dataset, in this case, the number of cameras. These values are derived from positions within our sample, which serve as markers to divide the data into four equal segments. Therefore, we adhere to the following rules:

- If P is not an integer, we interpolate between data points (average them out).
- If P is an integer, the quantile is the value at that position of the sorted list.

So:

- The first quartile (Q_1) divides the data so that 25% of all observations fall below it.
- The second quartile (Q_2), also known as the median, divides the data so that 50% of all observations are below it, effectively splitting the dataset into two equal halves.
- The third quartile (Q_3) has 75% of all observations are below it.

Let's put all the quartiles we calculated inside a friendly rectangle:

Q_i	Quantile	Value
Q_1	25%	3
Q_2	50%	4
Q_3	75%	4.5

Table 2.4: Antwerp's Streets: A Quartile perspective on surveillance.

Table 2.4 segments our dataset into four equal parts, delineated by the quartiles. This segmentation shows that 25% of the streets are equipped with 3 or fewer cameras, and 50% have 4 or fewer cameras. Notably, the third quantile (Q_3) is calculated at 4.5, illustrating that 75% of the streets have 4.5 or fewer cameras. Despite the discrete nature of our dataset, which does not allow for fractional positions, the quartiles values themselves can be fractional. Now, let's put some more letters in formula 2.5 to define a general formula for calculating quartiles:

$$P = \frac{(n+1) \cdot q}{4} \quad \text{where } q \text{ is the quartile number} \tag{2.8}$$

Visually this is what we are doing to the data when we calculate the quartiles:

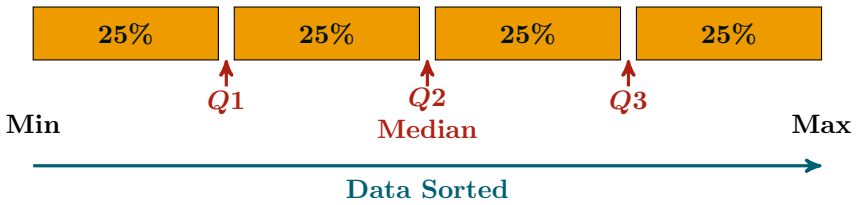


Figure 2.2: Four boxes and three quantiles, what else can we ask for?

In figure 2.2, the yellow rectangles symbolize segments of data that are arranged in ascending order, as highlighted by the big blue arrow. The labels 'Min' and 'Max' denote the lowest and highest observations in our dataset, which, in this specific case correspond to 2 and 5 cameras, respectively. Our buddies the Q's signify the quartiles, neatly segmenting the data.

There is one more particularity about the quartiles, more specifically involving the 25^{th} and 75^{th}. These two fellows not only are the first and third quartiles but we they also define the interquartile range:

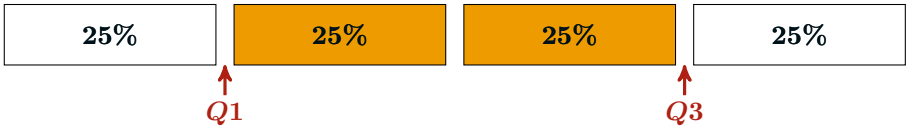


Figure 2.3: The yellow entrapment caused by the first and third $Q's$.

2.1.3.1 Tell Me Where I Need to Stand to Be Central: The Interquartile Range.

The "Interquartile Range" (IQR), defined as the range between the first quartile and the third quartile, is particularly valued in statistics for its robustness against outliers. This measure effectively encapsulates the central 50 percent of data, providing a resilient indicator of a dataset's central tendency and variability without being skewed by extreme values.

But why stop at just four segments? I wouldn't either. If four segments represent a quarter, then ten segments represent a tenth – and this is where the concept of deciles comes into play. The formula for calculating deciles is pretty much the same as that for quartiles, with the only twist being in the denominator:

$$P = \frac{(n+1) \cdot q}{10} \quad \text{where } q \text{ is the quantile number} \tag{2.9}$$

We can take our analysis a step further by creating percentiles, which involves dividing the data into 100 equal parts. Quantiles and deciles are crucial in machine learning. A common use case in the industry is predicting customer churn. Typically, we're given a dataset with behavioral data to create features. However, a challenge arises when we lack labels, the crucial piece of information needed to train a supervised learning algorithm.

To address this challenge, we can strategically generate labels using deciles. Consider a dataset where we track the number of days between a customer's last visit and the current date. The 9^{th} decile represents the lapse time for 90% of the customers in terms of their visit frequency. We can then set a threshold based on this fellow: if a customer's lapse time is in the top 10% (over the 9^{th} decile), we label them as having churned. This approach offers a pragmatic way to create meaningful labels, which are essential for developing robust models that accurately predict churn.

Yes, it is true that, just by looking at table 2.3, we can observed its quartiles. Remember though, this is an elementary example; if you are dealing with a dataset of considerable size, the observation technique won't be possible. So, we know that 50% of the streets have 4 cameras or less and also that 25% of them have less then 3 cameras. Let's explore the streets whose number of cameras are less than the median:

Street Name	Number of Cameras
Schupstraat	3
Pelikaanstraat	3
Vestingstraat	2
Herentalsestraat	3

Table 2.5: Surveillance below the median: Our chance to escape!

Now we can start planning the robbery properly; for that let's get a map of the building and the streets that surround it:

Figure 2.4: Map of the diamond centre.

Our target, the diamond center, is symbolized by, well.., the big ass diamond. The building is marked by three key reference points: a garage door, the main entrance, and a nearby police station. The circle with numbers in our map pinpoint potential parking spots for our getaway vehicle. Given their proximity to the police station, I recommend we rule out the streets surrounding it. Additionally, making an exit through the front door, while cinematic, is a scenario

best avoided in reality. Therefore, parking options at locations 1 and 2 are off the table. Location 4, situated on Rijfstraat, is less ideal due to its heavy camera surveillance and distance from the garage. This leaves us with location 3 as our prime choice. From here, we can plan our route accordingly:

$$\text{Appelmansstraat} \rightarrow \text{Vestingstraat} \rightarrow \text{Pelikaanstraat}$$

Two of the streets we've identified have a number of cameras that falls below our median value, effectively leveraging the insights provided by this measure of central tendency. We now have a viable parking spot and a strategic exit route!

For a heist of this magnitude and complexity, Notarbartolo assembled a crew of top-notch crooks, forming the infamous 'School of Turin'. The group included Speedy, Notarbartolo's close friend and an experienced, but also a somewhat paranoid thief. Then there was 'The Monster', a lock-picking maestro and a formidable getaway driver. 'The Genius' was the alarm specialist, capable of disarming even the most sophisticated systems. And lastly, 'The King of Keys' was unparalleled at forging keys. So far, we have me and you as the 'mathematicians' so it's time for us to start recruiting other members.

Turning our attention to table 2.1, specifically to the column listing camera brands, it appears that brand D is particularly prevalent. Let's tally up and see exactly how many cameras are from this dominant brand:

Camera Brand			
A	B	C	D
4	4	4	19

Table 2.6: We can spot a trend! D, you are in trouble.

In table 2.6 we have a count of the number of cameras per brand and, oh la la, can we spot a trend! There's a clear preference for cameras of brand D in our dataset. In statistics, we have a term for the value that appears most frequently in a dataset: it's called the 'mode'. Now, there are a few scenarios for which we might need the mode. Just as in the case of table 2.6, we can have a single mode,

where one value appears more often than any other. Alternatively, we can have multiple modes, which happen when multiple values show up with the same frequency. And lastly, there's the possibility of no mode at all, which occurs when no values repeat themselves.

2.1.4 The Mode: Math's Ultimate Popularity Contest.

The 'mode', another measure of central tendency, is calculated simply by identifying the value that occurs most frequently. Though it has its limitations, particularly in datasets where values don't often repeat, however, for our purposes, it provides a crucial insight. We could thus recruit an electrician who's an expert on brand D cameras. But let's be creative here, as things have changed a lot since 2003, so what about a hacker? Let's check how many brand D cameras we'll encounter along our planned getaway route to see if this recruitment is actually justified:

Street Name	Total Cameras	Total Cameras Brand D
Appelmansstraat	4	1
Vestingstraat	2	2
Pelikaanstraat	3	3

Table 2.7: The streets of Antwerp and their surveillance.

Well, 6 out of 9 cameras are the brand D, so if we can find a way to hack that system we will have an even better chance of escaping. These simple three measures of central tendency-mean, median, and mode-each have their unique advantages and drawbacks, offering different perspectives on a dataset. The mean is widely used for its mathematical properties and ease of computation, providing a quick sense of the overall dataset. However, it is sensitive to outliers, which can skew the results in datasets with extreme values. The median, representing the middle value of an ordered dataset, is less affected by outliers and gives a more accurate reflection of the 'central' value in skewed distributions. It is particularly useful in understanding any datasets where extreme values can distort the mean. The mode, the most frequently occurring value, is crucial for categorical data and for understanding the most common characteristics in a dataset.

While it is limited in representing numerical data's central tendency, it is invaluable for identifying the most prevalent category.

Utilizing these measures in conjunction allows for a comprehensive understanding of a dataset. The mean provides a quick summary, the median offers a balanced middle point, and the mode highlights the most common value. By considering all three, we can gain a deeper insight into the data's distribution, balance the influence of outliers, and cater to different types of data, be it numerical or categorical:

- **Mean:** Use the mean for data that is symmetric and lacks outliers.
- **Median:** Use the median for skewed data or when there are outliers.
- **Mode:** Use the mode when the most frequent observation is of interest or for categorical data. It is especially useful in describing the most typical case in a categorical dataset.

The Antwerp Diamond Heist, executed on the weekend of February 15-16, 2003, was meticulously planned for a time when success was most likely. This particular weekend was chosen strategically: the Diamond Centre was closed, and the Diamond District was quieter than usual. Most workers were observing the Sabbath, and a major tennis tournament in Antwerp further ensured that the streets would be less crowded than usual. Less traffic and fewer people in the area significantly increased the chances of pulling off the heist without drawing attention.

Leonardo Notarbartolo, the smart and calculated mastermind behind this heist, understood the importance of patterns when it came to their getaway. It's believed that he meticulously scouted traffic on key streets to decipher any regular patterns that could be exploited during their escape.

Having identified a trend in the usage of Brand D cameras, a critical element of our strategy revolves around gathering data on the functionality of these surveillance devices. During our observation

period, we've noticed an intriguing pattern: the red light, which indicates when the camera is recording, intermittently switches off.

We've found exactly what we needed: an edge. To leverage this potential vulnerability without risking our identities by lingering on the streets and observing the cameras' inactivity, we'll hire a hacker. This person can furnish us with vital data, pinpointing those moments when the cameras cease recording. Identifying these intervals could unlock crucial windows of opportunity for us:

00:00-02:00	
Weekend Day	**Inactivity Time**
1	22.88
2	27.53
3	27.29
4	32.08
5	24.22
6	33.91
7	37.47
8	19.65
9	32.13
10	33.38
11	26.81
12	28.01
13	29.34
14	28.51
15	28.45

Table 2.8: The times when the cameras can't see us from midnight to two AM.

02:00-04:00	
Weekend Day	**Inactivity Time**
1	28.88
2	20.44
3	26.86
4	32.41
5	30.5
6	32.62
7	32.41
8	34.47
9	33.95
10	34.49
11	36.16
12	37.58
13	38.32
14	11.69
15	11.67

Table 2.9: Some more of those good undetected times but now from two to four AM.

Inspired by the Nortarbartolo planning process, we collected data across 15 weekends, focusing on two specific time slots, 00:00-02:00 and 02:00-04:00, without distinguishing between Saturdays and Sundays, and this gives us a sample of data! Speaking of "sample," here we have an important distinction. When accessing an entire population (such as the number of cameras per street) is unfeasible, we revert to using samples. For instance, in our camera downtime analysis, a complete population would encompass all cases of inactivity from the inception of the cameras, and that is very hard, if not impossible, to have. So we've focused on a sample of inactivity periods during particular timeframes. There are various methods through which to collect samples:

1. **Random Sample**: Every member of the population has an equal chance of being selected. This approach was used in

our study, where random periods of camera inactivity were chosen without bias during the 00:00-02:00 and 02:00-04:00 timeframes.

2. **Stratified Sample**: Here, the population is divided into subgroups (strata) that share similar characteristics, and samples are drawn from each stratum. For our heist scenario, we could categorize cameras based on their location (e.g., inside vs. outside) and sample inactivity times from each category.

3. **Cluster Sample**: The population is divided into clusters (a group), and an entire cluster is selected for the study. In the context of our heist, suppose the cameras are grouped by streets. We could randomly select an entire street and analyze all camera downtimes within that cluster.

4. **Convenience Sample**: This method involves taking samples from a group that is easily accessible. For example, if we only analyze camera downtimes from the cameras easiest to access or monitor, this would constitute a convenience sample.

5. **Systematic Sample**: In this approach, every n^{th} member of the population is selected. If we were to study every fifth instance of camera inactivity recorded, that would exemplify a systematic sample.

The sample size becomes a critical factor when selecting a subset of the population for analysis. This is because the size directly influences the reliability and validity of our findings. For instance, a small sample size from a population can lead to errors in our estimates as we risk not having a full picture of the population causing us to overlook significant behaviors or trends.

This leads us to the concept of sampling error. Sampling error refers to the difference between the characteristics of the sample and those of the entire population. Understanding that sampling error doesn't imply a mistake in the sampling process is crucial. Instead, it's an inherent part of sampling that results from using a part of the population to estimate its overall characteristics. Recognizing

and managing this error is essential for drawing accurate conclusions from our data.

Throughout this book, we will go deeper into the concept of error, using examples and practical applications to illustrate them. Now back to our data analysis!

Our primary goal is to compare these two periods during the weekend. Accordingly, we present the data in tables 2.8 and 2.9, which detail the inactive time intervals of Brand D cameras located along our planned escape route. These tables show the total minutes each camera was inactive. To analyze this data effectively, we can begin by calculating both the mean and the median of these inactive periods:

$$\text{Mean}_{00:00-02:00} = \frac{1}{15}\sum_{i=1}^{15} x_i = 28.78$$

As before, we add up all 15 observations and divide the result by the same number. Next, we will compute the median, and, because we have an odd number of observations, the positions are calculated as:

$$P = \frac{15+1}{2} = 8$$

So, we need to sort the values of inactivity and select the element that is in position 8:

$$19.65, 22.88, 24.22, 26.81, 27.29, 27.53, 28.01, \mathbf{28.45}, 28.51, 29.34, 32.08, 32.13, 33.38, 33.91, 37.47$$

For the median value of the time frame 00:00-02:00 we have:

$$\text{Median}_{00:00-02:00} = 28.45$$

If we do the same exercise but now for the time intervals of inactivity observed over the 02:00-04:00 time period:

$$\text{Mean}_{02:00-04:00} = 29.50, \quad \text{Median}_{02:00-04:00} = 32.41$$

At first glance, it seems like the 02:00-04:00 slot would be the best for us to execute the heist, as not only is the mean of inactivity time superior, 29.50 minutes versus a value of 28.78 minutes of inactivity time observed on 00:00-02:00, but we also have a higher value for the median 32.41 against 28.45. But if we look at table

2.9 again, we can see that some observations have a recorded inactivity time of less than 20 minutes. Granted that the mean and the median are reasonable measures of central tendency that provide us with an overall view of the data, but if some time interval of fewer than 20 minutes is present in our dataset, it means that there is a likelihood of it happening. No matter how small this probability will be, we should investigate this phenomenon further, as it is crucial information for our goal.

Well, there is nothing better to start some mathematical endeavor with than some graphs. For these particular ones, as we are interested in frequency measures, we could create intervals representing an interval of inactivity, and then count the number of observations that fall inside these same intervals or bins. Then we could plot those counts in the shape of bars and visually understand how the data is behaving. Let's give it a try:

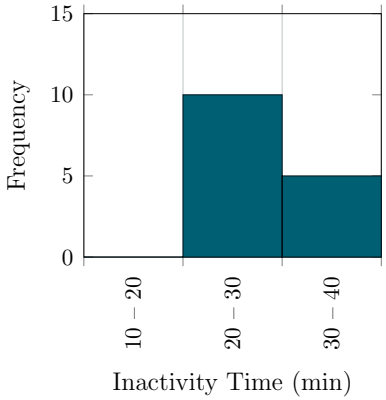


Figure 2.5: The bars that are not from a cell but for the inactivity times from 00:00 to 02:00.

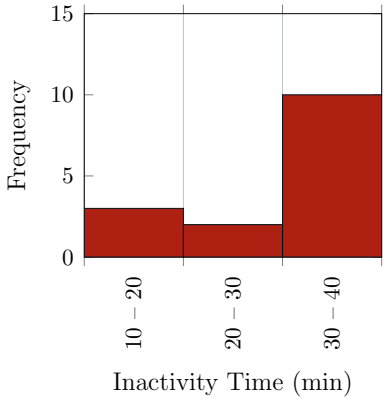


Figure 2.6: Gladly still not cell bars: The inactivity times from the latter time frame.

Figures 2.5 and 2.6 represent the frequency of times of inactivity grouped by beans (I meant bins, but come on, what a nice typo!) These plots, where we group data by given intervals, are called "histograms".

2.1.5 The Real Bar Exam: Mastering Histograms

Histograms are a handy technique and, quite frankly, something I use and abuse nearly as much as red apples (such an underrated fruit!) To start interpreting this new type of plot, let's turn our attention to the x-axis. For both histograms, we've established intervals of 10 minutes to capture periods of inactivity, creating three equal bins. Here's the process we followed to determine this specific 10-minute interval length:

$$\text{Bin Width} = \frac{\text{Top value} - \text{Lower Value}}{\text{Number of Bins}} \tag{2.10}$$

If we take a closer look at figures 2.5 and 2.6, we can see that the histogram on the left, the one with the blue bars, only has two of them, meaning we don't observe periods of inactivity between 10 and 20 minutes. This suggests we could have gone with a different Lower Value', for example, 20 rather than 10. However, our goal is to compare the two sets of data, 00:00-02:00 vs 02:00-04:00. In order to make these comparisons more straightforward and accurate, we opted to have both histograms with the same x-axis scale—this will mean that we are similarly grouping the data. So, to compute the bins for plots 2.5 and 2.6 these were the values we input in formula 2.10:

- **Top value** - 40
- **Lower Value** - 10
- **Number of Bins** - 3

So:

$$\text{Bin Width} = \frac{40 - 10}{3} = 10 \text{ minutes}$$

The actual data ranged from a minimum of 11.67 minutes to a maximum of 38.32 minutes. We selected a top value of 40 and a minimum value of 10 to ensure each bin represented a 10-minute interval of inactivity, facilitating easier analysis and visualization.

While such manual adjustments in bin settings can provide clearer insights for smaller datasets, they may become impractical or sub-optimal for very large datasets. In such cases, automated methods and computational tools are recommended to determine the most appropriate histogram bin widths based on data distribution and analysis goals. One example of such method is the Sturges' formula:

$$\text{Bin Width} = \log_2 n + 1$$

Sturges' formula uses a logarithmic scale to determine the number of histogram bins because it assumes that data can be efficiently organized into a number of categories that grows logarithmically with the size of the data. This makes sense because as we double the amount of data, we don't need to double the number of bins; instead, we only need one additional bin to maintain a similar level of detail, reflecting the efficiency of binary logarithmic division in data categorization.

However we are dealing with a small dataset to better grasp these concepts, so what about if we do some experimentation? I suspect that the shape of the histograms will change if we play around with the number of bins. Let's cut their size in half by increasing the count to 6:

$$\text{Bin Width} = \frac{40 - 10}{6} = 5 \text{ minutes}$$

Would you like to see the new histograms ?

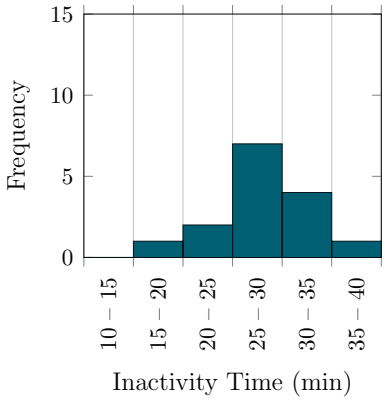


Figure 2.7: Shorter intervals but more bars?!?

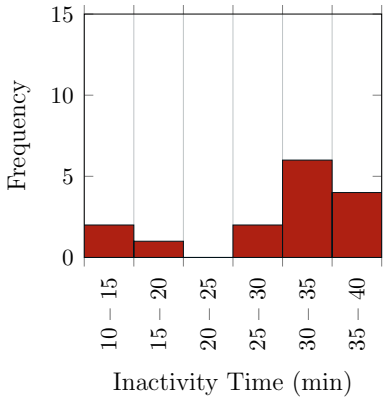


Figure 2.8: Hummm, I am a bit different than the blue counterpart.

The analysis of histograms in figures 2.7 and 2.8, which represent data distributions in 5-minute intervals, reveals a crucial insight for

our strategic planning. Although the datasets share similar central tendencies, indicated by their means and medians, they differ in their dispersion. Figure 2.7 shows a tighter clustering of data around the mean, suggesting a more consistent pattern of inactivity intervals. By contrast, the data points in figure 2.8 are more widely spread, indicating greater variability in the period for which the cameras are down.

Understanding these patterns is essential for identifying an optimal inactivity window, potentially ranging from 35-40 minutes on the 02:00-04:00 period, which could be advantageous for our planning purposes.

Greatness might be an overstatement, but quantifying this dispersion will certainly be useful. We can start by revisiting something we discussed while creating the histogram bin widths: the data minimum and maximum. By subtracting these values, we will obtain the range of the data.

2.2 Measures of the Spread of the Statistical Rumor.

That is right; the concept of 'range' introduces us to what we collectively term as 'measures of spread'. These statistical tools are crucial for understanding the dispersion within a dataset, helping to reveal the variability of data points around a central value. Rather than focusing solely on where data tends to center, measures of spread describe how data is stretched or squeezed, which is fundamental in assessing the reliability and predictability of statistical conclusions. As we proceed, we will study several measures that highlight different aspects of spread in a dataset, enhancing our analysis and interpretation skills. First the range!

2.2.1 Ra(n)ge Against the Data: Measuring the Extremes.

The 'range' is a measure of dispersion that indicates the difference between the highest and lowest values in a dataset. It helps to understand the spread of the data points:

$$\text{Range} = \text{Data Max} - \text{Data Min}$$

In our specific case:

$$\text{Range} = 38.32 - 11.67 = 26.65$$

While the range calculation of 26.65 minutes gives us a quick understanding of the overall spread of inactivity times, it also highlights significant variability. For strategic planning, this level of variability requires careful consideration, meaning we must dig deeper! For example, leveraging the fact that the mean is a measure of central tendency, we can set it as an anchor and then subtract it from each observation. This process will provide us with information on how far each data point is from the center. But before moving on with the calculations, we must consider something; only subtracting the mean and an observation might result in negative values; well, if we are after a measure of dispersion, this is not ideal. To get around this, we can use the absolute value of these differences:

$$\text{Abs Difference} = |x_i - \bar{x}| \tag{2.11}$$

Equation 2.11 represents the absolute differencence between a single data point x_i and the mean $\bar{x}$, however for us, an aggregate metrics will be more useful as we can better draw conclusions. Well one way to get there is to sum all the individual absolute differences.

2.2.2 Are you M(ean) A(bsolute) D(eviation)?!?!

The 'mean absolute deviation' (MAD) is a measure of spread that calculates the average of the absolute differences between each

data point and the dataset's mean. The formula is given by:

$$\text{MAD} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}| \tag{2.12}$$

Where the x_i's are the data points, $\bar{x}$ is the mean of the dataset, and n is the number of data points. Right, let's apply formula 2.12 to each of our time intervals:

$$\text{MAD}_{00:00-02:00} = \frac{1}{15}\sum_{i=1}^{15}|x_i - 28.78|, \quad \text{MAD}_{02:00-04:00} = \frac{1}{15}\sum_{i=1}^{15}|x_i - 29.50|$$

To compute these values one just has to input the entries of tables 2.8 and 2.9 and sum up all their absolute difference with the mean:

$$\text{MAD}_{00:00-02:00} = 3.42, \quad \text{MAD}_{02:00-04:00} = 6.39$$

The results are indeed intriguing! They reveal a more significant variability in the inactivity times during the later period (02:00-04:00) compared to the earlier one (00:00-02:00). This indicates that inactivity times during 02:00-04:00 are more widely dispersed from their mean, suggesting less consistency in the duration of inactivity intervals. Such insights are valuable for strategic planning. We could opt for a more conservative approach by choosing the 00:00-02:00 period, where we can expect inactivity intervals to be closer to the mean. Alternatively, if we're feeling adventurous, we might select the 02:00-04:00 period, which could offer intervals of greater length. However, it's important to remember that this choice carries a higher risk, as we might also encounter significantly shorter intervals.

While the MAD has proven helpful in evaluating the consistency of inactivity times, its methodology raises particular concerns. By calculating the absolute differences between individual values and the mean, this method treats all deviations equally, regardless of their magnitude. This uniform treatment can obscure the true impact of extreme values on the overall data analysis. For our purposes, this could be particularly problematic; low values of inactivity time need to be highlighted because failing to do so might mean missing critical risks or opportunities presented by these significant deviations. So how can we get around this? Well, what about if we instead of taking the absolute differences, we calculate the square

difference? We won't have negative values and the bigger the difference the more accentuated it will be!

$$\text{Squared Difference} = (x_i - \bar{x})^2$$

Now we just need to add up all the square differences to have another aggregated metric that will provide us with more information on how our data is spread. Yes you got it right, we just built another measure of spread! And this one is quite popular.

2.2.3 How Far Away From that Mean? The Sample Variance.

I am talking about the sample 'variance', a metric that measures the dispersion of data points around the mean, calculated by averaging the squared differences between each data point and the sample mean:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{2.13}$$

I feel like someone gave us a new toy, and I can't wait to use it! Too much? You are probably right; it's not even that cool of a formula, but still, let's explore it:

$$s^2_{\text{00:00}-\text{02:00}} = 19.52, \quad s^2_{\text{02:00}-\text{04:00}} = 67.27$$

When we calculate the mean, we use all the data points in our sample, so that guy is free of any constraints; oh, what a dream. So if we have a sample of size n, the sample mean has n degrees of freedom.

If we look at equation 2.13, we see that, unfortunately, this guy is not as free as the mean because one parameter is "fixed," the sample mean $\bar{x}$. This implies that the sample variance has one constraint; therefore, its degree of freedom is $n - 1$.

When calculating the sample mean, we divide the sum of all observations by the sample size, n. This process determines the average value of the dataset. However, when we compute the variance using this mean, dividing by n can introduce bias because it fails to

account for the degree of freedom caused by the use of the sample mean. To correct this bias, we use $n-1$ instead of n. This adjustment compensates for the lost degree of freedom inherent in the use of the sample mean in variance calculations:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{2.14}$$

This modification is particularly relevant in smaller samples where each observation has a significant impact on the mean and variance. For larger samples, the difference between dividing by n or $n-1$ becomes less significant. Thus, this correction ensures more accurate variance estimates, which is especially beneficial for smaller datasets.

Therefore, our revised variance calculations are now more robust and reflective of the actual variability within the data:

$$s^2_{00:00-02:00} = 20.92, \quad s^2_{02:00-04:00} = 72.08$$

In our scenario, understanding the precise variance in camera downtime is critical. An underestimated variance might lead to incorrect assumptions about the predictability of camera inactivity, potentially jeopardizing the heist. Using the adjusted variance gives us a more accurate picture of the risks involved, enabling more informed decision-making for the heist planning. We must keep our asses out of jail!

2.2.4 Is it Really that Standard to Have Such a Deviation?

Variance is a squared measure of dispersion, which does not allow us to have the deviations of inactivity time in minutes. However, taking the square root of that guy will solve this issue and provide us with a metric on a more interpretable scale:

$$s = \sqrt{s^2} = \sqrt{\frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2.15}$$

Equation 2.15 defines the sample 'standard deviation', a statistical measure representing the average amount by which an individual observation differs from the sample mean, and of course, our fourth measure of dispersion. Well, let's see what we have with respect to the 00:00-02:00 vs 02:00-04:00 timeframes:

$$s_{00:00-02:00} = \sqrt{20.92} = 4.57, \quad s_{02:00-04:00} = \sqrt{73.66} = 8.49$$

In our heist, selecting the 00:00-02:00 time slot implies a standard deviation of about 4.57 minutes from the mean, whereas if we pick the interval 02:00-04:00 it nearly doubles to about 8.49 minutes. What this means, is that if we pick a 00:00-02:00 timeframe in which to get rich, we can expect to be 4.57 minutes away from the mean of that same timeframe, 28.78.

We began by exploring measures of central tendency, namely the mean, median, and mode. These metrics provide valuable insights into the typical or average characteristics of our dataset. Next, we explored what information we could gather if we divided our dataset into intervals containing equal proportions of the dataset. They helped us to see how data is spread across the range and gave us information about the shape of the distribution. I am, of course talking about the quantiles and deciles. However, to gain a more comprehensive understanding of our data's behavior, we recognized the need for additional measures that could describe how data is dispersed, or spread out. This led us to examine measures of spread, such as range, MAD, variance, and standard deviation. These two fellows, the variance and the standard deviation offer a numerical measure of how much individual data points deviate from the mean, giving us a sense of the overall variability within our dataset.

Our dilemma remains the same: choosing the optimal time slot in which to carry out the diamond heist. On one hand, we have a time slot where periods of inactivity are closely clustered around the mean, as shown in figure 2.7. This indicates a more predictable pattern. By contrast, 2.8 represents a time slot with greater variability. This is evident from the standard deviation values: 8.49 for the 02:00-04:00 slot compared to 4.57 for the 00:00-02:00 timeframe. The 02:00-04:00 slot, though riskier due to its higher frequency of

both low and high inactivity periods (as depicted by the left and right extremes of its histogram), presents an intriguing option.

Now, do we choose to go rob the vault between 00:00 and 02:00 and play it safe? Or do we hope the hope that the thief god is with us (after a quick search, I can confirm that this god exists—Hermes) and go in the time slot of 02:00-04:00. Maybe we can somewhat "quantify" this luck...

Chapter 3

Probability and Theory: Partners in Decoding Uncertainty.

Yes! We have a theoretical framework to quantify randomness and uncertainty, known as probability theory. Randomness serves as a way to articulate our incomplete knowledge about specific events or instances, for example the scenario of entering the diamond vault on a weekend night. Will we encounter periods of high camera inactivity or not? This question shows our limited understanding of the situation and the uncertainty of the outcome, meaning that jail time is still possible. Using probability theory, we can translate this uncertainty into percentages, estimating how frequently we expect a particular scenario to occur. This is a tool with which we can gauge the likelihood of different outcomes based on our current knowledge.

Cool, so it's clear that we'll be working with these 'probability sisters,' where the first important fact is that they operate on collections of 'things.' In mathematics, these collections are called 'sets.

3.1 Now That We Are Set(s) In Motion, Bring Some Operations.

In probability theory, a set is a collection of objects, which are the elements of the set. The concept of a set is fundamental, as it provides a way to mathematically define and work with collections of outcomes. For instance, in the context of probability, a set

can represent all possible outcomes of a random experiment. Operations such as unions (combining two sets), intersections (common elements between sets), and complements (elements not in a set) are commonly used to analyze and understand probabilities of different combinations of events occurring.

Fancy wording, I know. Let's break this down with an example. As we're just beginning to explore the basic concepts of probability theory, let's consider the two scenarios for our times of inactivity: from 00:00 to 02:00 versus 02:00 to 04:00. We can define a set based on these events:

$$S = [0, 2[\cup [2, 4[\tag{3.1}$$

The set S, as defined by equation 3.1, encompasses all times within our two designated timeframes. Specifically, the interval $[0, 2[$ corresponds to the time slot from 00:00 to 02:00, while the interval $[2, 4[$ represents the time slot from 02:00 to 04:00. To indicate that a specific time is included in one of these intervals, we use the following notation:

$$1 \in S$$

That not so handsome e, the $\in$ symbol, means 'to belong'. Conversely, to express that an element does not belong to a given set, we just have to put a line over the $\in$:

$$6 \notin S$$

We refer to S, defined by equation 3.1, as the 'universal set' or the 'sample space.' Our pal S is the set with all possible events within our study, which in this case are the two-timeframes that define the sample space. We can also use the following notation to define a set:

$$S_{[0,2[} = \{x \,|\, x \in S, 0 \leq x < 2\} \tag{3.2}$$

$$S_{[2,4[} = \{x \,|\, x \in S, 2 \leq x < 4\} \tag{3.3}$$

Equations 3.2 and 3.3 represent another way of defining a set. We should read the symbol "|", as "given" or "such that". Because $S_{[0,2[}$ and $S_{[2,4[}$ are part of S, and are disjoint, meaning they do not overlap, but collectively exhaust the elements of S, we express this mathematically as:

$$S_{[0,2[} \subset S \quad \text{and} \quad S_{[2,4[} \subset S$$

These concepts of set and subset can be visualized with a type of plot called a Venn diagram:

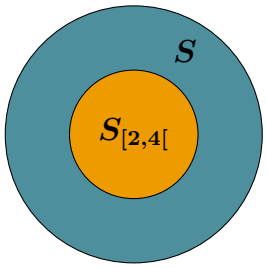


Figure 3.1: Is it Zen? No, it is Venn!

The Venn diagram visually demonstrates that all elements of subset $S_{[2,4[}$ are included in the universal set S, but S may contain additional elements not found in $S_{[2,4[}$ (the $S_{[0,2[}$ set). As $S_{[2,4[}$ is a subset of S we can also have a subset of $S_{[2,4[}$, let's call it $S_{[2,3[}$:

$$S_{[2,3[} = \{x \,|\, 2 \leq x < 3 \,\} \quad x \in S_{[2,4[}$$

Visually:

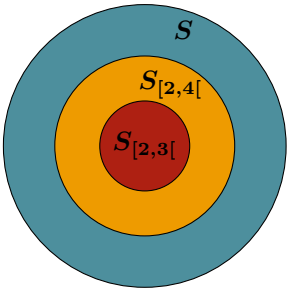


Figure 3.2: A less Zen, Ven diagram: 3 circles.

In our scenario, sets $S_{[0,2[}$ and $S_{[2,4[}$ are disjoint. Their union, however, forms the entire set S. This is why in our Venn diagram 3.1, we do not need to separately depict $S_{[0,2[}$ and $S_{[2,4[}$ within S; their combination inherently constitutes the whole of S. It is no secret that we have a notation for that union operation, and in fact, the union is just one of many we will explore.

3.1.0.1 Probabilistic Matrimony: The Union of Sets.

The 'union' of two sets A and B is another set containing all elements in A or B or even both. We represent the union by $\cup$ and

an example is:

$$S_{[0,2[} \cup S_{[2,4[} = S$$

Which is equivalent to:

$$[0, 2[\cup[2, 4[= S$$

Another example is if we were to union the sets $S_{[2,3[}$ and $S_{[0,2[}$:

$$S_{[0,2[} \cup S_{[2,3[} = [0, 3[$$

If we plot this:

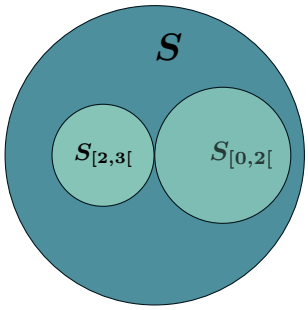


Figure 3.3: Isn't that nice? A set hug.

So mathematically, we say that $x \in S_{[0,2[} \cup S_{[2,4[}$ if and only if $x \in S_{[0,2[}$ or $x \in S_{[2,4[}$ and that $S_{[0,2[} \cup S_{[2,4[} = S_{[2,4[} \cup S_{[0,2[}$. The union of sets is not limited to two sets; we can have as many as we wish. Isn't that nice? I find it wonderful! So if we define n sets as $A_1, A_2, A_3, ..., A_n$ we can define their union as $A_1 \cup A_2 \cup A_3 ... \cup A_n$ which is equivalent to:

$$\bigcup_{i=1}^{n} A_i$$

Any more operations, you ask? Well, what about if we flip that $\cup$? 3,2,1 catch! $\cup$ ... $\subset$... $\cap$. Nice one!

3.1.0.2 Is that The Only Thing We have In Common? Intersection.

The symbol $\cap$ denotes the 'intersection' of two or more sets. The results of this operation are the common elements between them.

For example, if we again take set $S_{[2,4[}$ and $S_{[2,3[}$ and calculate the intersection of these two specimens, we find that:

$$S_{[2,4[} \cap S_{[2,3[} = [2,4[\cap[2,3[= [2,3[$$

Visually:

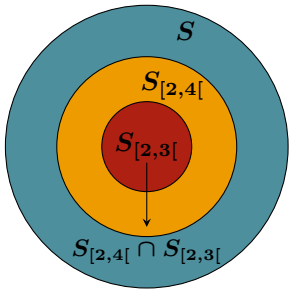


Figure 3.4: We have so much in common!

In graph 3.4, we have represented the universal set S and the sets $S_{[2,4[}$ and $S_{[2,3[}$. The red area in the diagram highlights the intersection of sets $S_{[2,4[}$ and $S_{[2,3[}$, which is precisely the set $S_{[2,3[}$. As we did with the union, we can have a more general formula to represent the intersection of any number of sets:

$$\bigcap_{i=1}^{n} A_i$$

Where A_i is a given set, the logic for more than two sets is the same as with two sets; the intersection results are what all the sets have in common, for example, if we have 3 sets:

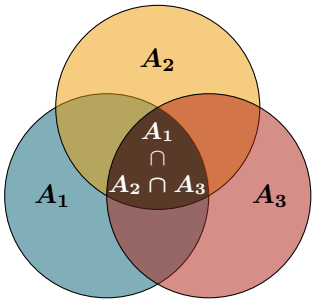


Figure 3.5: The communality of a few fellows.

The Venn diagram depicted in figure 3.5 illustrates the intersection among three specific sets: A_1, A_2, and A_3. This diagram effectively highlights the common elements shared by all three sets.

This shared region is represented by the darker area at the center of the diagram, visually emphasizing the elements that are common to A_1, A_2, and A_3. If the sets don't have anything in common, meaning that the interception is the empty set $\emptyset$ or $\{\}$, we say that they are mutually exclusive or disjoint:

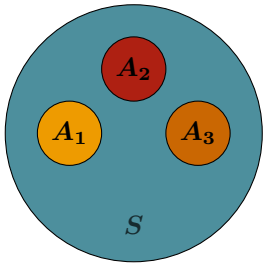


Figure 3.6: Are the sets upset with each other?

Next, we will cover three more concepts regarding set operations and properties. If, by now you are thinking, "Why do we need to bother with overlapping colorful circles?" Well, apart from being the foundation for all the probability theory we will study, this concept will help immensely when we are dealing with databases, so stick with me for a while longer. The next concept we will cover is the complement of a set:

3.1.0.3 The Other Half: Finding Wholeness with Set Complements.

We denoted the 'complement' of a given set $S_{[0,2[}$ by $S^c_{[0,2[}$ or $\overline{S_{[0,2[}}$. The complement is everything in the sample space S that is not in $S_{[0,2[}$. For example:

$$\overline{S_{[0,2[}} = [2, 4[$$

Visually:

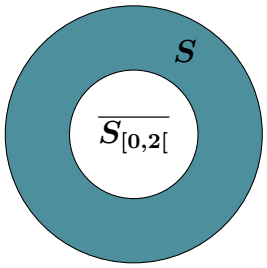


Figure 3.7: We have a hole, but is it in the set's soul?

In graph 3.7, we have illustrated the complement of set $S_{[0,2[}$, denoted as $\overline{S_{[0,2[}}$. The area within the white circle represents the portion of the universal set S not included in set $S_{[0,2[}$. This showcases $\overline{S_{[0,2[}}$ as comprising all elements of S that are outside the boundaries of set $S_{[0,2[}$, thereby visually capturing the concept of a set complement. We have two more concepts to go and then the coast is clear, after these two we can ... well, learn more about probabilities, what a day this will be! No? Tough crowd. Now, we will go into the concept of differentiation or subtraction:

3.1.0.4 Just Take It: Subtraction.

We can define the subtraction or differentiation of two sets as everything that is in one but not in the other one, for example:

$$S_{[2,4[} - S_{[2,3[} = [2,4[-[2,3[= [3,4[$$

Plot plot plot!

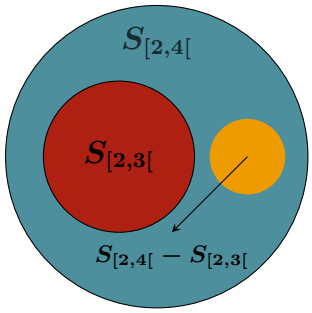


Figure 3.8: You are nothing without me! Oh, maybe you are...

In 3.8, the yellow circle represents everything that belongs to $S_{[2,4[}$ but is not part of $S_{[2,3[}$; therefore, it also represents $S_{[2,4[} - S_{[2,3[}$. The next concept has a cool name; it is the 'cartesian product':

3.1.0.5 Cross-Selling Sets: Exploring Cartesian Products.

Considering two sets A_1 and A_2, we define the 'cartesian product' $A_1 \times A_2$ as a set containing ordered pairs from A_1 and A_2. For demonstration purposes, let's define two new sets A_1 and A_2 such that:

$$A_1 = \{1, 2, 3\}, \quad A_2 = \{S, L, Y\}$$

So $A_1 \times A_2$ the cartesian product is then defined as:

$$A_1 \times A_2 = \{(x, y) \mid x \in A_1, y \in A_2\} \tag{3.4}$$

Equation 3.4 barks but it does not bite. It says we must construct pairs of elements where the first one, represented by x belongs to the first set A_1 whereas the second, the y is an element of the second set A_2:

$$\{(1, S), (1, L), (1, Y), (2, S), (2, L), (2, Y), (3, S), (3, L), (3, Y)\}$$

I told you that formula wouldn't bite; we just combined all elements of A_1 with elements of A_2. If the cartesian product was $A_2 \times A_1$ then the logic would be the same but we would have to combine the elements of A_2 with A_1:

$$\{(S, 1), (S, 2), (S, 3), (L, 1), (L, 2), (L, 3), (Y, 1), (Y, 2), (Y, 3)\}$$

Next, we address the elephant in the room... shit, it is behind you? Duck! Anyhow, why do we need all of this? Well, sets and their properties lay the foundation for grasping more complex statistical concepts. For instance, when we are dealing with probabilities, events are often represented as sets, and the likelihood of combined or mutually exclusive events is directly linked to set operations. Furthermore, and arguably more tangibly, the principles of set theory find a compelling application in databases and data manipulation. Our machine learning models need features, and creating those involves data manipulations. One of the techniques used for such manipulations is to make "joins," a process where we merge different tables, or sometimes the same table, based on common fields. For example:

Members Name	Role
Nortarbartolo	Get away driver
The Genius	Disable alarms
The king of keys	Vault Key

Table 3.1: The crew and their roles.

Members Name	Speciality
Nortarbartolo	Planning
The Genius	Alarm specialist
The king of keys	Forging keys

Table 3.2: The crew's expertise.

Say that we want to have all this information in one single table. One way to do this would be to operate based on the intersection of sets, meaning we combine everything these tables have in common based on the field "Member's Name". The new table looks like this:

Members Name	Role	Specialty
Nortarbartolo	Get away driver	Planning
The Genius	Disable alarms	Alarm specialist
The king of keys	Vault Key	Forging keys

Table 3.3: A full view of the crew's features.

This joining operation is a specific type called an "inner join", where the result includes only those elements that are common to both tables, based on one or multiple fields. A more commercial example would be if we wished to predict whether a customer would return to our e-commerce website to buy another product. Here, we could link web behavior with sales information. There are instances when this information is stored separately, so, if we can find a unique identifier per customer common in both datasets then, with the proper aggregation, we can have a final table that combines all the information.

These operations seem useful and, gladly, we are not limited to just "inner joins"; there are more. Below, we have a table with all the types of joins, meaning all the ways we can combine and manipulate data between tables alongside their analogous set operations:

Join Type	Set Operation
Inner Join	Intersection ($A \cap B$)
Left Join	Elements of First Set and Intersection ($A \cup (A \cap B)$)
Right Join	Elements of Second Set and Intersection ($B \cup (A \cap B)$)
Full Outer Join	Union ($A \cup B$)
Left Anti Join	Set Difference ($A - B$)

Table 3.4: Set operations.

We have established some foundations that will allow us to start quantifying uncertainty. Now it's time to get to work and use them to make a more informed decision about choosing the 00:00-02:00 or 02:00-04:00 time slot. As we are looking for a large period of inactivity, one strategy would be to select a time interval and understand how likely this exact interval is to occur in both periods. For example, say we want to understand the probability of us observing an

inactivity interval between 35 and 40 minutes. This means kicking of the probability section of the book, and for that, let's revisit our data samples for both periods of inactivity while introducing a new concept — 'cardinality':

$$|S_{00:00-02:00}| = 15 \tag{3.5}$$

Equation 3.5 defines the concept of cardinality, representing the total number of observations in a given set. To arrive at the number 15, we simply count the observations listed in table 2.8. The new notation is introduced because, when discussing set theory, we define the set $S_{[0,2[}$ to include all possible times between 0 and 2 (exact 2 excluded). To apply the same approach for $S_{02:00-04:00}$, we just follow the same methodology:

$$|S_{02:00-04:00}| = 15$$

Now, we aim to find out how many of these 15 observations fall within the 35-40 minute interval, meaning that we need the cardinality of the subset that contains only the interval 35-40 minutes of inactivity. Because the tables are up there in the book and we are looking for an interval that is 15 minutes long, let's create an auxiliary table:

Inactivity Interval	00:00-02:00 Frequency	02:00-04:00 Frequency
$]10, 15]$	0	2
$]15, 20]$	1	1
$]20, 25]$	2	0
$]25, 30]$	7	2
$]30, 35]$	4	6
$]35, 40]$	1	4

Table 3.5: Frequency count by inactivity time interval.

Table 3.5 is in essence the data we used to create histograms 2.7 and 2.8 in tabular form. Rather than using bars to show frequencies, this table presents the precise counts for each inactivity time interval. Cool, now to compute the likelihood of having an interval of 35-40 minutes in the time slot 00:00-02:00, we have to calculate the cardinality of this event:

$$|]35, 40]_{00:00-02:00}| = 1 \tag{3.6}$$

Equation 3.6 represents the number of observations that occurred in those specific time periods, which we can observe in table 3.5. Therefore, the probability for the interval 35-40 min is calculated as:

$$P([35, 40]_{00:00-02:00}) = \frac{1}{15} \approx 0.067 = 6.7\% \tag{3.7}$$

3.2 Chances Are We Need a Bit Of Probability In the House.

In equation 3.7, P stands for "probability". Probabilities can be expressed as decimals or percentages, and the sum of the probabilities for all subsets of the sample space S adds up to 1. They are calculated as:

$$P(A) = \frac{|A|}{|S|}$$

Where $P(A)$ is the probability of event A occurring, $|A|$ is the cardinality of the set A and S is the cardinality of the sample space. These cool fellows, the probabilities, have 3 axioms that are very simple, but also very important:

- **For any event $A, 0 \leq P(A) \leq 1$** - This says that a probability can't be negative, which makes sense. If we are trying to measure uncertainty, we can't have a negative number to quantify this measure. However this tells us that events very close or equal to 0 are either very unlikely or even impossible, meaning that $P(A) = 0$. Equally we can't have any event having more than 1 of probability of happening.

- **The probability of the sample space S is $P(S) = 1$** - Well if S contains all the possible outcomes of our random experiment, the outcome of each trial will always belong to S.

- **If $A_1, A_2, ..., A_n$ are disjoint events then $P(A_1, A_2, ..., A_n) = P(A_1) + P(A_2) + ... + P(A_n)$** - Let's be real here; from the 3 axioms, this is the one that seems most interesting. This guy reflects the additive nature of probabilities in the case of disjoint events. Since these events do not overlap, the total probability of any of the events occurring is the sum of their individual probabilities. This concept is behind ideas like probabilities distributions and the Bayes theorem. Both concepts will be covered in this book.

To explore the concept of disjoint events, let's pick two intervals of inactivity for the timeframe 00:00-02:00: $]10, 15]$ and $]35, 40]$. Are they disjoint? Do they overlap? They are disjoint, and therefore, they do not overlap. So, to compute the probability of any of those time intervals occurring:

$$P(]10, 15]_{00:00-02:00} \cup]35, 40]_{00:00-02:00}) = P(]10, 15]_{00:00-02:00}) + P\left(]35, 40]_{00:00-02:00}\right)$$

One of those, $P\left(]35, 40]_{00:00-02:00}\right)$, is the equation 3.7. So let's compute the missing bit, $P(]10, 15]_{00:00-02:00})$:

$$P(]10, 15]_{00:00-02:00}) = \frac{0}{15} = 0$$

For the final calculations:

$$P(]10, 15]_{00:00-02:00} \cup]35, 40]_{00:00-02:00}) = 0.067 + 0 = 0.067$$

As the probability of inactivity of the camera in the event $]10, 15]_{00:00-02:00}$ is 0, we can see that the probability of the union of those two events is approximately 6.7%. An intuitive way to read the union $\cup$ is "or". It is important to remember that the following formula:

$$P(A \cup B) = P(A) + P(B) \tag{3.8}$$

Is only valid if A and B are disjoint, which is the case for $]10, 15]_{00:00-02:00}$ and $]35, 40]_{00:00-02:00}$. Disjoint events mean no overlapping information; conversely, "joint events" imply some overlapping; it is clear that we are looking for a new formula to compute the probability of the union of joint events, so let's consider two specimens with these characteristics:

- **Event *A***: A period of inactivity of 25-35 minutes.
- **Event *B***: A period of inactivity of 30-40 minutes.

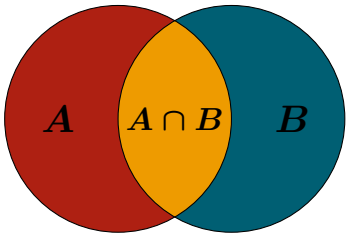


Figure 3.9: We have so much in commune.

Figure 3.9 represents an instance where A and B overlap, causing some elements of A to also be present in B and vice-versa. So, if we sum up the probabilities $P(A)$ and $P(B)$, we will double-count some particular instances, which will return garbage. Fortunately, we can sort out this incident by subtracting the probability of A intersected with B:

$$P(A \cap B)$$

If we wish to calculate the probability of the intersection of the events A and B we can start by calculating their intersection:

$$A \cap B =]25, 35]_{00:00-02:00} \cap]30, 40]_{00:00-02:00} = [30, 35]_{00:00-02:00}$$

Following we have:

$$P(A \cap B) = P([30, 35]_{00:00-02:00}) = \frac{4}{15} \approx 0.267 \approx 27\%$$

Cool, so since we are dealing with intersections let's consider the following events:

- **Event *C***: A period of inactivity of 10-15 minutes.
- **Event *D***: A period of inactivity of 35-40 minutes.

What will be the probability of the intersection of C and D be?

$$P(C \cap D) = 0$$

A probability of 0 comes up because the events are disjoint and have nothing in common. Joint, disjoint, and independent events

are fundamental concepts, so let's create a small summary for the first two and save the last one, the independence, for dessert (which will be covered shortly).

- **Disjoint Events:**
  - Disjoint events, also known as mutually exclusive events, are events that cannot happen at the same time.
  - If two events are disjoint, the occurrence of one event means the other event cannot occur in that specific instance.
  - The probability of disjoint events occurring together is always zero.
- **Joint Events:**
  - Joint events are events that can happen at the same time. They are not mutually exclusive.
  - If two events are joint, the occurrence of one event does not prevent the occurrence of the other.
  - Joint events can have an intersection, which is the scenario where both events occur together.

Right, where were we? Ah! Deducing a formula for the union of two joint events, A and B; thanks for the reminder. We said we needed to subtract the probability of the intersection, which means that the formula for this is:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{3.9}$$

Equation 3.9 is very similar to 3.8, which represents the union of disjoint events, but in this particular case, as we are dealing with the union of joint events, to avoid double counting, we subtract their intersection. We can now carry on with our calculations; how about if, this time, we prioritize B and go for $P(B)$ first?

$$P(B) = P(]30, 40]_{00:00-02:00}) = \frac{5}{15} \approx 0.3 = 30\%$$

Now, for event A we see that:

$$P(A) = P([25, 35]_{00:00-02:00}) = ...$$

Yeah, we do not have that time interval in our histogram... we can redo all the bins and then compute this probability:

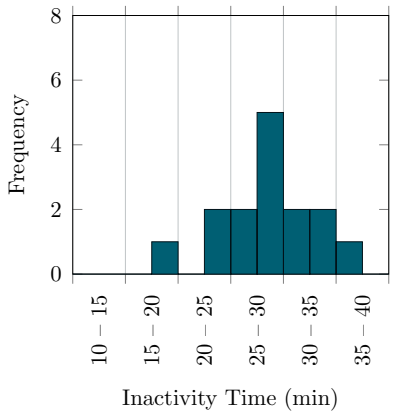


Figure 3.10: Really? More bars?!?

Sure, we can create a similar table to 3.5, in which we had all the frequencies for time intervals, but what happens, for example, if we want to know the probability of the time interval $[21, 32]$? Do we have to redo all those bins again? There must be a better way. One strategy is to make the number of bins tend to infinity, similar to what one does with integrals, going from rectangles to a continuous form. Well, this suggests some curves, and what better to accompany curves with than functions?

Since the last paragraph seems to set up a discussion about mathematical functions, it's a smart move to choose our ideal time interval for the heist before we jump into the math madness.

I believe we should play conservatively and minimize the probability of a low inactivity interval. For that, we will calculate the following probability for both time slots:

$$P(]10, 15] \cup]15, 20])$$

Starting with the 00:00-02:00 timeframe:

$$P(]10, 15]_{00:00-02:00} \cup]15, 20]_{00:00-02:00}) = \frac{0}{15} + \frac{1}{15} \approx 6\%$$

For the 02:00-04:00 slot:

$$P(]10,15]_{02:00-04:00} \cup]15,20]_{02:00-04:00}) = \frac{2}{15} + \frac{1}{15} \approx 20\%$$

If we pick the first time slot, the one from 00:00-02:00, we have a probability of 6% of having a period of inactivity between $[10, 20]$, whereas, with the interval 02:00-04:00, we have a likelihood of 20% for the same inactive times. It is equally valid to say that there is a higher chance of catching a time interval of inactivity of $[35, 40]$:

$$P([35,40]_{00:00-02:00}) = \frac{1}{15} \approx 6\%$$

For the 02:00-04:00 slot:

$$P([35,40]_{02:00-04:00}) = \frac{4}{15} \approx 26\%$$

Still, let's act cautiously and pick the time slot of 00:00-02:00, the one with less variance, whose values seem centered around the mean. Speaking of which, let's bring this histogram back:

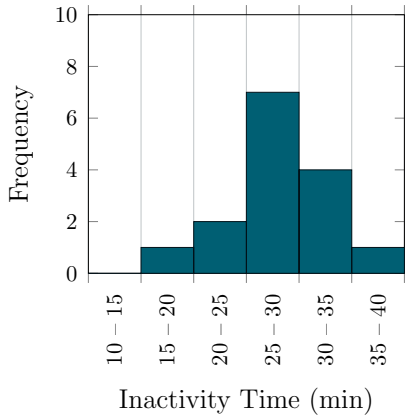


Figure 3.11: The histogram of our selected time interval.

Cool, so we have a timeframe, 00:00-02:00, and a few tasks to get on with. Firstly we need this probability:

$$P(A) = P([25,35]_{00:00-02:00})$$

That leads us to conclude that a function from which we can extract insights, and in this case probabilities, would be highly beneficial. Since functions map inputs to outputs, we need to define one

of these "input bad boys". Let's consider X to be a random variable that represents inactivity time.

A random variable is a function that helps us to quantify and study a group of elementary events of a sample space as we may not always be interested in the elementary events itself. It is a real valued function defined on the sample space and the event it represents. When we refer to X as a random variable representing the duration of inactivity intervals in minutes, we mean that X takes on different numerical values based on the length of these intervals, which are observed or measured as the experiment is conducted.

For example, if our camera has an inactivity period of less than 20 minutes, then $X < 20$. If the inactivity period is higher than 35 minutes, then $X > 35$. This makes X a random variable that directly quantifies the inactivity periods we observe. With X defined in this way, we can start discussing probabilities like $P(25 \leq X \leq 35)$ (Which is what we are after!) This approach transforms our raw observations into a format that can be analyzed using probability, giving us a deeper understanding of the inactivity patterns.

Notice that we don't have any $P(X = \text{"exact time length"})$ for this specific case. This is not an oversight; it's due to the types of random variables we have in probability and statistics:

- **Discrete Random Variables:** In the context of your heist, a discrete random variable can represent the number of security guards on duty at any given time. This variable is discrete, since you can count the number of guards (e.g., 2, 3, or 4 guards). Another example can be the number of alarms triggered during the heist. These are countable outcomes, and each has a specific probability, meaning it is possible to have $P(X = \text{exact number})$. Don't confuse countable with finite; we can have an infinite number of outcomes such as the number of attempts needed to crack a safe.

- **Continuous Random Variables**: On the other hand, a continuous random variable in our heist plan might be the time of inactivity of a given camera, which can take on any value within a range and is not limited to countable numbers. For

instance, it might be between 15.5 minutes and 16.75 minutes or any other interval of time measurement that the camera is malfunctioning.

Actually, in a continuous case, we can have the following probability $P(X = 15)$ that it is 0; therefore, it does not provide us with valuable information. Why? Well, let's play a game. Say we have all the numbers between 0 and 10, bearing in mind we are in a continuous setup. So we must consider an infinite number of them. Now, if I think of a number; what is the likelihood you will guess it? Well, it is one divided by infinity, which is 0. That is why, in a continuous setup, we work with intervals.

The distinction between discrete and continuous random variables is critical in probabilities. Discrete variables like the number of guards are countable and have distinct outcomes, each with a specific probability. By contrast, continuous variables like the inactivity time, can take on any value within a range and are represented by probabilities over intervals.

In our specific case, X is a continuous random variable that represents the inactivity periods in minutes; so its distribution or behavior has the shape of histogram 3.11. Cool, so now we need an equation to find outputs that, in this case, are linked to probabilities. If 3.11 represents the behaviour of X, the mathematical formula we are looking for would ideally return a curve that fits this same shape.

We can see that there is a lot of "mass" around the mean, which fades out as we move toward the tails for both sides, making it nearly symmetrical about the mean. When building the histogram, we created bins and then used the frequencies of the events to create each of those bars, meaning that this "mass" we speak of must be linked to probabilities. With this in mind, 3.11 suggests that the events surrounding the mean of 28.78 minutes are more likely to happen than the ones further away from it. Now, of course, it is that time of the day... equation time. (Hopefully you didn't think I meant doughnut time...). Now we need two things: a non-linear function that adjusts to the shape of that histogram and a tool that allows us to sum "masses" within a time interval. Remember we

concluded that our variable was continuous and that for this type of variable we need to compute probabilities on intervals. Well, calculus can come to the rescue! With calculus, we have a collection of functions and also a tool that provides us with areas under curves, the integrals!

If you are interested in the subject of calculus, check out Volume 2 of this series; if you have read it already, welcome back! So we need a symmetrical curve that decreases rapidly as x moves away from the mean; what about:

$$f(x) = e^{-x^2} \tag{3.10}$$

Let's plot this alongside the histogram:

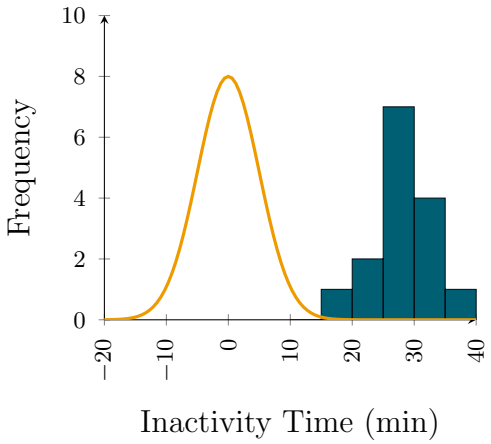


Figure 3.12: The first attempt at the function and it is a miss.

It would have been a great day if the first function we picked was a perfect fit. Nevertheless, we are off to a good start with e^{-x^2}. The most striking thing about figure 3.12 is that the yellow plot that represents equation 3.10 is to the left of our blue histogram. This means that we need to find a way to always center e^{-x^2} around a point, meaning we need to make the function's peak coincide with the histogram.

We previously stated that our data is centered on the mean, so this seems like a good point to place the peak of our function; this will guarantee that the function peaks at the mean, which is the behavior that we observe on the histogram. So, if we wish to "push" all the points to be around to the mean, we can subtract the

mean from all the points; making 3.10:

$$f(x) = e^{-(x-28.78)^2} \tag{3.11}$$

Let's plot the new curve 3.11 alongside the histogram:

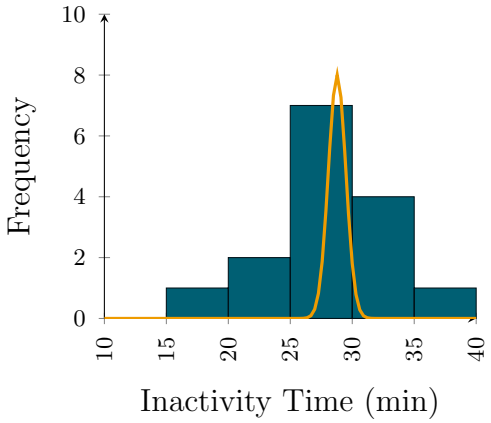


Figure 3.13: Some improvement but yet another miss: We need width!

We are certainly on the right track, but not quite there yet. Our theoretical function, referred to in equation 3.11, isn't wide enough, which hampers its ability to capture events as we move further from the mean——both to the left and the right. Therefore, we could use with some "spread"; fortunately, we have several options for this measure but let's explore the variance and the standard deviation.

The standard deviation quantifies the average distance of each data point from the mean. Consequently, a larger standard deviation indicates that data points are more widely spread from the mean, while a smaller one suggests they are more closely clustered.

Why consider incorporating this into our formula 3.11? Firstly, our choices were limited to means,median,modes, quantiles, range, mad,variances, and standard deviations. More importantly, using the standard deviation ensures that we have a metric that accommodates all distributions rather than having to select a specific number for each.

Now, let's explore how we can integrate this element into 3.11. We need the values further from the center to reflect higher deviations. Considering $(x - 28.78)^2$, these are the values where the difference is greatest. Since the standard deviation is always posi-

tive, we can either multiply or divide $(x - 28.78)^2$ by the standard deviation.

To decide which operation is more appropriate, consider an outlier, one of those values significantly distant from the mean. This would cause $(x - 28.78)^2$ to rise sharply compared to other values, and multiplying by the standard deviation would only exacerbate this effect. Such a distribution would be too wide, misrepresenting our data.

Conversely, dividing by the standard deviation scales down large deviations proportionally, ensuring that the function's shape accurately reflects the concentration of data points around the mean. This method makes our model sensitive to changes in the data spread, adjusting the width of the distribution to mirror the actual standard deviation of the dataset.

By normalizing the data, we bring all data points onto a comparable scale. This normalization is essential for comparing distributions across different datasets or ensuring that our model remains effective regardless of the inherent variability in the data. Thus, our equation now takes the following form:

$$f(x) = e^{-(\frac{x-28.78}{4.57})^2} \tag{3.12}$$

Let's see what 3.12 brings us:

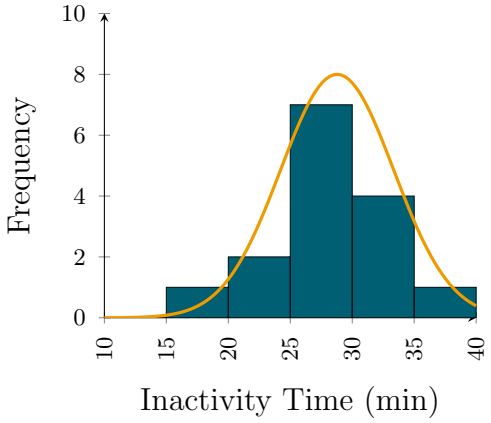


Figure 3.14: Did we just get it right?!?!

Yeah, baby! That is a much better module for our data as we capture the variance in our observations. We can say that our continuous variable X is represented by a function $f(x)$ such as 3.12.

Why X for random variable and x as argument of the function you ask? The distinction between X and x is important:

- X: This represents a random variable, which is a variable whose possible values are outcomes of a random phenomenon.
- x: This represents a specific value that the random variable X can take. It is used as an argument in the function $f(x)$.

Now with that matter clarified, what about if we check the curve on its own?

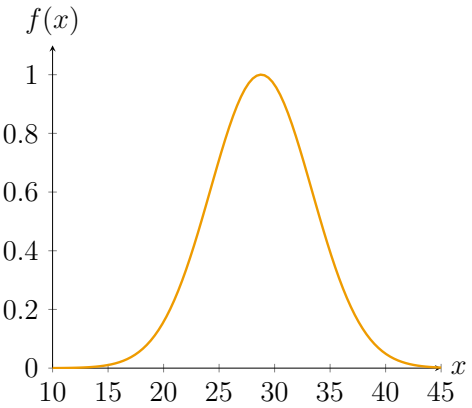


Figure 3.15: Look at that curve, just like a bell.

Graph 3.15 represents our curve, which we can create with the function 3.12. Still on the subject of a function, it is time to introduce the specific type that we are working with; it is called the "probability density function" (PDF) and comes with some peculiarities. It's kind of like that crazy uncle everyone has. He might ruin the family dinner, but he's your go-to guy if you are in trouble.

3.2.1 The Contours of Chance: Drawing Insights from Probability Density Functions.

How we extract information from a probability density function is different to any regular function. We can't just plug in values of x and expect to get direct probabilities as y's. This means the probability at a specific time, like 30 minutes, isn't simply $f(30)$.

If we try such calculations, we'll end up with the probability density, not the probability itself. In English? Imagine we are going to a concert and, because you're feeling generous, you've got us tickets to a particular spot in the stands, where we're above the crowd; thanks for the tickets!

Let's use this to understand why $f(x)$ doesn't directly give us the probability, but the probability density instead. As we arrive early, we watch people coming in and forming groups. By observing these groups, we can get an idea of how densely packed they are, just by seeing how much space they occupy. But to know how many people are in a given group, we need to count them. In our world, $f(x)$ tells us how dense the group is, but if we want the total number of people (or the probability), we need more than just $f(x)$. And that's where integrals come in.

An integral is like summing up very tiny rectangles; as the number of these rectangles approaches infinity, we get a precise calculation. This concept is pretty neat and exactly what we need here. So, to extract probabilities from $f(x)$, we integrate this function over a specific interval. And just like that, we turn the abstract density of a concert crowd into concrete numbers—or, in our case, probabilities:

For example:

1. **Density at $X = 30$**: The value $f(30) \approx 0.965$. Please do not confuse this with 96%; it is simply a density value of 0.965. To understand the significance of this value, let's compute the density for another value of X.

2. **Density at $X = 15$**: We find that $f(15) \approx 0.011$, which is much lower. This lower value suggests that outcomes near 15 minutes are considerably less likely in our distribution compared to $X = 30$. This is because the density around $X = 30$ is much higher.

Let's compute some integrals! Well firstly we need to verify a detail; all the probabilities must add up to 1. What this means is:

$$\int_{-\infty}^{+\infty} f(x)dx = 1 \tag{3.13}$$

No matter what shape our probability density function has, if we are dealing with a continuous variable, the integral of $f(x)$ between minus and plus infinity must be equal to one. Note that we specified that X must be continuous; this is because, if X were discrete, we would not need an integral, because we can simply count. We will go into more detail on the subject of discrete variables soon. So, what we have left to do, is to calculate the integral of our function 3.12:

$$\int_{-\infty}^{+\infty} e^{-(\frac{x-28.78}{4.57})^2} dx \tag{3.14}$$

The gentleman represented by 3.14 does come across as very striking to the naked eye, but we can make it friendlier by performing a change of variable. So, if we define u such that:

$$u = \frac{x - 28.78}{4.57} \tag{3.15}$$

Consequently, the integral 3.14 can be written as:

$$\int_{-\infty}^{+\infty} e^{-u^2} dx \tag{3.16}$$

One more extra step, and we are there. We can see a dx; however, there are no x's on our function 3.16, just u's. This means that we ideally need a du instead. No big deal, as we can find a relationship between dx and du by deriving 3.15 with respect to x:

$$\frac{du}{dx} \tag{3.17}$$

If we apply the chain rule to 3.17 we find that:

$$\frac{du}{dx} = \frac{1}{4.57}$$

$$dx = 4.57du$$

So our final formula that we will use to compute the integral is:

$$\int_{-\infty}^{+\infty} e^{-u^2} 4.57 du \tag{3.18}$$

This integral e^{-u^2} is an integral whose solution is well known, therefore we will skip the calculations and use this result; $\sqrt{\pi}$. With this new piece of information 3.18 becomes:

$$\int_{-\infty}^{+\infty} \sqrt{\pi} \cdot 4.57 du$$

$$= \sqrt{\pi} \cdot 4.57 \approx 8.10$$

Well, if that integral is supposed to represent a probability, we have one of around 810%; what a book! One of the probability axioms we studied says that probabilities should add up to 1 or 100%, which suggests that we must do a bit more work on formula 3.12.

Right, so the curve represented by 3.12 is either too tall, too wide, or both. To address the sharpness and potentially reduce the height of the distribution, one effective strategy is to widen the distribution. This is where the concept of variance becomes crucial. The variance, a measure of how spread out the values in a distribution are, directly influences the width of the distribution's curve. By adjusting the variance, we can control how steep or flat the distribution appears.

The equation 3.12 represents a type of function that we covered on the second volume of this same series called exponential decay, and in our case, the variance appears in the denominator of the exponent. Because of this, as the variance increases, the exponent decreases, leading to a broader and flatter curve. Conversely, when the variance decreases, the exponent increases, resulting in a narrower and sharper curve. Therefore, to make the curve wider, we need the variance to be larger, as it directly affects the spread of the distribution:

$$f(x) = e^{-\frac{1}{2}\left(\frac{x-28.78}{4.57}\right)^2} \tag{3.19}$$

The exponent $(\frac{x-28.78}{4.57})^2$ dictates how the probability density decreases as the values move away from the mean. So, if we have it without the $\frac{1}{2}$ it will imply less density around the value of the mean, which will result in a narrow and sharper peak and a less significant tail. That fraction would damper this effect, meaning we will have a smoother and wider curve, thus balancing its shape, making it

wider and more representative of the variability inherent in the natural data. This ensures that the tails of the distribution are not unduly truncated, accurately reflecting the less likely, yet possible, occurrences of values far from the mean. OK, so our new equation is:

$$f(x) = e^{-\frac{1}{2}(\frac{x-28.78}{4.57})^2}$$

Right, integration time! We must have a probability of one, which is the same as saying the total of the instances under that pretty curve is one:

$$\int_{-\infty}^{+\infty} e^{-\frac{1}{2}(\frac{x-28.78}{4.57})^2} dx \tag{3.20}$$

Now we can do the same change of variable so we define u the same way as in 3.15:

$$\int_{-\infty}^{+\infty} e^{\frac{-u^2}{2}} 4.57 du$$

Let's get to work:

$$\int_{-\infty}^{+\infty} e^{\frac{-u^2}{2}} 4.57 du = 4.57 \int_{-\infty}^{+\infty} e^{\frac{-u^2}{2}} du \tag{3.21}$$

That integral in 3.21 is very similar to 3.18 so we just need to adapt the solution for $\sqrt{2\pi}$. This is because of the 2 we see in the exponent fraction denominator:

$$\int_{-\infty}^{+\infty} e^{\frac{-u^2}{2}} = \sqrt{2\pi}$$

Therefore 3.20 becomes:

$$\sqrt{2\pi} \cdot 4.57 = 11.45 \tag{3.22}$$

Well, shouldn't we be getting closer to 1? Instead, we are going the other way. Rest assured, although the results don't seem encouraging, we are on the right path. We have normalized the sharpness (or how tall the curve is). However, we need to gain control of the width of the distribution. To overcome this little hurdle, we can use the same approach but, this time, normalize the spread of the

distribution. If we get the result of 11.45 from 3.22 and want it to be 1, one can use the following normalization term:

$$\frac{1}{\sqrt{2\pi} \cdot 4.57}$$

That would indeed get us the 1 we need. So our function becomes:

$$f(x) = \frac{1}{\sqrt{2\pi} \cdot 4.57} e^{-\frac{1}{2}(\frac{x-28.78}{4.57})^2} \tag{3.23}$$

Let's check it out:

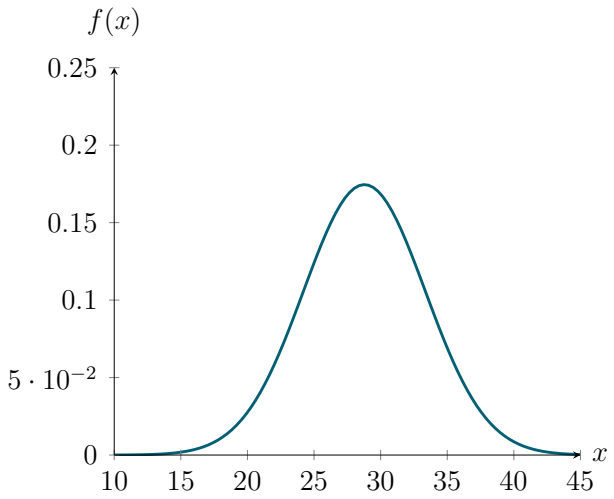


Figure 3.16: We'll go with flatter and wider; no it's not a TV, but our curve.

On graph 3.16, we have 3.23 represented by the blue line, and it looks like what we needed—a smooth, symmetrical curve. But we are not done yet. Since we have done all this work, we might as well finish with a general formula. This way, we can use it as a model for any dataset whose histogram of distribution is similar to the one we are studying:

$$f(x) = \frac{1}{s\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\bar{x}}{s})^2} \tag{3.24}$$

Would you look at that? We have letters for all the numbers on equation 3.23. This suggests that we can define an equation like 3.24 as a model for our data, given that we know two variables: the sample mean $\bar{x}$ and the sample standard deviation s. We have a name for equation 3.24: the PDF for the 'normal distribution'.

3.2.2 Is It Normal to Have Such a Distribution?

This is our first probability distribution and, happily, it is a "normal" one. The reason for this name is due to its frequent occurrence in numerous natural and social phenomena. For example, characteristics like height, blood pressure, and test scores in large populations tend to be distributed normally. You can also see this bad boy called the Gaussian distribution, in honor of Carl Friedrich Gauss, the gentleman who came up with it.

The Gaussian distribution is characterized by its bell shape curve as well as its symmetry. This particular "holy" shape represents how data points spread; most of the observations cluster around the mean, and, as you move away from the mean, the frequency of the observations decreases. This symmetry comes with some nice properties that we can explore:

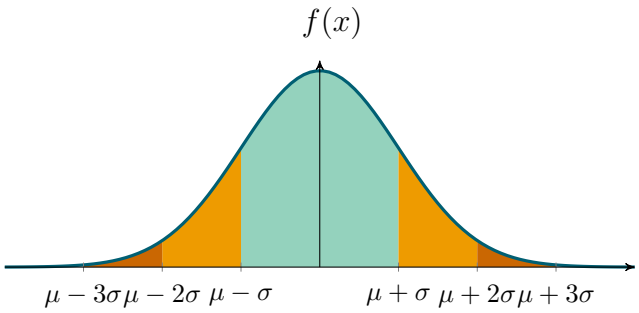


Figure 3.17: The mirror effect of the mean on our Gaussian curve

Figure 3.17 represents more than just a drawing I created for my girlfriend on Valentine's Day. (In any case, it unfortunately did not have the intended effect.) It symbolizes a normal distribution centered on the population mean μ. This distribution also includes a measure of spread, the standard deviation σ, which tells us how much a random observation is likely to deviate from the mean. The symmetrical nature of this distribution leads to some important insights. Chief among these is that a normal distribution is fully characterized by its population mean μ and population standard deviation σ. This can be succinctly represented as follows:

$$X \sim N(\mu, \sigma^2) \tag{3.25}$$

But wait a second, we define the equation 3.24 with the sample mean $\bar{x}$ and the sample standard deviation s and, now that we formally introduced the Gaussian probability density function, can we use the population mean μ and the population variance σ^2? Yeah this is true; as we can see from equation 3.25 the Gaussian distribution is defined with both μ and σ^2, so the correct PDF is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \tag{3.26}$$

When we started constructing our PDF equation, we used sample measurements. However, this shifted when we formally stated that the PDF was known as the Gaussian distribution, and we moved from the sample parameters to the population parameters. Well, we can't let this matter slide; is there a relationship between them? Why do we need both? Can one replace the other? These are all excellent questions, so we will be working hard on all those issues. We should be paying close attention to any change caused by any externalities driven by anomalies of conduct. Sound familiar? (Politics!) Don't worry; we will tackle them all. What better way to start than their definitions?

The **population mean $\boldsymbol{\mu}$** is the average value of a property across the entire population. It's an idealized figure, requiring data from every single individual in the population, which is often unfeasible. The population mean is a fixed value but typically unknown in real-world scenarios, meaning it must be inferred.

Conversely, the **sample mean $\bar{\boldsymbol{x}}$**, computed as $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ where x_i are the observed values and n is the sample size, represents an estimate of the population mean based on a subset of the population.

Both these means aim to provide an average value, but they differ in scope: the sample mean applies to a subset, while the population mean encompasses the entire population.

The same concept is applicable when comparing the sample standard deviation (s) to the population standard deviation (σ) or variance (s^2 and σ^2) . Alright, so we now understand how to distinguish between the "sample" and "population" when discussing statistical parameters.

Well, but how do they impact curve? I mean 3.25 suggests that, in order to define such curve, we need two parameters μ and σ; so what will they do to a Gaussian shape? The best way to answer these questions is with a graph:

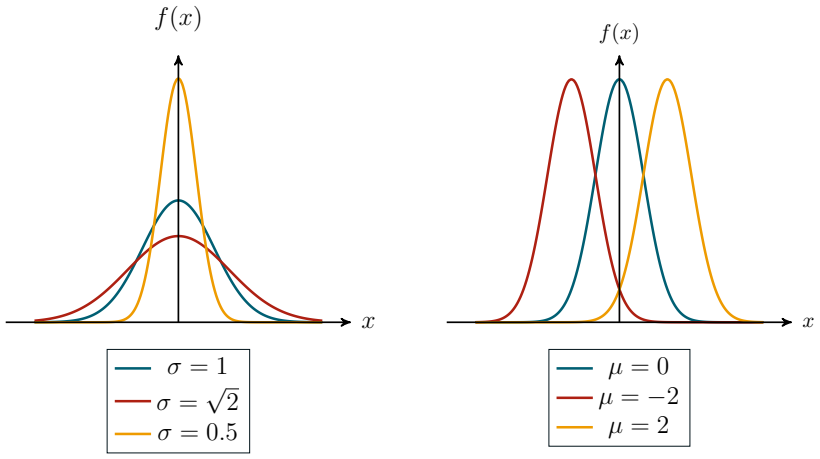


Figure 3.18: When σ spreads and μ shifts.

Figure 3.18 presents two panels of normal distribution curves to illustrate the effects of varying standard deviations and means while keeping other parameters constant. In the left panel, three normal distributions are depicted with the same mean of zero but different variances. The curves represent standard deviations of 1, $\sqrt{2}$, and 0.5, shown in blue, red, and yellow, respectively. This visualization helps in understanding how an increase in variance spreads the distribution wider and flattens its peak, whereas a decrease in variance results in a steeper, narrower peak. The right panel displays distributions with a constant standard deviation of 1 and means of 0, -2, and 2. These are also color-coded in blue, red, and yellow. This panel illustrates how the mean shifts the center of the distribution along the x-axis without altering its shape, emphasizing the role of the mean in positioning the central peak of the distribution relative to the origin.

In our deduction we ended up with equation 3.24 which is defined with sample parameters, but Gauss defined it with the population parameters. So, for our calculations to make sense, there must be some relationship between the sample mean, the population mean, the population standard deviation, and the sample standard deviation.

Let's go after it! For that, we will need to dive into a bit of abstraction. We have an equation to represent the behaviour of a particular variable called a random variable, a variable that is linked to chance. The reason we are trying to get an equation is, well, we are mathematicians, and, if we have equations, we can use a whole set of tools to do inference. These include integrals and derivatives. Going back to what we said about the population metrics, and specifically to the part where we stated these metrics are often unfeasible, I would suggest we treat μ and σ simply as parameters; if this still feels a bit foreign, consider the following analogy. In a linear regression model, we might encounter an equation like:

$$f(x) = a \cdot x + b \tag{3.27}$$

If we aim to identify the best estimates for a and b, one avenue to explore is linear regression. This methodology typically involves crafting a convex cost function dependent on both a and b and then employing optimization techniques like gradient descent to ascertain the optimal values for these parameters. This is an excellent strategy for estimating deterministic model parameters such as a and b. Now we must treat μ and σ as parameters and find a way to estimate them; however we are dealing with chance, so we must treat the estimates with the maximum likelihood of being the population's mean and standard deviation. We need statistics that maximize our chances of being correct, as our model's effectiveness hinges on the precision of these estimations.

We are about to embark on a journey to find the parameters with the highest likelihood of representing our data. The word "highest" suggests a maximization task, and to maximize anything, we need a function. The good thing is that we already have one—well, we actually just have one—and that leaves us with little wiggle room, so let's start with the PDF:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

In considering the application of this statistical approach, it is essential to clarify our objective. We are aiming to estimate parameters that best fit our data within a specified distribution, effectively measuring the likelihood that our observed dataset could occur

given our model parameters. By shifting our perspective—viewing our data as emanating from the distribution rather than fitting the distribution to our data—we can better quantify the likelihood of observing each data point within the distribution framework.

To achieve this, we must revisit the concept of the independence of events, a fundamental idea we previously set aside (the desert, remember?). Independence is critical for our analysis, particularly when considering whether observations are influenced by one another. For instance, consider whether an interval of inactivity lasting 12.3 minutes on one day influences an interval of 15.2 minutes on the subsequent day. The assumption of independence suggests that there is no such influence, allowing us to treat these observations as unrelated events in our likelihood calculations.

In probability theory, these events are called independent because the occurrence of one does not affect the probability of the other. The likelihood of both events occurring together is the product of their individual probabilities. This means if we have two independent events A and B, the probability of both occurring (their intersection) is computed as:

$$P(A \cap B) = P(A) \cdot P(B)$$

For continuous random variables, this principle extends to probability densities. If we have n independent observations $X_1, X_2, \ldots, X_n$, the probability density of observing this specific sequence is the product of the probability densities of each observation.

In our particular case, the events are independent because the likelihood of one given time interval does not affect the outcome of any other. Additionally, we are considering that they all come from the same probability distribution, meaning they are identically distributed. This leads us to conclude that we have a variable X that is i.i.d, (independent and identically distributed). Cool, so given that, for a continuous random variable, the probability density of a specific value is given by its probability density function, and considering that X is i.i.d, then to understand the likelihood of our data coming from a given model, we can multiply all the densities

to form a likelihood function L:

$$L(X) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2} \tag{3.28}$$

My eyes! My EYEEEEEESSS! Equation 3.28 is exactly why people open a mathematics book and think "NEVER!" Deep breath please; it is just a new symbol and we can make it better while explaining a new concept. Firstly let's deal with the new symbol $\prod$; it works much like $\sum$ but, instead of adding all the elements, we are multiplying them:

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + ... + x_n$$

$$\prod_{i=1}^{n} x_i = x_1 \cdot x_2 \cdot ... \cdot x_n$$

I don't like multiplications, they are hard to handle, like juggling chainsaws on a unicycle... well, maybe not that hard. Anyway we do have a good friend, something that turns multiplications into additions; this fellow was prominent in the second volume. Do you need one more hint? Fireplaces. No? The log! Remember?

$$\log_b(A \cdot B) = \log_b(A) + \log_b(B)$$

OK, so let's apply the log to both sides of equation 3.28:

$$\begin{aligned} \ln(L(X)) &= \ln\left(\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}\right) \\ &= \sum_{i=1}^{n} \ln\left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}\right) \end{aligned}$$

The ln instead of log is because we have an exponential, therefore the base e will help us simplify even further:

$$\ln(L(X)) = \sum_{i=1}^{n} \left(\ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

We can simplify it a bit more:

$$\ln(L(X)) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 \tag{3.29}$$

The likelihood function 3.29 in the context of a normal distribution, parameterized by μ and σ, is a function that quantifies the plausibility of the data given specific values of μ and σ. While it is based on the probability density function, it is not itself a probability distribution. Instead, it serves as a measure of how well the chosen parameters μ and σ explain the observed data. The interpretation of this function is straightforward: higher values of 3.29 indicate a greater probability that the observed data would occur under the specified parameters, serving as a quantitative measure of how well the model parameters fit the data.

If higher values of the likelihood function L represent a better fit of our parameters to the data, we must aim to maximize this function. To do so, we need to find the values of μ and σ that result in the maximum value of L. This process involves finding critical points where the likelihood function reaches a maximum. Critical points occur where the derivatives are zero. Since we have two parameters to estimate, μ and σ, we employ partial derivatives to determine these points. This means calculating the partial derivatives of L with respect to μ and σ, and setting them to zero. Solving these equations provides us with the estimates that maximize the likelihood function:

$$\frac{\partial}{\partial\mu}\ln(L(X)) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(-2)(x_i - \mu)$$

$$\frac{\partial}{\partial\mu}\ln(L(X)) = \sum_{i=1}^{n}\frac{(x_i - \mu)}{\sigma^2} \tag{3.30}$$

What we need to do now is check when 3.30 is zero:

$$\sum_{i=1}^{n}\frac{(X_i - \mu)}{\sigma^2} = 0 \tag{3.31}$$

As σ has no impact on these calculations we can express that expression 3.31 as:

$$\sum_{i=1}^{n} (x_i - \mu) = 0$$

That simplifies to:

$$\sum_{i=1}^{n} x_i - n\mu = 0$$

Two more steps:

$$n\mu = \sum_{i=1}^{n} x_i$$

$$\mu_{MLE} = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}$$

Would you look at that? The sample mean! So, the maximum likelihood estimate for the population mean μ seems to be the sample mean. I say "seems" because we are unsure if this point is a maximum; the only guarantee is that the sample mean is a critical point. Well, we have a good old friend that can provide us with this information — the second derivative!

$$\frac{\partial}{\partial \mu} \ln(L(X)) = \sum_{i=1}^{n} \frac{(x_i - \mu)}{\sigma^2} \tag{3.32}$$

Equation 3.32 represents the first derivative with respect to μ. Well if we are after the second derivative, we just need to derive 3.32 once more with respect to μ:

$$\frac{\partial^2}{\partial \mu^2} \ln(L(X)) = \frac{\partial}{\partial \mu} \left(\sum_{i=1}^{n} \frac{(x_i - \mu)}{\sigma^2} \right)$$

Since the summation is over n independent terms, we can differentiate each term separately:

$$\frac{\partial}{\partial \mu} \left(\frac{(x_i - \mu)}{\sigma^2} \right) = \frac{-1}{\sigma^2}$$

Therefore, the second derivative is:

$$\frac{\partial^2}{\partial \mu^2} \ln(L(X)) = \sum_{i=1}^{n} \frac{-1}{\sigma^2} = \frac{-n}{\sigma^2}$$

Thus, the second derivative of $\ln(L(X))$ with respect to μ is:

$$\frac{\partial^2}{\partial \mu^2} \ln(L(X)) = \frac{-n}{\sigma^2}$$

This result is negative, indicating that the log-likelihood function is concave with respect to μ, confirming that the critical point is a maximum. With this, we found our maximum likelihood estimation for μ; it is the sample mean. If we repeat the same process to σ:

$$\frac{\partial}{\partial \sigma} \ln(L(\mu, \sigma)) = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n} (x_i - \mu)^2 \tag{3.33}$$

Set 3.33 to zero to find the critical point:

$$0 = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n} (x_i - \mu)^2$$

Rearranging and solving for σ:

$$\frac{n}{\sigma} = \frac{1}{\sigma^3} \sum_{i=1}^{n} (x_i - \mu)^2$$

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2$$

Finally:

$$\sigma_{MLE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_{MLE})^2}$$

If we take a second derivative with respect to σ we will end up with a concave function just like with μ, meaning that we found yet again another maximum.

The conclusion of using the maximum likelihood estimation (MLE) method for fitting a normal distribution to our data is quite straightforward. To find the best values for the parameters μ and σ , we simply use their sample counterparts. This process involves constructing a likelihood function that represents the probability of observing our data under a normal distribution. By maximizing this function, we find that the optimal estimates for μ and σ are, respectively, the sample mean and the sample standard deviation. I guess that by now it comes with no surprise that the name of this method is "the maximum likelihood estimator".

3.2.3 Give Me an M, Give Me an L, Give Me an E... the Maximum Likelihood Estimator!

The concept of 'maximum likelihood estimation' (MLE) is used widely in statistics to estimate the parameters of a given probability distribution or statistical model. This method involves choosing parameter values that maximize the likelihood function, assuming these parameters are the most likely to produce the observed data. The likelihood function for a set of observations $X = (x_1, x_2, \ldots, x_n)$, given a model parameterized by θ, is defined as:

$$L(\theta; X) = \prod_{i=1}^{n} f(x_i; \theta) \tag{3.34}$$

In equation 3.34, $f(x_i; \theta)$ is the probability density function (PDF) of the observed parameters x_i, depending on whether the data is continuous or discrete (the functions for discrete data are not called density function. Don't worry we will cover this soon!). To simplify calculations, especially with products, the log-likelihood function is used:

$$\ell(\theta) = \log L(\theta; X) = \sum_{i=1}^{n} \log f(x_i; \theta)$$

The parameter θ which maximizes the log-likelihood function is found by setting the derivative of $\ell(\theta)$ with respect to θ to zero and

solving for θ:

$$\frac{\partial \ell(\theta)}{\partial \theta} = 0$$

Following we need to use the second derivative of $l(\theta)$ to confirm if the critical point is indeed a maximum. So, we now have a technique to estimate parameters for probability distributions called the maximum likelihood estimator. By the way, this is not exclusive to a normal distribution.

When deriving the normal distribution, we focused on the differences between the sample and the population mean, stating that, because we don't have access to the entirety of the data, we can't possibly calculate the population mean value, but... we have a function, and the knowledge that there is a link between the probabilities of events' integrals and their densities. This suggests that we can extract some information from a normal density function.

To understand how we can extract this, let's think about the arithmetic mean again. That rascal is calculated as a division of the sum of all values in a sample and the total number of observations. It represents the average contribution of each data point to the total sum.

That formula assumes that each data point contributes equally to the final result we call the mean, raising the question, "What if we assume otherwise?" This would mean that each parameter has a different contribution. Let's follow this rabbit hole.

If we are considering a different way to compute a mean, it does not make sense to call it the mean; let's call this parameter the 'expected value', which will give us an idea of what we can expect from a distribution. Now our focus shifts towards creating a formula that reflects such contributions.

To understand these so-called contributions, say it's your birthday and you love cake. If you want to celebrate your special day, one thing it will be necessary to know is the likelihood of me contributing to your birthday present by bringing eggs and helping bake the cake. Well, first, I need to show up and then have eggs! One way to quantify this is by multiplying the likelihood of my showing up by the number of eggs I bring. If I don't show up, there is a problem

with the eggs; showing up but bringing only one egg is still not the best contribution. This is a way we can quantify contribution. (As a sidenote, I would show up without eggs but with a big cake, so don't worry.)

So we have to multiply the probability of an event occurring by the event, yeah ... fancy names for a weighted sum. As we have a probability density function $f(x)$, this contribution will have the following shape:

$$x \cdot f(x) \tag{3.35}$$

3.2.4 Is Anything Beyond the Mean Expected to Have Any Value?

Cool so now we need to add up all those individuals of 3.35. Because we are in a continuous scenario and created a model where x can take any value from the domain, we must use integrals, the tools that allow us to compute infinite sums:

$$E[X] = \int_{-\infty}^{+\infty} x \cdot f(x)dx \tag{3.36}$$

Equation 3.36 has brought us some new notation, that $E[X]$. It stands for the 'expected value', and it is calculated by the infinite sum of the individual contributions of x to the new metric. While the mean focuses on summarizing a sample, the expected value extends this concept to probability distributions, especially continuous ones. It represents the average outcome we anticipate if an experiment (modeled by the distribution) is repeated many times. In mathematical terms, for a discrete variable (which we will cover shortly), it's a sum weighted by probabilities, and for a continuous variable, it's an integral of the variable multiplied by its probability density function.

Conceptually, we can compute the integral represented by 3.36 to find the expected value in a normal distribution. If the distribution is centered around the population mean, this implies that the population mean μ is the value we can expect on average from a

dataset that follows a normal distribution:

$$E[X] = \int_{-\infty}^{+\infty} x \cdot f(x)dx = \mu$$

It's important to note that, while the mean and the expected value can be the same for certain distributions, they are not universally interchangeable. For example, in a normal distribution, the mean and the expected value coincide, reflecting the distribution's symmetry. However, in skewed or non-standard distributions, the mean and the expected value derived from the probability distribution can differ significantly;

This is a just in case summary!

- **Mean**
  - **When:** The data distribution is symmetrical with no significant outliers.
  - **Reason:** The mean provides a true center in symmetric distributions but can be skewed by outliers.
- **Median**
  - **When**: The data is skewed, or there are significant outliers.
  - **Reason:** The median is less affected by outliers and skewed data, providing a more representative central value in such cases.
- **Mode**
  - **When:** The most frequently occurring value is of interest, especially in categorical data.
  - **Reason:** The mode identifies the most common item but may not provide insight into the data's overall distribution.
- **Expected Value**

- **When:** Dealing with probabilistic models or when outcomes have varying probabilities.
- **Reason:** The expected value accounts for the likelihood of different outcomes, providing a weighted average that is crucial in decision-making in situations of uncertainty.

So, all of a sudden, we have a new measure of central tendency, the expected value, and a new methodology to estimate parameters, called the maximum likelihood estimator, two more techniques that we can apply to various probability distributions! Even after these two successes, we must not let something slide... yes, the colors on plot 3.17. They represent relationships between the mean, the standard deviation, and the probabilities between the ranges that define the colors.

This means that we can finally use our PDF function and start computing probabilities via integration. The thing is that integration requires both an upper and a lower limit, but as we stand, the colors are bounded by generic expressions, for example, $\mu + \sigma$. Such expressions will make it hard to compute integrals, so let's define a normal distribution to allow us to attempt these calculations. What about this one?

$$X \backsim N(0, 1) \tag{3.37}$$

A normal distribution centered around the mean 0 where the variance and the standard deviation share the exact value of 1 was baptizsed as "standard normal distribution". With this, we can have a nicer looking x-axis, with some numbers instead of just Greek letters, which means, we can also define limits of integration!

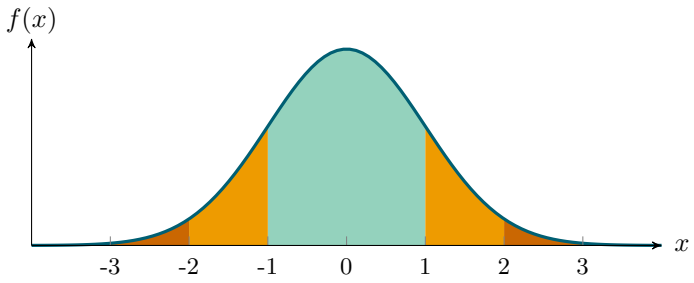


Figure 3.19: Finally something more standard.

Long live rock and roll! I mean, that graph 3.19 looks friendlier (But also, long live rock and roll). We now have integers delimiting the colored areas; for example, the ugly blue area is between -1 and 1. Now, we can integrate the PDF between each color border and get the probabilities of the colored areas. Starting with the blue:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

Because we defined our normal distribution by 3.37, to have its PDF we just need to replace μ with 0 and σ with 1

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

Now we can integrate:

$$\text{Area}_{blue} = \int_{-1}^{1} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx = 0.687$$

This result says that there is a probability of 68.7% that a given observation on a standard normal distribution is up to one standard deviation distance from the mean. Moving on to the yellow areas: We can either compute the integral between 1 and 2 to multiply it by 2, leveraging the symmetry characteristics of the distribution, or we can do it from -2 to 2. Let's go with option number 2:

$$\text{Area}_{blue+yellow} = \int_{-2}^{2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx = 0.95$$

We get a probability of 95%, which intuitively makes sense because we moved further away from the mean to both sides. Therefore, we have a bigger area of the curve covered. Lastly, we need to check what happens when we have include orange, yellow, and blue:

$$\text{Area}_{blue+yellow+orange} = \int_{-3}^{3} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx = 0.997$$

Excellent, 99.77%! Is it really that great? Especially for our heist planning? Graph 3.19 gives us a quick grasp of data distribution, particularly in the tails. For instance, if we're curious about the likelihood of encountering a time interval that's beyond three standard

deviations from the mean, a simple calculation (100%-99.7%) shows it's a mere 0.3%. This insight is crucial when considering the rarity of specific inactivity periods. Although valuable insights, the really excellent thing is that standard normal distribution represented by 3.37—that guy is a special cat.

Recall our debate over a specific time slot versus another one? We leaned on variance analysis and the probabilities of tail events because the distributions weren't centered on the same mean. This meant relying on multiple metrics to reach a conclusion. Now, imagine if we could align these distributions to a common mean for a straightforward comparison. That's where the distribution depicted by 3.37, comes into play. By scaling our datasets to have a mean of 0 and a variance of 1, we are allowing ourselves direct comparisons between datasets with different magnitudes of scale.

While centering involves subtracting the mean from each data point to align datasets along a common "center", scaling is equally crucial. We scale the data by dividing each centered value by its standard deviation. Here's why: different datasets can have varying levels of spread or variability. Without scaling, we only adjust for differences in location (mean), but not for differences in scale (variability). Dividing by the standard deviation normalizes this variability, bringing all datasets to a common scale.

This ensures that one standard deviation in any of the standardized datasets represents the same degree of variability, relative to their respective means. So we can do this with the following formula:

$$Z = \frac{x - \mu}{\sigma} \tag{3.38}$$

The outcome of 3.38 is what we call a Z-score. This is a dimensionless quantity that represents how many standard deviations a data point is from the mean. A Z-score of 1, for instance, means the data point is one standard deviation above the mean, while a Z-score of -2 means it's two standard deviations below the mean. Alright, let's bring back those two datasets and standardize them:

Day	00:00-02:00	Scaled data
1	22.88	−1.29
2	27.53	−0.27
3	27.29	−0.33
4	32.08	0.72
5	24.22	−1.00
6	33.91	1.12
7	37.47	1.90
8	19.65	−2.00
9	32.13	0.73
10	33.38	1.01
11	26.81	−0.43
12	28.01	−0.17
13	29.34	0.12
14	28.51	−0.06
15	28.45	−0.07

Table 3.6: The Z-scores for our first time slot.

Day	02:00-04:00	Scaled data
1	28.88	−0.07
2	20.44	−1.07
3	26.86	−0.31
4	32.41	0.34
5	30.5	0.12
6	32.62	0.37
7	32.41	0.34
8	34.47	0.59
9	33.95	0.52
10	34.49	0.59
11	36.16	0.78
12	37.58	0.95
13	38.32	1.04
14	11.69	−2.10
15	11.67	−2.10

Table 3.7: Some more Z-scores, now for the latest slot.

In tables 3.6 and 3.7 we have the scaled values for each timeframe, meaning that the inactivity entries are standardized, and now we can have a different angle on what time slot to pick. The first step is to visualize these scaled values:

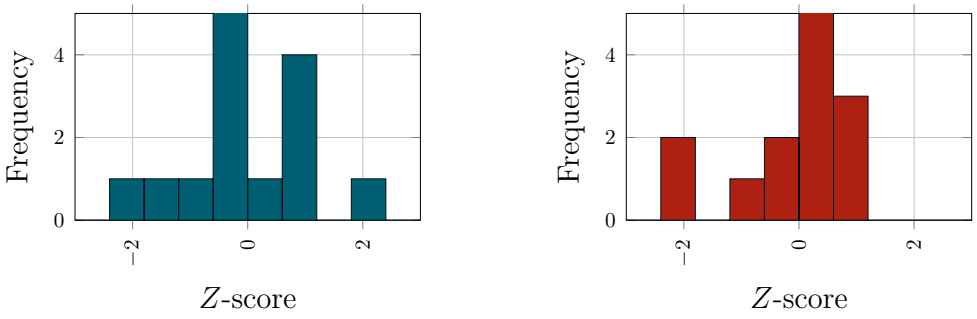


Figure 3.20: A side-by-side comparison of the Z-scores histograms.

The blue histogram represents the standardized data collected over the 00:00-02:00 period, whereas the red one depicts the same standardization for the 02:00-04:00 time slot. If we are interested in long periods of camera inactivity, we can spot a problem on the red histogram. That big red bar around the -2 scares the shit out of me. That means there is a good chance of a time interval being -2 standard deviations from the mean on the 02:00-04:00 time slot. Also, this same histogram seems more asymmetric when compared with the blue one. If we want to avoid surprises, judging by the histograms, the 00:00-02:00 time slot seems a better choice; the tails are less accentuated, and the data distribution seems more symmet-

ric. This confirms that our selection of the 00:00-02:00 for the heist was indeed the more conservative one.

We have put a lot of focus on the concept of tails and symmetry, and there is a good reason for that: we are trying to understand the likelihood of the events that are further away from the mean happening, as these are the ones that will make or break us. If these concepts are crucial, we must do better than just eyeballing graphs.

Do you have a white lab coat? Put it on, as we are about to turn scientific and deduce new formulas.

For this first one, I propose starting with something already familiar to us, the formula for variance. The reason for selecting the variance as a measure is because it provides information about the average squared deviation of the observations from the mean, which indicates the behavior of the distribution's tails. To understand the variance for a continuous random variable, consider that the expected value is given by:

$$E[X] = \mu$$

The variance measures the dispersion of the random variable around its mean, specifically, it is the expected value of the squared deviation from the mean. For a continuous random variable x with mean μ, variance is defined as:

$$\sigma^2 = E[(X - \mu)^2] \tag{3.39}$$

To express σ^2 in terms of an integral, we apply the definition of variance to the expectation. Given that the expected value of X is:

$$E[X] = \int_{-\infty}^{+\infty} x \cdot f(x)\, dx$$

Then, by analogous reasoning, the variance can be represented as:

$$\sigma^2 = \int_{-\infty}^{+\infty} (x - \mu)^2 \cdot f(x)\, dx \tag{3.40}$$

This formulation uses the integral to average the squared deviations of x from the mean μ, weighted by the probability density

function $f(x)$, thus quantifying the spread of the distribution around the mean. More generally if we have a function $g(x)$, then:

$$E[g(X)] = \int_{-\infty}^{+\infty} g(x) \cdot f(x)\, dx \tag{3.41}$$

In this equation 3.41 we transform the random variable x with a function $g(x)$. For $g(x)$, we can have any function we desire, but in our specific case:

$$g(x) = (x - \mu)^2$$

To calculate $E[g(X)]$, we integrate the product of $g(x)$ and the probability density function $f(x)$ over the entire range of x. This integral effectively accumulates the values of $g(x)$, each weighted by the probability of x occurring, as specified by its probability density $f(x)$.

With equation 3.41, we defined one of the six properties of the expected value we will study. This particular one, we call the expectation of a function. What follows might come across as boring, but the reality is that these six properties will be of immense help; every expectation of a continuous variable comes with an integral, and these guys can be tedious to calculate. If we have six properties we can leverage to manipulate expressions of expected values without touching integrals, I am willing to explore them, but you still don't look convinced. Well, neither is the judge in my court case. (I stole some oranges from a tree and am alleging I have a vitamin C deficiency). Let me build a case for you; for example let's expand 3.39:

$$\sigma^2 = E[X^2 - 2X\mu + \mu^2] \tag{3.42}$$

Go ahead and try to integrate that. Yeah, it is a massive integral. But I believe this was the right catalyst with which to grab your attention and focus you on the utility of these properties. In a perfect world, we could compute that expected value in parts, as this would simplify the integral calculations. Let's compute the expected value of those guys individually and see what comes out of that. Starting with $2X\mu$:

$$E[2\mu X] = \int_{-\infty}^{+\infty} 2\mu x \cdot f(x) dx$$

Because 2μ is a constant it follows that:

$$E[2\mu X] = 2\mu \int_{-\infty}^{+\infty} x \cdot f(x)dx = 2\mu E[X]$$

More generally:

$$E[cX] = cE[X] \quad \text{Where } c \in \mathbb{R} \tag{3.43}$$

We have stumbled upon another one of the six missing properties. Equation 3.43 is our second one. We can't use this result to simplify 3.42 yet because, to do so, we would have to split 3.42 into a sum of expectations, which we don't know how to do yet. However, considering we were on the hunt for these properties anyway, this is a step forward. I feel we may be on a winning streak, so let's check what happens to that fellow $E[\mu^2]$:

$$E[\mu^2] = \mu^2$$

What can we expect from a constant? Well it has to be that same constant, so:

$$E[c] = c \quad \text{Where } c \in \mathbb{R} \tag{3.44}$$

Yeah, three out of six! We found another property, and that is positive, but still, no progress toward simplifying 3.42. I will not lie; I knew these two properties wouldn't help much, but they were there for grabs. Now, let's get to one that will help us. We must break that subtraction inside the expectation 3.42; therefore, why not check what happens if we have two random variables X and Y along with constants a and b and wish to compute the following expectation?

$$E[aX + bY]$$

Summing or adding will produce the same result in terms of property, so let's expand the expectation of X and Y with integrals:

$$E[X] = \int_{-\infty}^{+\infty} x \cdot f_X(x)\, dx$$
$$E[Y] = \int_{-\infty}^{+\infty} y \cdot f_Y(y)\, dy$$

Cool, now let's create a relation between X and Y. We will call it Z; this guy is a linear combination with the following equation:

$$Z = aX + bY$$

Hey Z what can we expect from you?

$$\begin{aligned} E[Z] &= E[aX + bY] \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (ax + by) \cdot f_{X,Y}(x, y)\, dx\, dy, \end{aligned}$$

Expanding and separating the integral, we get:

$$\begin{aligned} E[aX + bY] = &\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} ax \cdot f_{X,Y}(x, y)\, dx\, dy \\ + &\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} by \cdot f_{X,Y}(x, y)\, dx\, dy. \end{aligned}$$

If we assume that X and Y are independent, it means that $f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y)$:

$$\begin{aligned} E[aX + bY] &= a \int_{-\infty}^{+\infty} x \cdot f_X(x)\, dx + b \int_{-\infty}^{+\infty} y \cdot f_Y(y)\, dy \\ &= aE[X] + bE[Y] \end{aligned}$$

So:

$$E[aX + bY] = aE[X] + bE[Y] \tag{3.45}$$

Before going any further, I was hoping you could give me the pleasure of announcing four out of six! Equation 3.45 is another property. This one we can leverage and take one step further in our quest for simplification, so:

$$\sigma^2 = E[X^2 - 2X\mu + \mu^2] = E[X^2] - E[2X\mu] + E[\mu^2]$$

Well, well, well... look what we can use now. After all, 3.43 and 3.44 will come in handy:

$$\sigma^2 = E[X^2] - 2\mu E[X] + \mu^2$$

We know that the expected value is the same as the population mean, so $E[X] = \mu$:

$$\sigma^2 = E[X^2] - 2\mu^2 + \mu^2 = E[X^2] - \mu^2$$

Finally:

$$\sigma^2 = E[X^2] - E[X]^2 \tag{3.46}$$

Equation 3.46 represents a way to express the variance as a function of expectations. We can't forget that we still have some rules to cover:

- For any two random variables X and Y:

$$E[X + Y] = E[X] + E[Y]$$

- If X and Y are independent random variables:

$$E[XY] = E[X] \cdot E[Y]$$

The proofs for those two remaining properties arise directly from the expansion of the formula with integrals; they are very similar to what we did with 3.45. Before moving on, there is a term we must talk about, that guy $f_{X,Y}(x, y)$. This is what we call a joint probability distribution.

A joint probability distribution for continuous random variables X and Y is described by a probability density function, denoted as $f_{X,Y}(x, y)$. This function does not provide probabilities directly but instead gives the density at a particular point (x, y). The actual probability that X and Y take values within specific ranges is found by integrating the PDF over those ranges. The joint PDF helps us understand how the values of X and Y are related and allows us to analyze the probabilities of different combinations of X and Y occurring together.

For example, say that we have decided on our roles for the heist. You will crack the safe and will be turning off the security system. If we defined your task as X and my task as Y, we can create a joint probability distribution; we will be able to understand not

just the individual chances of success for me and you but how their successes or failures are related, offering a complete picture of the possible outcomes for their heist.

We will cover this concept in depth later on in the book. For now, having an idea of the expression will be enough for it not to be a mystery.

Alright, we've covered six properties and introduced a new formula for variance, that handy measure that helps us understand how far away our observations tend to wander from the mean. Remember, we initially ventured into this territory because we were curious about asymmetry and the behavior of distribution tails.

A crucial aspect we're keen on unraveling is the direction of this asymmetry. Is our distribution leaning left or right? One intuitive way to figure this out would be to have a sign of skewness: a negative value for left-skewed distributions and a positive one for right-leaning distributions.

Here's the little snag: the square in our variance formula masks the direction of our skew. It treats deviations on either side of the mean equally. So, what do we do? We have two options: tone that exponent of 2 down to 1 or crank it up to 3. Opting for the cubic exponent, we get a formula highlighting those more significant deviations than a mere exponent of 1 would. So, let's take the plunge and go with 3, introducing a crucial metric in our statistical toolkit:

$$\gamma_1 = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] \tag{3.47}$$

We are nearly there! Because our data is already standardized, formula 3.47 for our particular case becomes:

$$\gamma_1 = E[(X-\mu)^3] \tag{3.48}$$

Because we are operating with Z-scores, we have $\mu = 0$ and $\sigma = 1$, so 3.48 becomes:

$$\gamma_1 = E\left[X^3\right] \tag{3.49}$$

We can expand 3.49:

$$\gamma_1 = \int_{-\infty}^{+\infty} x^3 f(x) dx$$

Good, we will use a computer to calculate that integral, for the blue and red histograms. This way, we can get to those skewness values:

$$\gamma_{1,\text{00:00-02:00}} = -0.0985, \quad \gamma_{1,\text{02:00-04:00}} = -1.366$$

- **00:00-02:00 Slots (Blue Histogram):** That value indicates a slight leftward skew, but it's close to zero, suggesting the distribution is almost symmetrical.
- **02:00-04:00 Slots (Red Histogram):** On the other hand, the skewness of the time slot of 02:00-04:00 indicates a more pronounced leftward tilt, suggesting the distribution has a longer tail on the left side.

Damn, mathematics works. These results confirm what we saw on the plot; that big red bar around the -2 made us reticent to pick that time slot, and the skewness formula 3.48 has now confirmed our suspicions.

3.2.5 Left, Right, Maybe Out of Sight: Skewness.

Skewness is a measure of the asymmetry of a probability distribution. It indicates whether and how much a distribution deviates from a symmetric bell curve.

- **Positive Skew (Right-skewed)**: In a positively skewed distribution, the right tail (higher values) is longer or fatter than the left tail. The mean and median will be greater than the mode. This type of distribution indicates a large number of lower values with a few higher outliers.

- **Negative Skew (Left-skewed)**: In a negatively skewed distribution, the left tail (lower values) is longer or fatter. The mean and median will be less than the mode. This suggests a large number of higher values with a few lower outliers.

- **No Skew (Symmetric)**: If the skewness measure is close to zero, the distribution is considered to have no skew, indicating that the tails on either side of the mean are balanced. This does not imply perfect symmetry or that the distribution resembles a normal distribution, as symmetric distributions can still exhibit characteristics like uneven tail lengths that do not affect skewness.

What a metric! Although powerful and insightful, the skewness left me wondering what would happen if we, for example, raised the variance formula to the power of 4. If a cubic created a formula with such utility, why not explore what happens with a power of 4?

$$\gamma_2 = E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] \tag{3.50}$$

By raising the deviation of each observation from the mean to the fourth power and normalizing it by the fourth power of the standard deviation, the statistic γ_2 highlights the impact of extreme values. This effect is more pronounced than in variance or skewness calculations, as values far from the mean receive disproportionately greater weight when raised to the fourth power. Consequently, a high value of γ_2 indicates a distribution with "heavier" tails, suggesting a higher likelihood of extreme values. While higher kurtosis is sometimes associated with a "sharper peak," this interpretation can be misleading. In reality, higher kurtosis primarily reflects the presence of more extreme outliers rather than a sharper peak or more clustering near the mean. Therefore, a high kurtosis value indicates greater occurrences of outliers, contributing to the heavier tails of the distribution.

3.2.6 Flat, Fat, and Fancy: The Kurtosis Style.

What about that? Should we keep going for 5, 6, and 7 exponents? Don't worry, we can stop here. So let's give this new fellow a name. This metric is called 'kurtosis' and, unlike variance, which measures general variability, or skewness, which assesses asymmetry, kurtosis is unique in its focus on the tails and the peak of the distribution, offering a nuanced view of the distribution's shape and the potential for extreme values within the dataset. Alright, let's see what we have in our data:

$$\gamma_2 = E\left[X^4\right] \tag{3.51}$$

As we have the data standardized, the kurtosis formula simplifies to 3.51 and now we can integrate:

$$\gamma_2 = \int_{-\infty}^{+\infty} x^4 f(x) dx$$

So, the kurtosis for our time slots is:

$$\gamma_{2,\text{00:00-02:00}} = 2.75, \quad \gamma_{2,\text{02:00-04:00}} = 3.28$$

- **00:00-02:00 Slots (Blue Histogram):** A kurtosis value of 2.75 for the 00:00-02:00 points towards a distribution that might be slightly less peaked than that of the 02:00-04:00. This suggests that inactivity periods during 02:00-04:00 are distributed in a way that extreme values (very long or very short periods of inactivity) are less pronounced, leading to a more even spread across the observed range.

- **02:00-04:00 Slots (Red Histogram):** The kurtosis value of 3.28 for the time interval 02:00-04:00 suggests a distribution with a relatively higher peak compared to the interval 00:00-02:00. This indicates a greater concentration of inactivity periods near the mean, combined with more pronounced occurrences of extreme values in the 02:00-04:00 interval than in the earlier period.

From these kurtosis values, we can infer that inactivity periods during the 02:00-04:00 timeframe have a distribution that's slightly more prone to exhibiting outliers or extreme values compared to a frame of 00:00-02:00.

Conversely, the slightly lower kurtosis for the 00:00-02:00 times indicates a broader spread of inactivity periods, with extremes being less common or less pronounced. This might suggest a more uniform pattern of inactivity throughout this same time slot, without the pronounced peaks and heavy tails indicative of outlier-dominated periods.

Well, we already said that skewness and kurtosis represent distinct measures that help us to understand the shape of a distribution; however there is one more thing we need to check. With skewness we have a highly intuitive measurement; a negatively skewed distribution has a longer left tail, and a positively skewed one extends more on the right. Understanding skewness is straightforward as it directly reflects the direction of the distribution's asymmetry. However, with the k fellow things are not that simple.

Because we are working with a normal distribution, kurtosis might initially seem to present a straightforward measurement of tail heaviness. However, this perception changes when we apply kurtosis to other distributions. Unlike the normal distribution, which is symmetrical and has a moderate level of tail weight, other distributions can exhibit more complex tail behaviors. In these cases, kurtosis doesn't simply measure tail weight but also the propensity for extreme values, which can significantly affect our interpretation of data, especially in risk-sensitive environments. This means that merely knowing that a distribution is leptokurtic (high kurtosis) or platykurtic (low kurtosis) is insufficient; we also need to come up with a frame of reference from which we can draw comparisons.

3.2.6.1 Probably Too Much of a Good Thing: Excess Kurtosis.

If we know the normal distribution so well, why don't we compare our values of kurtosis to the one for the normal distribution?

This concept is called 'excess kurtosis'. It is a statistical measure that quantifies how the tails of a given distribution compare to the tails of a normal distribution, which has a kurtosis of three. By subtracting three from the standard kurtosis of the distribution, excess kurtosis effectively sets the baseline at zero for a normal distribution. This adjustment simplifies comparisons: a positive excess kurtosis indicates a distribution that is leptokurtic, featuring heavier tails than a normal distribution and suggesting a higher likelihood of outliers. Conversely, a negative excess kurtosis describes a platykurtic distribution with lighter tails, implying fewer extreme values than a normal distribution:

$$\text{Excess Kurtosis} = \text{Kurtosis} - 3$$

Everything is pointing toward robbing those diamonds and between 00:00 and 02:00! Before moving on with further analysis, to be as sure as possible of which time slot to pick, there is a name for these three measures we just studied. Well, the name actually incorporates four measures. There is nothing to worry about as the missing fellow is a metric very familiar to us; it is the mean. The three remaining are: the variance, the skewness and the kurtosis. These four guys could form a rock and roll band called "The Variance Vanguards", but they also constitute the moments of a distribution.

3.2.7 Some Moments to Remember.

The term "moments" in statistics is derived from the concept of moments in physics, specifically from mechanics, which describes the distribution of mass in space and its resulting properties. In statistics, "moments" provide a mathematical way to tell the shape of a probability distribution, including its location, spread, skewness, and kurtosis, among other characteristics.

The formal definition of the n^{th} moment of a probability distribution about a point a is given by:

$$E[(X - a)^n] \tag{3.52}$$

Where X is a random variable, and E denotes the expected value. Now we must check that a individual and what values can it take. In theory, a can be any real value, but there are two choices:

- $\boldsymbol{a = 0}$: Moments about zero, also known as raw moments, are calculated without any shift. These moments provide insights into the overall scale of the distribution.
- $\boldsymbol{a = \mu}$ **(the mean of the distribution)**: Moments about the mean are called central moments. These are more commonly used in statistical analyses because they provide information about the distribution relative to its center.

We focused on the central moments of a probability distribution; the ones defined relative to the mean of the distribution. So for a random variable X with mean μ the n^{th} moment is defined by:

$$E[(X - \mu)^n]$$

The four central moments of a probability distribution for a random variable X with a mean of $\mu = E[X]$ are defined as follows:

1. **First Central Moment $\boldsymbol{\mu_1}$**:

   $$\mu_1 = E[(X - \mu)^1]$$

   However, since $E[X] = \mu$, the first central moment is always 0, as it essentially measures the deviation of the mean from itself.

2. **Second Central Moment $\boldsymbol{\mu_2}$**:

   $$\mu_2 = E[(X - \mu)^2]$$

   The variance quantifies the dispersion of the distribution around its mean.

3. **Third Central Moment $\boldsymbol{\mu_3}$**:

   $$\mu_3 = E[(X - \mu)^3]$$

The third central moment is related to the skewness of the distribution, indicating its asymmetry around the mean. Skewness is usually normalized by the cube of the standard deviation (σ^3).

4. **Fourth Central Moment μ_4**:

$$\mu_4 = E[(X - \mu)^4]$$

The fourth central moment relates to the kurtosis of the distribution, reflecting the "tailedness".

By simply observing the shape of the distribution depicted by our histogram, we could deduce quite a bit of mathematics. Then it came the temptation to create a function that was too big; we must always try to leverage our toolkit of mathematical techniques, so we came up with the concept of the probability density function. This is a precious form of mapping that receives observations as inputs and maps them to densities. Having this framework allows us to leverage techniques such as derivatives and integrals. Speaking about derivatives, these came into play immediately when we realized the probability density functions are not more than models to fit our data. Models come with parameters, so we required techniques to estimate them to maximize the likelihood of them representing our data in the best way possible. We introduced the notation of sample mean vs. population mean and used integration to conclude that the expected value of a probability distribution is the population mean.

Finally, we engaged on a journey to study the peaks and tails of the distribution and ended up nearly stumbling over the four moments of centrality. If we gather all the information from the concepts above, we are equipped with the knowledge we need to help us select the timeframe we wish to perform the heist. The sample we gathered on the 12-2 AM timeframe has a higher expected value, less variance, lower skewness, and a lower kurtosis value. However, because I would like to avoid jail time at all costs, there is something else we could do to solidify our decision to go on and rob those stones between 12 and 2 AM. Let's think about the expected value and what it means to us. It approximates the population means, a

key statistic for our decision; we always refer everything back to it. But approximation is synonymous with error and, therefore, risk, so what can we do to minimize it?

One strategy is to collect more data samples, but this time, let's get them only for the 00:00-02:00 period. So we spoke with our hacker guy and asked for 16 samples of 15 observations each for periods of inactivity:

Sample No.	1	2	3	4	5	6	7	8
Sample Mean	28.82	27.54	28.53	28.92	28.98	28.69	28.36	30.01

Sample No.	9	10	11	12	13	14	15	16
Sample Mean	28.08	28.61	29.58	30.52	27.94	30.21	29.20	29.50

Table 3.8: A sample of sample means. That sounds weird but better than a mean of mean samples, which just sounds mean.

Alright, let's create a histogram:

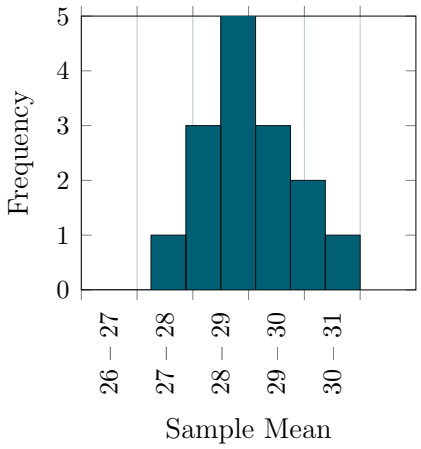


Figure 3.21: Is it giving us the finger?

The histogram 3.21 represents the behavior of our sample of sample means from table 3.8 and, I'll be damned! That bell again?!?! I guess the name normal makes sense. But this is not a result to leave untouched, so let's explore this finding. First and foremost, we consider the random variable to be the sample mean. This concept may seem a bit strange, but it will soon turn into a powerful resource, because, if we can have random variables for parameters, we can learn a lot about them and therefore make better-educated guesses. (I wish I had some similar techniques at my fingertips when I decided to drive my car after being at the bar and this nice policeman stopped me. Big, big mistake!)

Now, it appears that our X, representing the sample means, adheres to a normal distribution. With data from 16 samples, each with 15 observations, we've computed 16 distinct sample means. Calculating these means, a fundamental measure of central tendency, inherently minimizes the likelihood of encountering extreme values, unlike what we might observe in individual samples. Consequently, these means tend to cluster around a central value, effectively, the "mean of the means".

3.3 Is it Central to have some Limit(s) When talking About Theorem(s)?

Note that we haven't specified the distribution of our samples. While we're assuming a normal distribution for simplicity, we've only mentioned that the sample means tend to follow a Gaussian distribution, seemingly independent of the actual distribution of the samples. This point warrants further discussion, and indeed, it is covered by the 'Central Limit Theorem' (CLT). The CLT asserts that the sample means will approximate a normal distribution regardless of the underlying distribution, provided that the variables are independent and identically distributed (i.i.d). This is the classic interpretation of the CLT, which we will focus on in our study. There are other versions of the theorem that relax the i.i.d requirement, but we will not cover those.

Additionally, there is a critical condition concerning sample size that we must address soon. But first, let's explore the parameters for the mean and the variance of the normal distribution of the sample means. As we know, a sample mean is represented by:

$$X = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{3.53}$$

In equation 3.53, we denote X as the mean of a single sample. Applying this formula to our specific case, we will calculate 16 sample means, each derived from a different set of observations. The individual observations within a sample are represented by x_i, and

for each sample in our analysis, we have 15 such observations. However, if we extend this concept to a sample of samples, we use:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{3.54}$$

Where each X_i is itself a sample mean, representing the average of observations within that particular sample. This formulation allows us to analyze the overall behavior across multiple samples, where each sample mean X_i is treated as a random variable derived from its respective observations. If we are after a statistic that more accurately represents our population, and we're focused on a mean that characterizes the normal distribution of sample means, then we can calculate the expected value of 3.54:

$$E[\bar{X}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i]$$

Because we know that $E[x_i] = \mu$ for each i:

$$E[\bar{X}] = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

So, we have our first parameter of the distribution of the sample means, μ is represented by... well μ, so:

$$\bar{X} \sim N(\mu, ?)$$

We still need to find what will represent the variance of the distribution of the sample means. The strategy will be the same, but now we will compute the variance of 3.54:

$$\text{Var}\left(\bar{X}\right) = \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \tag{3.55}$$

Hmm, we are stuck here; we have a constant being multiplied by a sum of random variables. However, this is not our first rodeo, so we will come up with something that will allow us to carry on, and I do have an idea! What happens with the following?

$$\text{Var}(aX)$$

Where a is a constant and X is a random variable. I know that it might seem like I am digging an even bigger hole, but give me a chance here. Memory refresher of the day: "Variance" (which might seem like the setup for another poem, but it is just a formula):

$$\text{Var}(X) = E[(X - E[X])^2]$$

Right, let's squeeze that a in there:

$$\text{Var}(aX) = E[(aX - E[aX])^2]$$

With this, we can work baby! Given the expected value $E[aX] = aE[X]$:

$$\text{Var}(aX) = E[(aX - aE[X])^2]$$

We can factor out the constant a from the square, remembering that when we take a constant out of a square, it gets squared:

$$\text{Var}(aX) = E[a^2(X - E[X])^2]$$

Using the property that the expected value of a constant times a random variable is the constant times the expected value of the random variable, we get:

$$\text{Var}(aX) = a^2 E[(X - E[X])^2]$$

Recognizing that the expression inside the expected value is the variance of X, we arrive at the scaling property:

$$\text{Var}(aX) = a^2 \,\text{Var}(X)$$

With this result we can turn 3.55 into:

$$\text{Var}\left(\bar{X}\right) = \frac{1}{n^2}\text{Var}\left(\sum_{i=1}^{n} X_i\right) \tag{3.56}$$

We are very close. Now we must deal with that summation on the right. Let's start simple, and try to figure out what will be the result of this:

$$\text{Var}\left(X + Y\right)$$

We know that for a given random variable Z, we can express its variance as:

$$\text{Var}(Z) = E[Z^2] - (E[Z])^2$$

If we apply the definition to $X + Y$, where both X and Y are independent:

$$\text{Var}(X+Y) = E[(X+Y)^2] - (E[X+Y])^2$$

Let's expand $E[(X+Y)^2]$:

$$E[(X+Y)^2] = E[X^2 + 2XY + Y^2] \ = E[X^2] + 2E[XY] + E[Y^2]$$

With $E[X+Y]$ we are capable of making progress:

$$E[X+Y] = E[X] + E[Y]$$

$$(E[X+Y])^2 = (E[X] + E[Y])^2 = (E[X])^2 + 2E[X]E[Y] + (E[Y])^2$$

Finally;

$$\text{Var}(X+Y) = \big(E[X^2] + 2E[XY] + E[Y^2]\big) - \big((E[X])^2 + 2E[X]E[Y] + (E[Y])^2\big)$$

Given X and Y are independent, $E[XY] = E[X]E[Y]$. Therefore:

$$\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y)$$

Which also means that:

$$\text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}(X_i)$$

This result applied to 3.56 gives us:

$$\text{Var}\left(X\right) = \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}(X_i)$$

Now, that summation of variances can be replaced by $n\sigma^2$:

$$\text{Var}\left(X\right) = \frac{1}{n^2}\cdot n\sigma^2 = \frac{\sigma^2}{n}$$

And there we have it! If $\bar{X}$ represents the samples means we say that $\bar{X}$ follows a normal distribution with mean μ and variance $\frac{\sigma^2}{n}$:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \tag{3.57}$$

Let's do an intuitive verification of these results. The mean parameter indicates that if you were to take an infinite number of samples from a population and calculate each sample's mean, the average of these sample means would be equal to the population mean. This underscores the unbiased nature of the sample mean as an estimator of the population mean.

Now, for the variance we see that it indicates how much the sample mean is expected to vary from the true population mean. A smaller variance suggests that the sample mean is a more accurate estimate of the population mean. The variance decreases as the sample size increases, which indicates that larger samples tend to produce more precise estimates of the population mean. Therefore the value of $\frac{\sigma^2}{n}$ seems like a good choice:

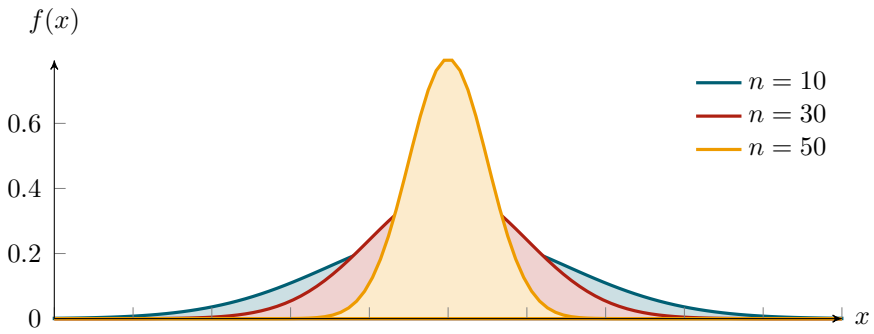


Figure 3.22: A cascade of normal distributions representing the sample of sample means.

In figure 3.22, we have represented three different examples of the distribution of the sample means. The n represents the sample size, and we have three distinct values: 10, 30, and 50, corresponding to the blue, red, and yellow curves. We can observe that, as n increases, the distribution becomes more narrow, indicating a smaller standard deviation value, which makes sense. The more sample means we can observe, the better the estimate of this parameter will be; consequently, we will have less variance.

When dealing with the sample mean distribution, we call the standard deviation the standard error of the mean (SEM). This measures how much the sample mean of $\bar{X}$ is expected to vary from the

population mean μ. We can represent this cowboy by:

$$\text{SEM} = \frac{\sigma}{\sqrt{n}} \tag{3.58}$$

Alright, it seems we have everything we need to define our normal distribution representing the sample means. For n we see that it is equal to 15 as each of the 16 samples has 15 observations. Now we need to compute the sample variance and the sample means of our data present in table 3.8:

$$s^2 = 0.695$$

$$\frac{\sigma^2}{n} \approx \frac{s^2}{n} = \frac{0.695}{15} = 0.046$$

For the sample mean of the sample means:

$$\bar{\mu} = 28.97$$

To get to those parameters estimations we just applied the formulas of the sample variance and the sample mean to the 15 entries of the table 3.8; with this, we can define our normal distribution for the sample means as:

$$\bar{X} \sim N(28.97, 0.046) \tag{3.59}$$

When considering only one sample of the inactivity intervals for the 00:00-02:00 slot, we had an expected value of 28.78. Given that we want reassurance for our time slot decision, we went back and asked for 16 more samples of 15 observations each. After that, we retrieved the sample mean of each of those 16 samples and created a histogram with the sample means. This shape was identical to a bell curve, so we went on to compute the parameters of this normal distribution and ended up defining the probability density function for a random variable $\bar{X}$ representing the sample means; this is the equation 3.57. We then computed, the parameters μ and $\frac{\sigma^2}{n}$ to finally end up with the values in 3.59.

The motivation behind all of this was to try and get a better estimate of the population mean, and the expected value changed slightly when we compared the one calculated with only one sample,

28.78, with the one obtained with the distribution of the sample means, 28.97.

One thing that leaves me slightly on edge is that we have predominantly dealt with estimates, and we know these fellows incorporate randomness and, consequently, errors. For example, we just defined a parameter called standard error. If errors and randomness are a concern, one way to deal with such occurrences would be to create a measure of confidence. If we have a way to quantify an error given by the standard error, we could generate an estimate of an interval instead of a point. In this context, it would be a range of inactivity values instead of just one, allowing us to make less risky decisions. And on top of that, if we can quantify how confident we are on this interval, that would be the icing on the cake, the cherry on top, the woogie boogie morly noogie... you got my point.

3.3.1 Make It to Go and in Brackets: an Order with Confidence and Intervals.

We aim to define a quantifiable range that gives us enough confidence to make better-informed decisions. That sounded a bit complex, right? Let's break it down into simpler terms. Since we're focusing on understanding the mean of a dataset and knowing that estimations aren't perfect (errors exist), it wouldn't be wise to rely on a single fixed value for the mean. Instead, we should consider a range of possible values, which is where the concept of a confidence interval comes in.

So, we create a range of values for the mean, underpinned by a confidence level. This isn't just any "range; it's one that is backed by a probability measure, say, a "95% confidence interval." What this means is that we can be 95% certain that the true mean of our data falls within this range. By calculating this interval, we end up with a range where, statistically, there is a 95% chance that the true mean is captured within it. This approach helps us to accommodate the natural variability in data and reduces the risk of surprises from those inevitable small variations. As a little side note, the reason behind the 95% confidence is just due to normal convenience. Other

commonly used confidence levels are 99% and 90%.

I foresee the need for at least two things to start this endeavor: eggs and butter. Oh, my bad, that was the cake recipe. What we actually need is the parameters we want to bound around an interval; this is the sample mean of the sample means and a measure of the width for the interval. So the standard error will be a good candidate:

$$\text{CI} = \bar{X} \pm \frac{\sigma}{\sqrt{n}} \tag{3.60}$$

That 3.60 is a nice simple formula, but it is missing something crucial: we have no parameter from which to incorporate a confidence level associated with the interval. It implies an assumption that the population mean will lie within this range but without a statistical basis for how confident we can be in this assertion.

Our interval is static, offering a fixed view that lacks the dynamism required to encapsulate the variability inherent in statistical data. What would be great is if we had a mechanism that scales our interval to reflect different levels of confidence, integrates a measure of likelihood, and incorporates the critical aspect of standard deviation. Sound familiar? The Z-score!

By its very design, the Z-score measures the number of standard deviations a data point is from the mean, providing a precise method with which to adjust our confidence interval dynamically. This brings to the table exactly what we were missing—a way to quantify our confidence level with an attached likelihood rooted in the distribution of our data, allowing us to say, with statistical backing, how confident we are that our interval accurately captures the true mean. Let's bring back that colorful plot:

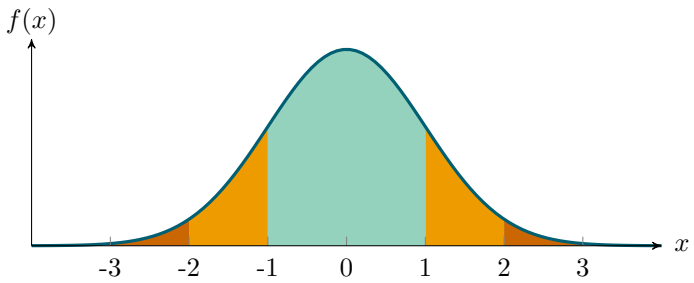


Figure 3.23: The standard normal distribution with some make up.

When we first tackled the subject, we approximated the areas under the curve of the standard normal distribution, $N(0, 1)$, leading to a simplified understanding that the combined areas of the blue and orange regions, captured by integrating the probability density function (PDF) from -2 to 2, accounted for approximately 95% of the total probability. This simplification led us to round the critical Z-score to ± 2, even though the precise value is ± 1.96.

This approximation helped us get an intuitive grasp of how probabilities are distributed under the curve of a standard normal distribution. By integrating the PDF, we can visualize and estimate the probability encompassed within a certain range around the mean. However, this method, while insightful for initial understanding, has limitations when precision is paramount, especially in statistical analysis where exact confidence levels are crucial.

Now, our exploration takes a nuanced turn as we seek not just an approximation, but the exact Z-score that delineates a specific probability, such as the 95% confidence level. To articulate this quest formally, we examine the probability:

$$P(-z \leq X \leq z) = 0.95 \tag{3.61}$$

Granted that we happened to know the z's for the likelihood of 95%, but what if we are after 93%? Or 71%? Our task now is reversed, with the expectation that we are looking for the probability, given an interval; now, we are trying to identify which interval has a certain probability value. Mathematically, this will mean we have to find the limits of integration, so that the result of that specific integral is the desired confidence:

$$\int_b^a f(x)dx = 0.71 \tag{3.62}$$

Now we must find a and b for that equation 3.62? I have to apologize. It is currently 6:15 am, and I am in no mood to solve that problem, so allow me to guide us toward a simpler path that will give us another wrench in our statistical toolkit. Let's refer back to 3.61; since simplicity is the word of the day, we will analyze half of that formula:

$$P(X \leq z) \tag{3.63}$$

So we see that:

$$P(X \leq z) = \int_{-\infty}^{z} f(x)dx \tag{3.64}$$

Given we still have an integral to calculate and are leaving behind the $P(z \leq X)$, this strategy doesn't seem to be going well. The crux of the issue lies in our reliance on the PDF, where probability is conveyed through the area under the curve, with densities mapped along the y-axis rather than direct probabilities. This representation necessitates using integrals to translate these areas into probabilities, which, while mathematically sound, is only sometimes the most intuitive approach for probability calculations.

So, what is the next step when an essential piece is missing? Exactly; we have to come up with a solution! Ideally, we would have a function where we input probabilities, which would return the z scores corresponding to this mass. Getting such mapping would involve computing the inverse function of the PDF, which is a complexity nightmare; however, that 'inverse' word gave me an idea; this will be one of those situations where we take one step backward to then, take two forward. With 3.63, we are looking for the mass of probability up until some value z, so why don't we create a function that does this for all the z's?

That function won't directly provide us with what we are looking for, but, if we have a framework where we input a z and get a probability, the inverse will allow us to input a probability mass and get a z in return, so answering questions such as 3.63 will be possible. It seems like we have a well-defined strategy, and, for a better chance of success, let's kick things off with a visualization:

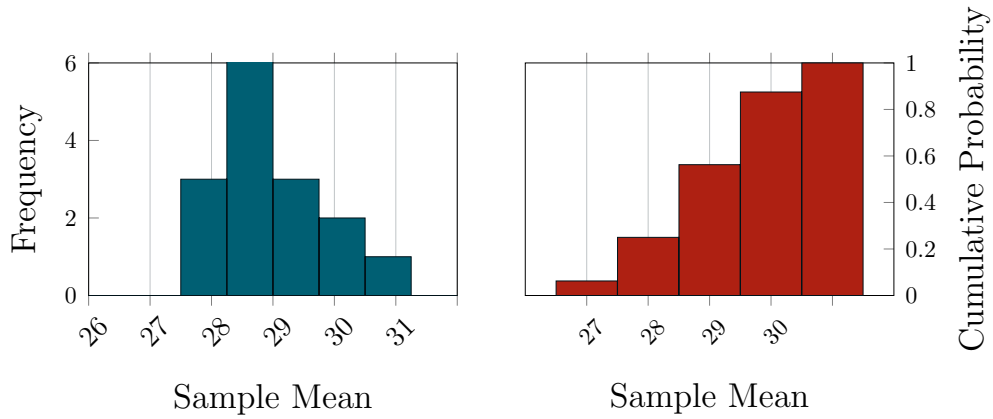


Figure 3.24: A PDF vs a CDF, what a duel!

On graph 3.24, in the left corner, dressed in an elegant shade of blue, the sample means "golden right jab" histogram. (Sorry. My fond memories of watching boxing with my grandfather might have sneakily danced their way into that explanation.) The left hand figure with the blue bars is the histogram of the sample means; it is the same representation as 3.21.

Moving to the right, we encounter a plot that maps out the cumulative probabilities, marked by red bars. Unlike the histogram, this graph doesn't just count occurrences; it accumulates them. For each point on the x-axis, the corresponding y-value on this plot doesn't just tell us about the likelihood of that specific value, but it also includes the probabilities of all preceding values. The easiest way to interpret this red-barred plot is to think of it as the result of stacking the blue bars of the histogram as we move from the left to the right. Because the blue bars are used to compute likelihood, if we add them up in sequence, we will have to end up with a bar of size one, representing the total probability of the sample space:

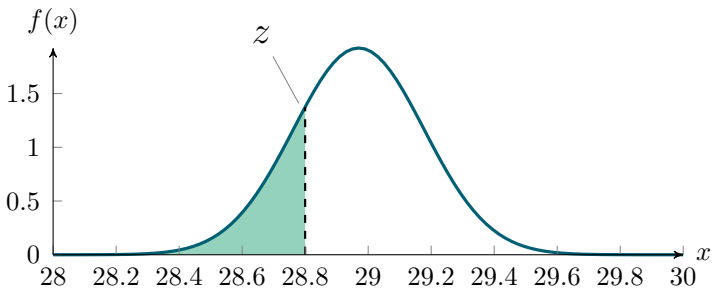


Figure 3.25: How much until a point?

In Figure 3.25, we depict a probability density function that assigns a density value $f(x)$ to an input x. The blue area corresponds to the probability we are aiming to obtain, as outlined in equation 3.63. Our objective is to preserve the representation along the x-axis while transitioning the y-axis from representing density to representing probability. To achieve this, we introduce a new function whose outputs are the cumulative probabilities, equivalent to the areas under the curve within specified intervals:

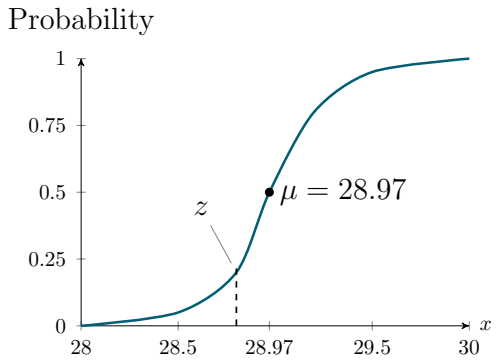


Figure 3.26: Our first sight of a CDF curve.

Figure 3.26 represents exactly what we need, a direct mapping between inputs $\bar{X}$ and probabilities y's. The equation and the name?

3.3.1.1 From Zero to One, Building Confidence: The Cumulative Density Function.

Allow me to introduce you to the "cumulative distribution function" (CDF). This is a fundamental concept in probability and statistics that describes the probability that a real-valued random variable X will take on a value less than or equal to x. It is a function $F(x)$ associated with every random variable X, and it is defined for all x in the real numbers. We actually have seen the equation for this function before:

$$F(x) = P(X \leq x) \tag{3.65}$$

Cool, but we are after the following probability:

$$P(-z \leq \bar{X} \leq z)$$

Let's represent it on the CDF plot:

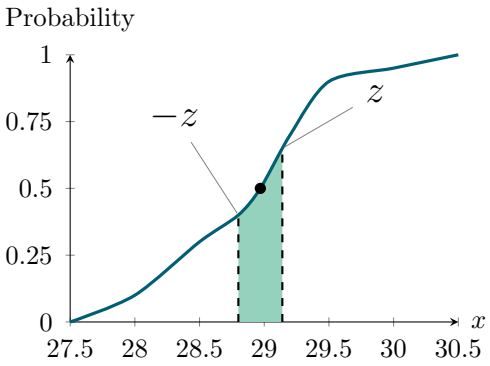


Figure 3.27: The desired interval of z's.

In graph 3.27, the area between the two dashed lines is bounded by the z-values we are looking for. This visually represents the probability that X lies between $-z$ and z. Essentially, we are determining the proportion of the probability distribution encompassed within these bounds.

Because the CDF gives us the probability that X is less than or equal to x, $F(z)$ will return the probability from the start of the distribution all the way up to z. For the same reason, $F(-z)$ will provide the probability from $-\infty$ to $-z$; therefore, to find the probability strictly between $-z$ and z we can do so with the following subtraction:

$$P(-z \leq X \leq z) = F(z) - F(-z)$$

We are halfway there, now we need to understand how to compute the inverse of the CDF. Before, you suggested that I take a walk because we now have computers and can do anything we want regarding calculations; back in the day, we relied on paper tables listing CDF probabilities for various significance levels. So we needed this function.

What about some applications of these functions? Remember the percentiles? The median, which is the 50^{th} percentile, is the same as $F(x) = 0.5$. Even for our heist, we could minimize risk by computing the cumulative probability that a given time interval of inactivity is lower than a specific time. Before sorting out the formula of the confidence intervals, let me list some properties of the CDF:

1. **Non-decreasing**: $F(x)$ is non-decreasing, which means for any two numbers a and b such that $a \leq b$, we have $F(a) \leq F(b)$. This is mathematically represented as:

$$a \leq b \Rightarrow F(a) \leq F(b)$$

2. **Right-continuous**: $F(x)$ is right-continuous. For every x, $F(x)$ equals the limit of $F(x + \varepsilon)$ as ε approaches 0 from the right:

$$\lim_{\varepsilon \to 0^+} F(x + \varepsilon) = F(x)$$

3. **Limits at Infinity**: $F(x)$ approaches 0 as x approaches negative infinity and 1 as x approaches positive infinity:

$$\lim_{x \to -\infty} F(x) = 0 \quad \text{and} \quad \lim_{x \to \infty} F(x) = 1$$

4. **Interval Probability**: The probability that X lies within an interval $(a, b]$ is given by the difference between the CDF values at b and a:

$$P(a < X \leq b) = F(b) - F(a)$$

5. **Total Probability**: The total probability captured by the CDF is 1, as it encompasses the entire probability distribution of X:

$$F(\infty) = 1$$

A moment to do a point of the situation:

- **Cumulative Distribution Function (CDF)**: $F(x)$, of a random variable X is defined as the probability that X will take a value less than or equal to x, i.e.,

$$F(x) = P(X \leq x).$$

What about a family photograph of the two sisters CDF and PDF so we can remember the good times later?

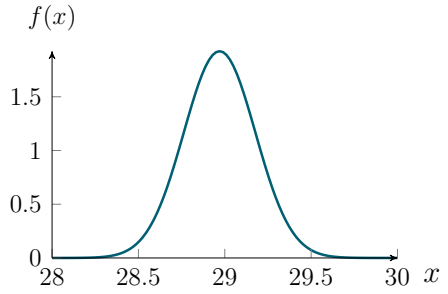


Figure 3.28: The family dinner photo of the PDF

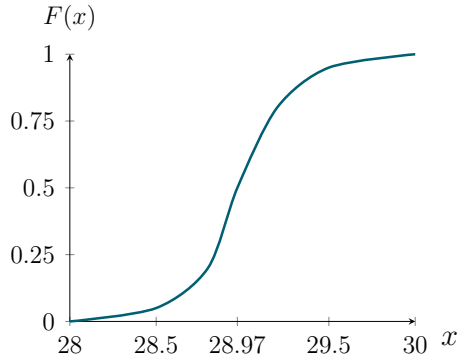


Figure 3.29: The family dinner photo of the CDF

These functions are so helpful that we could even say the right plot, the one representing the CDF would be an excellent candidate to represent the distance travelled on our morning bike ride ... wait... In fact, let's take a closer look into these fellows and assume that 3.29 represents our bike ride where x is the distance and y the velocity. Referencing figure 3.29, focus on the middle part of the graph. This segment represents the fastest part of our bike ride, where we covered a greater distance in less time compared to the other segments. It's like pedaling downhill: quick and covering lots of ground.

Now, if we shift our attention to the left hand graph in figure 3.28, notice that this corresponds to where the PDF peaks. This peak indicates the highest density or frequency of data points-similar to our fastest pedaling speed during the ride. Essentially, the PDF can be thought of as showing our velocity at different points in time, with higher peaks indicating higher speeds. The behavior in the tails of both graphs further supports our observations. In these areas, the PDF is lower, corresponding to slower speeds on our ride, just as the slope of the CDF becomes less steep, indicating less distance travelled during these slower intervals.

Understanding these graphs in the context of our bike ride helps to clarify their statistical significance: the PDF shows how frequently specific speeds (or data points) occur, while the CDF indicates the accumulation of these distances (or data points) over the course of the ride. If we recall our study of calculus, the derivatives are tools we use to represent velocities, so if we assume that

the PDF $f(x)$ is some "velocity" metric:

$$f(x) = \frac{dF(x)}{dx} \tag{3.66}$$

The formula 3.66 illustrates a fundamental concept in calculus: the derivative of a function at any point describes the slope of the tangent line to the function at that point. In the context of the CDF, $F(x)$, and the PDF, $f(x)$, this relationship becomes particularly revealing. The CDF $F(x)$ represents the cumulative probability up to a specific value x, akin to plotting the total distance traveled over time. The derivative of this function then represents the rate at which this cumulative probability changes with respect to x, essentially, the slope of the CDF at any point x. This slope tells us how densely packed the probabilities are around x, which directly corresponds to the PDF $f(x)$.

Even though the cumulative density function, or if you prefer its rap name CDF, was just introduced to us, I feel we are already on good terms with it. We understand that it will allow us to have a one-to-one relationship between inputs and probabilities with the caveat that these likelihoods are returned in probability masses. We also presented the relationship between the CDF and the PDF and all the properties of the CDF.

If you recall, before this odyssey, we were constructing a formula to provide us with a confidence interval for the mean of our population of sample means, and we got stuck because we needed a term that could represent our confidence level. With this, we explored the z's distribution and concluded it was an excellent candidate to provide us with this flexibility, so now let's incorporate it into our formula:

$$\text{CI} = \bar{X} \pm z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \tag{3.67}$$

When we multiply $z_{\frac{\alpha}{2}}$ by the standard error of the mean, $\frac{\sigma}{\sqrt{n}}$, we are essentially scaling the variability (standard error) of the sample mean to a level that reflects the desired confidence interval. This multiplication determines how far from the sample mean ($\bar{X}$) we must go on either side to achieve our specified level of confidence in estimating the population mean. The $z_{\frac{\alpha}{2}}$ value adjusts this range

to ensure that it encompasses the central portion of the distribution corresponding to our confidence level, effectively capturing the uncertainty in the estimate of the mean due to the sample variability. Essentially, the z-score controls how many standard deviations from the mean we are willing to accept in both directions, and we can adjust this using α, the significance level. Wait a minute. Who said anything about α's? Don't panic; the α in equation 3.67 represents the level of significance, which is essentially the complement of our confidence level within the context of probability. To illustrate, if our confidence level is 95%, then the level of significance, α, is calculated as the difference between 100% and the confidence level:

$$100\% - 95\% = 5\%$$

This means that α is 5%, indicating that we are allowing a 5% probability that our confidence interval does not contain the true parameter value due to sampling variability. But we don't even have α; we have $\frac{\alpha}{2}$. This relates directly to the structure of the normal distribution, which is symmetrical about its mean. By constructing a confidence interval for the mean, we are aiming to capture the probability distribution's central portion, representing our confidence level.

The normal distribution's symmetry implies that, to maintain our specified level of confidence centered around the mean, we must allocate the error rate α equally on both sides of the distribution. This ensures that the interval is balanced and accurately reflects our confidence level. Visually:

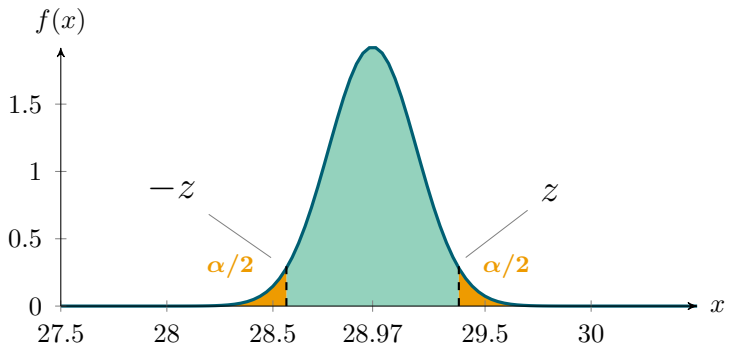


Figure 3.30: Squashed by two z's.

Figure 3.30 visually interprets why we need half of the α on the confidence interval formula 3.67. The area shaded in light blue delineates the central 95% of the probability mass, symmetrically distributed around the mean. Given that the total error margin is 5%, and to maintain the symmetry of the distribution and center the interval around the mean, this error is evenly divided between the two tails of the distribution. Consequently, each tail embodies a 2.5% segment, visually clarifying why we allocate $\frac{\alpha}{2}$ to each end in our calculations, the yellow regions.

We are ready to calculate something! First, let's collect the parameters we have already defined that are required for us to have a confidence interval:

$$\bar{x} = 28.97, \quad s^2 = 0.695, \quad n = 16, \quad \alpha = 0.05$$

The objective is to compute the 95% confidence interval for the mean using the formula:

$$\text{CI} = \bar{x} \pm z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \tag{3.68}$$

Oh wait a second, did your spot something? It seems we are not quite ready yet. Equation 3.68 asks for the population standard deviation and we don't have it. Granted we don't have the population mean either but that is OK as this is what we are building the confidence interval on. Well a good estimate for the population standard deviation is its sample counterpart s. OK, let's try that:

$$\text{CI} = \bar{x} \pm z_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} \tag{3.69}$$

However, because s comes from a sample, and samples differ from each other, substituting s for σ introduces more uncertainty into our confidence interval calculation. This additional uncertainty primarily affects the distribution's tails, making them heavier or more pronounced than those of the standard normal distribution. Therefore our z in the formula 3.69 is no longer adequate. The correct formula is therefore:

$$\text{CI} = \bar{x} \pm t_{\frac{\alpha}{2},n-1} \cdot \frac{s}{\sqrt{n}} \tag{3.70}$$

There is a new sheriff in town (what a bad expression given the context!). That t is representative of a new distribution called t distribution.

Before jumping into the properties of the t distribution let's clarify something, just in case. We started mixing in concepts of scores and distribution, namely Z-scores and Z distributions and now we introduced a new distribution t that will also have a t-score. Right, so a Z-score is a metric we use to standardize data, to make if centered around 0 so we can compare samples with different magnitudes. Now this Z-score follows a Gaussian distribution with a mean of 0 and variance 1:

$$Z = \frac{X - \mu}{\sigma}, \quad Z \backsim N(0, 1)$$

Similarly, in the context of the t distribution, the t-score (or t statistic) is used to assess how far a sample statistic (such as the sample mean) is from a hypothesized population parameter, expressed in terms of the standard error. The t-score is especially useful when the population standard deviation is unknown and must be estimated from the sample. The formula for the t-score resembles that of the Z-score but substitutes the sample standard deviation for the population standard deviation:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}, \quad t \sim t_{(n-1)} \tag{3.71}$$

The primary difference between the Z and t-score,—and by extension, the Z and t distributions, lies in their application and underlying assumptions. The Z-score and Z distribution assume that the population standard deviation is known and that the data follows a normal distribution. By contrast, the t-score and t distribution are used when the population standard deviation is unknown and is estimated using the sample standard deviation.

3.3.1.2 Too Much Variance In a Single Distribution.

The t distribution shares some similarities with the normal distribution; both are symmetrical and bell-shaped, reflecting the central

tendency of data. However, a distinctive feature of the t distribution is its heavier tails compared to the normal distribution. This characteristic implies a greater likelihood of encountering data points that are significantly distant from the mean. Such a property is particularly relevant when dealing with small sample sizes, as it accounts for the increased uncertainty inherent in estimating population parameters from these samples.

Upon revisiting the formula 3.71 for the t-statistic, we can observe that t follows a t distribution with a parameter $n-1$. This bad boy represents the degrees of freedom associated with our statistical analysis. We touched upon this subject when studying the sample variance, remember? We used it to adjust for bias in the estimation process, ensuring our sample variance is an unbiased estimator of the population variance. Similarly, in the t distribution, degrees of freedom serve to fine-tune the distribution's shape, accommodating the variability expected from estimating the population mean using a sample, which means that it will impact the shape of our distribution. Would you like to see how the t distribution looks? Me too!

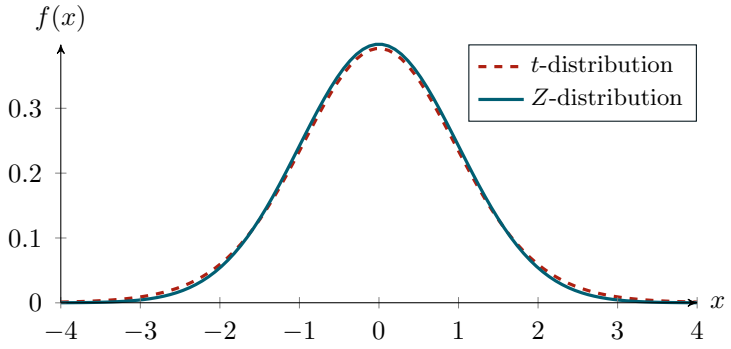


Figure 3.31: I am bit heavier on the tails, says the t to the Z.

In figure 3.31, we have the standard normal distribution Z dressed in blue, whereas the red dahsed line depicts a t distribution of 15 degrees of freedom (our $n = 16$, so $n - 1 = 15$). We can see that the dashed red curve has wider tails, better enabling it to capture events with less likelihood of happening. Now, it is true that the t distribution plot came from a PDF; however, this time, we will skip the deduction, as the calculations for it are very extensive, complex, and outside the book's scope. But, we will study the density equation

and, therefore the distribution:

$$f(x) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{v\pi}\,\Gamma\left(\frac{v}{2}\right)}\left(1+\frac{x^2}{v}\right)^{-\frac{v+1}{2}} \tag{3.72}$$

Where:

- v represents the degrees of freedom,
- Γ is the gamma function,
- and π is a portion of apple pie... No, it is the mathematical constant Pi.

There we have it, our PDF equation for the t distribtuion is that handsome fellow 3.72. From the parameters described above, v, representing the degrees of freedom is everywhere on that equation 3.72, so let's check its impact by creating three different plots from three t distributions, each with a different value for degrees of freedom:

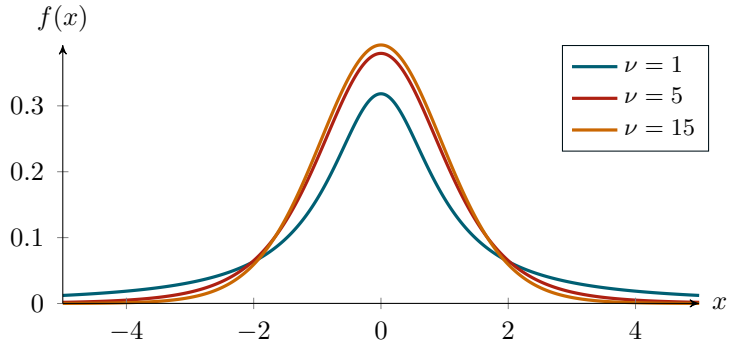


Figure 3.32: I will get normal if you keep increasing my degrees of freedom.

In graph 3.32, we have three curves, each coming from a t distribution with a different number for degrees of freedom. The blue fellow has only one degree of freedom, whereas the red line has a value of five, considering the same parameter. Finally, we have the orange curve, representing the t distribution with 15 degrees of freedom(the ones that fits our data).

Moreover, in plot 3.32, a pattern emerges where the width of the tails is inversely related to the degrees of freedom. The blue

curve, representing the lowest value of ν, has the widest tails compared to the red and orange curves, which have higher degrees of freedom. Consistent with this trend, the red distribution—having more degrees of freedom than the orange—also exhibits wider tails.

Calculating degrees of freedom is tied to sample size, resulting in wider tails for smaller samples. This reflects increased uncertainty, which is a reasonable assumption given the limited information available from smaller samples. This characteristic is exactly why we choose to follow up with the t distribution. Another pattern is that, as the degrees of freedom increase, the t distribution becomes more peaked, gradually resembling the standard normal distribution. This convergence is explained by the law of large numbers, which states that, with more degrees of freedom or larger sample sizes, the sample variance more accurately estimates the population variance, thereby reducing the dispersion of the t distribution.

Now, we need to understand how to incorporate this distribution into our confidence interval formula. The first thing we need to verify is around the significance level $\alpha = 0.05$. This fellow indicates the total probability of error we're willing to accept, corresponding to the areas outside the confidence interval in both the t distribution tails. Before, we had this represented by the critical z-scores; now, we must do the same analysis to find the necessary t-scores that bounds the central 95% of the probability distribution. The tactic will be the same as before; we split α evenly across the two tails of the distribution, resulting in $\frac{\alpha}{2} = 0.025$ for each tail:

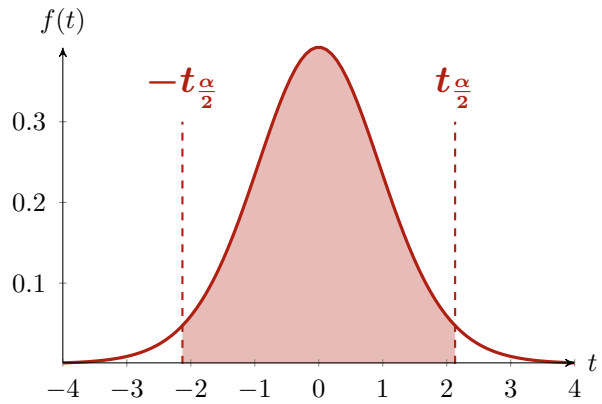


Figure 3.33: We will be very confident if you land in the shaded area.

The red shaded area in figure 3.33 represents a probability mass of 95% when we consider $\alpha = 0.05$ in a t distribution. Because our goal is to introduce the t distribution on the confidence interval formula, the next step is no surprise; we need an equation representing that same red-shaded area. What about starting with what we know? Because of the symmetry of the t distribution and the value of $\alpha = 0.05$, we know that the tails of the distribution, the white area under the curve, can be mathematically represented by:

$$P(T \leq -t_{\frac{\alpha}{2},\text{df}}) = \frac{\alpha}{2} \quad \text{and} \quad P(T \geq t_{\frac{\alpha}{2},\text{df}}) = \frac{\alpha}{2}$$

Well, we are dealing with a PDF, which means that the area under the curve always equals one. Therefore, the same area is equal to the sum of the red area with the two white areas of the tails; if we know the white areas' value, we just need to subtract it from 1 to get the red area:

$$P(-t_{\frac{\alpha}{2},\text{df}} \leq T \leq t_{\frac{\alpha}{2},\text{df}}) = 1 - \alpha = 0.95 \tag{3.73}$$

Equation 3.73 represents the red shaded which also the complement of the white one. This prompts the question: "What values of $-t_{\frac{\alpha}{2},\text{df}}$ and $t_{\frac{\alpha}{2},\text{df}}$ are needed such that the likelihood of t falling between these values is 95%?" Essentially, it seeks to determine the critical t-values that encapsulate the central 95% of the t distribution, leaving 5% in the tails (2.5% in each tail).

To answer such a question, we are back in a situation where we need some machinery that relates a variable, in this case t, with probabilities. Our statistical toolkit has two major players that can help us: the probability density function and the cumulative density function. The spotlight here is on the CDF because, unlike the PDF, which serves up probability densities, the CDF deals directly with probabilities. With the CDF, we input a z or a t and get a probability mass back. However, we wish to do the inverse, to input a probability mass and get a z or a t returned.

Well, the term 'inverse' solves our problem because, if the CDF takes a t-value and provides us with a probability, its inverse will receive a probability and spit out the corresponding t-value:

$$F(t) = P(T \leq t)$$

Then its inverse, denoted as $F^{-1}(p)$ for a given probability p, is the value of t for which

$$P(T \leq t) = p \tag{3.74}$$

This defined implicitly by:

$$F^{-1}(p) = t \quad \text{such that} \quad P(T \leq t) = p \tag{3.75}$$

3.3.1.3 Flipping Perspectives: The Inverse of the Quantile Function.

For any p within the range [0, 1], the inverse CDF, denoted as $F^{-1}(p)$, finds the corresponding t-value where the cumulative probability up to t is equal to p. This is particularly useful for finding percentiles or quantiles within a distribution, hence the function is commonly referred to as the "Quantile Function".

Now, we're ready to set up our confidence interval. Specifically, we need the limits of our interval that are represented by $-t_{\frac{\alpha}{2},\text{df}}$ and $t_{\frac{\alpha}{2},\text{df}}$. For this, we will use our quantile function:

$$t_{\frac{\alpha}{2},\text{df}} = F^{-1}_{t,\text{df}}\left(1 - \frac{\alpha}{2}\right)$$

If we input our values $\text{df} = 15$ and $\alpha = 0.05$:

$$t_{\frac{0.05}{2},15} = F^{-1}_{t,15}\left(1 - \frac{0.05}{2}\right) = F^{-1}_{t,15}(0.975)$$

$$t_{\frac{0.05}{2},15} = 2.131$$

This critical value, $t_{\frac{0.05}{2},15}$, can be determined using software or or (in the old days) consulting statistical tables. It represents the t-score at which the upper 2.5% of the distribution begins, and due to the symmetry of the t-distribution, the lower 2.5% ends at the negative of this value, $-t_{\frac{0.05}{2},15}$. Using these values, we insert them into our confidence interval formula to calculate the interval boundaries:

$$\text{CI} = 28.97 \pm 2.131 \cdot \frac{0.834}{\sqrt{16}}$$

Meaning that 95% confidence interval is:

$$\text{CI} = (28.97 - 0.444, 28.97 + 0.444) = [28.526, 29.414]$$

In our study, we talk about a 95% confidence interval for the mean of sample means. We can picture this as a range where we're 95% sure that the true average of the entire population falls. It's like making a bet where you're pretty confident you're right 95% of the time. This interval doesn't tell us where individual observations or even individual sample means might end up. Instead, it's about the big picture — estimating the overall average (the population mean) based on the samples we've taken. Think of it this way: if we keep taking more samples in exactly the same way and calculate this interval each time, about 95 out of 100 of these intervals will actually have the true population mean in them.

Upon establishing a clear understanding of what our confidence interval represents, it's imperative to revisit the assumptions that underpin its validity:

- **Independence**: The inactivity time interval is independent meaning that one observation does not impact any other one. This was concluded before, so this one is verified.
- **The data is normal distributed**: Well, this was nearly a leap of faith; we eyeballed a histogram and jumped to the conclusion that the data was normally distributed. Not only that, but we were even bold enough to announce the central limit theorem.

Confidence intervals are a powerful tool, providing us with a range of values from which we can make inferences about the population mean. However, they are not the only method we have to assess the likelihood of a particular value for the population mean. For instance, we might ask, "Is it possible that the population mean is 30 minutes of inactivity time?". To answer such questions, we need more than just a range—we need a structured approach to test specific claims. So we can explore this concept of testing, join me in hypothesizing, and since this is our original hypothesis, let's call it H_0:

- **Original Hypothesis H_0**: The population mean is equal to some hypothesized value. In this particular case, we are testing for 30 minutes of inactivity time:

$$H_0 : \mu = 30 \tag{3.76}$$

Equation 3.76 defines our original hypothesis, denoted by H_0, which states that the population mean $\mu = 30$. We can also refer to H_0 as the "null hypothesis".

When conducting a test, it is essential to have a frame of reference or a basis for comparison. Therefore, alongside the null hypothesis H_0, we should define an alternative hypothesis, denoted as H_1:

- **Alternative Hypothesis H_1**: The population mean differs from the hypothesized value:

$$H_1 : \mu \neq 30 \tag{3.77}$$

With H_0 and H_1 defined, we cover both potential scenarios, meaning either the population mean equals 30 minutes of inactivity time or this value won't happen. Now, it's time to choose a side. We need to find a scientific method to determine whether it's plausible for the population mean of our inactivity time to be 30 minutes. Remember, we are ensuring we can reach that diamond center between 00:00 and 02:00—those enticing diamonds are waiting! To resolve the debate between H_0 and H_1, because we are in the world of mathematics, we can't solely rely on other people's opinions; we need to create a suitable metric or statistic that can decisively indicate which hypothesis holds more robust merit.

Do you feel the excitement? Do you sense the anticipation? Yes, it's deduction time! Let's dive into defining our test statistic. Consider this: our calculated sample mean is 28.97. We're now faced with whether it's feasible for this metric to reflect a population mean of 30 at a given point. If we continue collecting data, will our sample mean inch closer to that 30 mark? Our test statistic must provide information on whether a deviation of this magnitude is within possibility. To quantify this, let's begin by setting up a

metric that effectively captures this potential shift:

$$\bar{x} - \mu \tag{3.78}$$

Leaving the equation as $\bar{x} - \mu$ in 3.78 presents a challenge: it doesn't consider the variability within our dataset. This raw difference alone isn't sufficient to determine whether the observed deviation is significant or negligible. The significance of a deviation can vary greatly depending on the spread of the data. To address this, we normalize the difference by dividing it by the standard error of the mean. This step scales the deviation, allowing us to assess the variability of our sample and providing a clearer picture of its significance:

$$\frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \tag{3.79}$$

Or if we are working the sample variance:

$$\frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \tag{3.80}$$

Yes indeed! We've circled back to the t-score (3.80). These statistic is handy and precisely what we need for this analysis. Given that we are working with the sample variance, it's appropriate to compute the t-score or t-statistic for our test:

$$t_{obs} = \frac{28.97 - 30}{\frac{0.833}{\sqrt{16}}} = -4.95$$

Indeed, -4.95 is quite a significant number—whether it's aesthetically pleasing is up for debate. The crucial question is how we use this number to assess the plausibility of H_0. If -4.95 represents the deviation we want to test, one might wonder how likely a value of this magnitude could occur. Fortunately, we can quantify this! We know that our t statistic follows a t distribution. Let's explore this avenue with the hope of determining the likelihood of observing a value as extreme as -4.95. Given that $n = 16$, our t distribution is characterized by $n - 1 = 15$ degrees of freedom:

$$t \sim t_{15}$$

Let's take a moment to assess where we stand. Our goal is to determine the plausibility of the population mean for inactivity time being 30 minutes. To achieve this, we have formulated two competing hypotheses, H_0 and H_1. These hypotheses help us evaluate the likelihood of observing a deviation from the sample mean of this specific magnitude (assuming that H_0 is true). After defining our test statistic to quantify such a deviation, our next task is to understand its behavior. Fortunately, we know that our test statistic T, with a value of $t_{obs} = -4.95$, follows a t-distribution. But what insights does this distribution provide? Primarily, it helps us gauge the frequency of observing t_{obs}-values—like our current $t_{obs} = -4.95$, our observed t.

Imagine I tell you that you'll win 100 dollars every time it rains, but you have to pay me the same amount if it doesn't. What's the first thing you'd do? You'd likely check the weather forecast to see how often it rains in your area. The more it rains, the better your chances of winning. Here, the situation is similar: instead of predicting rain, we're looking at how often we'd observe a t value like -4.95 under the assumption that H_0 is true (meaning the population means is 30). Instead of checking the weather forecast, we refer to the t distributions. Because this distribution assumes that H_0 is true, meaning that the population mean is 30, we used this exact value in the calculations of our t-score; we call it this distribution that assumes that H_0 is true, the 'null distribution'.

Fantastic, we've established a test statistic and identified the corresponding null distribution to gauge the frequency of observing such a statistic. But what are we missing? A referee, of course! Consider this: what if we find that $t_{obs} = -4.95$ occurs only 10% of the time? Or perhaps 7%? Fortunately, we have the perfect arbitrator to resolve this debate—our trusty α. By setting a significance level α, we not only decide whether to reject H_0 or don't have enough evidence to do so, but we also control the risk we're willing to take. In essence, α helps us define the boundaries of our acceptance and rejection regions.

Speaking of which, how should we structure these acceptance and rejection regions? Do we distribute α equally between the two

tails like we did with confidence intervals, or allocate it to just one side? Considering that our test aims to ascertain whether $\mu = 30$ or $\mu \neq 30$, the direction of the deviation doesn't particularly matter. Thus, we'll handle it much like a confidence interval, splitting α equally between the two tails. And what should α be? Let's set it at 0.05. With this in place, the values of t that delineate our acceptance and rejection criteria are already determined:

$$t_{0.025,15} = 2.131, \quad -t_{0.025,15} = -2.131$$

Visually:

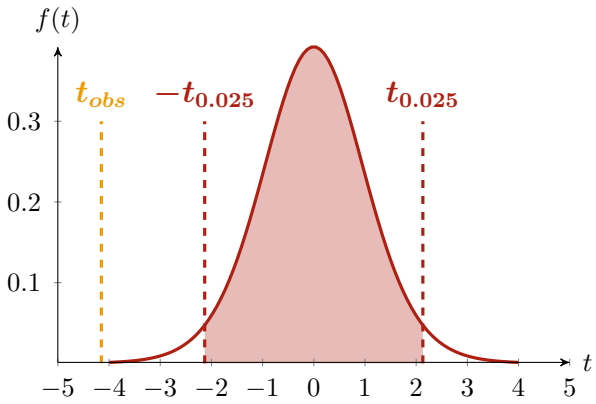


Figure 3.34: If you land in the red, we'll take you instead!

In Figure 3.34, we have represented the t distribution (null distribution) with 15 degrees of freedom, which our test statistic follows. The red shaded area, bounded by the dashed lines, represents the region where we fail to reject H_0, indicating that it is plausible for the population mean to be 30. Conversely, if our t-value, represented by the yellow dashed line, falls outside this red region, we don't have enough evidence to fail to reject H_0, and we must reject the null hypothesis in favor of H_1. Analytically, our criteria for acceptance and rejection are as follows:

- If $-2.131 \leq t_{obs} \leq 2.131$, we **fail to reject** the null hypothesis H_0.
- If $t_{obs} < -2.131$ or $t_{obs} > 2.131$, we **reject** the null hypothesis H_0.

Ermmmmm, isn't that 3.34 the same as 3.33, the one we used to construct our confidence interval? Yes, it is! Both feature the red-shaded region representing the range of values where we wouldn't reject the null hypothesis, aligning with the 95% confidence interval for the population means. This overlap isn't just a fluke; it stems from both the confidence interval and the hypothesis testing relying on the same distribution and significance level, facilitating a unified acceptance. Well, this relationship between the confidence interval and the t-statistics was great; I mean, because we had the critical values for our t's, we could create an acceptance rejection region for our null hypothesis distribution. Still, if you recall, we introduce the null distribution with the argument of quantifying how often we can observe a particular value of a test statistic.

I know, I know, we've just rejected the null hypothesis, so why dwell on it?' Let me explain a couple of reasons: Firstly, our current scenario of having a t distribution benefits from having a neat CDF function with an inverse, which might not always be the case, so we need a plan B for those instances. Secondly, quantifying how often we expect to see a test statistic as extreme as $t_{obs} = -4.15$ equips us with a standardized measure of evidence against the null hypothesis. This standardization isn't just a fancy term——it genuinely aids in harmonizing decision-making across various statistical tests. So, while we could march ahead after rejecting H_0, pausing to understand these statistical tools' full scope and utility is crucial for robust, consistent scientific inquiry. Well, let's define this probability of how often we observe a value as extreme as $t_{obs} = -4.15$ like:

$$P(T = -4.5) \tag{3.81}$$

Let's revisit 3.81. Firstly because the t distribution is continuous that guy is equal to zero, but also, are we solely interested in when we observe a t-value of -4.15? In our test context, the primary interest extends beyond just observing this specific value. Considering our significance level α, which defines our critical threshold for decision. If our t-statistic, say -4.15, lands in this region, we reject H_0. Furthermore, any value more extreme than -4.15 also leads to a rejection of H_0. Therefore, we're interested in the likelihood of obtaining a t-value as extreme as -4.15 or more extreme. Hence,

the correct expression to evaluate this probability would be:

$$P(T \leq -4.15) \tag{3.82}$$

We are making progress, but we must recognize the nature of our test. Our rejection criteria consider both tails of our symmetric distribution, and 3.82 addresses only the left tail region of rejection. It's crucial to remember that while $t = -4.15$ is a specific observed value, in a two-tailed test, the other tail, in our case, the right one, also has a rejection area. This means that we are also interested in the likelihood of observing a t-value as extreme or more extreme than 4.15 in the positive direction. Thus, to fully encompass our test's scope, we should consider both tails:

$$P(T \leq -4.15 \cup T \geq 4.15) \tag{3.83}$$

Given the symmetry of the distribution, knowing the probability of one tail allows us to infer the same for its counterpart. This leads us to simplify 3.83 as follows:

$$2 \cdot P(T \geq \mid -4.15 \mid) \tag{3.84}$$

Finally, it's important to contextualize these probabilities within the framework of the null hypothesis being true. To explicitly state this condition, we use the conditional probability notation. Let's introduce the "given" condition with the appropriate symbol $\mid$, therefore 3.84 can be expressed like:

$$2 \cdot P(T \geq \mid -4.15 \mid \mid H_0 \text{ is true}) \tag{3.85}$$

Probabilities such as the one represented by 3.85 are called "conditional probabilities," an idea we have yet to explore. They are a fundamental concept in probabilities and statistics, and this means we have to learn what they are. Is the 'conditioning' the same as conditioning humans with massive debt, so that they don't cause any undue disruption to the elites? Let's see...

After successfully executing the heist at the Diamond Center, Notarbartolo and his team made their escape in the early morning. They expertly navigated from the building using a stairwell to access a secluded garden, descended a ladder to a terrace, and exited

through a second-floor window into Pelikaanstraat, where their getaway car was strategically parked. They drove directly to a safe house, with their hearts pounding with the thrill of victory, to examine the loot and celebrate their well-planned success.

As we plan our operation, we face new challenges due to upgraded security measures. The streets around the the Diamond Center now have surveillance cameras and retractable bars that can block vehicle access. Our analysis has pinpointed the times when the cameras are most likely to be inactive and identified the most feasible escape route. This route traverses streets monitored by cameras and retractable bars, and so far, we have no way of making those bars come down.

One strategy can be to ignore our getaway route and park the car further from the center, outside the area protected by the bars. However, this would mean ignoring our covariance analysis, which suggests a reduced patrol presence during rainy days. Attempting to leave the center with stolen goods in the rain and running to a distant car could be be a recipe for disaster; imagine slipping and landing on our back with a bag full of diamonds.

Therefore, our ideal plan is to take advantage of the identified periods of camera inactivity and try to go on a rainy day. This means we must park our car inside the Diamond Center's garage, which is free of cameras. Such a strategy will allow us to leverage all the insights we have gathered so far: time of inactivity, best exit route, and fewer guards patrolling the streets on a rainy day. However, to achieve this, we must circumvent or deactivate the retractable traffic bars, so let's get to work.

Luckily, I know a guy in the city's traffic department who's agreed to help us remotely control these retractable bars. There is a caveat: this fellow is a bit eccentric and unreliable, two personality traits that we would prefer to avoid when planning a robbery. But if we can quantify how much we can rely on this person, perhaps we can use his help; for example, we can call him randomly and note if the bars came down:

Contact Attempt	Bars Down
1	1
2	0
3	1
4	1
5	0
6	0
7	1
8	1

Table 3.9: Can we trust our traffic department contact?

Well, I made eight calls to our friend Hélio, and table 3.9 showcases this data alongside a 0, indicating a missed opportunity to lower the bars, possibly due to our guy being busy feeding his pet tarantulas. On the other hand, the 1's represent the instances when we successfully established contact and he lowered the bars. Now, if we want to quantify the likelihood of success of having the bars down, we can start by defining an event A that represents "The bars being down":

$$P(A) = \frac{5}{8} \approx 62.5\% \tag{3.86}$$

We count the number of successes represented by the 1's and divide by the total number of contact attempts; nothing new. Considering the stakes, I have to say that that a probability of 62.5% is low and we must explore a different avenue. Perhaps picking up the phone while he is at work to remotely putting the bars down is challenging for him as people may be in his office, so let's contact him eight more times, but this time, we will add an extra piece of information; we will distinguish the means of contact: text or call:

Attempt	Success	Means of Contact
1	1	Text
2	1	Text
3	0	Call
4	1	Text
5	0	Call
6	1	Text
7	0	Call
8	1	Text

Table 3.10: Is it better with a text instead of a phone call?

With table 3.10, we've gained valuable insights, revealing a trend that success significantly varies with the contact method, especially favoring text messages. This finding is critical and prompts us to refine our approach to calculating the likelihood of success by integrating this new information.

Initially, we established a baseline probability of success using data from table 3.9. Now, we want to update our calculations with additional data from table 3.10, which differentiates success rates between text messages and calls.

The next step involves including a condition in our probabilities calculations: "What is the probability of success given that we used text for the contact?" In statistics, we call this a 'conditional probability' and this concept helps us assess the likelihood of event A occurring given the presence of another condition B, such as using text to communicate. For example, we can calculate $P(A|B)$, where B represents the condition of using text messaging and A a success:

$$P(\text{Success} \mid \text{Text}) \tag{3.87}$$

Given equation 3.87, our goal is to determine the conditional probability of achieving 'Success' given that the 'Text' method has been used. Essentially, this equation guides us to assess our chances of success under the specific condition that communication was made via text. To calculate this probability, we focus only on instances where 'Text' was the method of contact, specifically attempts 1, 2, 4, 6, and 8. Within this subset, we examine how many times 'Success' coincided with 'Text'. This process involves counting the instances of successful outcomes when the text was used and dividing this by the total number of text attempts:

$$P(\text{Success} \mid \text{Text}) = \frac{\text{Number of success and Text}}{\text{Total number of Text}} \tag{3.88}$$

Let's input some numbers in our formula and see what comes from it:

$$P(\text{Success} \mid \text{Text}) = \frac{5}{5} = 100\%$$

Indeed, by incorporating the new data into our analysis, we significantly improved our odds. This update confirmed our suspicions: our contact is fully on board with our plan, but favors texts over calls, probably for reasons of discretion (I promised him a diamond so...).

We know that probabilities have a complement, and conditional probabilities are no different. So if 3.88 results in 100%, what is the complement of this guy? It needs to be 0, but is it the probability of failure given we used text messages or the probability of success when we call him? If we defer back to the definition of probability, we know it relates directly to the original condition under consideration. If 3.88 gives us the probability of success given that we used text messages, and this probability sums up to 100%, then its complement is not about the method of contact changing (i.e., switching from text to call); instead it concerns the outcome under the same condition:

$$P(\text{Fail} \mid \text{Text}) = \frac{\text{Number of fails and Text}}{\text{Total number of Text}}$$

$$P(\text{Fail} \mid \text{Text}) = \frac{0}{5} = 0\%$$

3.3.2 Conditional Probabilities: Guiding for Collective Good.

There is quite a bit of text but few Greek letters, and we must keep the Greek gods on our side, so let's try to replace as much text as possible with mathematics. I'll go first——number of successes and Text '. We have a symbol for 'and'; it is $\cap$, so that becomes 'Number of success $\cap$ Text '. Should I do another one? Sure, 'Success,' from now on you are A; do you like it? Yes?!? We have a keeper. I am on a roll, so I will do one more. 'Text,' you are B, why? It is B for because! Let's see where this gets us:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \text{ when } P(B) > 0 \tag{3.89}$$

If A and B are two events in a sample space S, then the conditional probability of A given B is defined in equation 3.89:

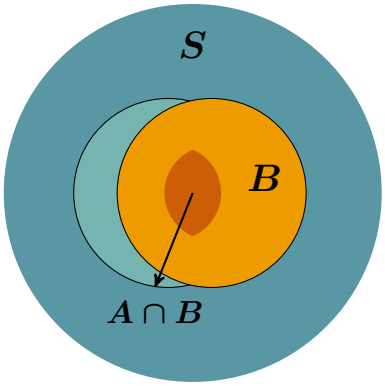


Figure 3.35: What is left after all the conditioning and intersections.

Assuming that event B has occurred, our consideration is confined solely to set B, illustrated by the orange and yellow areas in figure 3.35. Hence, for event A to transpire given B, it must reside within the intersection $A \cap B$, represented by the orange region in the figure. This illustrates the concept of conditional probability, which, as a form of probability, adheres to the fundamental axioms of probability:

- **For any event $A, 0 \leq P(A \mid B) \leq 1$.** This axiom states that the probability of an event A occurring, given that event B has already occurred, is always non-negative.
- **$P(B|B) = 1$.** This principle says that, if we know event B has occurred, the probability of B occurring given B has occurred is 1, or 100%.
- If **$A_1, A_2, ..., A_n$ are disjoint events, then**

$$\begin{aligned} P(A_1 \cup A_2 \cup ... \cup A_n \mid B) = {} & P(A_1 \mid B) \\ & + P(A_2|B) \\ & + \dots \\ & + P(A_n|B). \end{aligned}$$

The axioms in the box above are similar to the ones we learned when introducing probabilities. The framework of conditional probabilities underpins the foundation for Bayes' theorem. This theorem, a cornerstone of statistical inference, offers a powerful mechanism

for updating our beliefs or probabilities in light of new evidence. Its elegance lies in its ability to incorporate prior knowledge and dynamically adjust it as new data becomes available. It makes it a fantastic tool in machine learning, and we will study it in more depth later in this book. But let's not forget why we introduce the concept of conditional probabilities; we were trying to quantify how likely we are to observe a test statistics as extreme or more extreme than a specific value; in fact, let me bring back the equation we were working on:

$$2 \cdot P(T \geq | \, -4.15 \, | \, | \, H_0 \text{ is true}) \tag{3.90}$$

Now that we've established a methodology for computing 3.90, why not take the next step and formally define the concept that this equation represents? ?

$$p\text{-value} = 2 \cdot P(T \geq | \, t_{\text{obs}} \, | \, | \, H_0 \text{ is true}) \tag{3.91}$$

That is correct. We were discussing the concept of the p-value, a conditional probability that quantifies the likelihood of observing a test statistic as extreme or more extreme than the one actually observed under the null hypothesis.

3.3.3 If You Go To That Extreme, I Can Still See You, But I Don't Know If I Can Accept You: the *p*-value.

The term p-value refers to a metric that quantifies the probability of encountering a test statistic as extreme as, or more extreme than, the one observed, assuming the null hypothesis is valid. Given this understanding, let's compute the p-value for our test as represented in 3.90.

There is more than a way we can compute our p-value defined by 3.91. Given that the CDF of our null distribution is well-established for this particular case, we can utilize it to determine the p-value. However, we will skip the manual calculations as this topic of cumulative distribution was covered previously in the book, so if we go and use software to compute the probability described by 3.91 we

have that:

$$p\text{-value} = 0.0001745$$

What do we do with that now? Is that probability very low? Well, we know that our referee is the significance level α, meaning that if we find a relationship between it and the p-value, we have everything we need to decide what to do with our pal H_0.

The p-value quantifies the probability that the observed data, or something more extreme, would occur under the null hypothesis. When we set a threshold for this probability α, we define the maximum risk of error we are willing to accept when rejecting the null hypothesis. This metric α essentially acts as a benchmark for determining when the evidence against the null hypothesis is strong enough to be considered statistically significant. When the p-value falls bellow α, it suggests that the likelihood of observing the data assuming the null hypothesis is true is exceptionally low. For instance, setting $\alpha = 0.05$ implies a 5% risk tolerance for rejecting the null hypothesis incorrectly. If the p-value is less than 0.05, the observed data is so unusual under the null hypothesis that it occurs less than 5% of the time, prompting us to reject the null hypothesis.

For our test, the p-value is smaller than our $\alpha = 0.05$, therefore there is strong evidence to reject the null hypothesis H_0, which is exactly the conclusion we reached by just using the t-critical values.

Before we formally define our statistical tests, one more concept will be beneficial to explore. We previously tested the hypotheses $(\mu = 30)$ versus $(\mu \neq 30)$. Because the "direction" of the difference did not matter in the context, our rejection region split into two tails. Tests with this specific characteristic, where the decision regions split into two sides, are called' two-tailed tests'.

If you are wondering if tests where the rejection region is only on one side of the distribution exist, let me tell you that they do! These occur when we alter our alternative hypothesis H_1 from a non-equality, such as $\neq$, to an inequality, either $\leq$ or $\geq$, making the "direction" of the difference crucial. This change in the hypothesis affects the placement of the rejection area. It modifies the compu-

tation of the p-value—an essential point we must understand.

$$H_0 : \mu = 30 \quad \text{vs} \quad H_1 : \mu < 30 \tag{3.92}$$

$$H_0 : \mu = 30 \quad \text{vs} \quad H_1 : \mu > 30 \tag{3.93}$$

In equation 3.92, we test $H_0 : \mu = 30$ against the alternative hypothesis $H_1 : \mu < 30$, indicating that the test intent is to detect whether the population mean is less than 30. Conversely, in equation 3.93, the test involves $H_0 : \mu = 30$ against $H_1 : \mu > 30$, which aims to determine if the population mean is greater than 30.

Because in both of these tests represented by equations 3.92 and 3.93, we are just interested in one direction, meaning greater or lower, we base our rejection-acceptance criteria on only one tail of the null distribution as opposed to two tails as showcased on the test represented in figure 3.34. As you can imagine, we have names for these fellows; we called these tests 3.92 and 3.93 "one-tailed test", more specifically, 3.92 is a "lef side one-tailed test" whereas 3.93 is a "right side one-tail test". Let's visualize these new rejection-acceptance regions for the one-tailed tests:

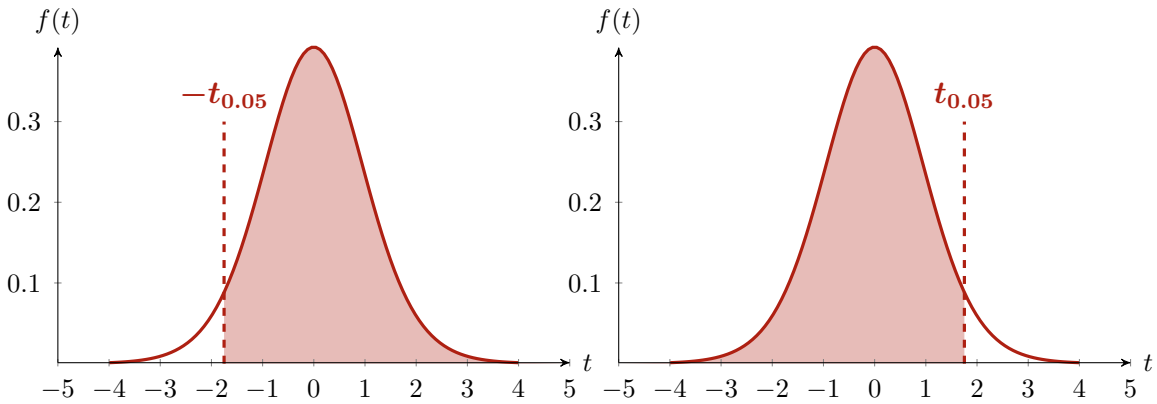


Figure 3.36: When rejection picks a side: left vs right one-tailed test.

The rejection criteria for one-tailed tests are specific to the direction of the test. For the left-tailed test represented by the left panel of figure 3.36, we reject the null hypothesis if the test statistic falls below a critical value in the lower tail of the distribution. For the right-tailed, the right panel of 3.36, the rejection occurs if the test statistic exceeds a critical value in the upper tail. In contrast,

considering both distribution extremes, a two-tailed test would reject H_0 if the test statistic falls either below the lower critical value or above the upper critical value.

To compute the t_α's, we can use the same formula as we did for the confidence intervals:

$$t_{0.05,15} = F_{t,15}^{-1}(1 - 0.05) = F_{t,15}^{-1}(0.95) = 1.753$$

$$-t_{0.05,15} = -1.753$$

Below are the acceptance-rejection criteria per test:

$$\begin{array}{ll} & \text{Reject } H_0 \text{ if } t_{obs} < -1.753 \\ \text{Left-tailed test:} & \text{Fail to reject } H_0 \text{ if } t_{obs} \geq -1.753 \\ & p - \text{value} = P(T \leq t_{obs} \mid H_0 \text{ is true}) \end{array} \tag{3.94}$$

$$\begin{array}{ll} & \text{Reject } H_0 \text{ if } t_{obs} > 1.753 \\ \text{Right-tailed test:} & \text{Fail to reject } H_0 \text{ if } t_{obs} \leq 1.753 \\ & p - \text{value} = P(T \geq t_{obs} \mid H_0 \text{ is true}) \end{array} \tag{3.95}$$

One-tailed and two-tailed tests cater to distinct research inquiries, each determining p-values in unique ways that reflect their specific objectives. One-tailed tests, which may be either left-tailed or right-tailed, concentrate on detecting an effect solely in one direction. A left-tailed test, for example, evaluates whether a parameter is less than a specified value and calculates the p-value as $P(T \leq t_{obs} \mid H_0 \text{ is true})$, representing the probability of observing a statistic as extreme, or more extreme, than the actual result under the null hypothesis. A right-tailed test, conversely, assesses outcomes that exceed a hypothesized value, employing $P(T \geq t_{obs} \mid H_0 \text{ is true}))$ for its p-value determination.

In contrast, a two-tailed test is utilized when a study necessitates the examination of deviations in both directions from the hypothesized value. This test computes its p-value by summing the probabilities of the test statistic occurring in both tails of its distribution, which effectively doubles the one-tail probability in scenarios where the distribution is symmetrical about zero. This more conservative

approach requires stronger evidence to reject the null hypothesis, making two-tailed tests particularly appropriate for exploratory research that does not anticipate a specific directional outcome.

The reality is that this seems like an important subject, and it is! We just introduced a concept called hypothesis testing, so what about if we formally introduce it?

3.4 The Only Match Where 0 Often Wins Against 1: Hypothesis Testing.

Hypothesis testing is a fundamental statistical method used to make decisions about a population parameter based on sample data; well, it extends to test more things than a population parameter, as we will see shortly. The process starts by formulating two competing hypotheses: the null hypothesis (H_0) and the alternative hypothesis (H_1).

- **Null Hypothesis H_0**: The null hypothesis posits that there is no effect or difference in the population parameter being studied, suggesting that any observed differences in sample statistics arise purely from random chance.
- **Alternative Hypothesis H_1**: The alternative hypothesis challenges the null hypothesis, proposing that there is a genuine effect or difference. It implies that the observed changes in the data are due to this effect rather than randomness. Failing to reject H_1 means rejecting H_0 as it indicates that the differences observed are statistically significant.

To decide between these hypotheses, a test statistic is defined, and an observed value for it is calculated from the sample data. This statistic, measures how far the sample statistic deviates from what is expected under H_0. The decision to reject or not reject H_0 hinges on the calculated p-value or the test statistic's position relative to a critical value:

- ***p*-value**: The p-value quantifies the probability of observing a test statistic as extreme as, or more extreme than, the value obtained under the assumption that H_0 is true. If the p-value is less than a pre-determined significance level (α), typically 0.05, H_0 is rejected. This suggests that the observed data are unlikely to have occurred by chance under the null hypothesis.
- **Critical region**: Alternatively, the decision can be based on whether the test statistic falls into the critical region, which is defined by the critical values that correspond to the chosen α. If the test statistic lies outside the range defined by these critical values, H_0 is rejected.

The p-value and the critical regions are equivalent methods and always yield the same result. Their calculations depend on the type of test we conduct: one-tailed test vs two-tailed test. However, the characteristics of hypothesis tests continue beyond this. Earlier, when we created a two-tailed test for the mean, we questioned whether our data indeed follows a normal distribution. When testing if the population mean could be equal to 30, we reverted to the assumption that the data is Gaussian. This is because we used a t-statistic. If we make assumptions about the underlying distribution, our hypothesis test is called parametric. On the other hand, if we don't assume any specific distribution, the test is non-parametric:

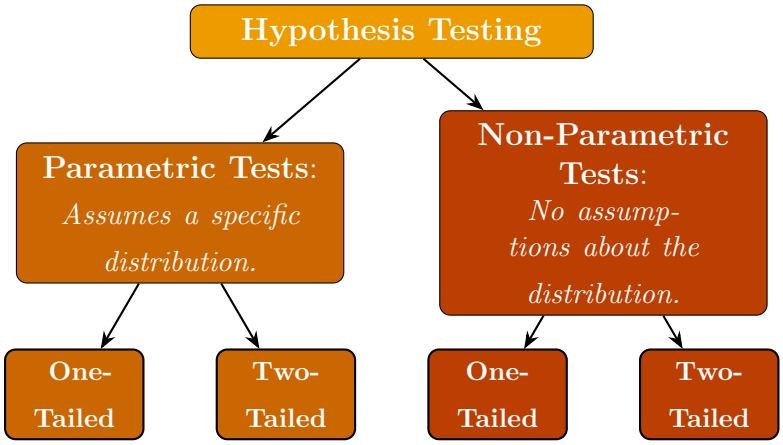


Figure 3.37: Show me our tails and your assumptions, and I will tell you what you are.

So, how we reject the null hypothesis determines whether we're doing a one-tailed or a two-tailed test. Assuming a specific distribu-

tion for the data (like the normal distribution) is called a parametric test. On the other hand, if we don't assume any particular distribution, the test is considered non-parametric.

All of the above is great, but whether our data is normally distributed still needs to be solved. Why not dive deeper into hypothesis tests? Hmmm, can we test our data for normality? Or are these tests so lame that we can only use them to test parameters?

They are versatile, and indeed, we can test different things amongst them if our data follows a particular distribution. Pretty cool, no? How about we explore a non-parametric test– -something we just introduced but haven't fully explored—that helps us determine the normality of our data? Great idea, you say? I thought so, too. First, let's define this bad boy:

$$H_0 : \bar{X} \sim \mathcal{N}(\mu, \sigma^2) \text{ vs } H_1 : \bar{X} \not\sim \mathcal{N}(\mu, \sigma^2) \tag{3.96}$$

Equation 3.96 defines our new test where H_0 states that our data follows a normal distribution, whereas H_1 is the party pooper that states otherwise. With our hypothesis defined, we now require a test statistics, the metric that will allow us to make inferences and confidently decide whether we reject or fail to reject our H_0. In addition, we also need a null distribution under H_0 and significance level to take this decision. The problem is that we don't have any test statistic to understand whether a variable follows a normal distribution or a null distribution under H_0; let's create one! We will draw some inspiration from the t-score, the metric we use to test a deviation from the sample mean. In that instance, the t-score was a great candidate for the test statistics because its formula represented what we needed to test. In addition, because we have t distribution defined, the null distribution under H_0 was for free. Now the case is different, we need a test statistics that somewhat allow us to understand if our sample follows a Gaussian distribution.

If our goal is to test the normality of our sample, let's focus on the characteristics of the Gaussian curve. This curve is characterized by its central peak, where the majority of the data density is concentrated, and by its tails, where this density gradually decreases, creating the familiar bell shape. A straightforward approach

to identifying a Gaussian curve is by examining whether the data distribution forms this bell shape.

I propose exploring this method of comparing distributions based on their shapes. By doing so, we can develop a technique not only to test for normality but also to determine if any two distributions, either normal or not, share similar shapes. This is important because the shape of a distribution reflects the underlying behavior of the data. Therefore, if two samples have identical shapes, they likely share similar distributions, indicating similar underlying processes or characteristics in the data. If we are going to compare shapes, what about if we draw some? Let's start with an arbitrary example with two distributions that visually clearly showcase a different shape and see if we can develop a methodology:

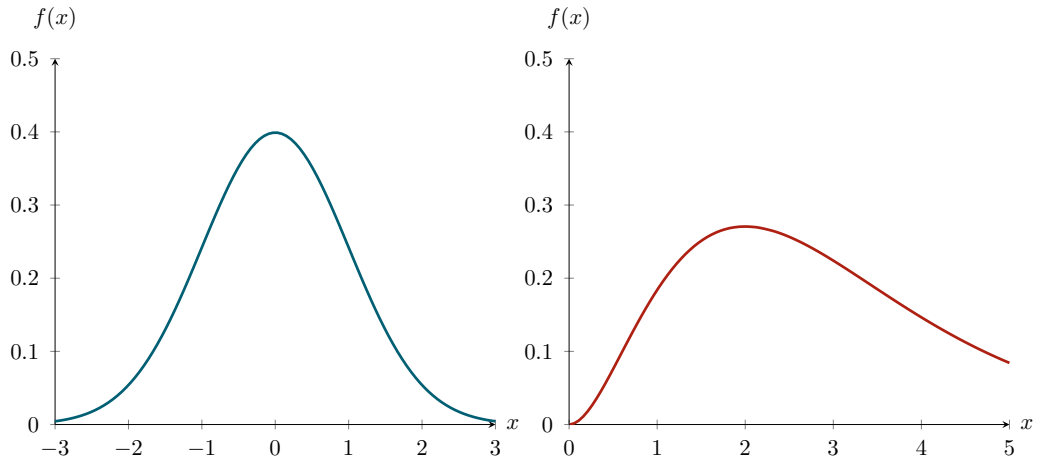


Figure 3.38: Sure, you're taller and blue, but does that really matter? We both pack the same density, after all.

The million-dollar question is: Are these distributions similar? No, you say? Well, you caught me off-guard, and I don't have that money to pay you, so what about if we settle with a piece of candy?

Those plots in 3.38 are distinct, no question about it. The left plot shows a symmetric distribution centered around zero, typical of a normal distribution. In contrast, the right plot depicts a skewed distribution, where the data is more concentrated towards the lower end. Now, what methodology can we create to detect this shape difference? Well, what if we go into our statistical toolkit and assess what information we can get about the shape of a distribution?

Let's start with the basics: measures of central tendency alone do not provide a complete picture of a distribution's shape. Although complementing them with measures of dispersion enhances our understanding of the data's behavior, this method necessitates analyzing multiple metrics. These measures are particularly effective for symmetric distributions where they offer a reasonable summary of the central tendency and spread. However, in scenarios where one of the distribution is skewed or exhibits significant tail behavior, such as the red distribution shown in the right panel of figure 3.38, the mean and standard deviation fall short. They fail to accurately reflect the data's asymmetry and actual spread, potentially leading to a skewed (no pun intended) perception of data point concentration and distribution.

Moving on to quantiles—these are incredibly insightful as they segment a distribution based on the data's rank order, effectively slicing the data into equal-sized parts. We might be on to something with our good friends Q's! If quantiles evenly divide the distribution, they can be a robust tool for comparing different distributions. Specifically, observing differences between the corresponding quantiles across distributions could highlight distinct characteristics between them. Conversely, similarity in quantiles would suggest a likeness in distribution patterns. Why not apply this concept directly to our analysis? Let's incorporate some of these quantiles into our curves, focusing on the most recognized ones within the interquartile range:

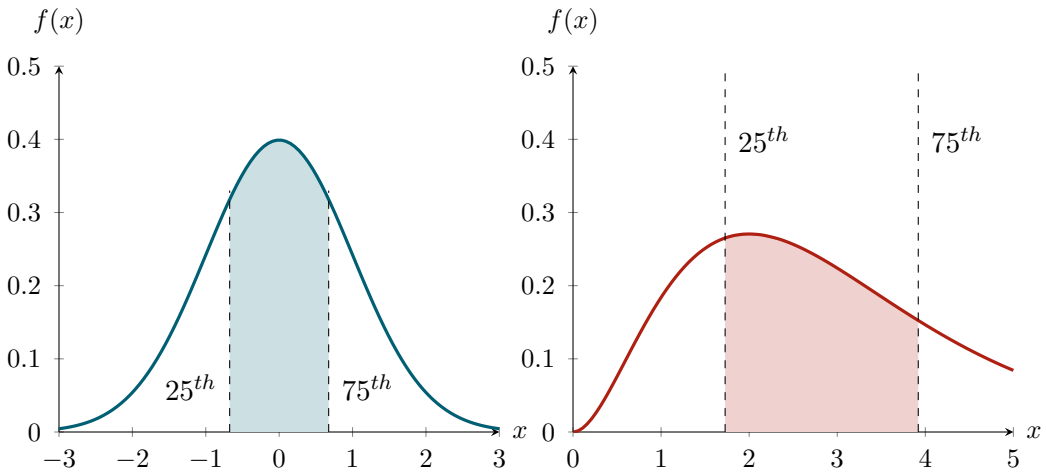


Figure 3.39: See blue, I told you this is not basketball. We carry the same density.

The interquartile range, which encompasses 50% of the data, is visually represented in the plots of figure 3.39. In the left plot, this segment is highlighted by a blue shaded area, and in the right plot by a red shaded area. Given that these regions—spanning from the 25^{th} to the 75^{th} quantiles—encapsulate the same probability mass, comparing the physical distances between the corresponding x-coordinates of these quantiles offers a practical method to assess the similarity in the shapes of the distributions.

If these interquartile ranges contain equivalent amounts of data but the spatial distance between the x-coordinates of the quantiles differs between the distributions, it suggests varying distribution densities. Specifically, if one distribution exhibits a shorter distance between these quantiles compared to another, it indicates a higher peak density in that distribution. This occurs because it must pack the same proportion of data into a more compact space.

Consider the analogy of a balloon: If you compress it, it narrows while still containing the same air. Similarly, we are examining whether the same effect—a compression in spatial spread without a loss in data mass—is observable in the distribution profiles by scrutinizing the distances between the x-coordinates. This comparison can reveal insights about the underlying distribution shapes, suggesting which one might have a steeper or more concentrated peak.

Looking at the two plots in 3.39, we can see that the blue and red distributions differ in shape. Let me get the values of those quantiles so we can compute their distance:

$$Q_{25}^{(\text{Blue distribution})} = -0.674, \quad Q_{75}^{(\text{Blue distribution})} = 0.674$$

$$Q_{25}^{(\text{Red distribution})} = 1.727, \quad Q_{75}^{(\text{Red distribution})} = 3.920$$

Now, let's compute the distances between those quantiles:

$$\text{Distance Blue distribution} = \mid (-0.674) - 0.674 \mid = 1.348$$

$$\text{Distance Red distribution} = \mid 3.920 - 1.727 \mid = 2.193$$

I think we're onto something! The distance between the quantiles of the red distribution, at 2.193, is larger than that of the blue distribution, which measures 1.348. This observation supports our hypothesis that distributions with higher peaks, such as the blue ones, necessitate shorter distances between quantiles to contain the same mass. Conversely, more spread-out distributions, such as the red one, exhibit greater distances between these critical points when considering identical densities. However, while this method gives us some insight, it still doesn't seem robust enough. For example, by focusing solely on the interquartile range, we're ignoring the wealth of information contained in the distribution's tails, which can be crucial for understanding the overall shape and behavior of the data.

Also, disregarding everything between the 25^{th} and 75^{th} quantiles feels like a leap of faith—much distribution is left unexamined. So, what if we consider more quantiles? This seems promising, as it would provide a more complete picture of the distribution's shape. However, more quantiles mean more distances, raising a new challenge: How will we compare all these distances we have to calculate?

Still, let's not abandon our reasoning just yet. We began this journey by suggesting that for two distributions to be identical, the distances between the x-coordinates of their corresponding quantiles should be similar or the same, meaning that if these distributions are the same, they must match. However, juggling multiple distances and comparing them directly seems like a pain, so we need a more streamlined approach.

Focusing solely on the x-axis may not be the best path forward, as comparing multiple distances will lead to a dead end. What if we reverse our approach? Instead of starting with the x-coordinates and trying to see if the distances match perfectly, what if we start from a stand of "perfection" and see how well these distances fit there? Now, what would represent "perfection" in this situation? It is to have the same distances between the corresponding quantiles of the two distributions. I know of a mathematical framework that can represent that type of "perfection"; a line:

$$y = mx + b$$

If we consider a line, the distance between consecutive points along this line is always consistent. To adapt this concept into our methodology, we must first define what line we'll consider. Quantiles must play a role, this is as certain as taxes. Moreover, we need points to define a line, at least two of them, further narrowing our focus. So, we must establish a line that involves information about the quantiles and need at least two points to define it.

What if the coordinates of these points were the quantiles themselves? Let's explore this idea: suppose we set up a couple of points where each of them is characterized by the corresponding quantiles from different distributions.

If we then define a line through these points, in theory, if the distributions are somewhat similar and we continue to define more points with their quantiles, these new points should either land on or very close to the line. This expectation stems from the notion that if a line is defined by the quantiles and we know that consecutive points on a line maintain equidistant, then, if the distributions are indeed the same, this property should hold true. To further explore this idea,we will define two points with the 25^{th} and 75^{th} quantiles as coordinates:

$$\text{Point}_1 = (Q_{25}^{(d_1)}, Q_{25}^{(d_2)}), \quad \text{Point}_2 = (Q_{75}^{(d_1)}, Q_{75}^{(d_2)})$$

$$\text{Point}_1 = (-0.674, 1.727), \quad \text{Point}_2 = (0.674, 3.920)$$

Above, we've identified two points based on quantile values from each distribution. For example, Point_1 is constituted by the 25^{th} quantiles of both distributions, represented as $Q_{25}^{(d_1)}$ and $Q_{25}^{(d_2)}$. Similarly, Point_2 follows the same rationale but uses the 75^{th} quantiles, denoted as $Q_{75}^{(d_1)}$ and $Q_{75}^{(d_2)}$. Now, let's plot the 25^{th} and 75^{th} quantiles from both distributions to visually assess their relationship:

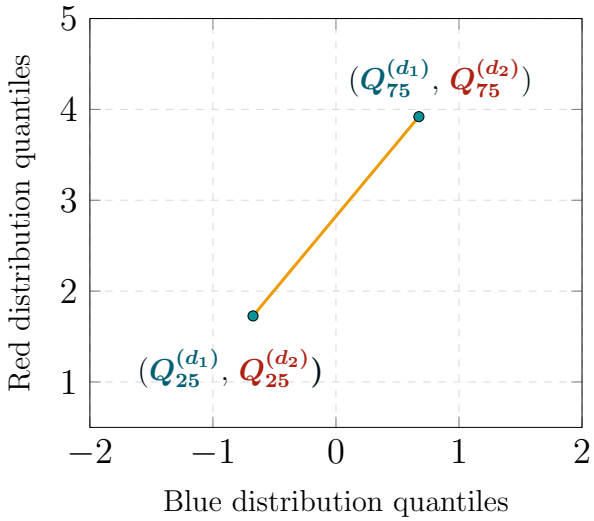


Figure 3.40: Two quantiles and a line, everything seems perfect!

In figure 3.40, we've marked two points in blue, representing Point_1 and Point_2 and therefore the coordinates for the points representing 25^{th} and 75^{th} quantiles of our distributions. These points are connected by a yellow line, illustrating the expected linear relationship if the distributions were perfectly aligned. However, this simplistic view might not reveal the whole story. Of course, if we only consider two points, we can always define a line that passes through these same points; after all, this is the definition of a line. So let's gather more information from both our distributions and create a third point; say that now we are also going to represent the coordinates of the 50^{th} quantile:

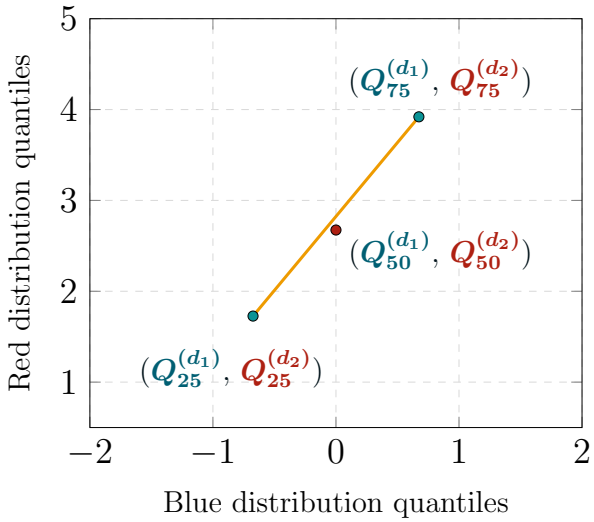


Figure 3.41: The red dot had to come and nearly ruin everything!

That point in 3.41 has coordinates corresponding to the 50^{th} quantiles for both the blue and red distributions. Even though the distributions are notably different, the point isn't that far from the line. Let's increase the number of quantiles and see what happens:

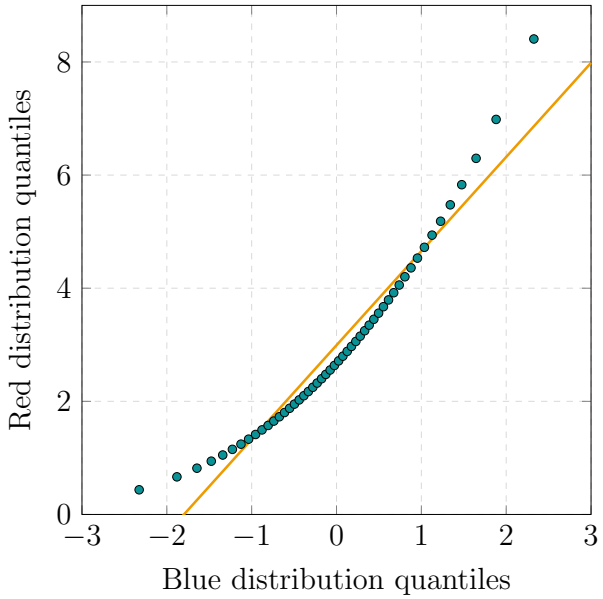


Figure 3.42: Not all of them are in line...

In figure 3.42, we can now see some divergences from this reference yellow line, as expected. When data points diverge from the line, it signals potential discrepancies between the blue and the red model. Such deviations from the line indicate that the blue distribution may differ in scale, center, or overall shape from the red counterpart. Points that lie far from the reference line, particularly in the distribution tails, may suggest the presence of skewness, heavier or lighter tails (kurtosis) than the blue distribution, or other anomalies in the red distribution.

3.4.1 A Quest through the Quagmire of Plots.

Figure 3.42 introduces what's known in statistics as a Quantile-Quantile (QQ) plot. A QQ plot is a graphical tool used to compare the distribution of a dataset against a theoretical distribution or another dataset. It plots the quantiles of one dataset against the quantiles of the other. If the two distributions are similar, the points on the QQ plot will approximately lie on a straight line. The slope and

position of this line provide insights into how the two distributions differ in terms of location, scale, and symmetry.

Let's create one of these QQ bad boys for our sample means. In the deduction of our method, we observed that, the more quantiles we plotted, the better we could understand the discrepancy between distributions. So, if we have a sample of size n, the maximum number of quantiles we can have is n, and in this particular case, $n = 16$. Therefore, we will create the same amount of quantiles for this sample. To start with, we sort the data in ascending order:

$$27.54, 27.94, 28.08, 28.36, 28.53, 28.61, 28.69, 28.82$$

$$28.92, 28.98, 29.20, 29.50, 29.58, 30.01, 30.21, 30.52$$

By sorting the sample, we establish a clear vantage point. The initial values in the array delineate the left tail, while the values at the end define the right tail. The distribution's peak emerges from the values in the array's center.

We are ready to compute the quantiles of each position; in this case, we will go with percentiles. The process is the same as before; we will divide the ranking by the sample size $n = 16$. It is essential to notice that, since we are working with a sample, it is common practice to introduce a correcting factor. This helps us account for the possibility that lower and higher values than those presented in our dataset can exist. For example, if we wish to compute the percentile of 30.52, we can use the following formula:

$$\frac{i^{th} - 0.5}{n} = \frac{16 - 0.5}{16} = 0.97$$

This implies that the value 30.52 is located at the 97^{th} percentile, meaning that 97% of the data falls below this point, offering a nuanced and possibly more accurate depiction of its standing within the distribution. The 0.5 value is a convention in many statistical methodologies and is employed to ensure consistency across different statistical analyses and data interpretations.

After sorting our sample and determining ranks, we proceed to map these to the corresponding percentiles of the standard normal distribution. The essential question we address at this stage is, 'Which Z-value in the standard normal distribution corresponds to a given percentile of our data?' This question is pivotal because, although percentiles convey relative positions within our dataset, to plot them in a QQ plot and scrutinize for deviations, we need exact coordinates: specifically, where these percentiles fall on the standard normal curve, represented on the x-axis.

The process hinges on the translation of these percentiles into Z-scores, using the Φ^{-1} function, the inverse of the CDF. This function is our bridge from percentile-based probabilities to the precise locations on the x-axis of our plot. By converting percentiles to Z-scores, we pinpoint the exact coordinates needed to draw our comparison line and thus visually assess how our data's distribution aligns with, or deviates from, the theoretical normal distribution:

$$\text{Rank} = 2$$

$$\text{Percentile} = \frac{2 - 0.5}{16} = 0.09$$

$$\Phi^{-1}(0.09) = -1.318$$

Given the calculation that places the second-ranked observation at the 9% percentile, our next step involves determining the corresponding value on the x-axis of the standard normal distribution. This value represents the point at which the cumulative probability is 9%. To find this, we utilize the inverse CDF of the Z distribution, which yields a Z-score of approximately -1.318. This specific Z-score is derived by consulting statistical tables or using software functions designed for this purpose. Continuing this process for all observations:

Sample Mean	Ranking	quantile	Inverse CDF (Z-score)
27.54	1	0.031	-1.862
27.94	2	0.0937	-1.318
28.08	3	0.1562	-1.010
28.36	4	0.2187	-0.776
28.53	5	0.2812	-0.579
28.61	6	0.3437	-0.402
28.69	7	0.4062	-0.237
28.82	8	0.4687	-0.078
28.92	9	0.5312	0.078
28.98	10	0.5937	0.237
29.20	11	0.6562	0.402
29.50	12	0.7187	0.579
29.58	13	0.7812	0.776
30.01	14	0.8437	1.010
30.21	15	0.9062	1.318
30.52	16	0.9687	1.862

Table 3.11: I need means, rankings, quantiles, oh and since you will be down there, bring some inverses of Z's.

With 3.11 we can now create a QQ plot:

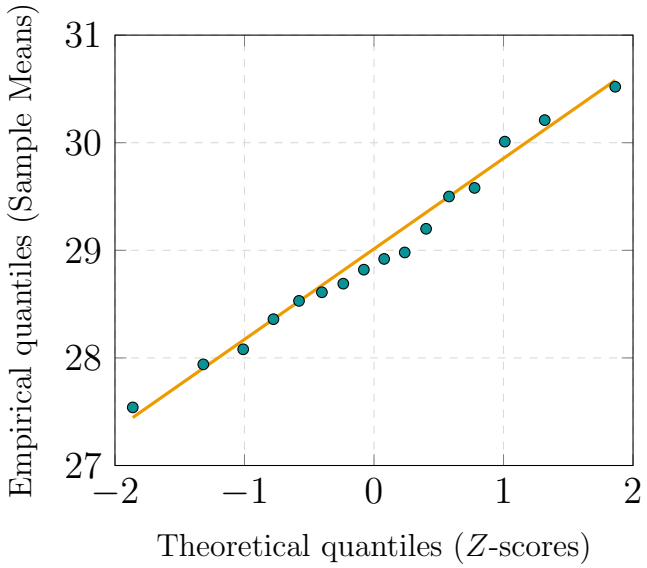


Figure 3.43: The good times when dots are on straight lines.

As before, we created a yellow line using the 25^{th} and 75^{th} quantiles to serve as our reference for assessing this linear relationship between the quantiles of both distributions. Because, on table 3.11 we don't exactly have the 25^{th} and 75^{th} we use the ones that are closest to them for the points to define the line:

$$\text{Point}_1 = (28.36, 0.218), \quad \text{Point}_2 = (29.50, 0.718)$$

In our plot 3.43 we have a notable alignment between the sample means and the theoretical quantiles of the standard normal distribu-

tion. The proximity of the data points to the trend line (highlighted in yellow) suggests a general adherence to normality. This observation is crucial for validating assumptions underlying many statistical tests that require the data to be normally distributed.

However, while the QQ plot serves as a powerful visual tool for assessing normality, it is not without limitations. One key challenge is the subjective nature of visual interpretation; what appears to be a close fit to one observer might not to another. Additionally, the QQ plot primarily illuminates deviations in the tails of the distribution, potentially overlooking subtler deviations from normality in the data's central region.

Moreover, the QQ plot does not provide a statistical test or quantitative measure of normality. It cannot, on its own, quantify the degree of deviation from a normal distribution. The truth is that this was no accident; I took advantage of the fact we were going to deduce a statistic to test normality and introduce us to a new statistical tool called the QQ plot.

Returning to our quest for a robust comparison method to build a test statistic, we find ourselves nearly back to ground zero. Having already leveraged quantiles, a valuable tool from our statistical toolkit, we must focus on the remaining candidates: The PDF and the CDF.

It is likely that both of these statistical frameworks will enable us to devise a test statistic. However, we mustn't overlook a crucial detail: we mentioned earlier our intention to explore a non-parametric test. This consideration is essential when selecting the appropriate tool for constructing our test statistic. Choosing a PDF to begin crafting our metric might lead us into a trouble, mainly because using a PDF implicitly assumes a specific distribution. This leaves us with the CDF as our viable option. Can we develop a non-parametric test using the CDF? Well, let's give it a try!

Let's start by revisiting the definition of the CDF; it maps any given value x to the cumulative probability of encountering a value less than or equal to x:

$$F(x) = P(X \leq x) \tag{3.97}$$

If we need to determine whether a distribution follows a normal distribution, we can move forward by establishing a comparison. To initiate this comparison process, let's define a CDF from a normal distribution with the same mean μ and variance σ^2. We'll consider this our theoretical distribution—a benchmark for subsequent comparisons.

Next, we turn to our data to develop a metric or a mathematical element that facilitates comparison with this theoretical CDF. At its core, a CDF represents the cumulative probabilities of observed events.

This means we can construct a CDF from our sample data by calculating the likelihood of each event and then cumulatively summing these probabilities, mirroring the method used for the theoretical CDF. This approach allows us to directly compare the empirical distribution of our data against the expected normal distribution. This new CDF has a name; it is the ECDF standing for 'Empirical cumulative distribution function'.

3.4.2 Function Frontier: The Empirical Cumulative Distribution.

To create an empirical cumulative distribution function (ECDF), we can leverage the foundational concept of the cumulative distribution function to calculate the cumulative probability up to a specific observation x. Therefore we execute as follows:

$$F_n(x) = \frac{\text{number of elements} \leq x}{n} \tag{3.98}$$

Equation 3.98 represents the Empirical Cumulative Distribution Function and is defined by:

- $\boldsymbol{F_n(x)}$: Is the ECDF value at a point x.
- $\boldsymbol{n}$: Is the total number of observations in the sample.
- $\boldsymbol{x}$: Is a value for which the ECDF is being calculated.

The ECDF is an essential statistical tool that offers a graphical representation of a dataset's distribution. It is constructed by plotting the value of each data point against its percentile rank within the dataset. This arrangement results in a step function that ascends at each observed value. Such a configuration enables the ECDF to illustrate the proportion of observations that are less than or equal to x, effectively displaying the cumulative probability for each observation.

Let's showcase how to do these calculations for one single x value, and then I will present a table with all the remaining computations. Say we pick $x = 28.08$; well, we have three values that are lower or equal to 28.08: 27.54, 27.94 and 28.08, so:

$$F_{16}(28.08) = \frac{3}{16} = 0.187$$

If we do similar calculations for all the remaining observations:

Sample No.	1	2	3	4	5	6	7	8
Sample Mean	27.54	27.94	28.08	28.36	28.53	28.61	28.69	28.82
ECDF	0.0625	0.125	0.187	0.25	0.3125	0.375	0.4375	0.5

Sample No.	9	10	11	12	13	14	15	16
Sample Mean	28.92	28.98	29.20	29.50	29.58	30.01	30.21	30.52
ECDF	0.5625	0.625	0.6875	0.75	0.8125	0.875	0.9375	1.0

Table 3.12: The Sample Means and Corresponding ECDF of 16 Random Samples.

We can visualize both of these cumulative distributions:

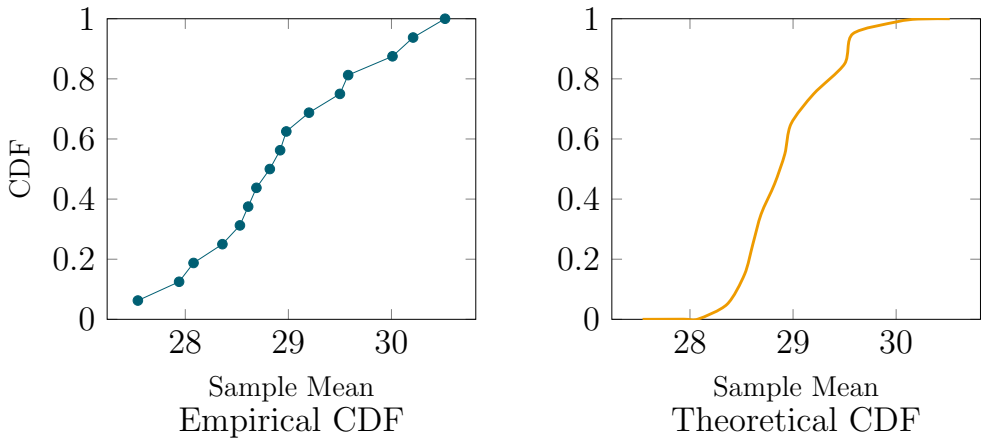


Figure 3.44: A clash between empirical and theoretical.

In graph 3.44, we have two panels: The left guy, in true blue, depicts our ECDF, the cumulative distribution we got from our dataset. On the left, the mellow yellow corresponds to a theoretical CDF from a normal distribution with parameters $\mu = 28.97$ and $\sigma^2 = 0.043$, which we estimated from our data sample of time of inactivity on cameras. At first glance, the distributions seem pretty similar to the untrained eye. However, our eyeballing days are over. Otherwise, we would have stuck with the QQ plot. The crux of hypothesis testing is developing a precise metric that facilitates comparison, and, on top of that we must understand the distribution of this metric.

Well, what about if we create a metric that gives a measure of comparison between both the CDF with the ECDF? One way to do it is to calculate the difference between the points that make both of these cumulative distributions:

Sample Mean	ECDF	Theoretical CDF	Difference
27.54	0.062	0.043	0.019
27.94	0.125	0.108	0.016
28.08	0.187	0.143	0.044
28.36	0.250	0.232	0.017
28.53	0.312	0.299	0.012
28.61	0.375	0.333	0.041
28.69	0.437	0.369	0.068
28.82	0.500	0.429	0.070
28.92	0.562	0.477	0.085
28.98	0.625	0.505	0.119
29.20	0.687	0.609	0.077
29.50	0.750	0.738	0.011
29.58	0.812	0.768	0.044
30.01	0.875	0.894	0.019
30.21	0.937	0.931	0.005
30.52	1.000	0.968	0.031

Table 3.13: Yellow marks the spot! The differences between distributions.

Table 3.13 comprises four columns: the sample mean observations, our ECDF values, the theoretical CDF values generated based on the assumed normal distribution, and finally, the "Difference" column, which captures the absolute differences between the corresponding ECDF and the CDF values.

The "Difference" column is particularly crucial because it aggregates the deviations of the ECDF from the CDF or the other way around. Now, we need to find a way to synthesize this information into a singular value that can serve as our test statistic; there are

several methods we could employ: for example, we can sum these differences or even sum the squares of these differences. However, our approach is hinted at by the row highlighted in yellow. This specific row shows the maximum difference between the points of the ECDF and the CDF, and this maximum gap is what we'll use to define our test statistic.

The name of this test is the "Jorge Brasil test" in honor of the mathematician Jorge Brasil, who discovered it... I am joking! The actual name of the test is the Kolmogorov-Smirnov test, as they were the brilliant mathematicians who devised the it.

3.4.3 The Day That Kolmogorov and Smirnov Tested the Waters.

The test we are introducing is called the "Kolmogorov-Smirnov" (KS) test; where the test statistic involves comparing CDF's. This metric is represented by the letter D. They decided to compare the supremum (the maximum difference) over all possible values of x between the ECDF and the theoretical one:

$$D_{obs} = \sup_{x} |F_n(x) - F(x)| \tag{3.99}$$

In equation 3.99, D represents the test statistics, $F_n(x)$ is the ECDF of the sample data, $F(x)$ is the CDF of the theoretical distribution, and $\sup_x$ denotes the supremum, that fancy word for maximum difference over all possible x's. Our attention now shifts to that D, where a large value suggests a significant difference between the distributions, while a small value of D shows that these distributions are similar. So, in our particular situation:

$$D_{obs} = 0.119$$

Visually:

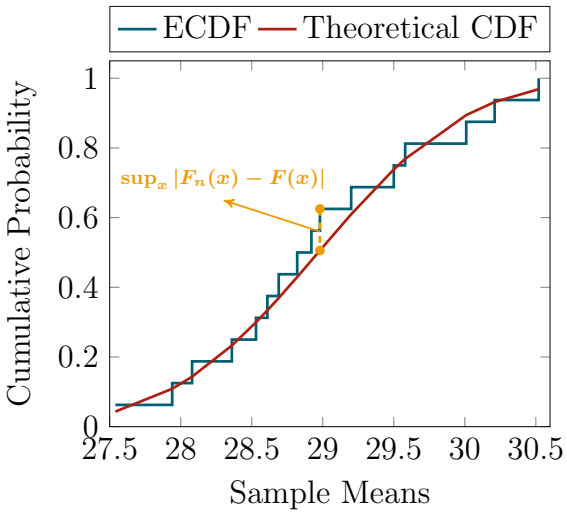


Figure 3.45: Where the ECDF and the CDF part ways.

Figure 3.45 visually represents the KS test idea. The blue line (looking like a staircase) represents our ECDF, whereas the curvy red line showcases the generated value of the CDF. The yellow setup, meaning the two dots and the dashed line, represents the observations with the highest absolute difference, our D. The question now is how sure we are that this supremum difference is significant and not just caused by variance:

$$H_0 : \bar{X} \sim \mathcal{N}(28.97, 0.046) \quad vs \quad H_0 : \bar{X} \not\sim \mathcal{N}(28.97, 0.046) \tag{3.100}$$

Equation 3.100 defines our hypothesis and if we want to make a decision between them, this is where once again the null distribution comes into play! If we have access to the PDF, representing how often a value of D is likely to occur, we can then make a decision on whether we reject or fail to reject H_0.

In this particular instance, we won't have a closed form solution for the PDF of D, just for the CDF. A "closed form" refers to a mathematical expression that can be computed and expressed using a finite number of standard operations; in essence, it was not possible to integrate the PDF. Well, if we don't have a PDF we must find another way to get to a null distribution; otherwise we can't complete our test. For that, we will use a technique called simulation.

3.4.3.1 We Need a Lot of Them: Monte Carlo Simulations.

Simulation is a technique used to mimic or replicate the behavior of real-world systems or processes through mathematical models and computer algorithms. This method enables the study and prediction of complex phenomena without the need for physical experiments. Simulations can operate under controlled conditions to test theories, explore system responses, or predict outcomes, making them a versatile tool across various disciplines. Essentially, by creating a model based on data, we can generate multiple samples from that model to perform statistical inference. There are several types of simulation, but the focus here will be on Monte Carlo simulations. These are a distinct type of simulation that relies on repeated random sampling to obtain numerical results. The term "Monte Carlo" is derived from the randomness associated with casino games in Monte Carlo, Monaco, reflecting the method's reliance on random sampling techniques. Don't do it! You will get wrecked; casinos know all about this.

If we wish to apply this technique to create our null distribution, these are the steps we must follow:

- **Estimate Parameters:** From your original data, calculate the sample mean $\bar{x}$ and sample standard deviation s. These estimates will be used to standardize your data. *Note:* While using these parameters can be practical, be cautious as using estimated parameters from the sample that you are testing can statistically invalidate the test. This is because they make the sample appear more normally distributed than it might actually be.

- **Generate Simulated Samples:** Perform a large number of iterations (e.g., 1000), generating new samples from a normal distribution with known parameters. For an unbiased simulation, use parameters that are not derived from the sample being tested, such as a standard normal distribution $N(0, 1)$.

- **Calculate the D_{obs} Statistic for Each Sample:**

  - Calculate the ECDF: For each simulated sample, compute the empirical cumulative distribution function (ECDF).
  - Calculate the CDF: Use the cumulative distribution function (CDF) of the standard normal distribution.
  - Compute the D statistic: Identify the maximum absolute difference between the ECDF of the simulated sample and the CDF of the normal distribution.

- **Build the Null Distribution:** Aggregate all the D_{obs} statistics from each simulation to form the null distribution. This distribution represents how the D statistic behaves under the null hypothesis that the data comes from the specified normal distribution.

I did this for us (all the code used in this book will be provided), and below is a plot of what this distribution looks like:

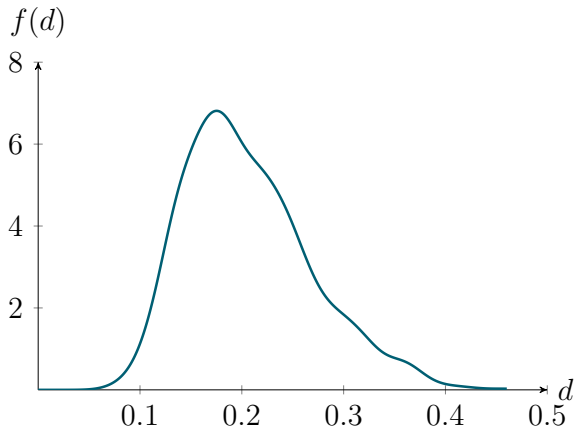


Figure 3.46: Our first Null distribution!

Figure 3.46 illustrates the PDF for our test statistic, D. Now that we have calculated our D_{obs} statistic, and have our null distribution, we need to establish criteria to either accept or reject the null hypothesis. Let's consider what D signifies. This statistic represents the maximum deviation between the empirical cumulative distribution function and the cumulative distribution function of the assumed theoretical distribution. Essentially, if a given value of D_{obs} is acceptable, then any smaller D_{obs} values are also acceptable, as they represent a smaller maximum difference.

So, the crucial question becomes: How likely is it that we will observe a D_{obs} as extreme as or more extreme than 0.119 under the null hypothesis? Yes, indeed, we have the p-value to answer this question, and with this concept comes something we need to define: one-tailed or two-tailed? What is extreme to your KS test? Come out and reveal yourself! Uff, probably this is just me scared of being caught inside that vault by the police. We must nail these tests!

Our statistic reflects the maximum difference between the subjects we are testing, so the higher this metric is, the less likely we are in the presence of the same distribution. Therefore, we are interested in one of the sides of our null distribution, making the KS test for this case a one-tailed one.

Before jumping to the calculations of the p-value we need to define our significance level α and once again, let's go with $\alpha = 0.05$:

$$p\text{-value} = P(D \geq D_{obs} \mid H_0 \text{ is true})$$

$$p\text{-value} = P(D \geq 0.119 \mid H_0 \text{ is true})$$

$$P(D \geq 0.119 \mid H_0 \text{ is true}) = \frac{P(D \geq 0.119 \cap H_0)}{P(H_0)}$$

Calculating the p-value can be done by analyzing our simulated data and counting the number of values that are above or equal to 0.119, and then dividing by the total number of values, or 1000:

$$p\text{-value} = 0.956$$

The p-value of 0.956, which exceeds our significance level $\alpha = 0.05$, suggests that there is no strong evidence to reject the null hypothesis. This indicates that, according to the Kolmogorov-Smirnov test, your data do not exhibit significant deviations from a normal distribution when considering the estimated mean and standard deviation. Consequently, the results of our analysis support the conclusion that the random variable X, representing the sample means, appears to follow a normal distribution.

I have a confession... we didn't need to use monte carlo simulation to create a null distribution so we could compute the p-value.

In the case of the KS test, we have a formula for a treshold:

$$P(D_n \leq D_{obs}) = 1 - 2\sum_{k=1}^{\infty}(-1)^{k-1}e^{-2k^2d^2} \tag{3.101}$$

This formula describes the probability that the KS statistic D_n does not exceed a particular value d, involving an infinite series that illustrates the underlying statistical theory. Although this expression is theoretically important, its complexity makes direct computation challenging. Fortunately, approximations and tables have been developed to simplify the use of this formula, enabling easier application of the KS test in practical scenarios:

$$D_\alpha \approx \sqrt{-\frac{1}{2n}\ln\left(\frac{\alpha}{2}\right)} \tag{3.102}$$

Equation 3.102 outlines the critical value D_α we seek and we can even go a step further and devise a formula for a p-value:

$$p\text{-value} = 2\exp\left(-2nD^2\right)$$

So, if we have that, why engage in simulation? Firstly, it introduces us to a technique—Monte Carlo simulation—that will recur throughout this book. Secondly, not all tests readily provide access to critical regions like the t-test or thresholds such as those depicted in equation 3.102. Fortunately, armed with our newly acquired knowledge, we are now well-prepared to construct a null distribution whenever necessary.

Our discussion aimed to understand how a test statistic is derived and how we can create a null distribution based on it. This understanding is fundamental to any hypothesis test. Now, whenever you encounter the various symbols representing statistics in tests—such as D, Z, and t—you'll recognize that these represent statistics capturing features characteristic of the test. Our data and objectives dictate what test to use. Each of these statistics is associated with a null distribution, a critical region, or a threshold that enables us to make informed decisions regarding our null hypothesis.

Before moving on to the next part of our heist plan, let's go over some specifications of the KS test:

- **Sample size**: Larger samples provide a more accurate approximation of the underlying distribution, making it easier to detect differences between distributions.

- **Data continuity**: The KS test is designed for continuous data. Its application to discrete data or data with many ties (identical values) may not be appropriate, as it can affect the test's accuracy and interpretation.

- **Two-sample and one-sample versions**: In our particular situation we used the KS test to compare an empirical distribution to a theoretical one, but we can also use it to establish whether two samples come from the same distribution. For that, we compare the two ECDF's, which is more used when comparing two dataset and not a dataset to a distribution.

- **Non-parametric**: It does not assume any specific form for the distribution of data, making it applicable to any continuous distribution. Because this characteristic is central to the KS test, it's important to emphasize that although we used it to test for normality, the KS test is versatile enough for other applications as well. For example, it can be used to compare empirical distributions of sample data against other theoretical distributions.

Wait a moment, we mentioned that the test doesn't assume any particular distribution, yet we used a normal CDF. What's happening here? Well, employing non-parametric tests does not prevent us from comparing our data against a theoretical distribution; it simply means that the test does not make assumptions about the underlying data behavior. We could have chosen any other distribution. The term 'non-parametric' indicates flexibility in this choice, as opposed to 'parametric' tests, which bind us to specific distribution assumptions.

This concludes our exploration of our first non-parametric test. While it's true that we began our chapter on hypothesis testing with a two-tailed parametric test, that scenario was relatively straightforward, the t statistic was already defined, and the null distribution

was provided. In contrast, our work here required us to derive these elements ourselves, offering a deeper understanding of the testing process. So what about if we do the same for a parametric test? The first thing to do with any test is to define the hypothesis—and it just so happens we can use the same ones as we had for the KS test:

$$H_0 : \bar{X} \sim \mathcal{N}(28.97, 0.046) \quad vs \quad H_0 : \bar{X} \not\sim \mathcal{N}(28.97, 0.046)$$

Now it is time to define our test statistic. Unlike the Kolmogorov-Smirnov test, which is non-parametric and thus has a test statistic that does not depend on any specific distribution, we now need to define a test statistic that specifically checks for characteristics typical of a normal distribution.

This distribution is best known for its hit song "Gaussian Jingle Bells: A Statistical Carol," but another unique thing about it is the symmetry of its curve. This symmetry suggests that for every data point on one side of the mean, there's a corresponding data point on the other side at approximately the same distance from the mean, creating the bell shape. The bell shape also implies that it is near the mean that we find most of the data points. As we look further away from the center, we see fewer and fewer data points in either direction, which characterizes the thinner tails, where less commonly observed values are present.

If we receive a sample of data and want to verify if it exhibits Gaussian characteristics, one effective method is to start by sorting the data from the smallest to the largest value. This arrangement allows us to easily assess the central tendency — we can check whether the median (or the average of the two central values in an even-sized dataset) aligns closely with the mean of the dataset. Next, we examine the distances of data points from the mean. Since the data is ordered, we expect to find the largest distances at the ends of the dataset and smaller, more frequent distances near the center. This pattern reflects the Gaussian distribution's characteristic bell shape, where data points are more concentrated around the mean and taper off towards the tails.

Another significant feature of the normal distribution is the size of its tails. Merely ordering the sample will not capture this behav-

ior adequately, so we need a method to incorporate this information along with the ordered sample. What if we create a coefficient, say a_i, that quantifies each observation x_i's contribution to the shape of the distribution? Consider this: in a Gaussian distribution, the data is predominantly clustered around the mean; hence, these central points contribute more significantly to the distribution's shape, so their a's will be higher than those in the tails. Essentially, we're talking about creating a weighted sum that represents the individual contribution of each observation x_i to the global shape of the distribution:

$$W = \sum_{i=1}^{n} a_i x_i \tag{3.103}$$

Equation 3.103 represents a weighted sum of the observations x_i. Before later discovering our assumptions are a whole load of bollocks, let's do a quick intuitive check. What do we expect from W if we have a sample that is a normal distribution and we apply 3.103 what will come? Strawberries and cream? Sure, that would be delightful; however, if the data genuinely follows a Gaussian distribution, most of the contribution to the shape comes from the central points, meaning their a's have higher values than those corresponding to the points in the tails. Thus, we expect a W value to be close to the mean.

What happens if the data is not normally distributed? Can we simply expect that W will deviate significantly from the mean? This prompts the question of what exactly "far" means. It seems we are on the right track, but we still need to refine Equation 3.103 to develop a robust statistic. For instance, 3.103 is not yet normalized, and we have yet to determine how to calculate the weights, $a's$. To gauge the extent of the deviation, we can take cues from the KS test, which uses simulations to generate numerous ECDFs and CDFs. This approach has proven effective in understanding the behavior of the D statistics. If we apply a similar strategy to generate a null distribution for our statistic W, we will have sufficient data to compute the p-value, thus providing comprehensive answers. However, our first step must be to devise a clever method for calculating those $a's$.

These coefficients must reflect the behavior of a normal distribution, meaning that if we analyze a bunch of these distributions, we can calculate the a's based on them. This way, we will have coefficients that are completely biased towards the distribution we want to check; don't worry, bias here is a good thing! For a dataset from a non-normal distribution, which might be skewed, heavy-tailed, the behavior of W can be markedly different. If the distribution is skewed, because the weights a are calculated based on normality, they will not compensate for skewness and W will be pulled towards the tail with more extreme values. Similarly, for a heavy-tailed distribution, W might overemphasize the tails if the weights are uniform or not sufficiently decreasing towards the ends.

Now comes the fun part; how will we compute those a's? Well, if we are set on computing a's based upon the normal distribution, why don't we create several samples from different normal distributions with $\mu = 28.97$ and $\sigma^2 = 0.046$:

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
Obs 1	**-0.47**	**-1.91**	**-1.42**	**-1.22**	**-1.96**
Obs 2	0.23	-1.72	-1.41	-1.06	-1.48
Obs 3	-0.23	-1.01	-1.15	-0.60	-1.33
Obs 4	**-0.14**	**-0.56**	**-0.91**	**-0.60**	**-0.30**
Obs 5	**0.50**	**-0.47**	**-0.5**	**-0.29**	**-0.12**
Obs 6	**0.65**	**-0.46**	**-0.23**	**-0.01**	**0.17**
Obs 7	0.77	0.24	0.07	0.38	0.20
Obs 8	1.52	0.31	0.11	0.82	0.21
Obs 9	**1.58**	**0.54**	**1.47**	**1.85**	**0.74**

Table 3.14: A colourful representation of the different parts of the bell curve.

In table 3.14, we present five samples, each comprising nine ordered observations drawn from a standard normal distribution Z. The reason we use a standard normal distribution is to keep everything in the same magnitude, so as to facilitate comparisons. The colors are the works of art of my friend's kid; he got hold of the book and colored the numbers that way. I thought it looked cute and left it. On a more serious note, the color scheme has some meaning, so let's start interpreting it:

Blue Observations (Tail Analysis): Each sample's first and ninth observations are pivotal for examining the distribution's tails. We expect these tail observations to exhibit mirror-like behavior across all samples in a perfectly symmetrical, normal distribution.

This symmetry, or the lack of it, can provide insights into the distribution's adherence to normality. A systematic analysis of these blue observations across samples helps us assess the distribution's tail behavior, crucial for normality tests.

Yellow Observations (Central Spread): The observations colored in yellow, specifically the fourth and sixth, are immediately adjacent to the median. Analyzing the variance between these observations within each sample gives us a measure of the distribution's spread around its center. For a normal distribution, we anticipate a consistent spread, indicative of the lack of the distribution's kurtosis and skewness. This analysis helps us understand how concentrated or dispersed the data is around the median.

Red Observation (Mean and Centrality): The mean, highlighted in red, is the central point of each sample. Its relationship with both the tails (blue observations) and the adjacent central points (yellow observations) provides a comprehensive view of the distribution's symmetry. Comparing the median to the tail observations and observing how it fits within the overall spread of the data can reveal deviations from normality, such as skewness.

And why stop there? We could do this pairwise analysis between all the observations for all the samples and capture a ton of information about how the points of the standard normal distribution vary with each other, COME ON!... oh crap, sorry for the overexcitement there; I just remembered that the only way we know how to compute some metric related to variance only works for one variable and now we need it for a pair. Nevertheless, let's explore this formula again and see if we can extract even more value from it:

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{3.104}$$

Where $\bar{x}$ represents the sample mean of X and n the sample size. But now, we are not just working with a random variable X, we are doing a pairwise analysis, so we also need another random variable Y:

$$s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{3.105}$$

Equally, $\bar{y}$ represents the sample mean of Y and n the sample size. As we are looking to capture the pairwise variance, we need to find a way to relate 3.104 and 3.105. Well, there are four ways we can do that: addition, subtraction, multiplication, or division—yes, that's right, elementary math, the best one! Well, summing and subtracting we can exclude, as that won't tell us anything meaningful about the relationship between the variances, as these operations would simply combine or differentiate the absolute variance values without providing insight into their interaction. The battle is really between multiplication and division. Dividing variances will provide us with a ratio indicating how one variance scales relative to another, rather than their interactive dynamics. So we are left with multiplying the fellows and this will evaluate the compounded effects of both these metrics, which is what we are looking for. So larger numbers in this operation reflect a strong relationship as both random variables exhibit a high deviation from their means:

$$H(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot (y_i - \bar{y})^2 \tag{3.106}$$

Equation 3.106 seems to somewhat reflect the behavior we are looking for but those squares might be an issue. The rationale behind squaring differences in variance calculation is to focus on the magnitude of deviation from the mean, disregarding whether these deviations occur above or below this average value. This approach suits scenarios where the direction of deviation, whether a data point lies to the left or right of the mean, is not a concern; our primary interest lies in quantifying the extent of this deviation.

However, the context shifts when we consider the relationship between two variables. Here, the direction of change holds significance because we are evaluating how two random variables move in relation to each other. For instance, if one variable increases while the other decreases, this inverse relationship is crucial to our understanding. Simply squaring the differences, as in variance, would obscure such directional relationships.

Well if the squares are a problem, let's get rid of them; by doing so, we preserve the directionality of the relationship between the two variables, enabling us to capture whether they move together (both

increasing or decreasing) or in opposite directions (one increases while the other decreases):

$$\text{Cov}(X, Y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y}) \tag{3.107}$$

Which is equivalent to:

$$\text{Cov}(X, Y) = E\left[(X - E[X])\,(Y - E[Y])\right]$$

3.4.4 It Takes Two to Tango, but Many Can Co-Vari(ance) .

With equation 3.107, we have a new concept, called the covariance; this fellow captures a linear relationship between X and Y. With $(x_i - \bar{x})$ and $(y_i - \bar{y})$ we have both X and Y deviations from their means captured. So, when we multiply these deviations and average them across all observations, the formula captures the essence of how X and Y change together:

- **A positive covariance**, $\text{Cov}(X, Y) > 0$ indicates that X and Y tend to increase and decrease together, suggesting a positive relationship.
- **A negative covariance**, $\text{Cov}(X, Y) < 0$ suggests that as X increases, Y tends do decrease, and vice verse, suggesting an inverse relationship.
- **A covariance near zero** suggests no linear relationship between X and Y.

You want an example of an application? Aye aye captain, move aside, as it is coming! As we've seen, the Antwerp Diamond Heist, carried out over the weekend of February 15-16, 2003, was meticulously planned to coincide with the Proximus Diamond Games, a prominent tennis tournament featuring stars like Venus Williams. This event not only captivated the city's attention but also diverted security resources and public focus away from the Antwerp Diamond

District. Additionally, the timing immediately after Valentine's Day could have further reduced vigilance, as people were distracted by holiday celebrations. (The police did go on to investigate Cupid as an accomplice).

Executing the heist over the weekend was a strategic move, taking advantage of the quieter period when the Diamond District's activity was at its lowest and ensuring that the theft would not be discovered until the return to work on Monday. This delay provided the thieves with a crucial window in which to make their escape, showcasing their deep understanding of the optimal timing to minimize detection and maximize their chances of a successful heist.

Because the police will be on high alert during these types of events, we need to come up with a different strategy on how to plan the day of our heist. Well, I know for a fact that most people do not enjoy being in the rain, especially the ones who are working. So I went out and collected a small dataset that showcased the amount of rainfall intensity vs. the number of police officers in the street surrounding the Diamond Center at any given moment. This data was collected over the weekend because it is the time frame we are most likely going to select:

Day	Rainfall intensity (X) (mm)	Security agents on patrol (Y)
1	0	20
2	5	18
3	10	15
4	3	19
5	8	16

Table 3.15: Maybe if it rains the guards will stay in.

It would be intriguing to explore how the number of agents patrolling the streets responds to changes in rainfall intensity. Do more agents patrol when it rains harder, or does the number decrease? To investigate this, we can apply the covariance to examine any potential relationship between these two variables. Given the data for rainfall intensity X and security agents on patrol Y, we can start by calculating their means:

$$\bar{x} = \frac{0+5+10+3+8}{5} = 5.2$$
$$\bar{y} = \frac{20+18+15+19+16}{5} = 17.6$$

Then, we compute the covariance between X and Y:

$$\begin{aligned}\text{Cov}(X,Y) &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) \\ &= \frac{1}{4}[(0-5.2)(20-17.6)+(5-5.2)(18-17.6)+ \\ &\quad (10-5.2)(15-17.6)+(3-5.2)(19-17.6)+ \\ &\quad (8-5.2)(16-17.6)] \\ &= -8.15\end{aligned}$$

Sooo, a negative covariance suggests that the number of agents and the rain intensity have a negative relationship, meaning that when the number of agents decreases, the rain intensity increases or vice versa. This is a good insight. We might have to sustain some rain, but fewer agents will be patrolling the streets! The motivation to comprehend this new tool for statistical analysis stems from the need to understand the pairwise variance of a sequential sample of points. It's crucial to remember that our goal is to derive a method to relate several variables to construct our test statistic. We have more than just two samples of points, as shown in table 3.14, so let's try to understand what happens when we have more than two variables.

Yeah, we need one more data point, but don't worry, I will take one for the team, go outside in the rain, and collect an extra one. It just so happens that I collected the time when I gathered those data points, so let's use it:

Day	Rainfall Intensity (mm) X	Security Agents on Patrol Y	Hour of Day Z
1	0	20	16
2	5	18	18
3	10	15	3
4	3	19	2
5	8	16	6

Table 3.16: Will you stay inside already?!?!

When our analysis expands to include three variables, X,Y, and Z, because the covariance formula can only accommodate a pair of variables, we can't simply apply it. One workaround will be to calculate the covariance for each pair of variables separately: $\text{Cov}(X,Y)$,

$\text{Cov}(X, Z)$, and $\text{Cov}(Y, Z)$, this way we can examine the linear relationship between each pair of variables:

$$\bar{z} = \frac{16 + 18 + 3 + 2 + 6}{5} = 9$$

Given the means of the variables, we can calculate the covariance and variance as follows:

$$\begin{aligned}\text{Cov}(X, Z) = \frac{1}{4}\,[&(0 - 5.2)(16 - 9) + (5 - 5.2)(18 - 9) + \\ &(10 - 5.2)(3 - 9) + (3 - 5.2)(2 - 9) + (8 - 5.2)(6 - 9)] \\ &= -15\end{aligned}$$

$$\begin{aligned}\text{Cov}(Y, Z) = \frac{1}{4}\,[&(20 - 17.6)(16 - 9) + (18 - 17.6)(18 - 9) + \\ &(15 - 17.6)(3 - 9) + (19 - 17.6)(2 - 9) + (16 - 17.6)(6 - 9)] \\ &= 7.75\end{aligned}$$

Cool, let's put all the covariances side by side so we can have an overall picture and try to draw some conclusions:

$$\text{Cov}(X, Y) = -8.15, \quad \text{Cov}(X, Z) = -15, \quad \text{Cov}(Y, Z) = 7.75$$

For our domain of analysis, one of the pairs, the pair (X, Z), is not of much interest. The covariance between the rainfall intensity and the time of day does not bring us any value in terms of planning the heist. Now, the opposite can be said when we analyze the pair (Y, Z). The covariance of this pair is positive, $\text{Cov}(Y, Z) = 7.75$, which means that, as the time of day decreases, the number of patrolling guards will follow the same trend. Equally, as the time of day increases, the number of uniforms roaming around will also get higher. Well, if we now do a joint analysis of $\text{Cov}(Y, Z)$ and $\text{Cov}(X, Y)$, we conclude that the best configuration for a heist would be a rainy day in the early hours of the morning.

Do you hear that? It's the sound of the covariance formula screaming for a matrix representation. We know that in a matrix each entry finds its place at the intersection of a specific row and column. Given the pairwise nature of covariance calculations, adopting a matrix format emerges as a natural approach. Let's explore this

concept:

$$\Sigma = \begin{array}{c} \\ X \\ Y \\ Z \end{array} \begin{array}{c} \begin{array}{ccc} X & Y & Z \end{array} \\ \begin{pmatrix} \text{Cov}(X,X) & \text{Cov}(X,Y) & \text{Cov}(X,Z) \\ \text{Cov}(Y,X) & \text{Cov}(Y,Y) & \text{Cov}(Y,Z) \\ \text{Cov}(Z,X) & \text{Cov}(Z,Y) & \text{Cov}(Z,Z) \end{pmatrix} \end{array} \tag{3.108}$$

Each entry of matrix 3.108 represents the pairwise covariance of the corresponding column and row variables. Now, there are a few entries we still have not calculated, namely, all those savages that make the diagonal:

$$\text{Cov}(X,X), \quad \text{Cov}(Y,Y), \quad \text{Cov}(Z,Z)$$

Let's use formula 3.107 and expand the matrix with one of those specific covariances that form our matrix diagonal, for example:

$$\text{Cov}(X,X) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (x_i - \bar{x})$$

We can simplify that a bit:

$$\text{Cov}(X,X) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

But that is the variance!

$$\text{Cov}(X,X) = \text{Var}(X)$$

It actually makes sense; how does X vary with X? This is exactly the variance of X. So, our diagonal will be constituted by the variance of each of the random variables. The only pairs missing are the pairs below the diagonal, but this will be the same as their correspondent counterpart, meaning $\text{Cov}(X,Y) = \text{Cov}(Y,X)$. To understand why, let's expand the formula for those two covariances:

$$\text{Cov}(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})$$

$$\text{Cov}(Y,X) = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y}) \cdot (x_i - \bar{x})$$

We know that the multiplication is a commutative operation therefore:

$$\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$$

And all of a sudden we are nearly ready to populate that matrix, the only thing missing are the variances for each of the random variables:

$$\begin{aligned}\mathrm{Var}(X) = \frac{1}{4}\Big((0-5.2)^2 + (5-5.2)^2 + (10-5.2)^2 \\ + (3-5.2)^2 + (8-5.2)^2\Big) = 15.7\end{aligned}$$

$$\begin{aligned}\mathrm{Var}(Y) = \frac{1}{4}\Big((20-17.6)^2 + (18-17.6)^2 + (15-17.6)^2 \\ + (19-17.6)^2 + (16-17.6)^2\Big) = 4.3\end{aligned}$$

$$\begin{aligned}\mathrm{Var}(Z) = \frac{1}{4}\Big((16-9)^2 + (18-9)^2 + (3-9)^2 \\ + (2-9)^2 + (6-9)^2\Big) = 56\end{aligned}$$

Now we are ready!

$$\Sigma = \begin{array}{c} \\ X \\ Y \\ Z \end{array} \begin{array}{c} \begin{array}{ccc} X & Y & Z \end{array} \\ \begin{pmatrix} 15.7 & -8.15 & -15 \\ -8.15 & 4.3 & 7.75 \\ -15 & 7.75 & 56 \end{pmatrix} \end{array} \tag{3.109}$$

More generically, if we let $X_1, X_2, \ldots, X_N$ be random variables, the covariance matrix Σ is defined as:

$$\Sigma = \begin{array}{c} \\ X_1 \\ X_2 \\ \vdots \\ X_N \end{array} \begin{array}{c} \begin{array}{cccc} X_1 & X_2 & \cdots & X_N \end{array} \\ \begin{pmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1,X_2) & \cdots & \mathrm{Cov}(X_1,X_N) \\ \mathrm{Cov}(X_2,X_1) & \mathrm{Var}(X_2) & \cdots & \mathrm{Cov}(X_2,X_N) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_N,X_1) & \mathrm{Cov}(X_N,X_2) & \cdots & \mathrm{Var}(X_N) \end{pmatrix} \end{array}$$

Why bother? Good question but there is a very good reason. The covariance matrix is... well, it is a matrix and this in itself means

that we can transition to the world of linear algebra, where we can leverage a set of techniques to derive further insights. Furthermore, the covariance matrix has distinctive properties: it is symmetrical and positive definite. Such matrices boast unique features, including positive and real eigenvalues. This principle underlies the popular Principal Component Analysis (PCA) algorithm, which decomposes the covariance matrix to identify patterns and simplify data complexity. For those interested in exploring this technique and the broader field of linear algebra, the first volume of the series offers an in-depth look.

Before moving on to compute our a's, there is something we should explore. We often normalize metrics so we can do comparisons with metrics calculated on different scales. So what will happen if we apply this to the covariance, meaning:

$$\frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} \tag{3.110}$$

This guy 3.110 refers to a well-known concept that is often used and abused, especially when someone is trying to look smart, we are talking about the correlation.

3.4.4.1 That Covariance is Going Out of Bounds: Correlation.

The 'correlation' ($\rho_{X,Y}$ or r) is a normalized version of covariance, scaled by the standard deviations of the variables involved, thus providing a dimensionless value that measures the strength and direction of their linear relationship. It is defined as:

$$\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

Where σ_X and σ_Y are the standard deviations of X and Y, respectively. Correlation values range from -1 to 1, where 1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear relationship. While variance and covariance deal with the spread and joint variability of data, correlation normalizes this relationship, making it independent

of the units of measurement and allowing for comparison between different datasets. Thus, correlation can be seen as a standardized form of covariance, not variance:

- **Covariance** is utilized to identify the direction of the relationship between two variables. It indicates whether increases in one variable coincide with increases (positive covariance) or decreases (negative covariance) in another, which is particularly relevant when the scale of the variables matters and they are measured in the same units. This helps to understand the magnitude of their relationship.

- **Correlation**, on the other hand, is used to quantify the strength and direction of a linear relationship between variables, independent of their units. This makes it ideal for comparing relationships across different scales and datasets. It is widely used in fields like finance, psychology, and research, providing a standardized way to assess the degree of relationship between variables.

Now is time to go back to our main task of computing our a's. For that we explore the structure of a covariance matrix, as we can efficiently encapsulate the relationships among the variables presented in table 3.14. Our goal is to methodically analyze the data row-wise, computing the covariance between all possible pairs of these variables. To illustrate this, let's consider Z_1, which encapsulates the following set of observations:

$$-0.47, -0.91, -1.42, -1.22, -1.96$$

For clarity, we designate each row of observations as Z_i, where i represents the row number. This notation allows us to construct a covariance matrix, Σ, which systematically captures the variance within each row and the covariance between every pair of rows. The

matrix is structured as follows:

$$\Sigma = \begin{array}{c} \\ Z_1 \\ Z_2 \\ Z_3 \\ Z_4 \\ Z_5 \end{array} \begin{array}{c} \begin{array}{ccccc} Z_1 & Z_2 & Z_3 & Z_4 & Z_5 \end{array} \\ \begin{pmatrix} \mathrm{Var}(Z_1) & \mathrm{Cov}(Z_1, Z_2) & \mathrm{Cov}(Z_1, Z_3) & \mathrm{Cov}(Z_1, Z_4) & \mathrm{Cov}(Z_1, Z_5) \\ \mathrm{Cov}(Z_2, Z_1) & \mathrm{Var}(Z_2) & \mathrm{Cov}(Z_2, Z_3) & \mathrm{Cov}(Z_2, Z_4) & \mathrm{Cov}(Z_2, Z_5) \\ \mathrm{Cov}(Z_3, Z_1) & \mathrm{Cov}(Z_3, Z_2) & \mathrm{Var}(Z_3) & \mathrm{Cov}(Z_3, Z_4) & \mathrm{Cov}(Z_3, Z_5) \\ \mathrm{Cov}(Z_4, Z_1) & \mathrm{Cov}(Z_4, Z_2) & \mathrm{Cov}(Z_4, Z_3) & \mathrm{Var}(Z_4) & \mathrm{Cov}(Z_4, Z_5) \\ \mathrm{Cov}(Z_5, Z_1) & \mathrm{Cov}(Z_5, Z_2) & \mathrm{Cov}(Z_5, Z_3) & \mathrm{Cov}(Z_5, Z_4) & \mathrm{Var}(Z_5) \end{pmatrix} \end{array}$$

What exactly do we want to achieve with this matrix? Ah, the a_i's! Yes, the a_i coefficients are a crucial piece of the puzzle for assessing the normality of sample data. The a's are designed to weight the contribution of each data point or certain statistical characteristics (like order statistics) to the overall test statistic. The objective is to enhance the test's ability to detect specific features or deviations in the data distribution, such as the presence of heavier tails or a non-normal skewness. If only we had a tool that could give us a measure of the amount of variance from a matrix... wait a second...

Recall from linear algebra that we have a powerful tool for identifying the direction of highest variance? The eigenvectors. We can decompose the covariance matrix Σ and selecting the eigenvector associated with the largest eigenvalue. Then, since variance is a key measure of spread and deviation from the mean, we can use the eigenvalues as the a_i coefficients naturally aligning with the goal of emphasizing areas of significant statistical fluctuation.

When we have the entries of the eigenvector, there are a few extra steps we must take before being ready to apply the a_i's to our test metric: the first step is that they need to be normalized to ensure that their combined square sums to one. Next, because the normal distribution is symmetrical, the contribution of each of the points weighted by the a_i coefficients must reflect this symmetry. Importantly, each coefficient a_i is calculated to be the negative mirror of a_{n+1-i}, i.e., $a_i = -a_{n+1-i}$. This asymmetrical arrangement is essential, as it amplifies the test's ability to detect skewness and other deviations from normality, particularly emphasizing the tails of the distribution. By structuring the coefficients this way, where each is the negation of its symmetrical counterpart from the opposite end of the sequence, the test not only ensures a balanced assessment across

the dataset but also enhances sensitivity to asymmetrical properties, which is crucial for robust statistical analysis of normality.

So, we have learned a new and valuable concept called covariance. Consequently, we get a methodology to compute the a's. We do not need to calculate these coefficients manually, as we can again use software or pre-calculate a's that we can access on probability tables. It is essential to mention that these were computed with more than five samples and nine observations. In our case, the small number of samples and observations was what we have been using to understand the main idea behind the test.

So our formula 3.103 seems to be getting more robust, but there is one thing that stands out. The a's were computed with samples of standardized distributions, but our formula does restricts our data to be standardized. The values of x_i are just the ordered sample of data. Well, we must normalize these bad boys and also formally introduce a new test. Please, meet the Shapiro-Wilk test:

3.4.5 A SW? South West? Ah the Shapiro-Wilk Test!

The Shapiro-Wilk test, assesses normality by comparing the sample distribution to a normal distribution, with adjustments for sample size. The test statistic is normalized by the sample variance to account for scale differences. Importantly, the numerator is squared, a strategic choice that emphasizes deviations from normality. This squaring magnifies the impact of outliers and tail behaviors, such as skewness and kurtosis, on the test statistic. By doing so, it highlights features that differ from those expected under a normal distribution, providing a robust measure of normality. The W statistic for this test is the normalized version of the weighed sum we defined previously:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_i\right)^2}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2} \tag{3.111}$$

Where:

- x_i are the ordered sample values, from the smallest to largest.

- $\bar{x}$ is the sample mean.
- a_i are the coefficients calculated from the means, variances, and covariances of the order statistics of a sample from a standard normal distribution. These coefficients essentially weight the ordered sample values in the numerator of W.

Well what about if we bring back our sample and conduct the test on it?

$$27.54, 27.94, 28.08, 28.36, 28.53, 28.61, 28.69, 28.82$$

$$28.92, 28.98, 29.20, 29.50, 29.58, 30.01, 30.21, 30.52$$

The denominator of formula 3.111 calls for a variance. First step, the sample mean:

$$\bar{X} = 28.97$$

Following:

$$\begin{aligned}\sum_{i=1}^{16}(x_i - \bar{x})^2 =&(27.54 - 28.97)^2 + (27.94 - 28.97)^2 + \\ &(28.08 - 28.97)^2 + (28.36 - 28.97)^2 + (28.53 - 28.97)^2 + (28.61 - 28.97)^2 + \\ &(28.69 - 28.97)^2 + (28.82 - 28.97)^2 + (28.92 - 28.97)^2 + (28.98 - 28.97)^2 + \\ &(29.20 - 28.97)^2(29.50 - 28.97)^2 + (29.58 - 28.97)^2 + (30.01 - 28.97)^2 + \\ &(30.21 - 28.97)^2 + (30.52 - 28.97)^2 = 10.42\end{aligned}$$

Now for the famous a's we must calculate them using the following relationship:

$$a_i = -a_{n+1-i} \tag{3.112}$$

Let's expand the formula 3.112 for a couple of i's, say $i = 1$ and $i = 2$:

$$a_1 = -a_{16+1-1}, \quad a_2 = -a_{16+1-2}$$

$$a_1 = -a_{16}, \quad a_2 = -a_{15}$$

If we follow that logic, we will end up with 16 coefficient a's with eight distinct absolute values. As in many cases in statistics, there are tables we can consult or software we can use to have the necessary values for the required coefficients. After some research on the internet, I got conflicting values for the the a's. Therefore,

I will provide the expansion of the formula required to compute W and then use software for the computations:

$$\left(\sum_{i=1}^{16} a_i \cdot x_i\right)^2 = (a_1 \cdot x_1 + a_2 \cdot x_2 + a_3 \cdot x_3 + a_4 \cdot x_4 + a_5 \cdot x_5 + a_6 \cdot x_6 + a_7 \cdot x_7 + a_8 \cdot x_8 + a_9 \cdot x_{(9} + a_{10} \cdot x_{10} + a_{11} \cdot x_{11} + a_{12} \cdot x_{12} + a_{13} \cdot x_{13} + a_{14} \cdot x_{14} + a_{15} \cdot x_{15} + a_{16} \cdot x_{16})^2$$

Our W is:

$$W = 0.935$$

We have arrived at our test statistic value, so now it is decision time, we need to decide whether we have enough evidence to reject or accept H_0... what H_0? Let me refresh our memories:

$$H_0 : \bar{X} \sim \mathcal{N}(28.97, 0.046) \quad vs \quad H_0 : \bar{X} \not\sim \mathcal{N}(28.97, 0.046)$$

Now we need our hypothesis test detective, the framework that provides us with the evidence; yeah, that is right, our pal, the null distribution. We will use the same methodology as we did with the analogous distribution of the D statistics to access it. Let's get 1,000 normal distributions of 16 observations each and compute their W's. By doing so, we will be able to plot a histogram:

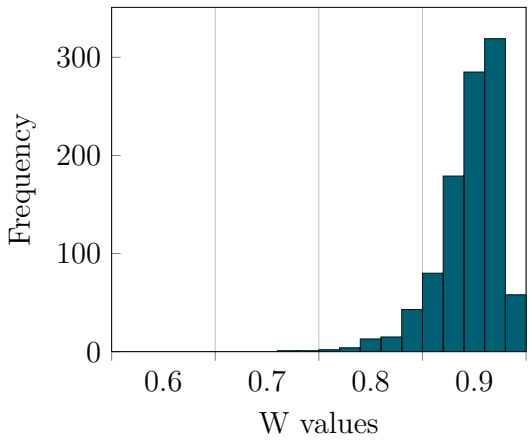


Figure 3.47: The empire state W.

Figure 3.47 is a histogram based on 1,000 different W statistics generated from our experiment. The data predominantly clusters

between 0.8 and 1, indicating that values approaching 0.8 or lower become increasingly rare for W statistics derived from normally distributed samples.

Do we accept or reject our value of $W = 0.935$? Once again we need to answer the question of how likely are we to observe a value as extreme as or more extreme than $W = 0.935$.

Let's start by understanding what characterizes extremeness in the case of the SW test and therefore the value of W. Is it moving to the right side and therefore towards 1? Or the opposite, meaning having a W closer to 0. The plot suggests that extremeness will occur if we have W's nearing 0 as the majority of the mass is around 0.9, but let's confirm this assumption by looking at the formula for W.

A W statistic close to 1 indicates that the sample distribution closely matches a normal distribution, as the numerator—representing a weighted sum of ordered sample values, approximates the sample's total variance in the denominator. Conversely, a W value nearing 0 signals a significant deviation from normality. This happens when the weighted sum of ordered sample values significantly deviates from the expected pattern of a normal distribution, leading to a numerator that is much smaller than the denominator. Such a scenario highlights cases of non-normality within the data, such as excessive skewness or kurtosis, which are critical for statistical analyses predicated on the assumption of normality.

Having a clear view on what extremeness means in this particular context the concept, we're now positioned to leverage a significance level, α, to make decisions about the null hypothesis. The p-value quantifies the probability of observing a W statistic as extreme as, or more so than, our specific value of $W = 0.935$. In the context of our analysis, this metric underscores the likelihood of such extremeness under the assumption that the null hypothesis is true.

Setting a significance level, such as $\alpha = 0.05$, provides a critical benchmark for this evaluation. A p-value less than α suggests that the degree of extremeness observed, embodied in our case by a W statistic significantly deviating from 1, would be unlikely to occur

if the null hypothesis were true. Consequently, observing a p-value below this threshold grants us statistical grounds to reject H_0, supporting the conclusion that our sample does not follow a normal distribution:

$$P(W \leq 0.935 \mid H_0 \text{ is true}) \tag{3.113}$$

Let's expand equation 3.113:

$$P(W \leq 0.935 \mid H_0 \text{ is true}) = \frac{P(W \leq 0.935 \cap H_0)}{P(H_0)}$$

But we just stated that H_0 was true for the purpose of the test so:

$$P(W \leq 0.935 \mid H_0 \text{ is true}) = P(W \leq 0.935)$$

That we can compute! We just have to go back to our sample of 1,000 generated W's and count the number of them that are smaller than or equal to 0.935 and divide that number by the total number of W's:

$$P(W \leq 0.935) = 0.554$$

So our p-value is higher than our level of significance, suggesting that we do not have sufficient evidence to reject the null hypothesis for our sample. In the context of the Shapiro-Wilk test, this outcome indicates that the distribution of our sample does not significantly deviate from a normal distribution, based on the criteria set by our chosen level of significance α. Therefore, we can conclude that the sample data is consistent with coming from a normally distributed population.

Just like we did with the KS test, let's go over the features of the SW test:

- **Sensitivity to sample size**: The SW test is highly sensitive to the sample size. It is most effective for small to medium-sized samples (typically $n \leq 50$), but it can be used for sample sizes of up to around 2,000. The sensitivity of the SW test means that it can detect even small departures from normality, making it particularly useful for rigorous normality testing in smaller datasets.

- **Parametric nature**: The Shapiro-Wilk (SW) test is a parametric test because it assumes that the data should follow a normal distribution. The test evaluates the goodness of fit of the data to this expected normal distribution.

- **Limitations and considerations**: While the SW test is powerful for detecting deviations from normality, especially in small samples, it can be overly sensitive in large samples, leading to rejection of the normality hypothesis for trivial deviations that may not impact practical analysis.

When applying the Kolmogorov-Smirnov test and the Shapiro-Wilk test to assess the normality of our dataset, we encountered similar outcomes. Both the KS test and the SW failed to provide enough evidence to reject H_0, suggesting our data follows a normal distribution.

Upon looking deeper into the KS test specificities, we hypothesized that the sample size might influence these divergent results. We discovered that, while the SW test is particularly adept at identifying deviations from normality in small sample sizes, the KS test's effectiveness is not as dependent on sample size, making it less sensitive in such contexts. Interestingly, this isn't the first time we've encountered the notion of sample size. When introducing the Central Limit Theorem (CLT), which we touched upon earlier, the concept of sample size was also present. This sample size shenanigans seems relevant; so let's explore it.

3.5 Let's keep it real, (Sample) Size Matters!

If we think about most of the formulas introduced throughout this book, it becomes apparent that nearly all of them incorporate an n, symbolizing sample size. This just highlight its impact on the various techniques we have discussed. Beyond the theoretical implications, the sample size also carries significant practical conse-

quences in commercial settings, particularly concerning cost and the potential for challenging discussions with product team colleagues.

Imagine being tasked with conducting a statistical test, only to realize that achieving the ideal sample size would exceed the timeframe promised for delivering results. I can assure you that, if you wish to work on a commercial setting with a job title that has the word "data" in it, you will most likely face that conversation, and the only thing I can say after having had multiple of those conversations is——best of luck, my friend.

Let's start with the confidence intervals and see how the sample size impacts them; and what better way to do than with an equation? After all, if we are worried about n we better see it!

$$\text{CI} = \bar{x} \pm z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \tag{3.114}$$

Equation 3.114 is the formula we originally deduced for a confidence interval, only to realize we were missing the population standard deviation σ. So we decided to go ahead and use the sample standard deviation instead, to come to the conclusion that the standard normal distribution was no longer the best choice. That brought us to the t distribution:

$$\text{CI} = \bar{x} \pm t_{\frac{\alpha}{2},n-1} \cdot \frac{s}{\sqrt{n}} \tag{3.115}$$

Given our sample size of $n = 16$, we opted for the t distribution with 15 degrees of freedom when constructing the confidence interval. The rationale for selecting the t distribution over the standard normal Z distribution primarily hinges on the wider tails of the t distribution. The latter reflect the greater variability expected when working with smaller samples, thus providing a more conservative estimate for the confidence interval. As the sample size n grows, the sample standard deviation s is likely to converge towards the population standard deviation σ. Well, graphically this must mean that the tails of the PDF of the t distribution should get closer and closer to the PDF of Z if we increase n and consequently the degrees of freedom of that t distribution bad boy:

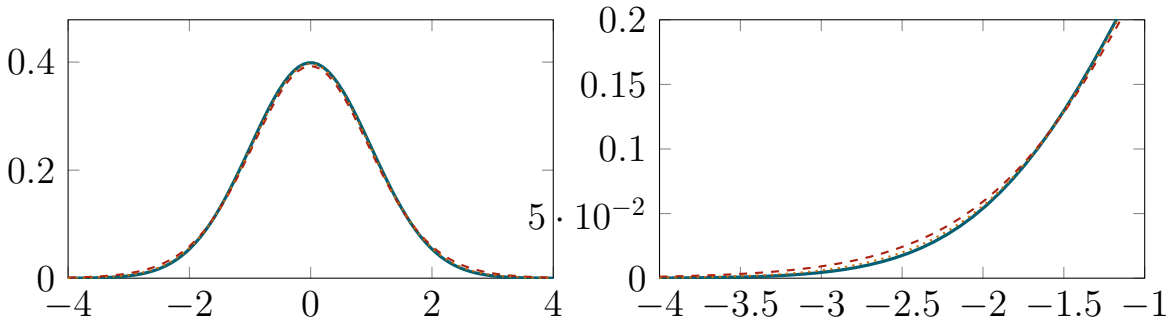


Figure 3.48: The curves on the left and the zoomed-in tails on the right.

In the graph 3.48, we see the representation of a little experiment I ran for us... popcorn with chocolate? It is not that, but, I have not one but two graphs! No bueno?

The left plot illustrates three curves, each representing a different probability density function. The standard normal distribution is depicted by the blue curve. The red and orange curves correspond to t distributions with 15 and 45 degrees of freedom, respectively.

Given the dense nature of the plot and our primary interest in the behavior of the distribution tails, there is a zoomed-in view of the left tail in the right panel. With a closer look we can see that has the degree of freedom from the t distribution increases the closer it gets to the normal Z distribution. Specifically, the tail of the t distribution with 15 degrees of freedom (the red curve) is further from the Z distribution compared to the t distribution with 45 degrees of freedom (the orange curve).

So, it is evident that the sample size n impacts our confidence interval, but how exactly? Increasing the sample size makes the sample standard deviation s closer to σ. Thus we are capturing more variance with the t distribution and, therefore, will be more confident in the interval, making it smaller. Remember that the smaller the range of our interval, the better it will be, as we have less values for our mean to take. To do a quick verification let's compute the width of two intervals, one for a sample size of 16 and another for one of 46:

$$\text{Width} = 2 \cdot \text{ME}$$

$$\text{Width} = 2 \cdot t_{\frac{\alpha}{2},n-1} \cdot \frac{s}{\sqrt{n}}$$

The width of interval is double the ME (margin of error). This is because of the $\pm$ in equation 3.115. So if we input our values for ME we can calculate the width of the interval for the specific case of a sample size of 16:

$$\text{Width}_{16} = 2 \cdot \left(2.131 \cdot \frac{0.834}{\sqrt{16}}\right) \approx 0.889 \tag{3.116}$$

The values used in 3.116 are the sample mean and sample standard deviation of the dataset of the sample means, the same values we use to construct our first confidence interval. It's commonly understood that the sample mean changes very little with an increase in sample size, exhibiting slow responsiveness to such changes. Moreover, as the sample size increases, the sample standard deviation tends to stabilize. This stability arises because a larger sample size captures more of the variance present in the data, providing a better estimate of the population standard deviation. For the sake of this experiment, let's assume that the mean and standard deviation remain consistent despite the increase in sample size. Under this assumption, the width of the confidence interval for a sample size of 46 would be calculated as follows:

$$\text{Width}_{46} = 2 \cdot \left(2.014 \cdot \frac{0.834}{\sqrt{64}}\right) \approx 0.419$$

This width is nearly half the size of what we would expect with a smaller sample size of 16! For comparison, let's examine the confidence intervals for sample sizes of 16 and 46 side by side. We know that $\bar{x} = 28.97$ now we just need to compute the margin of error:

$$\text{CI}_{16} = 28.97 \pm 2.131 \cdot \frac{0.834}{\sqrt{16}}$$

$$\text{CI}_{16} = (28.97 - 0.444, 28.97 + 0.444) = (28.526, 29.414)$$

Now for a sample of size 46:

$$\text{CI}_{46} = 28.97 \pm 2.131 \cdot \frac{0.834}{\sqrt{46}}$$

$$\text{CI}_{46} = (28.97 - 0.21, 28.97 + 0.21) = (28.55, 29.39)$$

Well, we have a narrower interval: not by much but hey, it is still an improvement! This interval implies a smaller range of possible values for the mean within our specified confidence level, which is generally more desirable for precision in estimates. We can actually do a bit of a switcheroo and define a sample size given a margin of error meaning that if:

$$\text{ME} = z_\alpha \frac{\sigma}{\sqrt{n}}$$

Then:

$$n = \left(\frac{z_\alpha \sigma}{\text{ME}}\right)^2$$

Hypothesis testing is another statistical technique that is influenced by changes in sample size. In hypothesis testing, we use sample data to make inferences about the population parameters, and then decide whether there is enough evidence to reject a null hypothesis in favor of an alternative hypothesis. The key metrics involved in this process include the p-value and the test statistic, both of which are affected by the sample size.

As the sample size increases, the test statistic becomes more sensitive to differences between the sample statistic and the null hypothesis value. This sensitivity is due to the decreased standard error associated with larger sample sizes. The standard error quantifies the variability of the sample statistic from one sample to another if the experiment were repeated multiple times. A smaller standard error means that the sample statistic, such as the sample mean, is more likely to be close to the true population mean. Consequently, a significant difference between the sample statistic and the hypothesized population parameter (under the null hypothesis) is more readily detected, resulting in a larger absolute value of the test statistic.

Additionally, a larger test statistic directly influences the p-value, which measures the probability of observing data as extreme as, or more extreme than, the data actually observed, assuming the null hypothesis is true. Lower p-values indicate stronger evidence against the null hypothesis, making it more likely that we can reject the null hypothesis when it is false.

However, it's also important to recognize that increasing the sample size can lead to tests that are too powerful in terms of detecting very small differences between the sample statistic and the population parameter, differences that may not be practically significant. This scenario underscores the necessity of not only considering statistical significance (as indicated by the p-value), but also evaluating the practical significance and the effect size of the findings. The question now is, how do we know if this is happening to our test statistic and p-value?

When we asked our contact in the traffic department to lower the street bars on two separate occasions, I observed that each time, a guard left the police station to verify the situation with the bar. This observation raises important tactical considerations. To ensure our entry into the garage is smooth and unobstructed, we need to carefully time our actions. Specifically, the interval between the lowering of the bars and the guard's response must be long enough for us to access the garage without detection. It's crucial that we calculate this timing precisely, to avoid any overlap that could lead to our operation being compromised. This analysis will help us to determine if we have a sufficient window in which to drive to the garage door unnoticed.

The time elapsed from Notarbartolo's relocation to Antwerp and the heist day was around 29 months. The planning of a heist of this magnitude requires method and patience. I hope you are in no rush, as it appears we must also embrace a similar mindset, because understanding the guard's response time to the lowered bars will necessitate a period of careful observation and data collection. To avoid arousing suspicion, we cannot afford to lower the bars too frequently. Therefore, we will run the following experiment: we'll signal our contact within the traffic department to initiate the lowering of the bars, and then meticulously record not just the time it takes for the guard to arrive at the scene, but more critically, the moment they first see the bars. This data collection will occur during two distinct time windows, 00:00-02:00 and 02:00-04:00, paralleling the time frames previously examined:

Observation	00:00-02:00 RT (min)	02:00-04:00 RT (min)
1	12.61	14.71
2	14.60	13.75
3	12.77	14.93
4	12.01	16.24
5	16.33	13.61
6	14.25	15.28
7	15.91	13.45
8	11.75	15.90
9	14.30	12.87
10	13.96	14.65
11	12.88	15.72
12	14.11	12.98
13	13.25	14.37
14	15.02	13.19
15	12.67	12.46
16	13.40	14.92

Table 3.17: Guard response times for lowering the street bars during different time windows.

In table 3.17 we have represented 16 observations of the response time (RT) for both time slots. Our task now is to check if there is any significant difference between the two time slots and, more importantly, if there is enough time for us to get to that garage door. What about if we plot two histograms?

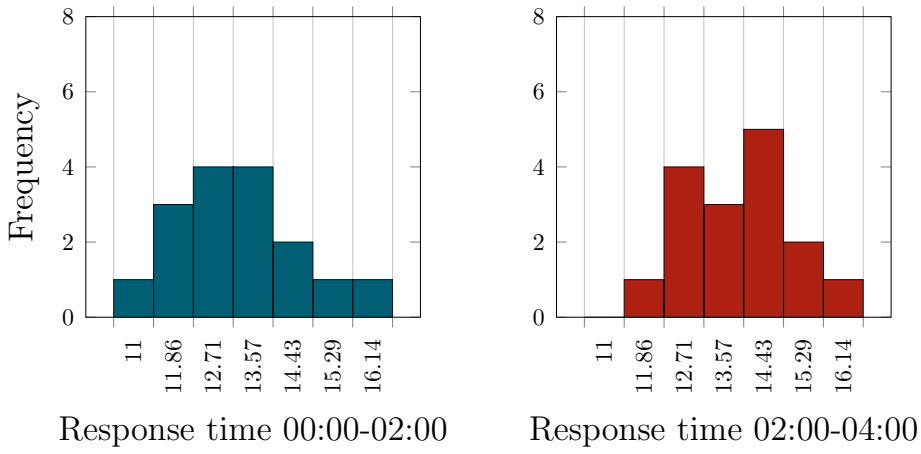


Figure 3.49: Comparison of guards response times in two time windows.

Both histograms in figure 3.49 show distributions in which the vast majority of density is clustered around a central location. This density then fades away as we move to the tails on both sides. Well, let's get this central point of tendency by computing the mean:

$$\bar{x}_{00:00-02:00} = 13.74 \text{ minutes}, \quad \bar{x}_{02:00-04:00} = 14.31 \text{ minutes}$$

As it happens, these means come with good news and a potential problem. Firstly, it seems we have enough time to drive to the

garage door no matter what time slot we pick. However, throughout the book, we have meticulously studied techniques that led us to conclude the best time slot for the heist was between 12 to 2 AM. And now, we're confronted with a predicament: data shows the timeframe from 2 to 4 AM has a higher mean response time for guards checking the street security iron bars. For our planning we must consider any edge we can get. Let's see by how much the slots differ:

$$\begin{aligned} \Delta_{\bar{x}} &= \bar{x}_{02:00-04:00} - \bar{x}_{00:00-02:00} \\ &= 14.31 - 13.74 \\ &= 0.57 \text{ minutes} \end{aligned} \tag{3.117}$$

In equation 3.117, we quantify the difference between the average times recorded during two distinct periods, 02:00-04:00 and 00:00-02:00, through Δ. The calculation 3.117, yields a difference of 0.57 minutes, which might not seem like much, yet every moment is significant in our context. However, before we rush and change everything about our strategic planning based on this observation, let's take a deep breath and reflect upon our data. We can deduce a probability density function to model the times of inactivity of the cameras that resulted in us leaning towards an assumption of normality.

Well, this particular situation, where we are trying to understand if there is any significant difference between time slots, should be no different. So, let's assume that variability is again present in our samples. In that case, it becomes evident that a metric such as Δ, without variance considerations, can lead us to wrong conclusions because we don't know if its value is due to chance or not.

This means we are required to enhance our equation 3.117 with steroids... sorry, with the variance... it needs to be with the variance. The histograms 3.49 suggest that both datasets are normally distributed, so let's start by defining the Gaussian distributions for both datasets, and we can see where does this takes us:

$$\begin{aligned} \bar{x}_{00:00-02:00} &= 13.74, \quad s^2_{00:00-02:00} = 1.73 \\ \bar{x}_{02:00-04:00} &= 14.31, \quad s^2_{02:00-04:00} = 1.35 \end{aligned}$$

Therefore if we define $X_{00:00-02:00}$ as the random variable representing the response time from the guard for the timeframe of

00:00-02:00 and, 02:00-04:00 sets up the same scenario but for the other timeframe, we see that:

$$X_{00:00-02:00} \backsim N(13.74, 1.73), \quad X_{02:00-04:00} \backsim N(14.31, 1.35)$$

Let's plot these two distributions alongside our difference Δ_μ:

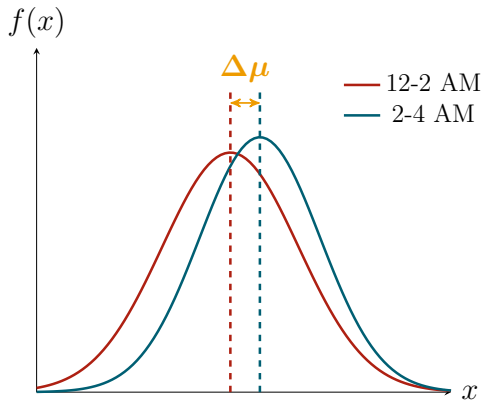


Figure 3.50: Randomly speaking how much do you differ?

In graph 3.50, we visualize the comparison between two normal distributions representing response times in two different time slots. The distinction in population means, denoted by $\Delta\mu$ highlights a difference we need to understand, if it is statistically significant:

$$\Delta\mu = \mu_{00:00-02:00} - \mu_{02:00-04:00} \tag{3.118}$$

Comparing 3.117 with 3.118 we generalize the formulae to the population's means; this is just for rigor, as we must define concepts for population parameters. And when we can't have them, which is, well... ALL THE TIME, we use approximations.

Equation 3.118 showcases the elegance of statistical analysis, but now we turn our attention to another critical aspect: variance. In subtracting means to arrive at $\Delta\mu$, we're essentially exploring the dynamics of subtracting entire distributions, each characterized by its own mean μ and variance σ^2. This process isn't just about the numbers; it's about understanding how these distributions interact with each other.

When we subtract two normal distributions, we're not only comparing their centers (their means) but also how spread out they are

(their variances). This operation gives us insight into the overall behavior of the differences between them. Starting with the mean, we recall a fundamental property of expectations: the expected value of the difference is the difference of the expected values. But what about the variances?

$$E[Z] = E[X] - E[Y], \quad Z = X - Y$$

Therefore:

$$E[Z] = \mu_x - \mu_y$$

So, the mean of the difference is the difference of means. Next, let's examine what happens to the difference of the variances?

$$\begin{aligned}
\mathrm{Var}(X-Y) &= E\left[((X-Y) - E[X-Y])^2\right] \\
&= E\left[(X - Y - (E[X] - E[Y]))^2\right] \\
&= E\left[(X - E[X] - (Y - E[Y]))^2\right] \\
&= E\left[(X-E[X])^2 + (Y-E[Y])^2 - 2\,(X-E[X])\,(Y-E[Y])\right] \\
&= E\left[(X-E[X])^2\right] + E\left[(Y-E[Y])^2\right] - 2E\left[(X-E[X])\,(Y-E[Y])\right] \\
&= \mathrm{Var}(X) + \mathrm{Var}(Y) - 2\mathrm{Cov}(X,Y)
\end{aligned}$$

Since X and Y are independent, their covariance, $\mathrm{Cov}(X, Y)$, is zero. Therefore:

$$\mathrm{Var}(X-Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$$

Meaning:

$$\mathrm{Var}(X-Y) = \sigma_X^2 + \sigma_Y^2$$

So, if X and Y are two normally distributed variables, we can say that:

$$X - Y \sim N(\mu_x - \mu_y, \sigma_X^2 + \sigma_Y^2) \tag{3.119}$$

With 3.119, we arrive at a probability distribution that models the difference between normally distributed variables. It is not a secret that we are setting up a hypothesis test to compare the differences between the means of two data samples. This means we need a test statistic; however, this does not change the fact that we need to include the variance. We have one that fits like a glove; what about a Z-score? It is a standard measure that is normalized

by the variance:

$$Z = \frac{\bar{x}_{00:00-02:00} - \bar{x}_{02:00-04:00} - \mu_{00:00-02:00} - \mu_{02:00-04:00}}{\sqrt{\sigma^2_{00:00-02:00} + \sigma^2_{02:00-04:00}}} \tag{3.120}$$

The Z-score, defined in 3.118 as the difference in sample means normalized by their variance, encapsulates a crucial statistical test measure. It embodies our test statistic, ushering us into the next necessary step: defining a null distribution to guide our decision-making process. Given that our metric 3.120 is assumed to follow a normal distribution and because we lack actual population parameters, we rely on sample estimates. This reliance on estimates rather than known parameters underscores the essence of hypothesis testing—estimating the unknown from the known:

$$\frac{\bar{x}_{00:00-02:00} - \bar{x}_{02:00-04:00}}{\sqrt{\frac{s^2_{00:00-02:00}}{n_{00:00-02:00}} + \frac{s^2_{02:00-04:00}}{n_{02:00-04:00}}}} \tag{3.121}$$

Now, with 3.121, we have a similar situation to the one we had with the confidence intervals, which is that we are likely to have introduced more variance by using the samples. And for that case, we have the t-score to come to the rescue:

$$t = \frac{\bar{x}_{00:00-02:00} - \bar{x}_{02:00-04:00}}{\sqrt{\frac{s^2_{00:00-02:00}}{n_{00:00-02:00}} + \frac{s^2_{02:00-04:00}}{n_{02:00-04:00}}}} \tag{3.122}$$

The question now becomes, what distribution does the t-statistic we have represent in 3.122 follow? Well, it will be a t distribution, but with how many degrees of freedom? The degrees of freedom represent the amount of information lost when estimating one or more parameters of the data. This metric is represented by the number of independent values that are free to vary when estimating parameters. For one sample, we concluded that the t distribution has $n-1$ degrees of freedom, where n is the sample size. But now we have two samples, so we must combine the amount of lost information; one way to do this is by adding up the degrees of freedom from both distributions:

$$(n_{00:00-02:00} - 1) + (n_{02:00-04:00} - 1) = n_{00:00-02:00} + n_{02:00-04:00} - 2$$

More generally for a two samples distribution the statistic t:

$$t \sim t_{n_1+n_2-2}$$

Where n_1 and n_2 are the sample sizes of the first and second samples, respectively. The only thing missing if we want to have our test fully defined are the null and the alternative hypothesis. However, before taking care of that, we must reflect a bit.

So far our approach to define a test for the difference of the means was similar to the ones we took with both the KS and SW tests. We started by defining a statistic, in the case of the KS test, it was one called D, showcasing the supreme difference between the empirical CDF and the theoretical CDF. From there, we defined the D_α threshold and used it to reject or accept H_0.

The approach with the SW test was very similar to the KS test, but we had a W instead of the D as the statistic; W was based on how consecutive points vary with each other, capturing the behavior of a normal distribution. The distribution of W was the null distribution; the difference was that, unlike in the KS test, where we had a threshold for D, with the SW test, we were not as lucky. Instead, to make our decision, we computed a new concept, the p-value.

These two tests had some similarities and differences: They were both based only in one sample and the decision of whether to reject H_0 or not was based on one side of the null distribution. This type of test, where we just consider one tail of the distribution to make a decision is called a "one-tailed test".

Now, we have a two-sample parametric test, as the assumption is that our data follows a normal distribution; the question now becomes whether we are constructing a one-tailed test or if we are dealing with a situation in which we make a decision based on considering both tails of the null distribution, therefore needing a two-tailed test. The reality is that, in this particular situation, we are the ones deciding which test to use, for example:

- **Null Hypothesis H_0**: The null hypothesis assumes that there is no significant difference in the mean response times of guards between the two time slots. In other words, any ob-

served difference in mean response times in our sample data is due to random chance. Mathematically, this can be expressed as:

$$H_0 : \mu_{00:00-02:00} - \mu_{02:00-04:00} = 0$$

Where $\mu_{00:00-02:00}$ and $\mu_{02:00-04:00}$ are the population mean response times for the 00:00-02:00 and 02:00-04:00 time slots, respectively.

- **Alternative Hypothesis H_1**: The alternative hypothesis challenges the null by suggesting that there is a significant difference in the mean response times between the two time slots. This implies that the observed difference is not due to random chance and reflects a true difference in the population:

$$H_1 : \mu_{00:00-02:00} - \mu_{02:00-04:00} \neq 0 \tag{3.123}$$

We have a statistic, a null distribution and two hypotheses, so we are ready... oh sorry, of course, Miss α, you are also invited and your value is 0.05. Satisfied? Great! Let's start by computing the t-statistic:

$$t = \frac{13.74 - 14.31}{\sqrt{\frac{1.73}{16} + \frac{1.35}{16}}} = -1.299$$

Having computed our t-statistic, the subsequent step involves analyzing its distribution through a PDF analysis. This will enable us to delineate the regions of acceptance and rejection within the PDF, thereby guiding our decision to either accept or reject the null hypothesis:

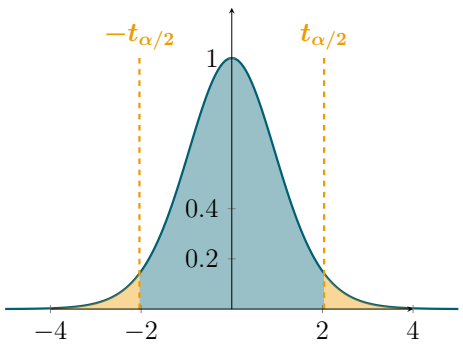


Figure 3.51: Is it a fantasy creature with two tails? It just a two-tailed t-test.

Figure 3.51 depicts the PDF corresponding to our t distribution with 30 degrees of freedom, the result of combining both sample sizes less two (16+16-2). In this visual, the yellow regions represent areas where we are challenging the null hypothesis (H_0), whereas the blue regions denote where the null hypothesis is not challenged.

The way we defined our hypothesis test specifically seeks to verify if there exists a statistically significant difference between $\mu_{00:00-02:00}$ and $\mu_{02:00-04:00}$, without a preference for whether one mean is greater or lesser than the other. This impartiality towards the direction of difference necessitates a two-tailed test.

A two-tailed test is employed when the null hypothesis does not predict the direction of the difference or effect, implying that an observed effect in either direction is relevant. Consequently, in this case, as the t distribution is symmetrical, we divide the α level, the predetermined threshold for statistical significance, equally between the two tails of the distribution. The symmetry of the t distribution facilitates this approach by allowing for an equal distribution of the significance level across both tails, thereby ensuring that the total area covered by the rejection regions (the two yellow areas) sums up to α.

I know the narrative was to check if the 00:00-02:00 hour slot mean was higher than the one of the 02:00-04:00; and we will get there, but first, let's at least explore the two-tailed test use case. Also, we are more than halfway there as the only thing left to compute is the p-value, which, by definition, is the likelihood of having a value as extreme as or more extreme than t, given that H_0 is true, so:

$$p\text{-value} = 2 \cdot P(T \geq |t_{\text{obs}}| \mid H_0 \text{ is true}) \tag{3.124}$$

Equation 3.124 outlines the calculation of the p-value for a two-tailed test. Once again, we will leverage software tools to perform these computations efficiently:

$$p\text{-value} = 0.204$$

With a p-value of 0.204, we find ourselves inside the blue region of 3.51, therefore we fail to reject the null hypothesis H_0. This outcome might initially seem like good news, because, if there is no

significant difference between the means, we don't really have the need to rethink the time slot in which initiate our heist. However, this conclusion is based on a sample-based estimation, raising the inevitable question of error.

Given that all our computations rely on samples and estimates, there's always a chance we might be making an incorrect decision. For instance, what if there is a discernible difference between the values of Δ, but our test didn't have enough evidence to reject H_0 confidently? Well, we need to check this!

With this specific test, we base our hypothesis testing on a null distribution, assuming a two-sample t distribution centered around zero under H_0. This distribution forms the foundation for deciding whether to reject or retain H_0. As such, it serves as an appropriate starting point for exploring how to quantify the likelihood of incorrectly retaining H_0 when, in fact, H_1 is true. One way to understand this situation is by plotting the PDFs of both H_0 and H_1 on the same graph. If we find that these curves overlap, this indicates that we might be making an error, because the overlapping region contains values that can belong either to H_1 or H_0, showing that outcomes are not clearly defined for either hypothesis.

However, we don't have a defined PDF for H_1 but let's see what we can do to come up with one. The alternative hypothesis, H_1, plays a pivotal role in our statistical analysis, which is designed to determine the significance of changes observed between the means. Specifically, if H_1 holds and we find that the difference, Δ, is significant, then the PDF associated with H_1 will not be centered at the same point as its counterpart under the null hypothesis, H_0.

This deviation is due to H_1 representing a scenario where an effect or difference shifts outcomes away from those expected under H_0. The PDF for H_1 reflects this shift, indicating that the observed data clusters around a new mean value, which might be either higher or lower, depending on the nature of Δ. This shift signifies a notable change in the distribution, showing that the center of H_1's PDF is positioned either to the left or right of H_0, highlighting the detected difference's direction.

This shift, also known as the effect size, can be either an uplift or a fall, directly correlating to the anticipated direction of the difference we are trying to detect. In the context of a two-tailed test, this modification acknowledges the possibility of significant differences in both directions. For this example, we will specifically employ an uplift to facilitate a more intuitive visual interpretation; let me warn you that we will therefore be exaggerating this uplift. The reason behind the decision to select a big uplift is so we can visually understand the concept. So, let's define the effect size to be 25%, to ensure the shift is noticeable in visual representations:

$$\text{Shift}_{H_1} = SD_{H_0} \cdot 0.25$$

This means that, if we move 25% to the right of the mean of H_0 we obtain the mean of H_1:

$$\mu_{H_1} = \mu_{H_0} + \text{Shift}_{H_1}$$

To shift a distribution's center, we adjust its mean by adding the effect size, moving the entire distribution left or right. To change the distribution's spread, we multiply its standard deviation by the effect size, altering its variability without shifting its center. Let's plot this:

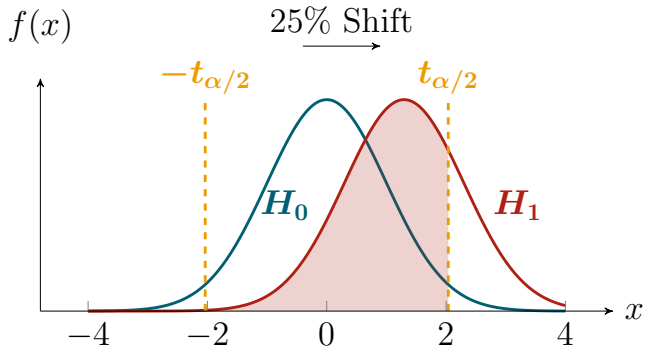


Figure 3.52: Red for the type II error!

In graph 3.52, the PDF of H_0 is depicted by the blue curve, while its counterpart H_1 is represented in vibrant cherry red colors. Let us focus on the region highlighted by the red shading. From our discussion on the two-tailed test, we recall that we should reject H_0 if our test statistic falls to the left of $-t_{\alpha/2}$ or to the right of

$+t_{\alpha/2}$. Therefore, the red shaded area, which is defined by everything between $-t_{\alpha/2}$ and $+t_{\alpha/2}$, represents the probability that we fail to reject H_0. Remember, the PDF and areas under the curve represent probabilities. However, observing a shift of 25% from H_0 to the right suggests that H_1 is probably true, which may indicate that we are incurring an error by not rejecting H_0. Any time I see the word "error" I get nervous, and that makes me want an equation:

$$P(\text{Failing to reject } H_0 \mid H_1 \text{ is True}) \tag{3.125}$$

Equation 3.125 represents the probability of failing to reject H_0 given that H_1 is true. In statistics, β usually denotes this exact type of error, and we also have a name for this fellow: "Type II error":

$$\beta = P(\text{Failing to reject } H_0 \mid H_1 \text{ is True}) \tag{3.126}$$

Equation 3.126 already provides some good insights, but can we do a bit more? For example, what is the likelihood of being right when we assume H_1 is true? Moreover, this will provide us with with the number of times we reject H_0 given that H_1 is true:

$$P(\text{Reject } H_0 \mid H_1 \text{ is True}) \tag{3.127}$$

But 3.127 is the complement of 3.126, so we can represent it as a function of β:

$$\text{Power} = 1 - \beta \tag{3.128}$$

We've got the power! That is indeed correct and not a typo; to the likelihood of rejecting H_0 given H_1 is true we call the 'power'. Let's also visualize this concept:

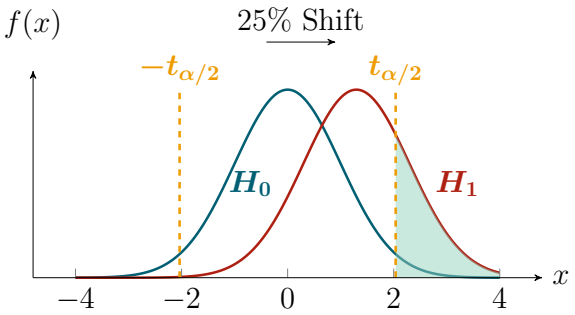


Figure 3.53: Light blue for power!

In figure 3.53, the light blue shaded area symbolizes our statistical power: this is the likelihood of correctly accepting H_1 when it does reflect the true state, especially in scenarios where we're anticipating an uplift. Well, if we have a Type II, we will definitely also have a Type I:

$$\alpha = P(\text{Reject } H_0 \mid H_0 \text{ is True}) \tag{3.129}$$

This one is easy; it is simply the confidence level, the error we are willing to accept when we reject H_0 and H_0 is true:

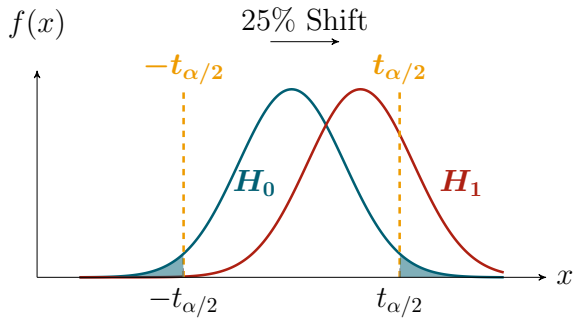


Figure 3.54: Finally a darker blue for the Type I error.

This representation 3.54 is not new to us; it is the same as the yellow regions of graph 3.51 representing the rejection areas. Right, let's do some tidying up and gather all this vital information in a table:

Decision \ True State	H_0 is True	H_1 is True
Fail to Reject H_0	Correct Decision (True Negative)	Type II Error (β)
Reject H_0	Type I Error (α)	Correct Decision (Power, $1-\beta$)

Table 3.18: Hypothesis Testing Outcomes

Now we move on to the calculations of all these fellows, considering an effect size of 25%, but first we need an α, not a male, but a significance level:

$$\alpha = 0.05$$

Kicking this off with the type II. This fellow requires us to compute the red shaded area of 3.52:

$$\beta = 0.758$$

Lastly the power:

$$1 - \beta = 0.242$$

In a nutshell: the Type I error, represented by α, is the probability of rejecting the null hypothesis H_0 when it is actually true. In our case, this is set at $\alpha = 0.05$, indicating that there is a 5% chance of incorrectly rejecting H_0. The Type II error, denoted by β, is the probability of failing to reject the null hypothesis when the alternative hypothesis H_1 is true. With a β of 0.758, this implies a high chance of 75.8% that we fail to detect the effect (i.e., failing to reject H_0 when we should). Finally the power of the test, calculated as $1 - \beta$, measures our test's ability to correctly reject the null hypothesis when the alternative hypothesis is true. With a power of 0.242, our test has only a 24.2% chance of correctly identifying an effect when it exists. The results might not be as encouraging as we had hoped, but remember we tried to test a massive uplift of 25%. If we had gone with some smaller values of the same metrics α, β and $1 - \beta$ probably would have been more use.

Through this process, we've not only engaged with a specific type of test, a two-tailed t-test, but have also been introduced to a new statistical measure, the t-statistic for 2 samples. Additionally, this has been an opportunity to learn about the errors that can occur due to the inherent randomness of systems. More importantly, we've seen how these errors can be quantified, offering us valuable insights into the nature of statistical testing and the interpretation of its outcomes. But what does all this has to do with the sample size? Remember, we embarked on this adventure while introducing the impact of the sample size on the different methods we have studied so far. What better way to understand the impact of anything other than an experiment?

For the experiment, let's walk backwards, no, not like a crab! The crab walks sideways; the lemur is the one that walks backward. The lemurs do not ring a bell? King Julian from Madagascar? "I'd like to move it move it?" We are not literally walking backwards though; we will start with the definition of power and, from there, walk our way backward until we can isolate a n representing a sample size. So, the power depends on the type II error, which is related to a H_1 distribution resulting from a shift in the distribution of H_0. We also know that H_0 is defined by:

$$t \sim t_{n_1+n_2-2} \tag{3.130}$$

With t defined as:

$$t = \frac{\bar{x}_{00:00-02:00} - \bar{x}_{02:00-04:00}}{\sqrt{\frac{s^2_{00:00-02:00}}{n_{00:00-02:00}} + \frac{s^2_{02:00-04:00}}{n_{02:00-04:00}}}} \tag{3.131}$$

In exploring the relationship between sample size and its impact on statistical power and Type II error, both equations 3.130 and 3.131 incorporate variables for sample sizes, underscoring their significance. This discussion begins by examining confidence intervals, which inherently provide a range where we can expect the true parameter value to lie, with a certain degree of confidence. This leads us to understand how increasing the sample size not only enhances the approximation of the sample standard deviation to the population standard deviation, but also refines our estimate of where the true parameter value is situated. Such precision also directly influences our hypothesis-testing process, especially in relation to the null hypothesis. When a confidence interval for a parameter, like the mean difference between two groups, does not include the value specified under the null hypothesis, it suggests there is evidence against H_0. Conversely, if the confidence interval includes the null value, this indicates the observed data do not provide a basis for rejecting H_0. This relationship shows how confidence intervals can inform our decision on whether to reject or retain H_0, tying back to our primary interest in statistical power and Type II error.

The convergence of the t distribution to the normal distribution with larger sample sizes thereby results in narrower regions of overlap between the null hypothesis H_0 and its alternative H_1 after the shift. Consequently, as the sample size grows, the standard error of the mean diminishes, yielding a more precise confidence interval that directly affects our hypothesis-testing outcome. The direct outcome of this process is a reduction in Type II error alongside an increase in the power of the test, illustrating the crucial role of sample size in the precision and reliability of hypothesis testing.

Should we see this in action? For illustrative purposes, we will maintain the same standard deviation and mean values while increasing our sample size from 16 to 64. This approach allows us to

isolate and highlight the effect of sample size on the power of our test and the probability of committing a Type II error, providing a clear demonstration of the critical role sample size plays in statistical analysis:

$$t \sim t_{64+64-2}$$

$$t \sim t_{126}$$

Now our t-statistics follows a t distribution with 126 degrees of freedom and the probabilities for the power and the type II error are:

$$\beta_{64} = 0.26, \quad 1 - \beta_{64} = 0.74$$

As opposed to our previous results:

$$\beta_{16} = 0.895, \quad 1 - \beta_{16} = 0.105$$

Comparing sample sizes of 16 and 64 reveals a notable decrease in Type II error (β) from 0.895 to 0.26 and an increase in test power from 0.105 to 0.74, respectively. This demonstrates that increasing the sample size significantly enhances the test's ability to detect true effects, reducing the likelihood of failing to reject a false null hypothesis and improving the test's overall reliability.

Before moving on, there are still two things I would like to cover: The first is the one-tailed test we could have defined in this particular situation, for example:

$$H_1 : \mu_{00:00-02:00} > \mu_{02:00-04:00} \tag{3.132}$$

Or, if we were expecting the opposite direction:

$$H_1 : \mu_{00:00-02:00} < \mu_{02:00-04:00} \tag{3.133}$$

Defining H_1 as 3.132 or 3.133 would mean we are interested in just one of the directions, indicating a one-tailed test. The main difference between a one-tailed test vs. a two-tailed one is where we construct the rejection areas:

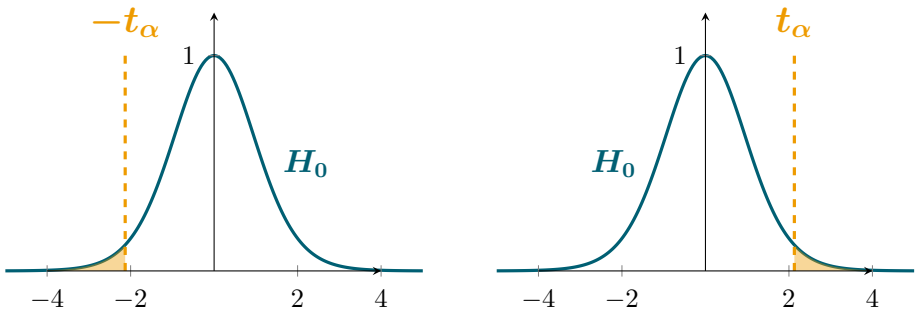


Figure 3.55: Either to the right or to the left, we must define where to reject this fellow.

The two plots depicted in 3.55 illustrate the application of one-tailed t-tests under the assumption of the null hypothesis, which presumes there is no difference between two groups. The first plot, a left-tailed test, assesses whether the mean of the first group, $\mu_{00:00-02:00}$, is less than that of the second, $\mu_{02:00-04:00}$.

The shaded area left of $-t_\alpha$ identifies the critical region. If the test statistic falls within this area, it suggests that $\mu_{00:00-02:00}$ is indeed lower than $\mu_{02:00-04:00}$, supporting the alternative hypothesis $H_1 : \mu_{00:00-02:00} < \mu_{02:00-04:00}$. Conversely, the second plot is a right-tailed test checking if $\mu_{00:00-02:00}$ is greater than $\mu_{02:00-04:00}$. Here, the shaded area to the right of t_α marks the rejection region for the null hypothesis.

A test statistic landing in this right hand tail supports the alternative hypothesis $H_1 : \mu_{00:00-02:00} > \mu_{02:00-04:00}$. Each plot in 3.55 clearly demarcates the critical regions where the null hypothesis would be rejected in favor of the corresponding directional alternative hypothesis, illustrating the focused aim of the one-tailed tests to detect an effect in a specific direction.

Well, if the rejection areas are different between the one-tailed test and the two-tailed test, this means that the Type II error will also be impacted. Consequently, the power of the test will also differ, let's see how:

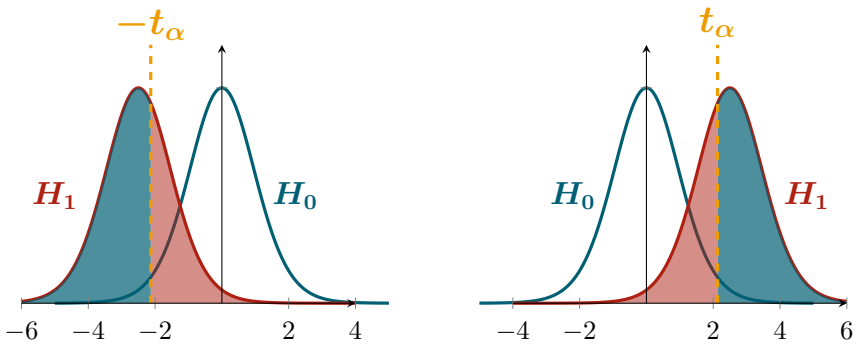


Figure 3.56: Does power really have one side?

In graph 3.56, we encounter two plots, each showcasing two overlaid PDF's. The blue curve represents the null hypothesis distribution, while the red curves indicate the shifted distributions under the alternative hypothesis.

The red area is designated for the Type II error, representing the scenario where we fail to reject H_0 despite H_1 being true. Conversely, the blue area symbolizes the test's power, essentially the statistical muscle needed to "unite and tell them to fuck off"... I wish it was, but this power is actually to highlight the test's capacity to correctly reject H_0 in favour of H_1.

Focusing first on the left-hand graph, which illustrates the left-tailed test corresponding to the hypothesis that $\mu_{00:00-02:00}$ is greater than $\mu_{02:00-04:00}$ (equation 3.132), we observe that the effect size under H_1 shifts the distribution to the left. In such a left-tailed test, H_0 is rejected if our statistic falls to the left of $-t_\alpha$. Thus, the area under the H_1 distribution to the right of $-t_\alpha$ marks the region of Type II error, where we fail to reject H_0 when H_1 is true.

As for the power, depicted by the blue region, it encompasses the segment of the H_1 distribution to the left of $-t_\alpha$. This area represents the probability of correctly rejecting H_0 when H_1 is true, showcasing the test's effectiveness. The same rationale extends.

The same rationale extends similarly to the right-tailed test, where, due to the rightward shift of the H_1 distribution in accordance with the hypothesis in equation 3.133, the power situates to the right of t_α, and the Type II error spans the area under the red curve to the left of t_α.

All this conversation about power, errors in one-tailed vs a two-tailed tests raises a pertinent question, what about the two methods we chose to introduce the whole topic of hypothesis tests? Remember the Kolmogorov-Smirnov and the Shapiro-Wilk tests? Do they also have a power, and type I and type II errors? The answer is yes; we did not cover this topic then because of the complexity of the PDFs of both test statistics and the challenges of visualizing the areas representing the probabilities that characterize those three metrics.

For the KS test, the effect size is conceptualized as the maximum distance between cumulative distribution functions, reflecting the largest discrepancy in distribution shapes rather than differences in central tendency. Similarly, the SW test's effect size relates to the overall deviation of a sample from normality, encompassing aspects of the distribution's shape such as skewness and kurtosis. These conceptualizations make intuitive visualization and direct interpretation of effect size more complex. Unlike the t-test, where effect size can be directly tied to practical differences between means, the KS and SW tests require a nuanced understanding of distribution properties and their impacts.

With the KS Test, power calculation is not straightforward due to its non-parametric nature. Power depends on the sample size, the significance level, and the extent of deviation from the null hypothesis. We can generate a large number of samples under the alternative hypothesis, apply the KS test to each, and calculate the proportion of tests that correctly reject H_0, giving an empirical estimate of power. The Type II error rate is inherently related to power and varies similarly with the factors mentioned.

For the SW Test, calculating power and Type II errors also typically involves simulation due to the complexity of its distribution under the alternative hypothesis. Like the KS test, we would generate samples under the assumption of non-normality (the alternative hypothesis), perform the SW test on these samples, and observe the proportion of tests that reject H_0, thereby estimating power. The selection of the alternative hypotheses in simulations needs careful consideration to reflect the specific aspects of non-normality (e.g.,

skewness or kurtosis) you are testing against.

Neither test has simple, closed-form solutions for power and Type II error, due to their dependencies on the sample distribution's specific characteristics. Therefore, we can use our friends' computers and statistical software to get the all these test metrics, as now we know what they are and how to interpret them.

The sample size is a central topic in statistics, it impacts a population's statistics and therefore all the methodology we wish to apply to better study the underlying distributions. Even the central limit theorem is not complete without specifying a sample size, remember? When introducing it, mentioned that something would be covered later; that was the sample size. There is a rule of thumb that we need 30 observations for the sample of the means to follow a normal distribution. This raises two questions: Firstly why 30 and not 47? And secondly, if the sample size is so important, how can we make sure the samples we have are large enough for us to make an inference?

In regards to the 30, it is what we call an empirical rule of thumb. Through extensive observation, statisticians have found that for many practical applications, a sample size of 30 is often large enough for the sampling distribution of the mean to approximate normality. Ours has 16 and we got both the KS and the SW tests accepting the null hypothesis. So it is just a rule of thumb, not a "must-have". What naturally arises from this is how to choose a sample size?

In a commercial context, imagine the website development team at a company is considering a change to a button in the homepage. Specifically, they're contemplating altering the color of the checkout button to see if it positively impacts sales. The team approaches you to design an experiment to evaluate this change. Essentially, they're asking for the necessary duration of the test, or, in statistical terms, the sample size needed to ensure the experiment's reliability. This scenario requires us to almost reverse-engineer our hypothesis test to determine this crucial piece of information.

The power of the test becomes a critical consideration in this

setup. We need to ensure a high degree of confidence in our decision; if we choose to adopt the new button color based on our test results, we want to minimize the risk of screwing up (Yeah I have been there, it is not pretty).

Let's establish a relationship between the sample size and the power of the test. By invoking the Central Limit Theorem, we understand that a sample of the sample means will follow a normal distribution, providing a solid foundation for the distribution underlying our null hypothesis. So we decide to collect an average number for the daily sales. Now, to advance our test setup, we need to define two critical parameters: the significance level and the test's power.

Cool, so now we are ready for our two hypotheses:

- **Null Hypothesis H_0**: The average daily sales for variant B are less than or equal to the average daily sales for variant A. In terms of means, this can be written as $\mu_B \leq \mu_A$.
- **Alternative Hypothesis H_1**: The average daily sales for variant B are greater than the average daily for variant A ($\mu_B > \mu_A$), which corresponds to a one-tailed test.

Nothing new here, just another one-tailed test. Lastly, we need an effect size, let's once again call it Δ:

$$\Delta = \frac{\mu_B - \mu_A}{\sigma}$$

As we are going to be working with the standard normal distribution, we standardize the effect size also; this way everything is measured in standard deviations. The next step is to define a null distribution; for that, we will leverage the fact that we are going to need a sample size higher than 30, meaning that our average daily sales will follow a normal distribution. So we know that the Z-statistic is defined as:

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

But now we have two samples! No problem, we know what the parameters of the distribution that represents the difference between two normal random variables, as we deduced it when building our t-statistic for the test on the responses time. So:

$$Z = \frac{\bar{x}_B - \bar{x}_A - (\mu_B - \mu_A)}{\sqrt{\frac{\sigma^2}{n_A} + \frac{\sigma^2}{n_B}}}$$

Assuming equal sample sizes $n_A = n_B = n$ and that $\mu_B - \mu_A$ represents the expected difference under H_1, this equation simplifies to:

$$Z = \frac{\bar{x}_B - \bar{x}_A - \Delta\sigma}{\sigma\sqrt{\frac{2}{n}}} \tag{3.134}$$

Here, $\bar{x}_B - \bar{x}_A$ is the observed difference between the sample means. The goal now is to link equation 3.134 to the concept of statistical power. To refresh our understanding, the power of a test $(1 - \beta)$ represents the probability of correctly rejecting the null hypothesis H_0 when the alternative hypothesis H_1 is true. For a specified power, there is a corresponding Z_β such that:

$$P(Z > Z_\beta) = \beta$$

When considering a significance level α, the critical value Z_α is defined by:

$$P(Z > Z_\alpha) = \alpha$$

To determine the sample size n necessary to achieve a desired power, we consider the total Z required to exceed the critical value:

$$Z_{\text{observed}} = Z_\alpha + Z_\beta$$

Substituting into the equation, we get:

$$Z_\alpha + Z_\beta = \frac{\Delta\sqrt{n}}{\sigma\sqrt{2}}$$

Solving for n:

$$n = \frac{2 \cdot \sigma^2 \cdot (Z_\alpha + Z_\beta)^2}{\Delta^2} \tag{3.135}$$

Where:

- Z_α is the Z-value corresponding to the significance level α.
- Z_β is the Z-value corresponding to the desired power $(1-\beta)$.
- Δ is the effect size (difference in means divided by the standard deviation).
- The factor of 2 accounts for the assumption that the sample sizes for both groups are equal.

This formula 3.135 accurately reflects the relationship between the significance level, power, effect size, and required sample size for a one-tailed test. So we can select a sample size n and given a power value; for example if we wish to have a power of 80% we just need to compute β:

$$1-\beta = \text{Power}$$
$$\beta = 1 - 0.8 = 0.2$$

An example? Really? Alright, alright. Let's define some hypothetical parameters:

$$\mu_a = 402, \quad \mu_b = 405, \quad \sigma = 2.7, \quad \alpha = 0.05$$

For the effect size:

$$\Delta = \frac{405-402}{2.7} = 1.11$$

Now we need to compute $Z_{\alpha=0.05}$ and $Z_{\beta=0.2}$:

$$Z_{0.05} = 1.645, \quad Z_{0.2} = 0.84$$

So n is:

$$n = 2 \cdot \left(\frac{2.7 \cdot (1.645 + 0.84)}{1.11} \right)^2$$
$$n \approx 73$$

So we need roughly 73 samples on each variant to have a test with 80% power. If our test was two-tailed instead, we would just have to adjust equation 3.135 accordingly:

$$n = \left(\frac{Z_{\alpha/2} + Z_\beta}{\Delta} \right)^2 \sigma^2$$

The only difference is with $Z_{\alpha/2}$ because a two-tailed test reflects the need to detect effects in both directions and accounts for the distribution of probabilities across both tails of the normal distribution. In a two-tailed test, we're interested in deviations from the null hypothesis in either direction, meaning that significant results could occur if the true mean is either greater than or less than the hypothesized mean hence the $Z_{\alpha/2}$.

The formula for calculating the required sample size varies based on several key factors, including the nature of the hypothesis test (one-tailed vs. two-tailed) and the specific statistics being analyzed. Each statistic has its own associated null distribution, leading to distinct approaches for determining the necessary sample size. It's crucial to recognize that the primary objective here is not to memorize an exhaustive list of formulas but to grasp the significance of statistical power and the importance of sample size in the context of experimental design.

Understanding what statistical power means (the probability of correctly rejecting the null hypothesis when it is false) is fundamental to conducting robust analysis. Similarly, recognizing why sample size plays a critical role in experiments allows us to design experiments just like the one described above, where we could see several techniques in action: the CLT, hypothesis testing, the test power and the sample size.

So far, this is where we stand in our meticulous planning. Starting with measures of centrality, we crafted an ingenious escape route. The median revealed the path with fewer cameras, while the mode pinpointed a trend in the brands of cameras deployed along that same route. We then reached out to one of our "associates" and received two data samples spanning different time periods, each detailing moments of camera inactivity. This led us to employ histograms, setting the stage for our next move: calling upon our trusted allies, the functions, but with a twist. This time, they returned densities, introducing us to an expectation metric that proved pivotal in choosing the optimal time slot for our heist based on the expected camera downtime.

With our time slot pinpointed, we gathered additional samples

for that period, a decision that allowed us to invoke the Central Limit Theorem. This theorem came with the added perk of confidence intervals, which naturally led us to cumulative density functions. As a result, we were provided with a range for our average camera inactivity time, a crucial piece of intel.

But, as true professionals (currently in the thieving business, but soon-to-be mathematicians), we couldn't simply rely on assumptions. We needed to verify that our sample of inactivity times truly followed a normal distribution. This investigation introduced us to a variety of tests and, along the way, the concept of covariance and correlation, where we found that rainy days tend to see fewer police patrols, a handy bit of knowledge. However, our plan had a weak link: those damn street security bars. We employed conditional probability to assess our odds of having the bars descend during our operation, depending on how we contacted our insider. This exploration brought us to the realm of p-values, Type I and II errors, and the power of a test. Finally, we needed to ascertain whether the average reaction time of a guard, upon noticing the lowered bars, differed between our selected sample time slots for camera inactivity. So at the moment, we have a time slot, a way to put the bars down, and even the ideal weather conditions, what we don't have is a way to enter the garage.

Notabartolo and his crew meticulously planned their heist, ensuring every detail was accounted for, including their entry and exit strategy through the garage door. The chosen location for the garage door was pivotal—close enough to the police station to benefit from the presumption of safety, yet positioned where it could remain unnoticed by law enforcement personnel. Their strategic selection exploited the advantage of seeming ordinariness and the natural sounds of the area to disguise their movements.

To gain access, the team skillfully cloned the garage door's remote control. By positioning themselves near the garage, they employed a device capable of capturing the door's frequency. This method allowed them to replicate the remote's signal, thereby enabling them to open or close the garage door at will, seamlessly integrating their entry and exit into the fabric of routine activities,

all while under the veil of normalcy and undetected by the nearby police. That took 30 minutes to do; imagine the sweat!

We decided to park our getaway vehicle inside the garage, making access to it a crucial element of our strategy. However, after the School of Turin heist, the center's security team tightened the security protocols for the garage entry. The door now incorporates sophisticated security features: it recognizes the license plates of vehicles authorized by the Diamond Center or requires a security code entered on a newly installed keypad.

These new measures present us with two possible entry methods: replicate the license plate of an authorized vehicle or obtain a security code used by the guards on the keypad. It's a pivotal decision, and again we find ourselves drawing inspiration from the tactics Notarbartolo and his crew used.

One crucial step in cracking the safe was to circumvent the code required at the safe's door keypad. To achieve this, Nortabartolo provided his access card to the diamond center to the "Genius." The plan involved the Genius installing a camera directly aimed at the keypad. This setup would allow the crew to review the footage later and, hopefully, decipher at least one of the codes.

We plan to adopt a similar strategy. Unlike Notarbartolo, who had free reign over the premises due to his office inside the diamond center, one of our challenges is installing this camera next to a door that guards heavily patrol; however, if we have information on those patrol patterns, maybe we can find a window in which to install a camera. How long do you think we need to install a camera and flee the scene? One hour? Yeah I think the same.

I took on this task and stationed myself to discreetly observe and record the frequency of keypad usage over 24 hours, collecting essential data to aid our mission:

Hour	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Number of visits	2	4	3	0	3	4	2	2	0	5	5	6

Hour	**13**	**14**	**15**	**16**	**17**	**18**	**19**	20	**21**	**22**	23	**24**
Number of visits	4	2	5	5	4	1	1	6	3	2	0	1

Table 3.19: A scary sight: the number of guard visits per hour.

In table 3.19, we have documented the hourly guard visits over a 24-hour period. Our objective is to identify a window of time in which to discreetly install a camera. According to table 3.19, hours 4, 9, and 23 recorded no guard visits at the keypad, suggesting these as optimal times for our installation. Given that one hour is ample time to set up the device, and with no guards scheduled during these slots, this task seems feasible.

However, this approach assumes that our data sample from a single day sufficiently represents typical guard patterns, which may be a stretch. Guard visits might vary significantly from day to day due to unknown factors or changes in security protocols. Given these limitations, we must treat each hour as an independent event; let's refer to them as trials. A trial will be a single experiment or observation that has exactly two possibilities. This means that occurrences in one hour do not influence what happens in another, making it challenging to pinpoint an exact hour for camera installation.

The thing is, we have a safe to crack, and I can't risk going outside to gather any more data. I nearly got caught last time. Therefore, we must make the best of the dataset we have. OK, let's get to work! Our goal is to develop a metric that can boost our confidence in timing our move to install the camera. For that, we need to extract useful information from our data, which is divided into hourly slots.

Considering our task, we don't really need to know the exact number of guards that visited. If it takes around an hour to install the camera, the presence of just one guard during that time could ruin our plan.

This realization allows us to simplify our data. For instance, we can label the hour slots where at least one guard appeared as a 'success', indicating a higher risk for us. Conversely, the slots where no guards appeared are considered 'failures' on their part, which are ideal opportunities for us. We will mark these opportunities with a '0' and the success with an '1':

Hour	1	2	3	4	5	6	7	8	9	10	11	12
Successes	1	1	1	0	1	1	1	1	0	1	1	1

Hour	13	14	15	16	17	18	19	20	21	22	23	24
Successes	1	1	1	1	1	1	1	1	1	1	0	1

Table 3.20: Successes and failures from the ones we need to avoid.

Table 3.20 closely resembles table 3.19, with a critical difference in how we represent our data. Instead of showing the actual count of visits per hour, we mark a time slot with a '1' if at least one guard visited during that period. Conversely, for a slot where we did not observe any guard, we mark it with a '0'. What about if we start simple? So we can understand the probability of success, we need a metric that will represent the chance of a guard coming at any given hour:

$$P(\text{Success}) = \frac{\text{Number of successful trials}}{\text{Total number of trials}} = \frac{21}{24} \approx 87.5\%$$

It turns out that, over a 24-hour period, we have an 87.5% chance of having at least one guard come to the keypad. So, we can define our probabilities of success and failures as follows:

- Probability of success (a guard came) p, is $\frac{21}{24}$.
- Probability of failures (no guard was present) q is $\frac{3}{24}$ or $1 - p$.

The likelihood of encountering a period when no guard is at the keypad is quite slim—only about 12.5%. And just so we're clear, you're the one who'll be going down there to install the camera. It's your turn to take one for the team! While the average number of guard visits per hour gives us a basic understanding of activity levels, it doesn't reveal the full complexity of their patterns or the specific risks associated with different times of the day. For instance, simple averages don't allow us to calculate the likelihood of consecutive periods with no guards, a critical metric that could make you more confident when you're down there putting the device in place. Another valuable piece of information would be the probability of a given number of successes,for example, what's the likelihood that we observe 10 visits to the keypad on any given day?

Essentially, if we had a framework where we could input data, marking each hour as a success or a failure, and receive probabilities

in return, we would be significantly better equipped to quantify our chances of successfully installing the camera without getting caught. We are setting ourselves up to have one of those probabilistic functions, and that means we have to start somewhere to deduce it. So what about if we compute the likelihood of any two events, for example, let's quantify the probability of having a success followed by a failure:

$$\begin{aligned} P(\text{Success than Failure}) &= P(\text{Success}) \cdot P(\text{Failure}) \\ &= \frac{21}{24} \cdot \frac{3}{24} \\ &\approx 0.109 \end{aligned} \tag{3.136}$$

Equation 3.136 represents the probability of having one success followed by a failure. Let's keep exploring this; what about if we get one more success in? Meaning we will calculate the probability of two successes and one failure:

$$P(\text{Success than Success than Failure}) = P(\text{Success}) \cdot P(\text{Success}) \cdot P(\text{Failure})$$

Which is equivalent to:

$$P(\text{Success than Success than Failure}) = P(\text{Success})^2 \cdot P(\text{Failure})$$

Now, if we use our p and $1-p$ notation to represent the probabilities of success and failure:

$$P(\text{Success than Success than Failure}) = p^2(1-p) \tag{3.137}$$

This raises a question; isn't it true that $p^2(1-p) = (1-p)p^2$? It sure is! We considered our trials to be independent, so the commutative property of the multiplication stands. For example, if we wish to represent the probability of two success followed by two failures we can do so by:

$$P(\text{Success than Success than Failure than Failure}) = p^2(1-p)^2$$

Neat! More generically, if we have k success and t failures our formula becomes:

$$p^k(1-p)^t \tag{3.138}$$

Let me put it out there; equation 3.138 is incomplete. I know this is not the news we were looking for. However, this is a chance for us

to point out the situation before completing 3.138. Our goal is to obtain metrics that will give us an idea of the possibility of installing a camera undetected. The strategy: start simple and consider only successes and failures instead of the number of visits. With that, we started our quest to establish a probability function by exploring some probabilities and ended up with 3.138. And then came Jorge, the party pooper, saying our model was incomplete; however, let's still try to make a case for it.

We considered the trials to be independent, which conceptually makes sense; we have one day of data, so assuming dependencies between trials, will require knowing more about patrolling routines, and at the moment, we don't. But also because of this independence, we have a problem; for example, if we wish to test two successes followed by one failure, with 3.138 that has the same expression as one failure followed by two successes:

$$P(\text{Success than Success than Failure}) = p \cdot p \cdot (1-p) \tag{3.139}$$

$$P(\text{Failure than Success than Success}) = (1-p) \cdot p \cdot p \tag{3.140}$$

Equation 3.139 and 3.140 have the same result; in fact we can even have one more way of getting that same value:

$$P(\text{ Success than Failure than Success}) = p \cdot (1-p) \cdot p \tag{3.141}$$

We can stay here all day, creating combinations like this over the 24-hour period of our observations, but the question that we need to raise is, "Does this make sense?" To maintain some mathematical coherence, we've overlooked the order of trials, opening up many possibilities for considering fixed numbers of successes and failures. And just like in real life, everything is a trade-off. For instance, equations 3.139, 3.140, and 3.141 demonstrate that sequences that are ordered differently, yield the same probability values, and can manifest at any time throughout our data collection period. Thus, to accurately calculate the likelihood of observing a given number of successes and failures, we must consider all possibilities.

This means we need to account for all those combinations, so, let's adopt a strategy of simplification before generalization. For the sake of clarity, we'll denote a success by S and a failure by F.

Instead of tackling the full 24-hour model upfront, let's first explore how we might achieve 2 successes within a more manageable framework of just 3 trials. This step-by-step method not only simplifies our notation but also lays the groundwork for understanding the dynamics of consecutive successes versus isolated ones. As we are only working with 3 trials our sequences will only have a length of 3, with two of their elements being S's. The notation for combinations of 3 trials and 2 successes is:

$$\binom{3}{2} \tag{3.142}$$

In expression 3.142, the top number represents the number of trials, whereas the bottom one represents the number of successes. Now, we need to find an expression for that bad boy. We can start by exploring all the possible permutations with three trials:

$$3! = 6 \tag{3.143}$$

But does this make sense? I don't think so because we are overcounting the S's. The factorial is blind to the trials; it just knows we have three of them, so we have no distinction between S's. This means:

$$S_2S_1F$$

$$S_1S_2F$$

$$S_2FS_1$$

$$S_1FS_2$$

$$FS_1S_2$$

$$FS_2S_1$$

This is what that poor factorial man sees. Please don't blame him; he is old, and his vision is not what it used to be, but we can help him. Because S_1 and S_2 are just successes to us, if we consider them to be the same, we end up with half of the trials:

$$SSF$$

$$SFS$$

$$FSS$$

We can overcome this by dividing 3.143 by the number of successes:

$$\binom{3}{2} = \frac{3!}{2} = 3 \tag{3.144}$$

But what happens if we have one more S and one more trial?

$$\binom{4}{3} \tag{3.145}$$

Well now we will have 4!, meaning 24 possible sequences and a lot of S's, more specifically 3! of them, so our original formula 3.144 is not capturing this second factorial on the denominator and we must change it:

$$\binom{3}{2} = \frac{3!}{2!} = 3 \tag{3.146}$$

It just so happens that $2! = 2$, which is why it worked for 3.144, but as we can see if we increase the number of successes to 3, the formula won't hold. We are nearly there but still missing essential details; what will happen when we increase the F's? The top factorial will be equally 'blind' to them, so we must also discount this matter. So if we have three trials and two successes, it leaves room for one failure, meaning the $F = \text{trials} - S$:

$$\binom{3}{2} = \frac{3!}{2!(3-2)!} = 3 \tag{3.147}$$

We have to use the factorial for the same reason we put it on the two in the denominator of 3.147. So, if we wish to generalize 3.147 for n trials and k success:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{3.148}$$

Formula 3.148 is known as the binomial coefficient; it accurately reflects the number of distinct ways to achieve k successes out of n trials, genuinely considering the essence of combinations where the specific order of outcomes within each combination does not matter.

With this, we are ready to define our equation! If we define X as a random variable that represents the number of success (k) and consider our equation 3.138 we have that:

$$P(X = k) = \frac{n!}{k!(n-k)!}p^k(1-p)^t \tag{3.149}$$

3.6 Are We Really Going To Make Everything About 0's and 1's? The Binomial Distribution.

However, because the number of failures is equal to the number of trials minus the number of successes, we can replace the exponent t with $n - k$:

$$P(X = k) = \binom{n}{k}p^k(1-p)^{(n-k)} \tag{3.150}$$

Well, and there we have it, a new probability distribution called the "binomial distribution" The binomial distribution is a framework that allows us to understand the outcomes of binary, or two-result, processes. This distribution applies to scenarios where an experiment, or trial, is repeated a fixed number of times, with each trial having two possible outcomes: success and failure.

The distribution allows for calculating the probability of observing a specific number of successes across these trials, given that the probability of success on an individual trial is known and constant.

There's a fundamental difference between the binomial distribution and the normal distribution; it hinges on the type of data we're analyzing. The binomial distribution deals with discrete variables, allowing us to count successes and failures directly within our dataset. This contrasts with the continuous variables associated with the normal distribution, such as time intervals, where measurements can take on any value within a range.

This distinction leads to several key changes in how we approach our analysis. First, the function we use to describe the binomial

distribution is not called a probability density function but a probability mass function. The PMF 3.150 reflects the discrete nature of our data, assigning probabilities to specific, countable outcomes.

Here's something else you might appreciate: no integrals are required with the binomial distribution! Because our trials are countable, we rely on summations rather than integrals, which are used to calculate areas under curves for continuous distributions like the normal. (Integrals represent the limit of infinite sums, so they are necessary for dealing with continuous intervals where, theoretically, an infinite number of values exist.)

Cool, so now we need the formulas for at least the expectation and the variance of a discrete distribution. The concept is exactly the same, meaning that the expected value of a discrete random variable also represents the theoretical mean or average of the variable if an experiment is repeated an infinite number of times. It provides a measure of the central tendency of the possible outcomes, weighted by the probabilities of those outcomes:

$$E[X] = \sum_{i=1}^{n} x \cdot p_i \tag{3.151}$$

Equation 3.151 represents the expected value $E[X]$ of a discrete random variable X, where p_i denotes the probability of each outcome x, and the summation extends over all possible outcomes of X. Is it time for the dice example? I think it is! Suppose you're playing a game where you roll a fair six-sided die, and you gain a number of dollars equal to the number showing on the die. Let's calculate the expected monetary gain from one roll of the die.

Random Variable $\boldsymbol{X}$: The outcome on the die can be 1, 2, 3, 4, 5, or 6.

Probabilities $\boldsymbol{p_i}$: Since the die is fair, each outcome has an equal probability of $\frac{1}{6}$. Using the equation for the expected value $E[X]$, which is:

$$E[X] = \sum_{i=1}^{6} x \cdot p_i$$

We can calculate $E[X]$ as follows:

$$E[X] = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6}$$

$$E[X] = \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) = \frac{1}{6} \cdot 21 = 3.5$$

Thus, the expected value of one roll of the die, or the average gain per roll, is 3.5. This means that, on average, you would expect to gain 3.5 per roll of the die in the long run.

This concept parallels the calculation of the expected value for a continuous random variable, where an integral over the probability density function is used instead of a summation. Furthermore, this approach of replacing integrals with summations extends to calculating all moments of a distribution, not just the expected value. The core ideas and properties, such as those pertaining to variances and expectations, apply to both discrete and continuous cases.

Equation 3.151 is the general formula for the expectation, so let's explore it in the context of the binomial distribution to see if we can get to the point where the formula is more pleasant to the naked eye. First, let's define our X:

- X follows a binomial distribution with parameters n (the number of trials) and p (the probability of success in each trial).

Mathematically:

$$X \sim B(n, p)$$

If X represents our random variable that follows a binomial distribution, then we can define each trial's outcome as X_i. In our example, X quantifies the total number of successes out of n trials, and X_i represents the outcome of each trial throughout the day. This means that each X_i can either be a success (1) or a failure (0), contributing to the total count of successes that X summarizes:

$$X = X_1 + X_2 + \ldots + X_n$$

We have $X_i = 1$ if the i^{th} trial was successful and $X_i = 0$ otherwise. So far, I like it! We have n trials, but out of curiosity, let's

check the expected value for a single trial, just any random i:

$$E\left[X_i\right] = 0 \cdot (1-p) + 1 \cdot p = p \tag{3.152}$$

In equation 3.152, the 0 (Failure) and the 1 (Success) represent the x_i's, meaning our possible outcomes and the variables p and $(1-p)$ are, respectively, the probabilities of these outcomes. Did you think the good news ended with us not having to develop new intuitions behind the four-momentum metrics? In that case, you will be pleasantly surprised, as both the properties of the expectation and variance also hold for the discrete case. What better way to celebrate this endeavor than using one of these properties?! What about linearity?

$$\begin{aligned} E[X] &= E[X_1 + X_2 + \ldots + X_n] \\ &= E[X_1] + E[X_2] + \ldots + E[X_n] \end{aligned} \tag{3.153}$$

If we use the results of 3.152 in 3.153 I feel we will be on our way to something:

$$E[X] = p + p + ... + p$$

Because we have n trials:

$$E[X] = np \tag{3.154}$$

Equation 3.154 tells us that the expected number of successes in n independent Bernoulli trials (a fancy name for an independent experiment with two outcomes: success or failure) , each with success probability p, is np. This result is intuitive; if we expect a proportion p of trials to be successful, then in n trials, you would expect np successes on average. Should we quickly do the variance?

$$\text{Var}(X_i) = E[(X_i - \mu)^2] \tag{3.155}$$

That μ there represents the population mean, so for a given trial:

$$\mu = E[X_i] = p$$

Now we can replace the μ in equation 3.155 with a p:

$$\text{Var}(X_i) = E[(X_i - p)^2]$$

Next we can expand it:

$$\begin{aligned}\text{Var}(X_i) &= (0-p)^2 \cdot (1-p) + (1-p)^2 \cdot p \\ &= p^2 \cdot (1-p) + (1-2p+p^2) \cdot p\end{aligned}$$

Simplify the expression by combining like terms:

$$\begin{aligned}\text{Var}(X_i) &= p^2(1-p) + p(1-2p+p^2) \\ &= p^2 - p^3 + p - 2p^2 + p^3 \\ &= p - p^2\end{aligned}$$

Therefore, we conclude that:

$$\text{Var}(X_i) = p(1-p)$$

Because we defined X as a sum of Bernoulli trials:

$$\text{Var}(X) = \sum_{i=1}^{n} \text{Var}(X_i) = \sum_{i=1}^{n} p(1-p) = np(1-p) \tag{3.156}$$

Now that we have our binomial distribution fully defined what about if we check its PMF plot? For that, we use our parameters $p = 0.875$ and $n = 24$ in the binomial equation:

$$P(X = k) = \binom{24}{k} 0.875^k (0.125)^{(24-k)} \tag{3.157}$$

Equation 3.157 represents the Binomial probability mass function for our dataset. The parameter k denotes the total number of successes within a 24-hour period, which aligns with the number of hours we want to assess for the likelihood of a guard's presence. Now for the plot:

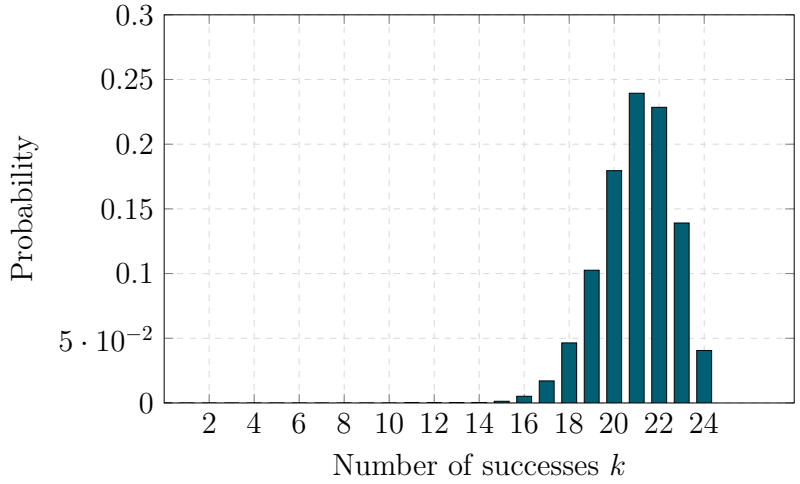


Figure 3.57: Our first picture of a PMF.

In Figure 3.57, we have plotted the PMF for our binomial distribution, as defined in equation 3.157. On the x-axis, we represent the number of successes, which varies from 0 to 24. Unlike the case of a continuous variable, where the y-axis represents densities, here, with a discrete variable, it represents probabilities. This distinction is crucial for correctly interpreting the graph.

Are we letting this slide? I mean, if that 3.57 isn't a bell shape, I don't know what is! Is this just coincidence? Each trial in a binomial distribution is a Bernoulli trial. As the number of these trials increases, the sum of successes (described by the binomial distribution) starts to resemble a normal distribution due to the Central Limit Theorem. This theorem states that the sum of a large number of small, independent random variables tends to form a normal distribution, regardless of the original distribution's shape. For the binomial distribution, this means that, as the number of trials grows, the distribution of the number of successes becomes bell-shaped, centering around its mean np and spreading out with a variance $np(1-p)$. This bell shape arises from the aggregation of numerous small, independent effects (each Bernoulli trial's outcome), showcasing the classical characteristics of a normal distribution.

Surely we can't drop knowledge like this and not learn how to leverage it. Well, if we can approximate a binomial distribution with a normal one, it means that Z-scores are back on the menu! One statistical technique that leverages these fellows is confidence

intervals. In this particular instance of a binomial distribution, the mean gives us the likelihood of success. That is our parameter p, so we need an estimate for this fellow:

$$\hat{p} = \frac{x}{n}$$

Where x represents the number of successes observed in our sample of size n, remember that p is the population probability of success, a parameter we don't know, so we use our best estimate for it, which we call $\hat{p}$, and this guy is also our statistic. Now, we need three things: limes, rum, and ice... ah, those three are for later... For now, we need a significance level α, a margin of error, and a way to introduce likelihood into our margin of error. Great... I guess...

The significance level is our choice, so there is no need for extra work here. For the standard error, we can use a similar approach to the previous confidence interval, the one we built around the sample means, and in that case, we used the variance: we built the previous interval for the case of a normal distribution, so let's compute the variance of $\hat{p}$:

$$\text{Var}\left(\frac{x}{n}\right) = \frac{\text{Var}(x)}{n^2} = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n} \tag{3.158}$$

For the margin of error, we need to define an estimation of the standard error SE, which is the square root of 3.158:

$$\text{SE} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Now, for the parameter that will capture randomness, we can use the Z distribution:

$$\hat{p} \pm \text{ME} = \hat{p} \pm z \cdot \text{SE}$$

So:

$$\left(\hat{p} - z \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \quad \hat{p} + z \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

Let's quickly compute the confidence interval for our $\hat{p}$. We have that $\hat{p} = 0.875$ and $n = 24$, and assuming a 95% confidence interval

(hence $z \approx 1.96$). To compute the confidence interval we need the margin of error ME, so let's start by calculating the standard error SE:

$$\text{SE} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.875 \cdot (1-0.875)}{24}} = 0.0675$$

For the margin of error ME :

$$\text{ME} = z_\alpha \cdot \text{SE} = 1.96 \cdot 0.0675 = 0.1323$$

Therefore our confidence interval comes as:

$$\begin{aligned} \text{CI} &= (\hat{p} - \text{ME}, \quad \hat{p} + \text{ME}) \\ &= (0.875 - 0.132, 0.875 + 0.132) \\ &= (0.742, 1.007) \end{aligned}$$

So:

$$\hat{p} \in [0.742, 1.007]$$

Granted that, for our particular task, the confidence interval of $\hat{p}$ does not provide us with a very valuable insight, the reason that the deduction of the formula was done is to showcase that not only the normal distribution and the mean have confidence intervals. This is a technique we can apply whenever we feel it will help us.

Right, now let's get some metrics that will quantify our chances of installing the device without being detected, remember, if you get caught, you don't know who I am. What about if we start with the expectation and the variance? Given the binomial distribution parameters $n = 24$ and $p = 0.875$, we can compute the expectation (mean) and the variance as follows:

$$E[X] = n \cdot p = 24 \cdot 0.875 = 21$$

Thus, the expected number of successes (the mean) is 21. This means that, in a 24-hour period we expect to have 21 different hours where the guards visit. Now, for the variance:

$$\text{Var}(X) = n \cdot p \cdot (1-p) = 24 \cdot 0.875 \cdot (0.125) = 2.625$$

Therefore, the variance, indicating the variability of the number of successes around the mean, is 2.625. Still not very encouraging, so let's dig a bit deeper. What about the probability of three

consecutive failures?

$$P(k \text{ consecutive failures}) = (1-p)^k \tag{3.159}$$

Equation 3.159 represent a formula to for us to compute the probability of consecutive failures. It comes after the property of independence where we multiply the probability of failures the number of times we wish to check. In our case, our $k = 3$:

$$P(3 \text{ consecutive failures}) = (0.125)^3 \approx 0.002$$

Ouch, 0.2%. Before carrying on with our analysis, we must make sure we are clear about this concept. Why didn't we use the PMF equation to compute 3.157 to compute the likelihood of consecutive failures? Well, this is due to the word consecutive. The PMF of the binomial distribution will provide us with the likelihood of having a number of successes or failures in a 24-hour period, regardless of their sequence.

It is important to understand that the binomial distribution does not give us likelihoods of consecutive trials but the probabilities of them occurring regardless of their order. Having made this important distinction, let's move on with our plan.

Every successful heist relies on a touch of good fortune, and it seems we're at the point where we too must place our hopes in being favored by Lady Luck. Notarbartolo also counted on it when executing his heist. At that time, two concierges frequently patrolled the diamond center. However, during all of Notarbartolo's reconnaissance missions, he had never encountered these two during the late-night hours. This wasn't a guarantee they'd never be there, but, as they say, no guts, no glory. I appreciate your commitment, and I'm thankful for it. How many guard-free hours would make you feel comfortable to proceed with installing the camera? Would at least 13 hours without guards at the keypad suffice? Let's figure out how to quantify this. Right now, we need to find a way to understand how likely that figure is to happen.

Our binomial framework provides us with the probabilities of successes, but as we fixed our number of trials to $n = 24$, failures are the complement of successes. This means that, for us to have

13 failures, 11 successes must occur (since 13+11=24). Let's start exploring this concept by computing the likelihood of 11 successes. Cool let's bring back our binomial equation 3.157, so we can derive insights that will enhance your chances of successfully installing the camera:

$$\begin{aligned} P(X = 11) &= \binom{24}{11} \cdot 0.875^{11} \cdot (0.125)^{13} \\ &= \frac{24!}{11!(13)!} \cdot 0.875^{11} \cdot 0.125^{13} \\ &\approx 1.045 \times 10^{-6} \end{aligned} \tag{3.160}$$

Equation 3.160 provides the probability for exactly 11 successes, which is also the value for the likelihood of 13 failures. But in this situation we would be better off with a cumulative metric of probability rather than the likelihood of precisely 11 successes (or 13 failures). Because if we are comfortable with 13 failures, then 14, 15, and so on will improve our chances of success:

$$P(X \leq 11) \tag{3.161}$$

As we are working with a random variable X that is discrete, equation 3.161 expands to:

$$F(k, n, p) = P(X \leq k) = \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i} \tag{3.162}$$

Where $F(k, n, p)$ represents the probability of observing k or fewer successes out of n trials, with p as the probability of success on each trial. Now, if we are interested in the probability of X being less than or equal to 11, we look at:

$$P(X \leq 11) = F(11, 24, 0.875) \tag{3.163}$$

We can work a bit on equation 3.163:

$$F(11, 24, 0.875) = \sum_{i=0}^{11} \binom{24}{i} (0.875)^i (0.125)^{24-i}$$

Resolving the calculations:

$$\begin{aligned} F(11, 24, 0.875) = &\binom{24}{0}(0.875)^0(0.125)^{24} + \\ &\binom{24}{1}(0.875)^1(0.125)^{23} + \ldots + \binom{24}{11}(0.875)^{11}(0.125)^{13} \\ &\approx 1.17 \times 10^{-6} \end{aligned}$$

You can't say it is not better than our previous result 3.160. The reality is that the chances of 11 successes or higher and, therefore, more than 13 failures are minimal, not to say zero. Before we give up on the whole plan and assume that it is not possible to install the camera, let's visualize the CDF to see if we can reach an agreement:

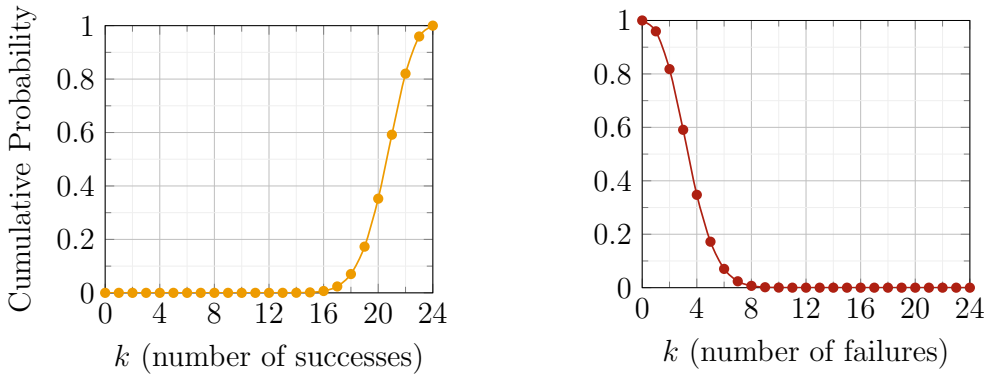


Figure 3.58: No, it is not another bell shape, rather a comparison between CFD's.

In figure 3.58, we have not one but two plots; what a great day! On the left, the yellow fellow represents our CDF, where k represents our success. The values of probabilites on the yellow curve were calculated with equation 3.162. Now, on the right, the red curve represents the complement of its yellow counterpart, meaning that it is the CDF but for a situation where k represents failures, and the computations is done with the formula:

$$1 - \text{CDF}$$

As we can observe, we are unlikely to have more than four failures on a given day. Having four or more failures has an approximate probability of 35%. You got this! Just as with confidence intervals, we can also apply these results in hypothesis testing, using a Z-test to draw inferences about outcomes involving Bernoulli trials.

Installing the camera is just the initial step in our task of entering the garage. After the installation, it's crucial to ensure that we can clearly identify at least one valid code entered by the guards on the keypad. However, I suspect that not all codes may be decipherable. Factors such as the guards inadvertently obscuring the keypad or adverse weather conditions might obscure or blur the video feed. Given these variables, it's important to estimate the number of clear images we can realistically expect on a typical day. This will help us assess if the risk of proceeding with the installation is justified.

Our previous data, which recorded counts of code entries rather than clear-cut successes and failures, provides a foundational basis for these estimates. Having spent considerable time understanding the binomial distribution, we should utilize this knowledge to estimate the daily number of codes visible in the video feed. The challenge lies in effectively applying this statistical method to predict clear visual captures under variable conditions.

To get to the binomial we moved from hourly counts to hourly trials, where each was classified by a success or a failure. But an hour is a collection of minutes, so what about we divide an hour into several different minutes and create multiple successes and failures per hour? For example, say we pick hour 11, where we observe five visits:

Minute	Guard Presence
6	0
12	1
18	0
24	1
30	0
36	1
42	0
48	1
54	0
60	1

Table 3.21: When we split a hour in minutes.

Table 3.21 represents the hour slot 11 divided into 10 portions of 6 minutes. We can observe that we have five successes and five failures, and, because we are not giving any importance to the order of events, remember that a fundamental assumption was not to infer anything regarding hourly patterns, we can use the binomial distribution. Because of that, with 3.21 we are back in a situation

where we have a collection of independent trials that can either be a success or a failure, and that is a setup we know how to work with, as we have the binomial distribution. So, if we split all the time slots, we can finally deduce the framework we seek. The crux of it relies on the way we divide the hour into different slots.

Do we divide all the hours into ten different trials? Why 6 and not 20, or even 100? What if, in some instances, the count soars to 7,000? Do we do it uniformly? Or do we create different portions for each hour? Well, I know a guy who can solve all of those problems for us. It is time to bring it back; I am talking about the eight with the wobbly knees. If we are concerned with the number of trials needed to accommodate as many successes as required, why don't we solve this by making those trials as large as possible? And what is larger than infinity? Well, I can think of some celebrities' egos, but in our world, nothing is larger than infinity. Thus, by contemplating an infinite number of trials, we address our current dilemma for uniformity, equipping ourselves with a mathematical model ready to tackle any similar dataset involving counts. So:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

And we are after:

$$\lim_{n\to\infty} \binom{n}{k} p^k (1-p)^{n-k} \tag{3.164}$$

It's an ugly ass limit, this 3.164, but let's tackle it by making a couple of assumptions: The first one is that when we make the number of trials n tend to infinity then, the probability of success p decreases. Essentially, this balance allows us to model large-scale phenomena with a degree of realism, as it moderates the impact of an infinitely increasing number of trials by proportionately reducing the chance of success in each, maintaining a realistic framework for analysis despite the huge scale. The second assumption is around the rate at which p balances n, we will consider it to be constant and defined by λ, so:

$$\lambda = np \Rightarrow p = \frac{\lambda}{n} \tag{3.165}$$

OK, let's expand our limit 3.164 and replace p with the result 3.165:

$$\lim_{n\to\infty} \frac{n!}{k!(n-k)!} \left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k}$$

Now we can work that a bit:

$$\lim_{n\to\infty} \frac{n!}{k!(n-k)!} \frac{\lambda^k}{n^k} \left(1-\frac{\lambda}{n}\right)^n \left(1-\frac{\lambda}{n}\right)^{-k}$$

$$\frac{\lambda^k}{k!} \lim_{n\to\infty} \frac{n!}{n^k(n-k)!} \left(1-\frac{\lambda}{n}\right)^n \left(1-\frac{\lambda}{n}\right)^{-k} \tag{3.166}$$

If we refer back to the Lord of all Curves "Calculus", there are a few results, we can extrapolate to simplify our limit:

- $\lim_{n\to\infty} \left(1-\frac{\lambda}{n}\right)^n = e^{-\lambda}$.
- $\lim_{n\to\infty} \left(1-\frac{\lambda}{n}\right)^{-k} = 1$ for finite k.
- The term $\frac{n!}{n^k(n-k)!}$ approaches 1 as n becomes very large.

3.6.1 In French It Is a Fish, for Us a Mathematical concept: The Poisson Distribution.

If we use all the results above, 3.166 simplifies to:

$$P(X=k) = \frac{e^{-\lambda}\lambda^k}{k!} \tag{3.167}$$

Equation 3.167 represents the PMF of the Poisson distribution. The Poisson distribution is a statistical tool used to model discrete variables; more precisely, the probability of certain events occurring within a fixed interval, assuming these events happen independently at a consistent rate. It's beneficial in machine learning when you want to predict the frequency of discrete events over time or space. For instance, it can model the number of visitors to a website, helping manage server resources by predicting traffic patterns. In natural language processing (NLP), it estimates the occurrence rate of

specific words or phrases in text, which is crucial for automating tasks like document classification and keyword analysis. Well, let's start deriving information from this fellow. What better way to start than looking at what we can expect from this distribution? Meaning, what is $E[X]$?

$$E[X] = \sum_{k=0}^{\infty} k \cdot P(X = k) \tag{3.168}$$

We can now expand 3.168 with the probability mass function formula:

$$E[X] = e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!}$$

Notice that $\frac{\lambda^{k-1}}{(k-1)!}$ is the derivative of $\frac{\lambda^k}{k!}$ with respect to λ. Thus, we can rewrite the sum as:

$$E[X] = e^{-\lambda} \lambda \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!}$$

By re-index the sum (letting $j = k - 1$), it now starts at $j = 0$:

$$E[X] = \lambda e^{-\lambda} \sum_{j=0}^{\infty} \frac{\lambda^j}{j!}$$

Recognize that the sum is the series expansion of e^{λ}:

$$E[X] = \lambda e^{-\lambda} e^{\lambda} = \lambda$$

Oh, that is quite a result. The expected value is λ but we also know that the expectation represents the mean, so:

$$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

But, if λ is the only parameter that we need to define the Poisson distribution and the mean can estimate this fellow, this tells us we have everything we need to define a Poisson distribution that models

the number of times a guard will visit the keypad per each hour over 24 hours:

$$\hat{\lambda} = \frac{1}{24}\sum_{i=1}^{24} x_i = \frac{70}{24} = 2.92$$

Having λ we can now define our Poisson PMF:

$$P(X = k) = \frac{e^{-2.92}(2.92)^k}{k!} \tag{3.169}$$

Equation 3.169 defines the probability of a certain number of visits k occurring in a given hour. However, for our present task, we are more interested in understanding how often we can capture a guard interacting with the keypad over the entire 24-hour period. Well, if λ represents the expectation of visits over one hour, we can estimate the number of visits by multiplying it by 24, meaning that we can expect $24 \cdot 2.92 = 70.08$ codes in a day.

In our particular situation, the Poisson cumulative distribution won't be much help, because questions like, "What is the likelihood of having more than four guards visits in one hour?" are not really what we are looking for; we are not interested in inputting success k's to understand their cumulative likelihood. However, since we are on a roll with the equations and just defined the PMF, why not define the CDF as well? So, for a Poisson distributed random variable X, representing the number of events less than or equal to some value k:

$$P(X \leq k) = e^{-\lambda}\sum_{i=0}^{k} \frac{\lambda^i}{i!} \tag{3.170}$$

Equation 3.170 represents the CDF for the Poisson distribution where λ is the average rate at which events occur per interval. This function sums the probabilities of observing 0 up to k events, providing a way to calculate the likelihood up to a certain number of events happening within a given timeframe based on the average rate λ. For example say we wish to calculate the probability of observing four or less guards:

$$\begin{aligned} P(X \leq 4) &= e^{-2.92}\left(\frac{(2.92)^0}{0!} + \frac{(2.92)^1}{1!} + \frac{(2.92)^2}{2!} + \frac{(2.92)^3}{3!} + \frac{(2.92)^4}{4!}\right) \\ &\approx 0.8285 \end{aligned}$$

Thus, the probability of observing 4 or fewer guard interactions in one hour is approximately 82.85%.

We completed our task successfully; after all that, we have a daily estimate of around 70 codes and a new probability distribution, the Poisson one. However, are we confident on that λ? What about if we build a confidence interval around λ? It would provide us with good information on what to expect regarding keypads inputs. I like it!

The strategy again involves evaluating the potential to apply the CLT to our analysis of λ. To effectively leverage this theorem, we require a sample of means, and we actually have a good candidate to helps us get there, Yeah, we see you λ! We can do this in a way that is analogous to our approach to deducing the Poisson mass function, where we subdivided an hourly interval into smaller segments to analyze successes and failures. By dividing each hour into smaller intervals, such as six minute slots, and then calculating the rate of visits for each of hourly slots, we effectively create multiple instances of hourly rates λ. For example, by calculating λ for each 10-minute segment and then aggregating these to formulate an hourly rate, we can generate 24 distinct λ values for a full day, assuming we consider each hour independently. As λ is estimated by the sample means formulae, 24 of them constitute a sample of means, with each λ representing an hour's average rate. This approach allows us to apply the principles of the CLT, provided the sample size is sufficiently large and the assumptions of independence and identical distribution are met.

Let's choose a confidence level, $\alpha = 0.05$, and determine our margin of error, which, according to the Central Limit Theorem, is the product of the Z-score and the standard error. With that in place, we realize that we have a missing piece, the variance. Because the standard error is derived from its square root, we need to compute this guy to complete our confidence interval:

$$\text{Var}(X) = E[X^2] - (E[X])^2 \tag{3.171}$$

Mister $E[X^2]$, it is time for you to go to work :

$$E[X^2] = \sum_{x=0}^{\infty} x^2 \frac{e^{-\lambda}\lambda^x}{x!} \tag{3.172}$$

With 3.172, we face one of those equations that could easily make us vomit; I mean, we have everything: !'s, λ's, $\sum$'s, e 's, the whole crew is there. However, those times of fearing equations are long gone now; looking at 3.172, we can use some mathematical sprinkles to turn it into something more familiar to us. If it weren't for the x^2, we would be in the presence of the formula of our friend expectation, meaning that perhaps with a bit of manipulation, we can get the equation for $E[X]$ and thereby simplify 3.172. As what we are after resembles the expectation formula, we need to have an X instead of an X^2, so what about if, instead of x^2, we have $x(x-1)+x$? This seems promising:

$$E[X^2] = \sum_{x=0}^{\infty} (x(x-1)+x) \frac{e^{-\lambda}\lambda^x}{x!}$$

Expanding that:

$$= \sum_{x=2}^{\infty} x(x-1) \frac{e^{-\lambda}\lambda^x}{x!} + \sum_{x=1}^{\infty} x \frac{e^{-\lambda}\lambda^x}{x!}$$

The term on the left simplifies to λ^2. Such a result comes from our calculus knowledge and recognizing that after factoring λ^2 from λ^x and adjusting the factorial in the denominator, it results in the series expansion of e^{λ}, which cancels with $e^{-\lambda}$. For the guy on the right, this fellow is simply the expected value $E[X] = \lambda$, thus:

$$E[X^2] = \lambda^2 + \lambda$$

Substituting $E[X^2]$ and $E[X]$ back into the variance formula 3.171, we get:

$$\text{Var}(X) = (\lambda^2 + \lambda) - (\lambda)^2$$

$$\text{Var}(X) = \lambda$$

Oh, the variance equals λ, which means that the expected value, which is also λ, equals the variance. Well, if you were still looking for

a positive in regards to the Poisson distribution, if that remarkable fact wasn't enough to convince you, god damn it. Nevertheless, we now know that by calculating one parameter, λ, we get information on the expected frequency of events and the variability around this expectation. And with this result, we are ready to have an equation for our standard error:

$$\text{SE} = \sqrt{\frac{\hat{\lambda}}{n}}$$

Where $\hat{\lambda}$ is the observed rate per unit of time (or other relevant interval). So our confidence interval is:

$$\text{CI} = \hat{\lambda} \pm z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{\lambda}}{n}}$$

$$\text{CI} = 2.92 \pm 1.96 \cdot \sqrt{\frac{2.92}{24}}$$

That results in $\hat{\lambda}$ being within the following range:

$$\hat{\lambda} \in [2.23, 3.6] \tag{3.173}$$

Nice! We are 95% sure that the hourly visits to the keypad will be in the interval represented by 3.173. However, we are not looking for an hourly metric; for our specific case, it is better to have a daily number of codes to understand how long we will have to have the camera on to collect a sample. Don't worry; all this work was not in vain.

Since the Poisson distribution assumes that events occur independently and at a constant average rate, scaling λ to a different time frame maintains the integrity of these assumptions. So, we can expect to collect 70.08 codes; but what about the hourly confidence interval? Can we also extrapolate on that? In theory, we can, but we must assume the event rate is constant throughout the day. Such belief means that the rate of occurrences (events per hour) does not change from hour to hour. This uniformity allows for linear scaling of the rate from an hourly to a daily basis, without adjusting for time-of-day effects or fluctuations in rate across different hours, so:

$$\hat{\lambda}_{24} \in [53.52, 86.4]$$

Cool, so we can expect between 54 and 86 keypad inputs from the guards on a given day. We did a pretty good job deducing equations and extracting a few results, but we also committed one mortal sin: we did not stop for even one second to inquire if this model fits our data. Well, does it? I know, I know. I guided us this way, so I should be the one answering that question. Let's start with a visual inspection by comparing our PMF with the data's histogram:

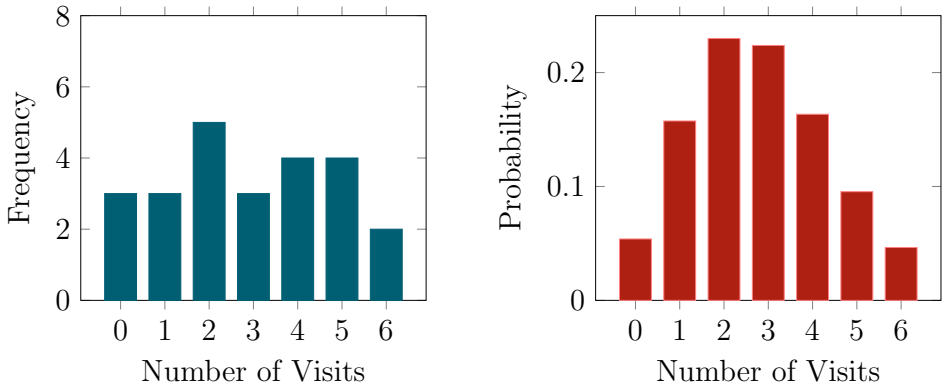


Figure 3.59: Two bar plots that should be similar...

In figure 3.59, on the left hand panel, we have a histogram representing the frequencies of the number of visits derived from our data, whereas, on the right, we have the PMF of a Poisson distribution with a parameter $\lambda = 2.92$, the one we use to fit the distribution to our dataset. The figures on the panels do not look the same, implying that we are off to a bad start. However, our eyes are not the last call on a statistical decision, as we don't yet have a lens that identifies statistical significance, so let's not reject this model yet. Right, an analytical method is needed.

It isn't time to embark on a journey of more equations; we just did that. Therefore, I suggest we leverage the peculiar feature that characterizes the Poisson distribution; its variance matches the expected value λ. If the sample variance is way off from the sample expected value, this suggests our model is not the best to fit our data:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = 3.56$$

Hmmm, it is not that far from our $\lambda = 2.92$ but also not quite the same. It is a very inconclusive result, and unfortunately, we

must go one step further to understand whether the Poisson model fits our data well. Yeah, you guessed it; we will construct another hypothesis test and that means we have to derive more equations, but isn't that the reason we are here?

As usual, two things are crucial when dealing with the classical match of H_0 vs H_1: a test statistic and its distribution, which we call the null distribution. From the tests, we know the only statistic we can use is D from the KS test, as it does not make any assumptions on the underlying distribution, but that is not how we roll; around here, we can take advantage of any chance we get, so let's continue this trend and create a new test.

While we will not employ the Kolmogorov-Smirnov test in our analysis, its principle of juxtaposing empirical data against a theoretical model serves as a valuable inspiration. In our case, as we are modeling frequencies, why not create a statistic about them? For that, let's refer back to table 3.20 and count the number of times we observed a certain number of visits; for example, let's see how many times we have 0 visits:

$$O_0 = 3 \tag{3.174}$$

We define the observed frequency by O, and subscript 0 refers to the observation from which we count frequencies. This means that in the instance represented by 3.174, we have 3 hours in which we observed 0 visits from the guards. So if we extend this to all the number of visits we have in table 3.19:

Visits	O_0	O_1	O_2	O_3	O_4	O_5	O_6
Observed Hours	3	3	5	3	4	4	2

Table 3.22: Observed Frequencies of Guard Visits to a Keypad Over 24 Hours.

To create a table like 3.22, we need to go to table 3.19 and select all the distinct frequencies of visits observed: these are 0, 1, 2, 3, 4, 5 and 6. Next, we count the times they occur per day and define this count by O_i, where i corresponds to the number of visits. The new metric O_i characterizes the empirical distribution, and, because we are building a statistic that works based on comparison, what will help us is having a similar metric to O_i but derived from the

theoretical distribution:

$$P(X = k) = \frac{e^{-\lambda} \cdot \lambda^k}{k!}$$

Because we started with the counts represented by O_0, it makes sense to consider the next step to find a way to compute the amount of 0 visits. Based on our theoretical distribution, we will compare two measures of the same magnitude. We currently have no formula for such frequencies, in reality, we only have an equation that allows us to compute probabilities. So, let's start by calculating the likelihood of having 0 visits and see if, from there, we can get to frequencies:

$$P(X = 0) = \frac{e^{-2.92} \cdot 2.92^0}{0!} = 0.053$$

When incorporating $k = 0$ into our probability mass function, we calculate the likelihood of observing no events within a specific interval, given our Poisson distribution with a rate parameter $\lambda = 2.92$. This process yields the probability of an interval containing zero visits, yet our analysis pivots on the comparison of frequencies as captured in table 3.22, which enumerates observed frequencies over a 24-hour span.

To translate this probability into a frequency we can use for comparison, we will utilize the notion of expected frequencies. With the probability of observing no visits in any given hour standing at approximately 0.053, and given our 24-hour observation window, we can extrapolate this probability into an expected frequency for the entirety of the observed period. Multiplying the probability for a single hour by the total number of hours furnishes us with the expected frequency of encountering zero visits within our theoretical distribution over the complete 24-hour timeframe:

$$\begin{aligned} E_0 &= 24 \cdot P(X = 0) \\ &= 24 \cdot \frac{e^{-2.92} \cdot 2.92^0}{0!} = 1.294 \end{aligned}$$

We represent the expected frequency by E, and subscript 0 again refers to the value we are analyzing. Theoretically, we expected to observe around 1.294 instances of 0 visits. If we extrapolate all the E's:

Visits	E_0	E_1	E_2	E_3	E_4	E_5	E_6
Expected Frequencies	1.29	3.83	5.59	5.45	3.97	2.32	1.13

Table 3.23: Expected Frequencies of Guard Visits to a Keypad Over 24 Hours.

A general formula for the E's present in table 3.23 is:

$$E_i = n \cdot P(X = i) \tag{3.175}$$

With E_i denoting the expected frequency for category i over n intervals, we have established both an empirical and a theoretical framework based upon frequencies. And now we are ready to begin constructing a metric for comparison. A straightforward approach is subtracting the observed frequencies O_i from their corresponding expected frequencies E_i and summing these differences across all categories:

$$\sum_{i=0}^{n}(O_i - E_i) \tag{3.176}$$

Although this method points us in the right direction, it presents potential challenges, including the possibility of negative values and discrepancies in magnitudes between O_i and E_i. To circumvent the issue of negative differences, we square the outcome of each subtraction. This will not only eliminate negative values but also accentuate more significant discrepancies. To account for varying magnitudes and ensure a uniform comparison, we introduce a ratio by dividing the squared difference between O_i and E_i by E_i, thereby normalizing our comparison metric:

$$\chi^2 = \sum_{i=0}^{n}\frac{(O_i - E_i)^2}{E_i} \tag{3.177}$$

We will call the i in equation 3.177 a category; in our particular case, we have seven of them, representing 0,1,2,3,4,5 and 6 visits. Because our data sample is small, we used one category per frequency observed. With larger datasets, we often group observations within specific categories, in a similar way to what we did with a histogram.

I believe we know how to compute all the players in 3.177: O_i is the observed frequency for category i, E_i is the expected frequency

for category i. That leaves us with the task of naming this statistic; I suggest we name it "Dick Dastardly", does that ring a bell? (If not, please go and check out the Wacky Races cartoons!)

3.6.2 You Are Nothing More Than a Fancy X! The Chi-Square Test.

Actually, this statistic 3.177 is called chi-square (χ^2), the test name is the chi-square test, and the null distribution is the chi-square distribution. Yeah, very original, but I guess it was a popular name back in the day. With the test statistics defined, the next step to complete our testing framework is to understand the behavior of the null distribution. Well, since the chi's are in vogue, let's define the chi-square distribution.

3.6.2.1 OK, I Take It Back, You Are Usefull: The Chi-Square Distribution.

The chi-square distribution is used in statistics to understand how the sum of the squared values of several normal random variables behaves. The number of these variables is called the degrees of freedom, denoted by k. Depending on how many variables (or degrees of freedom) there are, the shape of the chi-square distribution will look different:

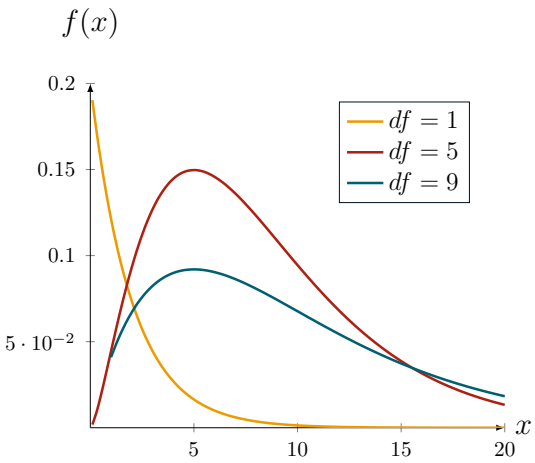


Figure 3.60: A salad of Chi-Square distributions with different degrees of freedom.

In figure 3.61, the chi-square distribution is illustrated for three degrees of freedom *df*: 1, 5, and 9, represented by yellow, red, and blue colors, respectively. The yellow curve with $df = 1$ exhibits a sharp peak near the origin and rapid decay, highlighting a high probability of low χ^2 values and significant skewness. As *df* increases to 5 (red), the curve transitions, showing a rise and gradual fall, indicating a shift towards less skewness and a more probable moderate range of χ^2 values. With $df = 9$ (blue), the curve becomes smoother and shifts rightward, reflecting a preference for slightly higher χ^2 values and approaching a more symmetrical shape, illustrating the Central Limit Theorem's influence on the distribution's form.

Mathematically, if we define X as a random variable and we want to express that fellow follows a chi-square distribution, we can do so with:

$$X \sim \chi^2$$

The probability density function for the chi-square distribution is given by the equation:

$$f(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{(k/2)-1} e^{-x/2} \tag{3.178}$$

Where:

- k represents the degrees of freedom in the chi-square distribution.
- $\Gamma(k/2)$ is the Gamma function evaluated at $k/2$.

We will skip the deduction of the PDF 3.178 because those calculations are complicated, extensive, and outside of the book's objective. The goal is not for us to be calculation machines but rather to understand the concept behind deriving densities, and we have gone through a few examples of how to get to those equations. From now on, we will be studying distributions by analyzing the equation of the PDF and not how to get there.

The $x^{(k/2)-1}$ component captures the behavior of squared deviations, scaling with the degrees of freedom k, which represent the number of independent variables in the analysis. This term ensures

that the distribution accurately reflects the variance in data as it becomes more spread out, with increasing k. Meanwhile, $e^{-x/2}$ introduces an exponential decline, acknowledging the decreasing probability of encountering extreme deviations from the expected frequencies as these deviations grow larger.

This new fellow Γ is the 'Gamma function'; it is here just as a normalization factor; it makes sure that our PDF integral is equal to one when we compute it on the whole domain. But how?

The Gamma function, denoted as $\Gamma(n)$, extends the idea of factorials to all real and complex numbers except for negative integers. For any positive integer n, the Gamma function is defined as $(n-1)!$, which means $\Gamma(n) = (n-1)!$. For example, $\Gamma(5) = 4! = 24$. What makes the Gamma function truly special is its ability to work with non-integer values. It fills in the gaps between the integers, providing a continuous curve that connects the dots defined by the factorial function.

To go out guns blazing with the chi-square test it's essential to grasp how degrees of freedom (df) are calculated, as our null hypothesis testing relies on the chi-square statistic 3.177 that is dependent on df.

Degrees of freedom in the context of the chi-square test are calculated based on the number of categories minus one $(k-1)$. Here's why: We are dealing with k categories, meaning that if we freely select $k-1$, the remaining one is automatically fixed by the total count of all categories combined. Thus, the degrees of freedom are $k-1$, reflecting that only $k-1$ category counts are independent variables, with the last one being dependent on the sum of the others.

However, the complexity increases when considering constraints that extend beyond the number of categories. For instance, when the expected frequencies E_i are modelled using a distribution such as the Poisson distribution, defined by a parameter λ, we encounter additional restrictions. Unlike the straightforward subtraction used to account for the total sum constraint, the impact of λ and other parameters of the distribution, denoted as p, must be considered differently. This does not alter the $k-1$ calculation directly but

emphasizes that the underlying distribution parameters also influence the degrees of freedom.

$$df = k - 1 - p$$

Where:

- $\boldsymbol{k}$: Is the number of categories or distinct groups in your data.
- $\boldsymbol{p}$: Is the number of parameters estimated from the data used to calculate the expected frequencies. For many tests, p might be 0, if no parameters are estimated directly from the data. When fitting to a specific distribution like the Poisson or binomial ones, p is typically 1 because you estimate a parameter like the mean (λ for Poisson) from the data.

For our specific case, we have seven different categories, meaning $k = 7$ and one parameter λ so $p = 1$ therefore:

$$df = 7 - 1 - 1 = 5$$

So, our test statistics follow a chi-square distribution with five degrees of freedom:

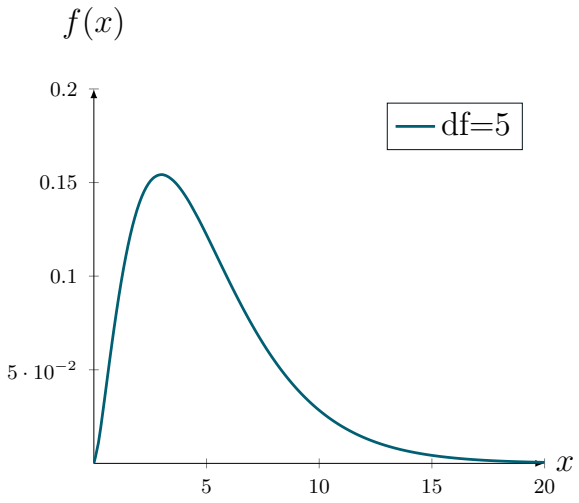


Figure 3.61: A Chi-Square distribution with 5 degrees of freedom.

Figure 3.61 is a chi-square distribution with five degrees of freedom (χ^2_5), and it also defines our null distribution, which will help

us to accept or reject the null hypothesis; speaking of which, let's define our hypothesis:

$$H_0 : \text{Data} \sim \text{Poisson}(\lambda)$$

Where λ is the rate parameter of the Poisson distribution, representing the average number of events in a fixed interval:

$$H_1 : \text{Data} \not\sim \text{Poisson}(\lambda)$$

With a significance level $\alpha = 0.05$ and our χ^2 statistic, we will be on our way to decision time, so let's get to work and calculate it:

$$\begin{aligned}
\chi^2 &= \sum_{i=0}^{n} \frac{(O_i - E_i)^2}{E_i} \\
&= \frac{(O_0 - E_0)^2}{E_0} + \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} \\
&+ \frac{(O_3 - E_3)^2}{E_3} + \frac{(O_4 - E_4)^2}{E_4} + \frac{(O_5 - E_5)^2}{E_5} \\
&= 5.30
\end{aligned}$$

Now, with the help of a computer we calculate the p-value:

$$p\text{-value} \approx 0.38$$

Given a significance level $\alpha = 0.05$, the p-value of is greater than α. This indicates that there is not a significant difference between the observed frequencies and those expected under a Poisson distribution with $\lambda = 2.7$. Consequently, the data is consistent with the expected model, suggesting a good fit. Therefore, the dataset passes the χ^2 test for how well it fits. Before moving on, let's review some key aspects of the χ^2 test:

- **Non-parametric Nature**: The chi-square test does not assume a data distribution, meaning it is a non-parametric test. This quality broadens its applicability to a variety of data types and distributions.
- **Requirement for Categorical Data**: It is designed explicitly for categorical (nominal or ordinal) data rather than continuous data. The data are grouped into categories, and the

test assesses the frequencies of occurrences within these categories.

- **Degrees of Freedom (*df*):** A pivotal concept in the chi-square test, the degrees of freedom, is determined by the number of categories under examination and any parameters estimated from the data. The df value influences the shape of the chi-square distribution used to interpret the test statistic.
- **Sensitivity to Sample Size**: The reliability of the chi-square test increases with larger sample sizes, otherwise the test might not be able to detect significant differences for smaller samples.

Something about those items worries me a bit. Yeah, the test is sensitive to the sample size, and we can't say that our sample contains an extensive number of observations; after all, we ran our test in 24 of them before we concluded that the sample size directly impacts the test's powers and, therefore, our ability to reject H_0, in situations where it is false. So, let's use a computer to calculate the power of this same test:

$$\text{Power} = 0.102$$

Ouch! The low power of the test hints at its limited capability to detect deviations from the Poisson model, potentially caused by the small sample size or the minimal effect size inherent in our data and analysis setup. Think about the methodology again; a better strategy could have been to pre-calculate the necessary sample size to achieve a desired power level, similar to our previous examination of how sample size affects hypothesis testing. Alternatively, we could switch to a more powerful statistical test that aligns better with the nature of our data. Collecting additional samples could also enhance the test's sensitivity. However, in our current scenario, we will proceed with the Poisson model defined by $\lambda = 2.7$, which predicts an average of about 70 codes per day based on our estimates.

As it stands, we've added three new distributions, a novel type of variable, and another test to our toolkit. Our current objective is to gain access to the Diamond Centre's garage. Inspired by Notabartolo's genius, our method to get a code that allowed us to enter the

garage was to install a camera facing the keypad that controls garage entry. Our goal is to capture video footage to help us identify the digits entered by the guards. To achieve this, we monitored the keypad over 24 hours, taking hourly samples of the guards' interactions with the device. This exercise led us to an important realization: attempting to discern patterns in the guards' routines could get our asses into an orange jump suit.

Despite these risks, our creativity helped us devise a way to extract valuable insights from the data. We introduced a new variable type—discrete—and developed a corresponding mathematical function, the Probability Mass Function (PMF). Our initial foray into using the PMF aligned with the binomial distribution, often associated with success or failure. Unfortunately, the results were disheartening; the likelihood of installing the camera undetected was discouragingly low.

This realization prompted us to expand our analysis to estimate the daily count of images captured, a metric crucial for determining whether we could collect enough data to decipher the garage code. This expansion brought two new distributions into play—the Poisson and the chi-square—along with the chi-square test. Our findings suggest that we might capture around 70 images per day. This number gives me hope that we might successfully capture the necessary code if we leave the camera in place for a few days. However, the possibility of getting caught still concerns me.

I value your dedication to this mission, but I must confess, the prospect of visiting you behind bars is less than appealing. We need to find a way to improve our chances, which we initially calculated using the binomial distribution. Let me explore alternative strategies that might enhance our odds of successfully installing the camera without detection.

Chapter 4

The Evidence Update Machine: Bayes.

In Antwerp, following in the footsteps of Notabartolo who had seamlessly blended in to gather intelligence, I took to the local area, hoping to learn the ropes. My journey led me to a grocery store where I met Rita, conspicuous for her vibrant scarf adorned with the insignia of K. Berchem Sport, the beloved local football club. Despite their position in the second amateur division, Rita spoke of the fervent support the team enjoys within the community.

“Every match day,” she proudly declared, “I display a flag from my window. It’s my small act to cheer on the team.”

Intrigued, I probed further about the tradition, uncovering a curious blend of dedication and occasional missteps. Rita confessed, “Well, sometimes I get so caught up with the excitement from our local fan club meetings that I put up the flag a day early, especially when a big game is on the horizon. It’s rare, but it happens!” She added that this does not happen very often, maybe 10% of times.

Learning about Rita’s dedication to K. Berchem Sport and how she and the entire neighborhood never miss a game led me to an intriguing question: ”Even the guards?” Rita replied, ”Well, I’m not sure, but I believe some of them do watch. Surely, if there’s a match, they would.” To test this hypothesis, we could check the game schedule and observe the patterns of the guards coming to the door during the same 24-hour period.

While I was digging online for the team’s match schedule, I hit

a dead end. We need a new strategy to figure out when matches are happening without the risk of being spotted around the stadium. What about Rita's flag? It is a simple yet ingenious sign that could be the key. By watching for her flag, we can stay clued in to match days from a distance, sidestepping the need to venture anywhere near the football stadium.

During our chat, Rita mentioned that matches occur weekly, more specifically on Saturdays. Considering her insights and assuming that some guards might attend these games, or at least keep track of them via their phones, we can infer a reduction in keypad checks on match days. We must integrate this valuable information with our existing knowledge to adjust our prior assumptions effectively. With the deduction of the binomial distribution, we learned that the probability of failure, which represents the likelihood that, if we pick a random hour of the day, we will see no guards at the keypad is:

$$P(\text{Failure}) = 0.125$$

Based on Rita's information, we know that a match is scheduled for Saturday, though the exact time remains unknown. To calculate the probability of the match being in progress at a randomly chosen hour, we can make some reasonable assumptions. Since the match occurs on Saturdays and typically lasts about two hours, we can focus on a 12-hour period during the day, rather than the full 24 hours, as it's highly unlikely the match would take place during late night or early morning hours. With this approach, we can proceed to calculate the probability accordingly:

$$P(\text{Match}) = \frac{2}{12} = 0.16$$

Another thing we can extract from the chat with our friend is the probability of a match happening given the flag is out:

$$P(\text{Match} \mid \text{Flag}) = \frac{9}{10} = 0.9 \tag{4.1}$$

It is not 100%: as she said, sometimes she puts the flag out on the day before the match. Now, in an ideal world what would be great to understand is?

$$P(\text{Guard not coming} \mid \text{Match}) \tag{4.2}$$

If we have a high probability of not having guards patrolling when a match is on the go, our likelihood of failure can change, and we must understand by how much. For that, let's start by expanding 4.2:

$$P(\text{Guard not coming} \mid \text{Match}) = \frac{P(\text{Guard not coming} \cap \text{Match})}{P(\text{Match})} \tag{4.3}$$

At the moment, we don't have information to compute that intersection; we don't really know what happens to the guards when the match is happening. However, we can make some inferences. Rita told us everybody loves the team, so we can use that information to turn 4.1 into:

$$P(\text{Match} \mid \text{Flag}) \rightarrow P(\text{Match} \mid \text{Guard not coming}) \tag{4.4}$$

We will treat the presence of Rita's flag as a reliable proxy, inferring that a match is probably occurring if guards are notably absent. This assumption rests on the possibility that, while some guards might attend the match, others will remain on duty for security. Given Rita's proven reliability about match days, we will trust her flag as a valid indicator for our planning. We can apply the formula for a conditional probability to 4.4:

$$P(\text{Match} \mid \text{Guard not coming}) = \frac{P(\text{Match} \cap \text{Guard not coming})}{P(\text{Guard not coming})} \tag{4.5}$$

Our goal is to have a way to compute 4.4, and for that, we will start by looking at the intersections:

$$P(\text{Match} \cap \text{Guard not coming}) = P(\text{Guard not coming} \cap \text{Match}) \tag{4.6}$$

The order of the intersections is irrelevant, so equality 4.6 holds. Looking at 4.5, we know the values for $P(\text{Match} \mid \text{Guard not coming})$ and $P(\text{Guard not coming})$ so, given that the probability of both intersections is the same, we can isolate the one in 4.5 and use this result in 4.3:

$$P\big(\text{Match} \cap \text{Guard not coming}\big) = P\big(\text{Guard not coming}\big) \cdot P\big(\text{Match} \;\big|\; \text{Guard not coming}\big)$$

Using this result in 4.3:

$$P(\text{Guard not coming} \mid \text{Match}) = \frac{P(\text{Guard not coming}) \cdot P(\text{Match} \mid \text{Guard not coming})}{P(\text{Match})} \tag{4.7}$$

Time to input some numbers in 4.7 and see if we improve our odds of installing a camera unnoticed:

$$P(\text{Guard not coming} \mid \text{Match}) = \frac{0.125 \cdot 0.9}{0.16} = 0.70$$

Thank you, Rita, for the info; we now know if we go to install a camera when the flag is outside the window, our chances of not encountering a guard go from approximately 12.5% to 70%, which is excellent. This technique, represented by 4.7, where we update our beliefs with new evidence, is called Bayes' theorem.

4.0.1 Bayes' Theorem: The Law That Defies Legislation.

If I create a histogram representing the counts of each mathematical technique I used to analyze or perform inference on data, Bayes' Theorem would have the biggest bar of them all by a mile. I mean look at this formula:

$$P(A \mid B) = P(A) \cdot \frac{P(B \mid A)}{P(B)} \tag{4.8}$$

So elegantly simple, yet profoundly impactful because it allows us to update our beliefs about the probability of an event A happening, given the occurrence of another event B. The beauty of Bayes' theorem lies in its ability to weave together prior knowledge $P(A)$ and new evidence $P(B \mid A)$ to refine our predictions, encapsulating the dynamic nature of learning and decision-making. We already know all the players in 4.8 but let me make a formal introduction:

- $P(A)$: The **prior probability** represents our initial belief about the probability of event A occurring, before considering the new evidence B. This can be based on historical data, earlier studies, or subjective judgment.
- $P(A \mid B)$: The **posterior probability** represents the probability of event A occurring given that event B has occurred. This updates our belief about A in light of new evidence B.

- $P(B \mid A)$: The **likelihood** represents the probability of observing event B given that event A has occurred. This is based on our model or hypothesis about the relationship between A and B.
- $P(B)$: The **marginal probability** of B. It represents the total probability of observing event B under all possible circumstances. This ensures that the posterior probability is a true probability value, acting as a normalizing constant.

I know, I know more mathematical vernacular, but at least these ones make us sound smart! From the items above, the first three names are very intuitive, and they represent notions that we were already familiar with. However, that marginal guy is a brand new concept.

Marginal probability is a fundamental concept in probability theory that refers to the probability of an event occurring, regardless of the outcome of other variables. In our scenario, $P(B)$ was the probability of a match occurring, irrespective of whether a guard was coming to the keypad or not, which indicates that to compute $P(B)$ we do it with A and $\neg A$ (not A), so:

$$P(B) = P(B \mid A)P(A) + P(B \mid \neg A)P(\neg A) \tag{4.9}$$

Equation 4.9 breaks the probability of an event B down into a weighted sum, accounting for every conceivable scenario that could precipitate B. If A occurs, we must consider B's occurrence conditional on A, thus $P(B \mid A)P(A)$. Conversely, recognizing that B must happen even in the absence of A, the formula also included this alternative, encapsulated by $P(B \mid \neg A)P(\neg A)$. If we consider our equation 4.7, Bayes' law reflecting the likelihood of a guard not coming given a match is happening, the marginal probability $P(\text{Match})$ is:

$$\begin{aligned} P(\text{Match}) = {} & P(\text{Match} \mid \text{Guard coming}) \cdot P(\text{Guard coming}) \\ & + P(\text{Match} \mid \text{Guard not coming}) \cdot P(\text{Guard not coming}) \end{aligned}$$

They are kind of cute, all these Bayesian formulas. Still, they haven't yet fully substantiated the claim I made at the beginning

of this chapter, where I said that Bayesian methods are my most widely used methodology. Let's build a stronger case by revisiting our experience with the χ^2 test. We found ourselves in a situation where we needed more evidence to reject our null hypothesis, as the power of our test was less than reassuring. Upon reviewing the specifications of the χ^2 test, we discovered that this technique is not the best for identifying nuances in small datasets. Typically, a hypothesis test provides a binary answer—yes or no—without much detail about what we are estimating.

Our test against a Poisson distribution defined by a λ yields a straightforward yes or no. But what if we could obtain more nuanced information about λ? Assessing the likelihood of λ being above or below a specific value would not be bad now that I think about it. In statistical terms, we want a distribution for λ from which we can extract more insights... hold on, a distribution of a parameter that defines a probability distribution? That is right, welcome to the world of Bayes!

$$P(\lambda \mid \text{data}) = \frac{P(\lambda) \cdot P(\text{data} \mid \lambda)}{P(\text{data})} \tag{4.10}$$

Equation 4.10 represents our posterior distribution for λ, which will provide us with a more nuanced view on that parameter, instead of just having it derived deterministically. To get to that same distribution, we need to define both the prior $P(\lambda)$ and the likelihood $P(\text{data} \mid \lambda)$; this brings something we haven't seen before, that 'data' parameter:

$$\text{data} = (x_1, x_2, ..., x_n)$$

Those x's are just variables; their meaning changes depending on the use case. As we are defining rates of hourly visits and we did so by using counts of daily hours, our x's will represent those same counts. Cool, having defined our variable data, we are now ready to move on to find a posterior distribution for our parameter λ. To do so, we will tackle the likelihood first:

$$P(\text{data} \mid \lambda) \tag{4.11}$$

In Bayes' law, the likelihood represents the probability of observing event B given that condition A has occurred. Translating this to

our equation 4.11 we can say that the likelihood function quantifies the probability of observing the data point given a set of parameters. Specifically, when we select a particular distribution to model our data, we are essentially hypothesizing about the underlying process that generates the data. We chose the Poisson distribution for this, so this bad boy seems like the perfect candidate here:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{4.12}$$

Equation 4.12 represents the PMF of the Poisson distribution for a single event k. However, when dealing with a dataset comprising multiple data points, we need to extend this concept to encompass all observations. This leads us to consider how the probabilities associated with individual events relate to one another. Does the occurrence of one event (e.g., x_0) affect the probability of subsequent events (e.g., x_1 or any other x)? In the context of the Poisson distribution, the answer is no. This is because the Poisson model assumes that events occur independently within the given interval. For example, having five visits on one day does not influence the count on the next day, assuming no information is provided about external factors like guard schedules or shifts. This assumption of independence is pivotal as it allows us to simplify the joint likelihood of observing the entire dataset. Mathematically, this is represented by:

$$P(\text{data} = (x_1, ..., x_n) \mid \lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} \tag{4.13}$$

Here, the Poisson distribution models the process generating the data (number of visits) given a specific rate (λ). It does not express prior beliefs about λ but quantifies how likely we are to observe the actual data given a particular λ. With our likelihood defined, we are now ready to determine our prior:

$$P(\lambda) \tag{4.14}$$

The prior distribution, $P(\lambda)$, represents our beliefs about λ before observing any data. It is independent of the specific data observed and remains unchanged unless new information is integrated. This can be defined by previous studies, expert knowledge, or can

even be assumed (non-informative) if little is known about λ beforehand. Well, we have none of those but fear nothing as we have a way of getting out of this uncomfortable situation; no we are not calling the Ghostbusters. Instead, we will rather study a concept called a conjugated prior.

In Bayesian statistics, a conjugate prior is a prior distribution that, when paired with a specific likelihood function, results in a posterior distribution that belongs to the same family as the prior. This characteristic greatly simplifies calculations and facilitates straightforward updates to the posterior as new data becomes available. Yes, that means exactly what you might be thinking; someone actually worked through various combinations of distributions to identify pairs that would yield a posterior of the same family, facilitating easier computations... we must thank these people!

However, it's not all smooth sailing. There is a trade-off. While conjugate priors offer computational benefits, they can sometimes limit the choice of the prior. Such limitations result from the necessity of selecting from a specific family of distributions that might only sometimes perfectly align with prior knowledge or beliefs about the data. Of course, we are free to choose any distribution we want for your prior, but this freedom comes at a cost, just as in real life. If we opt for a non-conjugate prior, the posterior might be hard to integrate, and we will probably need to employ numerical integration methods like Markov Chain Monte Carlo (MCMC) techniques, which we will explore later. For now, let's stick with the concept of conjugate priors because they work well despite the trade-offs and can be extremely useful in many situations! What is it the conjugated prior called? I thought you were never going to ask; it rhymes with Mamma... Gamma! We have already been introduced to the Gamma function, but now we will explore the Gamma distribution.

4.1 The Gamma Ray Laser Beam... Oh, it's another PDF, LAME! The Gamma Distribution.

The "Gamma distribution" is a versatile and widely used continuous probability distribution that is particularly useful in Bayesian statistics and various fields that involve modeling the time until an event occurs or the total magnitude of events over a specified interval. The PDF of the Gamma distribution is defined as:

$$f(x;\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \tag{4.15}$$

For $x > 0$, $\alpha > 0$, and $\beta > 0$, where:

- α is the shape parameter.
- β is the rate parameter.
- $\Gamma(\alpha)$ is the Gamma function evaluated at α.

Just like the normal distribution, the Gamma PDF has two parameters, namely α and β, that will affect the shape of our Gamma curve:

- **Shape Parameter ($\boldsymbol{\alpha}$)**: This influences the shape of the distribution. As α increases, the distribution's peak becomes more pronounced and shifts rightward, making the distribution more symmetrical for larger values
- **Rate Parameter ($\boldsymbol{\beta}$)**: Inversely affects the scale of the distribution. Higher β results in a steeper peak and a narrower spread, indicating a higher concentration of the variable x around a lower range of values.

Two more important metrics that allow us to have a better understanding of how the PDF behaves are the expectation (mean) and variance; they are determined directly by its parameters:

$$E[X] = \frac{\alpha}{\beta}, \quad \text{Var}(X) = \frac{\alpha}{\beta^2}$$

There is only one thing missing, and this is the family photo! Let's check how those parameters impact the shape of our curve. For that, we will fix an arbitrary value of α and make β vary. Then we will do the same, but for β, fixing it and letting our α vary:

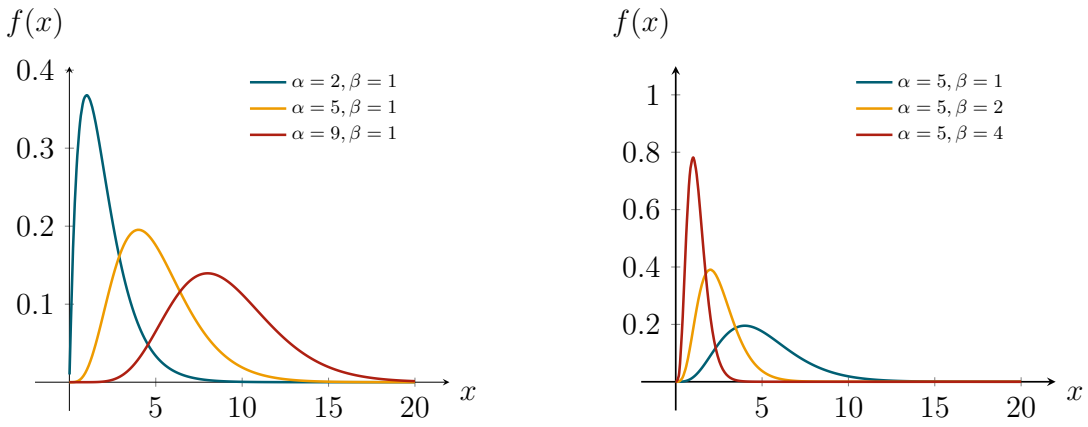


Figure 4.1: Miss α and mister β how do you impact Gamma?

In figure 4.1, we have two graphs showcasing the impact of both parameters α and β in the Gamma curve. The one on the left explores variations in the shape parameter α while maintaining a constant rate parameter $\beta = 1$. It reveals how increasing α progressively shifts the distribution's peak rightward and broadens its shape. Specifically, the blue PDF for $\alpha = 2$ peaks earlier and declines steeply, indicative of a distribution with lighter tails. As $\alpha = 2$ increases to 5 (yellow) and then to 9 (red), the peak not only shifts right but also flattens, suggesting a distribution capable of accommodating a wider range of values with significant probability.

The right hand one maintains a fixed shape parameter $\alpha = 5$ but varies the rate parameter β. The increase in β from 1 to 4 accelerates the decline of the PDF post-peak, effectively compressing the distribution and heightening the peak. This illustrates the rate's influence on controlling the spread and peak of the distribution. While variations in α primarily affect the distribution's skewness and the location of its peak, alterations in β adjust the steepness of the peak and the rapidity of decline, illustrating a clear delineation in roles between the two parameters. It's yet another distribution we can add to our collection and use in the future. So our prior

$P(\lambda)$ can be expressed by a Gamma distribution:

$$P(\lambda) = \text{Gamma}(\alpha, \beta) \Leftrightarrow P(\lambda) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$$

Right, with this, our posterior equation becomes:

$$P(\lambda \mid \text{data}) \propto P(\text{data} \mid \lambda) \cdot P(\lambda) \tag{4.16}$$

Not a bad equation that 4.16, plus it comes with a present. We have a new symbol, $\propto$, which denotes proportionality. This symbol is significant in Bayesian statistics as it simplifies the expression of complex relationships. Specifically, in this context, it indicates that the posterior distribution is proportional to the product of the prior distribution and the likelihood function.

We say that the posterior is proportional to this multiplication rather than equal because we often omit the normalization factor, commonly referred to as the marginal likelihood or evidence. This factor is integral to ensuring that the posterior distribution sums to one, thereby qualifying as a valid probability distribution; however, when we work with conjugate priors that are mathematically convenient and result in a posterior distribution of the same family as the prior, the normalization factor can sometimes be implicitly calculated or conveniently ignored during intermediate steps. This simplification arises because conjugate priors are designed to mesh neatly with the likelihood function, simplifying the mathematical operations involved. OK, so now we can input our equations for the prior and likelihood into 4.16:

$$P(\lambda \mid \text{data} = (x_1, ..., x_n)) \propto \frac{\beta^{\alpha}}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} \cdot \prod_{i=1}^{n} \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} \tag{4.17}$$

That 4.17 looks worse than a mugshot on my wall, which was taken after a night on the neat vodka; it is a friend, not me... However, there is hope as I see some common e's and λ's in the middle of those nasty multiplications. We will tackle the e's first:

$$e^{-\beta\lambda} \cdot e^{-n\lambda} = e^{-(\beta+n)\lambda} \tag{4.18}$$

The term $e^{-n\lambda}$ results from the product of n terms $e^{-\lambda}$. Moving to the λ's:

$$\lambda^{\alpha-1} \cdot \prod_{i=1}^{n} \lambda^{x_i} = \lambda^{\alpha-1+\sum_{i=1}^{n} x_i} \tag{4.19}$$

If we combine 4.18 and 4.19:

$$P(\lambda \mid \text{data}) \propto \lambda^{\alpha+\sum_{i=1}^{n} x_i-1} e^{-(\beta+n)\lambda} \tag{4.20}$$

Equation 4.20 bears a strong resemblance to the PDF of a Gamma distribution, albeit with notable modifications in the exponents of the terms involving λ and e. Specifically, the exponent of λ in this equation is $\alpha + \sum_{i=1}^{n} x_i - 1$, whereas, in the standard Gamma distribution, it is simply $\alpha - 1$. Additionally, the term involving e in our equation has an exponent of $-(\beta + n)\lambda$, which differs from the $-\beta\lambda$ found in the traditional Gamma distribution. So if we define $\alpha^{'}$ and $\beta^{'}$ like:

$$\alpha^{'} = \alpha + \sum_{i=1}^{n} x_i, \quad \beta^{'} = \beta + n$$

We can define a Gamma distribution with parameters $\alpha^{'}$ and $\beta^{'}$:

$$P(\lambda \mid \text{data}) \propto \Gamma(\alpha^{'}, \beta^{'}) \tag{4.21}$$

Because $\alpha^{'}$ and $\beta^{'}$ define the PDF of a new Gamma distribution, we don't have to worry about normalization, since all the well-defined probability distributions integrate to 1, the Gamma being of them. This results in us having a posterior $P(\lambda \mid \text{data})$ that is already normalized. Alright, there is just one thing left to do: input some values into 4.21 and check our PDF for λ. Oh, we need $\alpha^{'}$ and $\beta^{'}$, but these fellows are dependent on α and β. They define our prior, so let's start there:

$$P(\lambda) \sim \Gamma(\alpha, \beta)$$

Now we are stuck; without an α and a β, we can't have a prior; therefore, the dreams of a posterior are over. Before disappointment consumes us, let's consider what information the prior brings. The prior represents our initial beliefs; in this particular case, it will reflect what we know about the parameter λ, and we are saying a Gamma distribution reflects it.

When we introduced the prior, we said that one way to define it is with domain knowledge, well, we know that λ represents an expectation, so, what if we centre the Gamma distribution around that same parameter? With this, at least the expected value of the Gamma will be λ, which, let's face it, is far better than a random guess:

$$\lambda = \frac{\alpha}{\beta} \tag{4.22}$$

In equation 4.22 we define λ to be equal to the expected value of the Gamma distribution. However, we are still stuck; out of α, β and λ we only know the latter. One thing we can do is define α, β such that it's ratio is λ, for example, if we defined a scaling factor t:

$$\alpha = \lambda \cdot t, \quad \beta = t$$

If we divide α by β, the t's will cancel out, giving us λ; however, we know we have a new variable t, and our goal was to have fewer unknown terms. Worry not though, as we are moving forwards. For instance, now we have both α and β defined as a function of t and λ, meaning that, if we understand what that t does to our Gamma curve, we can define it in such a way that fits our needs. Regarding any probability distribution, two metrics provide a lot of information: the expected value and the variance. If we explore the expected value, the t's will cancel out. We just did that, so let's examine the variance:

$$\frac{\alpha}{\beta^2} = \frac{\lambda \cdot t}{t^2} = \frac{\lambda}{t} \tag{4.23}$$

GOTCHA! The formula 4.23 is defined by only λ and t; not only that, it shows how t adjusts the concentration of the distribution around its mean, and that is great because it will allow us to select one of those bad boys:

- **Higher values of t**: These result in a lower variance, indicating greater confidence in the estimate of λ. This produces a narrower distribution, reflecting a strong prior belief in the estimated rate.
- **Lower values of t**: These increase the variance, suggesting less certainty about λ. This results in a broader distribution,

accommodating a wider range of possible values for the rate parameter.

For example, we can start with $t = 1$, which results in $\alpha = \lambda$ and $\beta = 1$. With the parameters defined like that, we set the mean at λ and gave a variance of λ, making the distribution more informative, as we have less variance. Remember this is our prior belief about the parameter λ; it represents how we "think" λ behaves. Next we will update these beliefs with the likelihood, the piece of the puzzle that brings out all the information in the data. Well, let's define this prior:

$$P(\lambda) = \Gamma(2.92, 1) \tag{4.24}$$

And just like that, we have our prior. Let's plot it!

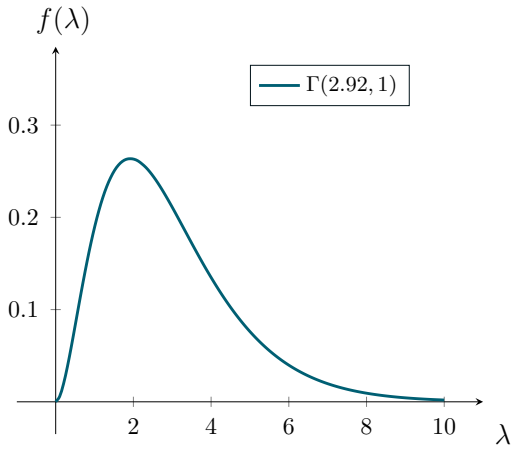


Figure 4.2: Where our believes are.

When examining figure 4.2, we observe that the likelihood of observing zero visits per hour is practically nil, aligning with our expectation that, even on quieter days, the keypad will receive some visits. The distribution shows a significant probability mass between 1 and 4 trips per hour. This concentration within a moderate range suggests that we wouldn't expect more than 5 visits in any given hour under typical conditions. Beyond this range, the probability sharply declines, indicating that guards rarely visit the keypad more than four times in an hour. Now, we are ready to compute our posterior parameters $\alpha^{'}$ and $\beta^{'}$:

$$\alpha^{'} = \alpha + \sum_{i=1}^{24} x_i$$

First we need to sum up all of our observations:

$$\sum_{i=1}^{24} x_i = 70$$

This means that our new $\alpha^{'}$ parameter is:

$$\alpha^{'} = 2.92 + 70 = 72.92$$

Now for the missing guy $\beta^{'}$:

$$\beta^{'} = \beta + n = 1 + 24 = 25$$

So our posterior for λ is defined by:

$$P(\lambda|\text{data}) \propto \Gamma(72.92, 25) \tag{4.25}$$

Well, let's plot it:

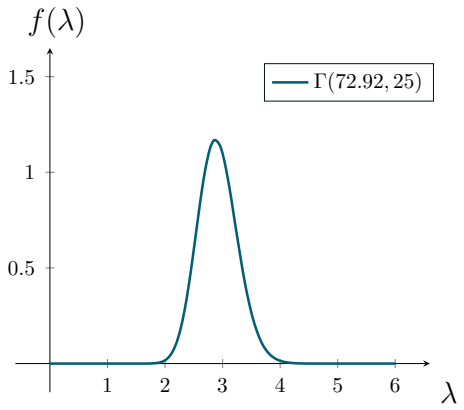


Figure 4.3: When we let our belief be updated with new evidence: The posterior.

Figure 4.3 represents the posterior distribution for our parameter λ. This distribution is noticeably narrower than the one representing our prior beliefs, as depicted by figure 4.2, This is an indication that the new evidence incorporated through the likelihood has provided us with a more precise estimate of λ, so let's check what we can expect from this fellow:

$$E[\lambda] = \frac{\alpha}{\beta} = \frac{72.95}{25} = 2.918$$

4.1. The Gamma Ray Laser Beam... Oh, it's another PDF, LAME! **The Gamma Distribution**.

All of this mathematical deduction to end up with an expectation? Never! We have a probability density function and that, allows us to extract more detailed information about λ compared to hypothesis testing and confidence intervals. For example, we can calculate the probability of $\lambda > 3$:

$$P(\lambda > 3) \tag{4.26}$$

Equation 4.26 asks for the CDF, and we are yet to define the one for our new specimen, the Gamma distribution, so here we go:

$$F(\lambda) = \frac{\gamma(\alpha, \beta\lambda)}{\Gamma(\alpha)} \tag{4.27}$$

Where:

$$\gamma(\alpha, z) = \int_0^z t^{\alpha-1} e^{-t} dt$$

Besides not having a closed solution, equation 4.27 is quite scary. However, we can use computer software to do these calculations; the main thing is to understand what a CDF represents:

$$P(\lambda > 3) = 1 - P(\lambda \leq 3)$$

Where:

$$P(\lambda \leq 3) = F(3)$$

Nothing new, just the old maneuvering of complements, as the CDF provides the probability up until a certain point, including it. After asking our computer for the calculations, this is what we get:

$$P(\lambda > 3) = 0.389$$

It's not that great, but there is a 39% probability of having a daily visit rate higher than three. However, this was to showcase where the power of Bayesian statistics lies. We have a PDF and a CDF that allow us to direct probabilistic statements about parameters, such as λ. We can check the expected value for λ, its variance, and how likely the parameter is to be over a certain threshold, just as we did with $P(\lambda > 3)$. These probabilities are directly derived from the posterior distribution and are intuitive for decision-making. By

contrast, the χ^2 test primarily assesses whether observed frequencies significantly differ from expected frequencies under a specified hypothesis, which is a more indirect method of inference that does not provide probabilities for parameter values.

Another feature of the Bayesian framework is that it is inherently adaptive. As new data becomes available, the posterior distribution can be updated, which is crucial in fields where data accrues over time or initial sample sizes are necessarily small. By contrast, the χ^2 test provides a one-time assessment based on a fixed dataset, so it must be re-evaluated entirely with each new data addition. Furthermore, in small sample sizes, traditional methods like the χ^2 test may have limited power to detect actual effects, leading to a higher risk of Type II errors (false negatives). The Bayesian method, by utilizing prior distributions, mitigates some of the issues related to small samples by effectively 'borrowing strength' from prior knowledge. This technique can lead to more robust and stable estimates, even when the data alone are insufficient for reliable inference.

Not convinced yet? I must admit, I'm pretty biased towards the Bayesian framework, and for good reason-it fundamentally changed my outlook on life. My second job as a data scientist, I was as a junior, working alongside a senior data scientist. We were part of a brand-new team, starting our roles on the same day. When I arrived at the office, I saw our names displayed on a large screen at the entrance. There it was, "Jorge Brasil - junior data scientist" next to a senior data scientist whose name hinted he was from the Middle East. My initial, unguarded thought was, "Oh man, this job will be boring..." I never truly believed Middle Eastern people were uninteresting, but I guess I had unconsciously absorbed some stereotypes.

I couldn't have been more wrong. My manager, who was also from the Middle East, proved to be incredibly supportive and anything but dull. As for my senior colleague, he truly changed my life. I probably learned more about mathematics applied to machine learning in the eight months we worked alongside each other than I had in my degree. He didn't just share his knowledge of maths; he also shared stories from his culture, enriching my understanding of

a world far beyond my experience. He boosted my confidence, encouraged me to persevere in my career, and, notably, was a fervent advocate for Bayesian thinking. So, after that lesson, my beliefs needed to be updated with new evidence. This experience was so impactful that, from that day onwards, I started approaching my life much like Bayes' Law suggests. Ideas and beliefs are temporal and always subject to updates when new evidence presents itself.

I have another example that further demonstrates the Bayesian framework's effectiveness. Remember our contact from the traffic department? I recently ran into him on the street and seized the opportunity to strike up a conversation, hoping to gather some new evidence. He invited me for coffee at his favorite café, conveniently located right across from the door where the keypad is situated.

During our chat, he brought up Rita, an old friend of his. Interestingly, he mentioned that she had taken a job in Paris and would be moving in two weeks. This news was significant because it meant losing our indirect method of determining whether it was a match day. However, all was not lost. Since he frequents the same café daily, I asked if he had noticed any patterns in the guards' visits to the keypad. He couldn't provide exact numbers or confirm if their visits fluctuated by the time of day, but he did note that there were noticeably fewer people around on weekends. With Rita gone, this became our only hope of finding a timeframe in which to go and install the camera, so I went down and collected two data samples, one for a weekday and another one for the weekend when Rita's flag was not out:

Hour	1	2	3	4	5	6	7	8	9	10	11	12
Successes	1	1	0	1	1	0	1	0	1	0	0	1

Hour	13	14	15	16	17	18	19	20	21	22	23	24
Successes	1	0	1	0	1	0	1	1	1	0	0	1

Table 4.1: Successes and failures of keypad visits during the weekdays.

Hour	1	2	3	4	5	6	7	8	9	10	11	12
Successes	0	0	0	1	0	0	1	0	1	0	0	1

Hour	13	14	15	16	17	18	19	20	21	22	23	24
Successes	0	0	1	0	1	0	0	1	0	0	1	0

Table 4.2: Successes and failures of keypad visits during the weekend.

We have collected two data samples at different times, as shown in table 4.1 for a weekday and table 4.2 for a weekend. Each table presents data for every hour of a single day, represented by binary entries. Just as before, in these tables, a '1' signifies a success, meaning that at least one guard appeared at the keypad, and a '0' indicates that no guards were present. This data setup is analogous to the way we used the binomial distribution in previous analyses. However, our current focus is on comparing these two distinct periods, to determine the optimal time for installing a camera without being noticed. What will we do? Well, I think the cat is out of the bag. Honestly, I always wanted to use that expression, and the cat we are talking about is the binomial distribution, so let's start by computing the likelihood of success p for each of the tables:

$$p_{\text{ Weekdays}} = \frac{14}{24} \approx 0.583, \quad p_{\text{Weekends}} = \frac{8}{24} \approx 0.333$$

With both of those p's we can define our two binomial distributions:

$$P(X = k)_{\text{Weekdays}} = \binom{24}{k} (0.583)^k (0.417)^{24-k}$$

$$P(X = k)_{\text{Weekends}} = \binom{24}{k} (0.333)^k (0.666)^{24-k}$$

With the goal of avoiding jail time, one avenue we can explore to understand the best timeframe for installing the camera is to do a hypothesis test comparing both p's, just as we previously did with the μ's. This time, we don't only want to accept or reject a hypothesis; we would like to quantify the idea that, if one time slot is better than the other one, we want to know how often this happens. As this is not possible with traditional hypothesis testing, let's resort to the Bayes framework. Because we are on the hunt for a comparison for the p's, an excellent way to start would be to define a posterior for these sisters:

$$P(p \mid \text{data}) \propto P(p) \cdot P(\text{data} \mid p) \tag{4.28}$$

We need to define a prior and a likelihood to have an expression for the posterior 4.28, a task that is no longer foreign to us; so let's start with the likelihood. The likelihood depicts the new evidence

from our dataset that represents successes and failures, so it makes sense that our likelihood is a binomial function.

It is true that, before when introducing the Bayes' law, we used the knowledge we got from the binomial to determine our prior, but this is because the evidence came from the conversation with Rita and not from the data sample of the hourly visits. Therefore, we used it as prior knowledge. Here, the setup is different; we will use the information from our data samples 4.1 and 4.2 as evidence to update our prior beliefs. So, we must discuss what to do for dinner... oh, sorry, wrong forum; the discussion here is around what prior distribution to select, and, once more, we will recur to the conjugated prior.

4.2 Beta? Is This Not The Final Version Of The Distribution?

Please welcome the Beta distribution, the conjugated prior to the binomial distribution. It is a continuous fellow that varies from 0 to 1, making it a great candidate for module rates like our p. Let's explore the Beta distribution before jumping into the model, starting with the PDF:

$$f(x;\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} \tag{4.29}$$

Where $B(\alpha,\beta)$ is the Beta function that normalizes the distribution, defined by:

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

Just as with the Γ the parameters α and β have a critical role in the shape of the curve:

- $\boldsymbol{\alpha}$: Is the shape parameter associated with the success count in a Beta distribution. It directly influences how the distribution accumulates probability density near the lower boundary of 0.
- $\boldsymbol{\beta}$: Is the shape parameter associated with the failure count in the context of a Beta distribution. It influences how the distribution accumulates probability density near the upper boundary of 1.

The expected value and the variance of the Beta distribution have the following expressions:

$$E(X) = \frac{\alpha}{\alpha + \beta}, \quad \text{Var}(X) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

The shape and behavior of the Beta distribution depending on the values of α and β. This summarized as follows:

- $\boldsymbol{\alpha > \beta}$: The distribution is skewed towards 1, with a peak closer to the right hand end of the interval. This indicates a higher probability density near 1.
- $\boldsymbol{\alpha < \beta}$: The distribution is skewed towards 0, with a peak closer to the left hand end of the interval. This shows a higher probability density near 0.
- $\boldsymbol{\alpha = \beta}$: The distribution is symmetrical about 0.5, which is especially evident when both parameters are greater than 1. This symmetry means that the distribution has its mode (peak) at the middle of the range, and is equally likely to take values below or above 0.5.

Time to put some of these α's and β's to work and visualize how they behave:

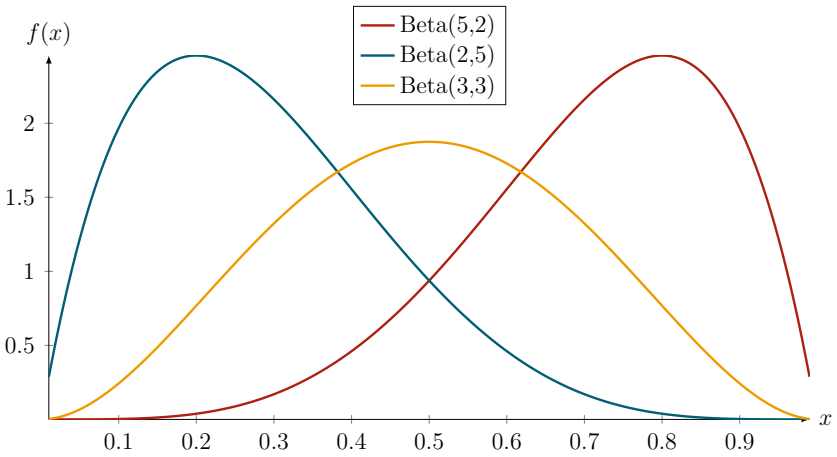


Figure 4.4: Are the Betas conspiring?

Figure 4.4 displays the Beta distributions for three distinct parameter configurations. The red curve, characterized by parameters $\alpha = 5$ and $\beta = 2$, illustrates a distribution that is heavily skewed towards the upper boundary of the interval, indicating a greater likelihood of outcomes near 1. This skewness is characteristic of scenarios where occurrences classified as 'successes' significantly outnumber 'failures'. Conversely, the blue curve, with parameters $\alpha = 2$ and $\beta = 5$, depicts a distribution that leans towards the lower boundary, or 0. This orientation suggests a dominance of 'failures' over 'successes', as reflected by the higher β value relative to α. The third curve, with the yellow jacket, utilizes balanced parameters $\alpha = 3$ and $\beta = 3$, resulting in a symmetrical distribution centered precisely at the midpoint of the interval, 0.5. This symmetry indicates an equal probability of outcomes occurring at either end of the scale. This is typical of situations where successes and failures occur with similar frequency.

There is a special case of the Beta distribution that is also worth taking a look at. It is the one when $\alpha = \beta = 1$. Let's replace these values on our PDF 4.29 and see what comes out of it:

$$f(x; 1, 1) = \frac{x^{1-1}(1-x)^{1-1}}{B(1,1)}$$

Where $x^0(1-x)^0 = 1$, simplifying the equation. The Beta function $B(1,1)$ is calculated as:

$$B(1,1) = \frac{\Gamma(1)\Gamma(1)}{\Gamma(1+1)}$$

Since the Gamma function $\Gamma(1) = 1$ and $\Gamma(2) = 1$, the Beta function $B(1,1)$ equals 1. Therefore, the PDF becomes:

$$f(x;1,1) = 1$$

This means that whatever value of x we input in our $f(x;1,1)$, it's density is always 1; consequently, every event will have the same likelihood of outcome:

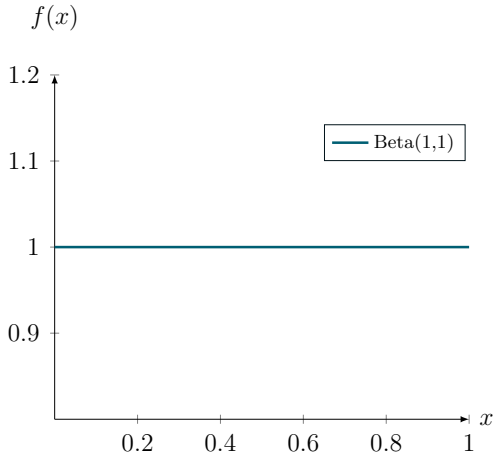


Figure 4.5: That is indeed uniform, I mean, it is a line!

What do you think about that PDF in graph 4.5? It is simple, that is for sure, but it also makes sense; if every x between 0 and 1 has the same density of 1, the PDF of a Beta(1,1) has to be a straight line. We do have a name for a distribution whose PDF plot looks like 4.5; it is called the uniform distribution.

4.3 No Favorites: The Impartiality of Uniform Distribution.

The uniform distribution is a basic model in statistics that assumes an equal probability of occurrence for all outcomes in a continuous range. It is a continuous distribution, characterized by the following probability density function (PDF) over an interval $[a, b]$:

$$f(x;a,b) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases} \tag{4.30}$$

Please do not misinterpret the fact that $B(1,1) = 1$ as suggesting that all events under this distribution have a probability of one. The PDF in probability theory represents density, not probability. Consequently, the area under the curve, which forms a rectangle in the case of a uniform distribution, must sum to one. This ensures that the function correctly defines a probability distribution. Equation 4.30 describes a completely flat probability curve between the limits a and b, ensuring that every point within this range is equally likely, adhering to the principles of a continuous uniform distribution. The expectations and variance for a uniformly distributed random variable are defined based on the uniformity of this range, giving a straightforward calculation for these statistical measures. For our expectation and variance, I will just showcase the formulas as we have done this several times already, and the process is the same:

$$E(X) = \frac{a+b}{2}, \quad \text{Var}(X) = \frac{(b-a)^2}{12}$$

Although defined as continuous, the uniform distribution can also be used to model discrete events. An example of this, is the famous example of rolling a fair dice, where each number has an equal chance of $\frac{1}{6}$. We got sidetracked a bit by this distribution, but it was just there for grabs; after all it is another one that might be useful, for example in neural networks, it's common to initialize weights randomly, to break symmetry during training. Using a uniform distribution across a small range ensures that no single neuron initially dominates the output, promoting effective learning.

OK, so we have our prior defined by a Beta distribution and based upon our data, the likelihood model is also established, which is a binomial distribution. We now need to do some calculations on 4.28 to come up with the formalae for our posterior distribution that will represent our parameter p. We know that, because we are using the conjugated prior; the posterior is from the same family as the prior, meaning that it will have to be a Beta, but let's verify this:

$$P(p \mid \text{data}) \propto P(p) \cdot P(\text{data} \mid p) \tag{4.31}$$

Substituting the likelihood and prior in 4.31 this expression gives:

$$P(p \mid \text{data}) \propto \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)} \cdot \binom{n}{k} p^k (1-p)^{n-k}$$

Simplify by combining powers of p and $1-p$:

$$P(p \mid \text{data}) \propto p^{k+\alpha-1}(1-p)^{n-k+\beta-1} \tag{4.32}$$

Equation 4.32 shows that the posterior distribution is still in the form of a Beta distribution (due to the powers of p and $1-p$). Therefore, we can recognize it as:

$$P(p \mid \text{data}) = \text{Beta}(k+\alpha, n-k+\beta) \tag{4.33}$$

Where k represents our success and n is the number of trials.

Now, with 4.33 defined, we are one step away from having a posterior probability distribution for our parameter p. We just need to select values for α and β. As these guys define the Beta distribution representing our prior, they should reflect our initial beliefs about $P(p)$.

We know that $P(p)$ represents the probability of success, specifically the likelihood of observing guards at the keypad (remember we define a success as a guard visit). Considering the environment, a Diamond Center secured by guards, it seems reasonable to assume we will observe guards at the keypad more often than not, indicating more successes than failures. Therefore, choosing $\alpha > \beta$ will skew our distribution towards one, reflecting our expectation of more frequent guard appearances:

$$\alpha = 4, \quad \beta = 2$$

The more extreme the ratio between α and β, the more pronounced the skewness of the prior. In our situation, we assume there will be more guards showing up than not, but we expect to see some times when they don't show up, so we want some skewness but nothing extreme. The reality is that if we don't use numerical methods (MCMCs) to select these parameters. We are just relying on empirical knowledge, so it will be a matter of trial and error:

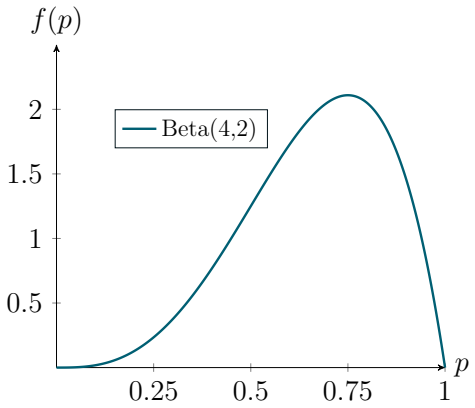


Figure 4.6: Our prior Beta, a nice curve indeed.

The curve on graph 4.6 represents our prior, defined by $\alpha = 4$ and $\beta = 2$. One thing that stands out is the concentration of density mass around the value of 0.75, indicating that we foresee more successes than failures; speaking of which, let's quantify what we can expect in terms of the probability of success:

$$E\left[p\right] = \frac{4}{2+4} = \frac{2}{3} = 0.66$$

Our prior distribution expected value is approximately 66%, indicating a success rate of being visited by a guard in a given hourly time slot when we try to install the camera. With the prior distribution established, we now have everything we need to create both our posteriors representing p_{Weekends} and p_{Weekdays}. Let's start with the weekdays. Our datasets reveal that, on the weekdays, we had 14 out of 24 hourly slots where we observed at least one guard, constituting the same number of successes. This finding, which allows us to define $k = 14$ and $n = 24$, directly impacts our understanding of the dataset:

$$P(p_{\text{Weekdays}}) = \text{Beta}(14+4, 24-14+2) = \text{Beta}(18, 12) \quad (4.34)$$

Now, we need to apply the same logic to define p_{Weekends}. In terms of hours where we see guards coming, in this particular slot, we see eight; therefore, $k = 8$. When it comes to α, β, and n, they will have the same values as the ones we used to define the posterior for p_{Weekends}. Remember that we must consider the same prior, as we are trying to see which time slot is better. Therefore,

we must update the same beliefs with different evidence to have fair comparisons, meaning that α and β must be the same:

$$P(p_{\text{Weekends}}) = \text{Beta}(12, 12) \tag{4.35}$$

Nice! So now that we have both our posteriors, how about we look at them?

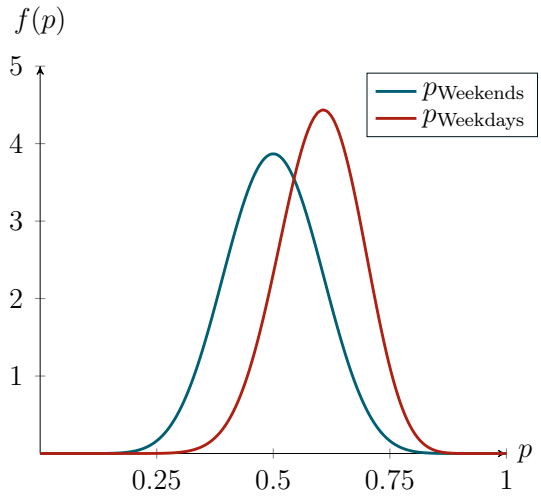


Figure 4.7: A majestic oil painting of a mountain landscape? Ah, not quite, just the two Betas.

In figure 4.7, we display both the weekend and the weekday posteriors. The blue curve represents the distribution we get when we update our prior beliefs with evidence from the sample of data we collected on the weekend. By contrast, our red buddy represents the posterior but is updated with the data collected over the weekday. Well, they differ! Visually, we can see that it will be beneficial for us to go install the camera over the weekend as there are fewer successes. Remember, we are looking for failures from the guards, these are the instances where they don't show up; let's quantify this by calculating the expectations for each case:

$$E\left[p_{\text{ Weekdays}}\right] = \frac{18}{30} = 0.6, \quad E\left[p_{\text{Weekends}}\right] = \frac{12}{24} = 0.5$$

We have a higher expected value for the weekday time slot, meaning that we are expected to have more visits on that same time frame. Therefore we will have better off on the other time frame, namely the weekend. But now we have distributions that showcase

the behavior of our success rate, and with that comes randomness, so the question naturally arises: "How often is the rate of success on weekdays higher than the one on weekends?" Mathematically, we can write that like this:

$$P(p_{\text{Weekdays}} > p_{\text{Weekends}}) \tag{4.36}$$

To calculate 4.36 we need to deduce the probability involving both posterior distributions:

$$P(p_{\text{Weekdays}} > p_{\text{Weekends}}) = \int_0^1 \int_0^y f_{\text{Weekdays}}(x) f_{\text{Weekends}}(y)\, dx\, dy \tag{4.37}$$

That double integral in 4.37 is nasty, but fortunately, we have a closed-form solution for it, so I will save us from the calculations and present not only the formula but the computed result also:

$$P(p_{\text{Weekdays}} > p_{\text{Weekends}}) = \sum_{i=0}^{\alpha_{\text{Weekdays}}-1} \frac{B(\alpha_{\text{Weekends}} + i, \beta_{\text{Weekends}} + \beta_{\text{Weekdays}})}{(\beta_{\text{Weekdays}} + i) B(1 + i, \beta_{\text{Weekdays}}) B(\alpha_{\text{Weekends}}, \beta_{\text{Weekends}})}$$

I told you it was not pleasant to the naked eye, but it is not hard to compute; it is a summation where we have to calculate densities and then deduce parameters. With the help of a computer, I got the likelihood for us:

$$P(p_{\text{Weekdays}} > p_{\text{Weekdend}}) = 0.94$$

That is an excellent result; not only do we have a higher success expectation for weekdays than weekends, 0.6 vs. 0.5, meaning that guards are more likely to come during weekdays, but we also expect that this same rate, 94%, of the time is higher on the weekday case.

With this methodology, we created an alternative to a two-sample hypothesis test, giving us two methods to choose from if we wish to perform an A/B test, another name for conducting inference over two samples.

These posterior distributions will come in handy more often than not. For example, it is very common for e-commerce companies to make changes to their websites to achieve improved performance. One of the key metrics they focus on is the conversion rate. This rate is crucial, as it represents the proportion of users out of the total

number of, for instance, visitors who successfully complete a desired action, by making a purchase. Understanding and improving this metric is a fundamental part of optimizing e-commerce performance:

$$\text{Conversion Rate} = \frac{\text{Number of conversions}}{\text{Number of visits}}$$

Suppose the team decides to test changing the color of the logo. They present the new version to one cohort of users over a period while another group continues to see the old one. Naturally, they want to determine if the changes yield better results, and the conversion rate is the metric to analyze. Our task is to conduct the analysis to determine if the logo's new version outperforms the old one. Our first task is to estimate the test time, which means defining the sample size. Because we are testing the old logo versus the new one, we opt for the two-sample hypothesis test way and calculate the sample size based on a specific power value we feel comfortable with:

$$\text{Power} = 0.8 \Rightarrow n = 10000$$

$$\text{Daily visits} = 500 \Rightarrow \text{Test days} \approx \frac{10000}{500} = 20$$

All these numbers are hypothetical and serve merely as an example. Our calculations suggest we require around 20 days to collect our samples for power value. When we communicate this to the team, they say we must have the results in 10 days. We can't afford to have the test running for any longer.

Now, what do we do? Compromise on the power? Instead, we can turn to the Bayesian framework! Unlike traditional approaches, Bayesian statistics do not require a fixed sample size; instead, they allow continuous updating as new data becomes available. This adaptability makes it possible to assess test outcomes dynamically, and adjust decisions on the fly without waiting for the total sample size to be collected.

Furthermore, Bayesian methods provide a probabilistic understanding of the results, which includes an explicit quantification of uncertainty. This phenomenon is beneficial when making decisions under constraints, as it allows stakeholders to weigh the potential

risks and benefits of early actions based on the strength of the evidence accumulated to date. Because we have a probability distribution, we can understand things like variance and derive better insights. It is not a competition between Bayes and frequentists, well at least not for me, but it is another methodology we know exists and can use, so let's list their main differences:

- **Data Interpretation:**
  - **Frequentist Framework:** Views data as a fixed outcome from repeatable random samples. Results are interpreted over the long run of repeated experiments.
  - **Bayesian Framework:** Treats data as observed from a realized sample, with parameters considered as random variables. Results incorporate prior beliefs and data.
- **Sample Size:**
  - **Frequentist Framework:** Requires a fixed sample size determined before the experiment begins, based on power analysis.
  - **Bayesian Framework:** Allows for flexible sample sizes; analysis and decisions can be updated as new data comes in.
- **Decision Making:**
  - **Frequentist Framework:** Decisions are based on rejecting or not rejecting hypotheses through p-values.
  - **Bayesian Framework:** Provides probabilities of hypotheses, offering a more nuanced understanding of outcomes with explicit quantification of uncertainty.

It seems we have a timeframe for you to install the camera, and the selected slot is the weekend. Yes, we'll have to accept that a 50% probability is sufficient to proceed with the installation. Additionally, we've estimated the daily frequency of code captures to be about 70 per day, allowing us to calculate how many days we'll

need the camera operational. Don't worry about removing it; I'll take care of that.

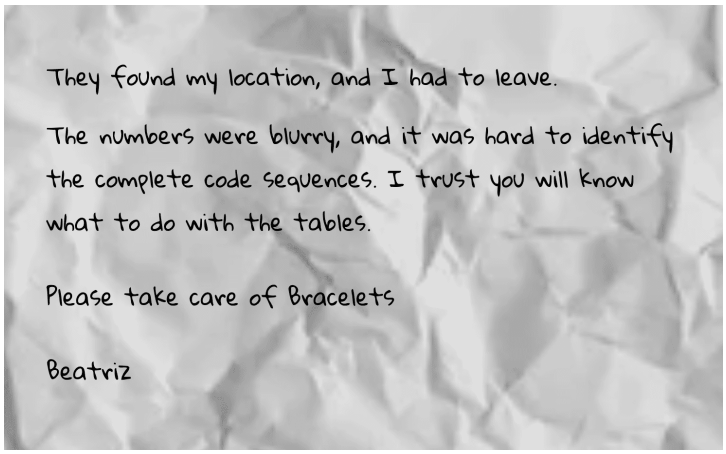
The time has come for us to move into the next stage of our plan. Yes, you will need to go down and install the camera. First, let's assess our current situation. Using binomial and Poisson distributions, we explored the feasibility of successfully installing a camera without encountering a guard. The good thing is that we have introduced more mathematics, which we always welcome.

However, these results were unsatisfactory, prompting me to seek further evidence. This search led us to our new framework, Bayes' Law, which opened up a new world of possibilities where our prior beliefs get updated with new evidence. With Bayes came a posterior, which also introduced us to the Gamma and Beta distributions, a consequence of the conjugate prior definition, and, more importantly, an alternative to hypothesis testing. Now I wish you the best of luck with the camera installation... speak soon.

... A few days later ...

I am glad to see you again, and I can confirm that the camera installation was a success-we've obtained the images from the videos-great job! However, we may be facing a minor setback. After retrieving the video feed from the camera you installed, I forwarded the recording to a colleague who specializes in video treatment and editing. She was supposed to give us one or more codes that could open the garage door with.

Despite multiple attempts to contact her over the past few days, I received no response, which led me to visit her apartment. When I arrived, I found the place eerily empty, with the front door unlocked, suggesting she had to leave in a hurry. Unfortunately, the original footage was nowhere to be found. The only item worth bringing back was this note:



Quick question: did you clean the camera lenses? No?!?!? It is in the past, so let's move on from it. Now, not only do we have some work to do on the recognition of those digits, but I have also adopted a cat. Bracelets is Beatriz's cat. She also left three tables for us, each of them with some notes:

Pixels (x_1)	V-Symmetry (x_2)	H-Symmetry (x_3)	Loops (x_4)	Class
210	8	9	1	0
185	7	7	2	0
200	6	8	0	0
195	7	9	1	0
205	9	6	1	0
225	4	6	1	9
230	5	7	0	9
215	3	5	2	9
220	4	8	1	9
240	5	5	0	9

Table 4.3: A collection of features and their corresponding numbers.

Table 4.3 contains numbers alongside some features that characterize them. The table is larger, but table 4.3 showcases a small portion of the original table that includes only two digits, 0 and 9. Here is a description of the features:

- **Pixels**: A numerical feature that represents the total sum of the grayscale values of the pixels in the image of the digit. Higher values generally indicate more filled or darker images, which can vary depending on the thickness and size of the digits in the images.

- **V-Symmetry (Vertical Symmetry)**: This numerical feature measures the symmetry of the digit image when split

vertically (down the middle). The scale from 0 to 10 quantifies how symmetrical the digit is along the vertical axis, with higher numbers indicating greater symmetry. For example, digits like '0' and '8' would score higher due to their rounded and symmetrical shapes.

- **H-Symmetry (Horizontal Symmetry)**: This is similar to vertical symmetry but measures the symmetry across the horizontal axis (from top to bottom). This is also on a scale from 0 to 10, where a higher score indicates a more symmetrical digit. This feature helps differentiate digits that might be symmetrical vertically but not horizontally, like '7'.

- **Loops**: A numerical count of the closed loops present within the digit. This is a straightforward count where digits like '0' have one loop, '8' has two, and others like '1' and '7' have none. This feature is particularly useful for distinguishing between numbers like '0' and '8'.

- **Class**: Represents the digit being classified. This field contains categorical data specifying the digit each row pertains to (0 or 9).

Table 4.3 is the first in our series; it presents a collection of features and classes that represent digits. In essence, it showcases a different way of characterizing numbers:

Entry Nº	Pixel (x_1)	V-Symmetry (x_2)	H-Symmetry (x_3)	Loops (x_4)
1	230	5	6	0
2	210	8	7	2

Table 4.4: Features from our numbers that were extracted from the video.

Table 4.4 is an example of the second table Beatriz left us. Hmmm, we have the same features as in table 4.3. However, now they seem to be the features representing our actual numbers from the video, as the title suggests. Beatriz is smart; there must be some relationship between 4.3 and 4.4, but before exploring this,

let's check the third table:

$$\begin{aligned} \text{Start} \to \text{Entry 1}, \quad & \text{Entry 2} \to \text{Entry 3} \\ \text{Entry 4} \to \text{Entry 5}, \quad & \text{Entry 6} \to \text{Entry 7} \\ \text{Entry 8} \to \text{Entry 9}, \quad & \text{Entry 10} \to \text{Entry 11} \\ \text{Entry 12} \to \text{Enter}, \quad & \text{Entry 13} \to \text{Entry 14} \\ \text{Start} \to \text{Entry 15}, \quad & \text{Entry 16} \to \text{Enter} \end{aligned} \tag{4.38}$$

This is the note she left as an explanation for 4.38:

Each code sequence I observed on the video initiates with a guard pressing the "Start" key and terminates with an "Enter." Sorry, Jorge, I couldn't finish my work, and I have to leave you with a pairwise representation of the input digits. The entries map to the features of table 4.4; for example, "Entry 1" in 4.38 corresponds to the row whose value of the column "Entry Nº" is 1.

It appears we don't have the individual digits, let alone the complete sequences of codes. Nevertheless, I have full confidence in Beatriz's capabilities, and I'm convinced that we can find the necessary codes using these tables. Our first step should be to address the issue of the missing digits. Once we have those, we can proceed to decipher the full sequences.

Both tables 4.3 and 4.4 share the same set of features, with the key difference being that 4.3 includes the corresponding class labels. Why not utilize the first table to create probability distributions of the classes based on these features? We can then apply these distributions to 4.4 to estimate the likelihood of each entry belonging to a particular class based on its features. Essentially, by establishing a posterior distribution for each class, we can input features from the unlabeled entries into these distributions. The class that yields the highest likelihood can then be attributed to each entry. Well, a posterior calls for Bayes law, so let's start with it:

$$P(\text{Class} \mid \text{Features}) = \frac{P(\text{Class}) \cdot P(\text{Features} \mid \text{Class})}{P(\text{Features})} \tag{4.39}$$

In equation 4.39, we have represented a probability of class given the features, which will define our classes' posteriors. For example,

we can calculate the probability of the number being a 9 given the features:

$$P(\text{Class} = 9 \mid (x_1, x_2, x_3, x_4) = \frac{P(\text{Class} = 9) \mid P((x_1, x_2, x_3, x_4)|\text{Class} = 9)}{P((x_1, x_2, x_3, x_4))} \tag{4.40}$$

To get to equation 4.40, we substituted the 'Features' with the corresponding columns from table 4.3, using the x notation to represent them. This equation uses the feature set $\{ x_1, x_2, x_3, x_4 \}$ to define a class or identify a number, mirroring our earlier approach, whereby we developed a posterior for the parameter λ. However, this scenario introduces a slight variation: we are now dealing with four variables, each exhibiting distinct behaviours.

Our first step is to assess whether these variables are independent. Confirming their independence is crucial, as it will affect how we handle the joint distribution. For instance, consider whether an increase in the number of loops (x_1) influences the vertical symmetry (x_2). We can reasonably assert their independence if we determine that such changes in one feature do not affect the others. Assuming independence holds after analyzing all features, we can simplify our model by writing the likelihood function that considers each feature's contribution separately:

$$\begin{aligned} P((x_1, x_2, x_3, x_4) \mid \text{Class} = 9) = P(x_1 \mid \text{Class} = 9) \cdot P(x_2 \mid \text{Class} = 9) \\ \cdot P(x_3 \mid \text{Class} = 9) \cdot P(x_4 \mid \text{Class} = 9) \end{aligned}$$

For example, $P(x_1 \mid \text{Class})$ represents the probability of observing x_1 given a specific class. Before carrying on with the deduction of equations, let's explore this likelihood by plugging in some numbers from table 4.3. As we have two different classes in our example table, if we consider the feature x_1:

$$P(x_1 \mid \text{Class} = 0)$$

$$P(x_1 \mid \text{Class} = 9)$$

Sound! So we have a probability distribution for x_1 by each of the present classes, meaning that we can go to the table 4.3 and see how these behave, for example if we wish to understand $P(x_1 \mid \text{Class} = 0)$

we need to collect the values of x_1 that correspond to a 'Class' equal to 0:

$$[210, 185, 200, 195, 205] \tag{4.41}$$

If we can transition from array 4.41 to formulated equations, we re-enter the familiar world of mathematics where we can employ all the techniques we've learned so far. Indeed, each class will have four distributions corresponding to the number of features. This implies that choosing different distributions for each feature will significantly complicate our calculations due to the need to multiply all those PDFs. So, before going all out and cramming every distribution we know into one big-ass equation, let's pause and consider our objectives.

We want to construct a posterior distribution for each class, in order to effectively distinguish the numbers. Ideally, we're looking to identify patterns that clarify the differences between classes. For this approach to be effective, the feature values across the dataset should exhibit variability, but within the confines of each class, they should ideally cluster around the mean. For instance, if we consider the array of samples for $P(x_1 \mid \text{Class} = 9)$:

$$[225, 230, 215, 220, 240] \tag{4.42}$$

If we compare 4.41 with 4.42, it is evident that their inputs differ; however, if we check their variance individually, we can see that the numbers don't change much when the array is considered in isolation, which means the values are kind of centered on the mean. So, we will assume that this behavior is present in each feature per class, and we have the perfect candidate for module observations centered on the mean; welcome back, Mister Gaussian distribution! So we can say that:

$$P(x_1 \mid \text{Class} = 0) \sim N(\mu_0, \sigma_0^2)$$

Where μ_0 and σ_0^2 represent the population mean and variance of the entries of feature x_1 that correspond to a class of 0. Now, if we wish to generalize our likelihood to all features and all classes:

$$P(x_i \mid \text{Class}) = \frac{1}{\sqrt{2\pi\sigma^2_{\text{class},i}}} \cdot e^{\left(-\frac{(x_i - \mu_{\text{Class},i})^2}{2\sigma^2_{\text{Class},i}}\right)} \tag{4.43}$$

With $i = 1, 2, 3, 4$. This equation barks, but it does not bite; looking at it, we have the density of the normal distribution where the x_i represents our features and the Class variable is our numbers. For example, if $i = 1$ and Class is equal to 0, we will have the PDF for $P(x_1 \mid \text{Class} = 0)$:

$$P(x_1 \mid \text{Class} = 0) = \frac{1}{\sqrt{2\pi\sigma^2_{\text{Class}=0,1}}} \cdot e^{\left(-\frac{(x_1 - \mu_{\text{Class}\ =0,1})^2}{2\sigma^2_{\text{Class}=0,1}}\right)}$$

With the likelihood established, our next logical step is to deduce the prior $P(\text{Class})$; for this particular one, we are going to take the route of simplicity, as we can estimate the initial probability of each class with our data:

$$P(\text{Class}) = \frac{\text{Number of instances of the class}}{\text{Total number of instances}} \tag{4.44}$$

Groovy! Now, if we multiply the prior 4.44 by the likelihood 4.43 we will end up with a formula for our posterior:

$$P(\text{Class} \mid x_1, x_2, \ldots, x_n) \propto P(\text{Class}) \cdot \prod_{i=1}^{n} P(x_i \mid \text{class}) \tag{4.45}$$

Where:

- $P(\text{Class})$ is the prior probability of the class.
- $P(x_i \mid \text{Class})$ is the likelihood of feature x_i under the Gaussian assumption:

$$P(x_i \mid \text{Class}) = \frac{1}{\sqrt{2\pi\sigma^2_{\text{Class},i}}} \cdot e^{\left(-\frac{(x_i - \mu_{\text{Class},i})^2}{2\sigma^2_{\text{Class},i}}\right)} \tag{4.46}$$

Because 4.45 is beautiful, it would be rude not to input some values from our dataset to see that thing working. For that we need the two usual suspects: the likelihood and the prior, and this time, we will start with the latter:

$$P(\text{Class} = 0) = \frac{5}{10} = 0.5, \quad P(\text{Class} = 9) = \frac{5}{10} = 0.5$$

We have ten observations in table 4.3 where five correspond to a class zero, and the other five correspond to a number nine; therefore, both priors are 0.5. Now it is time for the likelihoods, and because they are normal distributions, we need to compute their parameters μ and σ^2:

Feature	Class 0 (μ, σ^2)	Class 9 (μ, σ^2)
Pixels (x_1)	(199.0, 170.5)	(227.5, 116.9)
V-Symmetry (x_2)	(7.4, 0.8)	(4.0, 0.5)
H-Symmetry (x_3)	(7.8, 1.7)	(6.2, 1.7)
Loops (x_4)	(1.0, 0.8)	(0.8, 0.7)

Table 4.5: Parameters for Gaussian naive Bayes likelihoods per feature.

In table 4.5 we have represented the parameters μ and σ^2 per features per class. To compute them, we go to each feature per class and calculate the correspondent sample mean and variance. And all of a sudden we have everything we need to classify some entries!

$$s_1 = [230, 5, 6, 0]$$

The sample s_1 represents the first row of table 4.4, corresponding to the features of the first number extracted from our video feed. The goal is to identify if this entry is more likely to be a zero or a nine:

$$P(\text{Class} = 0 \mid (x_1 = 230, x_2 = 5, x_3 = 6, x_4 = 0) \tag{4.47}$$

$$P(\text{Class} = 9 \mid (x_1 = 230, x_2 = 5, x_3 = 6, x_4 = 0) \tag{4.48}$$

Let's first tackle the likelihood that s_1 is a 0, by expanding the equation 4.47 using Bayes' law:

$$P(\text{Class} = 0 \mid (x_1 = 230, x_2 = 5, x_3 = 6, x_4 = 0) = \\ = \frac{P(\text{Class} = 0) \cdot P((x_1 = 230, x_2 = 5, x_3 = 6, x_4 = 0) \mid \text{Class} = 0)}{P((x_1 = 230, x_2 = 5, x_3 = 6, x_4 = 0))}$$

We can now use the properties of independence:

$$P(\text{Class } 0 \mid x) \propto P(\text{Class } =0) \cdot P(x_1 = 230 \mid \text{Class } =0) \cdot P(x_2 = 5 \mid \text{Class } =0) \\ \cdot P(x_3 = 6 \mid \text{Class} = 0) \cdot P(x_4 = 0 \mid \text{Class} = 0)$$

Now we can use formula 4.46 to compute all of those individual conditional probabilities, for example:

$$P(x_1 = 230 \mid \text{Class } = 0) = \frac{1}{\sqrt{2\pi \cdot 170.5}} \cdot e^{\left(-\frac{(230-199)^2}{2 \cdot 170.5}\right)} \\ \approx 0.0018 \tag{4.49}$$

To get to equation 4.49, we use the parameters in table 4.5 and apply them to the general formula of the likelihood 4.5. To compute $P(x_1 = 230 \mid \text{Class} = 0)$, we replace x_1 with 230 and use the corresponding mean ($\mu = 199.0$) and standard deviation ($\sigma = 6.52$) for x_1 in Class 0. We then repeat this process for x_2, x_3, and x_4, using their specific values and the respective mean and standard deviation for each variable in Class 0:

$$P(x_2 = 5 \mid \text{Class } = 0) = \frac{1}{\sqrt{2\pi \cdot 0.8}} \cdot e^{\left(-\frac{(5-7.4)^2}{2 \cdot 0.8}\right)} \approx 0.0122$$

$$P(x_3 = 6 \mid \text{Class } = 0) = \frac{1}{\sqrt{2\pi \cdot 1.7}} \cdot e^{\left(-\frac{(6-7.8)^2}{2 \cdot 1.7}\right)} \approx 0.1180$$

$$P(x_4 = 0 \mid \text{Class } = 0) = \frac{1}{\sqrt{2\pi \cdot 0.8}} \cdot e^{\left(-\frac{(0-1)^2}{2 \cdot 0.8}\right)} \approx 0.2387$$

Cool! So for the probability that the observation is a zero:

$$\begin{aligned} P(\text{Class} = 0 \mid (x_1 = 230, x_2 = 5, x_3 = 6, x_4 = 0)) &= 0.0018 \cdot 0.0122 \cdot 0.1180 \cdot 0.2387 \cdot 0.5 \\ &= 3.13 \times 10^{-7} \end{aligned}$$

It does not seem like these features correspond to a zero as the probability we calculated is very low, but let's verify the likelihood of it being a class nine. The process is the same:

$$\begin{aligned} P(\text{Class } = 9 \mid x) \propto P(\text{Class } = 9) \cdot & P(x_1 = 230 \mid \text{Class } = 9) \cdot P(x_2 = 5 \mid \text{Class } = 9) \\ & \cdot P(x_3 = 6 \mid \text{Class } = 9) \cdot P(x_4 = 0 \mid \text{Class } = 9) \end{aligned}$$

Now for the likelihoods:

$$P(x_1 = 230 \mid \text{Class} = 9) = \frac{1}{\sqrt{2\pi \cdot 116.9}} \cdot e^{\left(-\frac{(230-227.5)^2}{2 \cdot 116.9}\right)} \approx 0.0359$$

$$P(x_2 = 5 \mid \text{Class} = 9) = \frac{1}{\sqrt{2\pi \cdot 0.5}} \cdot e^{\left(-\frac{(5-4.0)^2}{2 \cdot 0.5}\right)} \approx 0.2076$$

$$P(x_3 = 6 \mid \text{Class} = 9) = \frac{1}{\sqrt{2\pi \cdot 1.7}} \cdot e^{\left(-\frac{(6-6.2)^2}{2 \cdot 1.7}\right)} \approx 0.3024$$

$$P(x_4 = 0 \mid \text{Class} = 9) = \frac{1}{\sqrt{2\pi \cdot 0.7}} \cdot e^{\left(-\frac{(0-0.8)^2}{2 \cdot 0.7}\right)} \approx 0.3019$$

If we multiply all those likelihoods and the prior we will have the probability of the sample representing a number nine:

$$\begin{aligned} P(\text{Class} = 9 \,|(x_1 = 230, x_2 = 5, x_3 = 6, x_4 = 0)) \\ &= 0.5 \cdot 0.0359 \cdot 0.2076 \cdot 0.3024 \cdot 0.3019 \\ &= 0.00034 \end{aligned}$$

It is decision time! Now, we must pick the highest probability between the two we have just computed:

$$P(\text{Class} = 0 \mid (x_1 = 230, x_2 = 5, x_3 = 6, x_4 = 0)) \approx 3.13 \times 10^{-7}$$

$$P(\text{Class} = 9 \mid (x_1 = 230, x_2 = 5, x_3 = 6, x_4 = 0)) \approx 0.00034$$

Before deciding between classes 0 and 9, we must address something. We used Bayes' theorem to compute the probabilities that the features belong to a given class, and, in our small dataset example, we just have two of them: 0 and 9. However, remember that the complete dataset contains ten classes corresponding to the 10 digits. We will be inclined to think that, if we compute all the individual probabilities of a given set of features, such as s_1, to be to each of the digits, they will add up to 1. But, these probabilities are different from the final probabilities you might expect. If you do the exercises of calculation the likelihoods for all ten corresponding classes they won't add up to 1. Instead, they are more like raw scores that compare the probability of each class given the data. Each score is computed by multiplying the likelihood of the observed features, assuming they belong to a specific class, by their prior probability.

These calculated scores don't add up to one because the formula needs to include a step to normalize these values across all classes. Normalization would adjust the scores so their total would equal one, but it's skipped in Naive Bayes because it doesn't change which class has the highest score. Ups, I unintentionally dropped the name of the technique we just deduced! Nevertheless, before formally introducing this guy, let's officially decide if we are in the presence of a 0 or a 9. For that, we will use the calculated scores directly to pick the most likely class, skipping unnecessary calculations that don't impact the final decision:

$$P(\text{Class} = 9 > P(\text{Class} = 0)$$

$$s_1 \to 9$$

Therefore, s_1 is a nine! Now the formal intro.

4.4 Little Naive Bayes Thought He Could Be a Classifier.

The Naive Bayes classifier is a probabilistic machine learning model that's used for classification tasks. It's based on Bayes' theorem and makes an assumption of conditional independence among predictors. Despite its simplicity, Naive Bayes can yield very good results and is surprisingly effective. The probability of a given class is described by:

$$P(\text{Class} \mid x_1, x_2, \ldots, x_n) \propto P(\text{Class}) \cdot \prod_{i=1}^{n} P(x_i \mid \text{class})$$

Where:

- $P(\text{Class})$ is the prior probability of the class.
- $P(x_i \mid \text{Class})$ is the likelihood of feature x_i.

And the general formula for the class selections is:

$$\text{Predicted Class} = \arg\max_{\text{class}} P(\text{class}) \prod_{i} P(\text{feature}_i \mid \text{Class}) \quad (4.50)$$

Equation 4.50 is a general equation that allows us to select the class with the highest probability value, just as what we did to end up with 9.

In exploring the Naive Bayes classifier, we started with a simplified scenario, focusing on recognizing two digits: zero and nine. The reason for just having two classes was to ease the learning, allowing us to thoroughly understand the underlying mechanics of the model without the added complexity of multiple classes. Please note that we did not cover the subject of module evaluation; our goal was to create a methodology leveraging the Bayesian framework to identify

a class that gives a feature. Performance metrics of machine learning models will be the topic of a new book that has yet to be written.

By concentrating on two digits, we could perform manual calculations to understand how Naive Bayes computes probabilities and makes predictions based on data. In our scenario, where we are aiming to recognize numbers, we need 10 classes, but if we know how to apply this model for two, we can use it for as many labels as we want. However, with more classes, the volume of data needed grows significantly, as the model must learn to distinguish between a greater number of possible outcomes based on the same features. In real-world scenarios, you often don't have a chance to get more data, so you must work with what you have. But, in case we need to estimate a sample size, just as we had to do to collect these images, we couldn't just leave a secret camera installed by the door forever. So, one way to do so is with the following formula:

$$n = \left(\frac{Z^2 \cdot p \cdot (1 - p)}{E^2} \right)$$

Where:

- n is the sample size.
- Z is the Z-score corresponding to your desired confidence level (e.g., 1.96 for 95% confidence).
- p is the estimated proportion of success (e.g., the expected proportion of one class over another, which can initially be assumed to be 0.5 if no prior data is available).
- E is the margin of error you are willing to accept (e.g., 0.05 for 5%).

For example say we are OK with a Naive Bayes classifier to distinguish between classes with 95% confidence and a margin of error of 10% (0.1):

- $Z = 1.96$ (Z-score for 95% confidence)
- $p = 0.1$ (Assuming uniform likelihood across 10 classes)

- $E = 0.1$ (10% margin of error)

$$n = \left(\frac{(1.96)^2 \cdot 0.1 \cdot (1 - 0.1)}{(0.1)^2}\right) = \left(\frac{1.96^2 \cdot 0.1 \cdot 0.9}{0.01}\right)$$
$$= \left(\frac{0.3528}{0.01}\right) = 35.28$$

Rounding up, you would need at least 36 observations per class.

Naive Bayes is a robust, probabilistic classifier renowned for its simplicity and effectiveness, especially in classification scenarios. We'll use this model to recognize all the digits from the table that Beatriz left us. However, recognizing the digits is just part of the task. Remember, Beatriz didn't leave us complete sequences of codes; she provided a pairwise structure that we must map onto table 4.4 with our numbers features, the one without the classes. Here's our approach: We will utilize the Naive Bayes classifier to analyze all features of our numbers and then predict their classes. Once we have identified the classes for all entries, we will use the 'Nº' column to map these features to their corresponding pairwise representations. This will enable us to transform table 3, replacing 'Entry X', where X represents a specific number. Rest assured, I will handle this process for us:

$$\text{Start} \to 1 \quad 4 \to 1$$

$$2 \to 1 \quad 1 \to 3$$

$$3 \to 2 \quad 0 \to 4$$

$$4 \to \text{Enter} \quad 0 \to 8$$

$$\text{Start} \to 3 \quad 2 \to \text{Enter}$$

The arrays above represent the small sample of table 4.38. It is structured in pairs and provides us with the following digit given that one was observed, so Start $\to 1$ reads that on a given code sequence, we see a 1 after a Start. Well, just when we were thinking that we just got screwed with this incident, luck has come along to present us with some valuable information. I am talking about the Start

and Enter instances we have; those indicate the start and end of the code sequence. Now, how can we get from a pairwise representation to a code? We don't know if we have different codes or even if they have different lengths. But if we have the following number, given the previous one. This vernacular should sound familiar; it is similar to the conditional probability. Let's explore this to see if we can get at least one code:

$$P(X_2 = 1 \mid X_1 = \text{Start}) \tag{4.51}$$

With the context provided in equation 4.51, this involves calculating the probability of observing the digit 1 as the second entry given that the sequence starts with the term Start. Mathematically, this is calculated as the ratio of the number of sequences where 1 immediately follows Start to the total number of sequences that begin with Start. This can be represented as:

$$P(X_2 = 1 \mid X_1 = \text{Start}) = \frac{\text{Number of sequences with Start followed by 1}}{\text{Total number of sequences starting with Start}}$$

We only have one instance of a Start followed by One in our sample of data:

$$\text{Start} \rightarrow 1$$

And there are only two instances that initiate with Start:

$$\text{Start} \rightarrow 1, \quad \text{Start} \rightarrow 3$$

So:

$$P(X_2 = 1 \mid X_1 = \text{Start}) = \frac{1}{2}$$

Following this logic, if we want to understand the likelihood of having a third digit, for example, equal to 3, we will have the following:

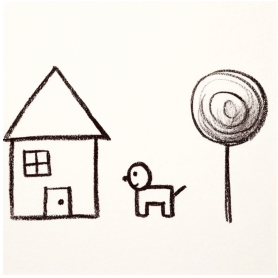
$$P(X_3 = 3 \mid (X_1 = \text{Start}, X_2 = 1)) \tag{4.52}$$

Generically we can define 4.52 as:

$$P(X_3 = i \mid (X_1 = \text{Start}, X_2 = j)) \tag{4.53}$$

In this setup, i and j can each be any digit from 0 to 9 or Enter, totalling 11 possible options each. Initially, we might assume the

code is just three digits long. But if the code turns out to be four digits, we have to change 4.53 by adding a X_4. Thisd significantly increases complexity. Furthermore, the actual length of the code remains unknown, which could lead to further complications if we explore even longer sequences. Each additional digit exponentially increases the number of possible combinations, quickly making our model increasingly complex. Considering the possibility of a five, or six-digit code, we soon realize this approach may be impractical. Yeah ... this is the wrong path to follow as, very soon we will be blowing up computers with all this complexity; right, back to the drawing board:



I might have taken that sentence too literally, but now that I have shown you my art, what do you think? Yes, that is a dog! Anyway, back to our method of identifying sequences. The goal of equation 4.52 is to predict the probability of 3 after Start $\rightarrow 1$:

$$\text{Start} \rightarrow 1 \quad 1 \rightarrow 3$$

From this sequence, we might conclude that a 1, following a Start, could influence the appearance of a 3. To further explore this, we should examine occurrences where a 1 leads to different digits and whether previous inputs, like Start, influence outcomes after 1. For example:

$$2 \rightarrow 1 \quad 1 \rightarrow 3$$

$$\text{Start} \rightarrow 1 \quad 1 \rightarrow 3$$

Regardless of whether 1 was preceded by 2 or Start, the outcome 3 follows 1. This repetition suggests that the next digit, 3, depends only on the immediate predecessor, 1, rather than the entire sequence leading to 1. Well, this is no mere suggestion, but something more powerful, that we call the Markov property.

Chapter 5

Just Tell Me about Yesterday: Insights from Markov Models.

The Markov property is a concept in probability theory and statistics. It refers to the memoryless nature of specific stochastic processes, where the future state of a system depends only on its present state, not on the sequence of events that preceded it. Oh, baby, there is a lot to unpack here, as we have new terminology to meet. Firstly, it would be rude to carry on without some formal introduction to the "stochastic process". It is a simple concept with a fancy name, just like this item I recently saw for sale, "Ambulatory Support Mechanism." That is a walking stick.

5.1 Chance in Motion: Navigating Stochastic Processes.

A stochastic process is a mathematical framework for modeling systems or phenomena that evolve randomly over time. In this context, a "system" refers to a set of elements or variables that interact or behave in a non-deterministic manner. Here, our system is a security keypad where the state changes as different keys (digits or commands) are entered. The process consists of a collection of random variables indexed by time or space, each representing a possible outcome at a given time point or position. In our model, each random variable X can take values from any digit between 0 to 9, or Start and Enter. The time or space component relates to

the subscript we introduce to represent the digit's position in our code sequence; for example, we know that our X_1 must be Start. However, X_2 can be any digit, and X_3 can be any digit plus the key Enter. With this process comes the definition of state, which refers to a particular configuration or condition of our system at a given time. For example, if $X_2 = 1$, the state of our system at time $t = 2$ is equal to 1. Accompanying this notion of state is a state space, representing all the possible outcomes that our system can take at a given time t, in our case:

$$0, 1, 2, 3, 4, 9, \text{Start}, \text{Enter} \tag{5.1}$$

With the caveat that $t = 0$ must be Start and Enter must come after $t > 1$. After carefully analyzing all the pairs in the whole dataset, I noticed that the only digits we find are 0,1,2,3,4,9. The numbers 5,6,7, and 8 are absent from our sample, so we will assume no codes contain these last four chaps. That was a big mistake by the guards when generating these codes, as this will reduce the number of possibilities and, therefore, the complexity of our system.

In terms of a stochastic process, it extends our knowledge of ways to model random variables. Before, we would define a random variable X with a probability distribution. Now, we have a framework called a stochastic process that allows us to create a dynamics where we can understand systems where there is a need for a sequence of random variables. In our case, our system is the code for the garage door, and our random variable X follows a distribution that we assume to be uniform, meaning that every digit and Enter have the same likelihood. So, the key characteristics of a stochastic process include:

- **Random Variables $\boldsymbol{X}$**: Each element of the process represents a key entered at a specific point in the sequence. For example, X_1 might be Start, X_2 could be any digit from 0 to 9, and X_3 might be Enter or another digit.
- **Index Set**: This set defines the positions in the sequence of key entries on the keypad. In our case, it is discrete, consisting of integers where $t = 1, 2, 3, \ldots, n$, with n depending on the length of the code entered.

- **State Space**: This is the set of all possible values the random variables can take. For our keypad system, the state space includes Start, all digits from 0 to 9, and Enter. This reflects a finite state space where each state corresponds to a potential key entry at any given time t.

If you are wondering how the Markov property and stochastic processes are linked, let's clarify that before moving on. Stochastic processes are generic; they can assume different characteristics, the Markov property being one of them. If we define such a process to be governed by the Markov property, we can view the progression of states as a sequence of probability distributions. What characterizes each state X_t in the process is that they are random variables with their own probability distribution. The essence of the Markov property is that the probability distribution for the next state X_{t+1} is conditioned solely on the current state X_t. Thus, each step in a Markov process represents a new state, and the process unfolds as a new distribution is explicitly influenced by the state immediately prior:

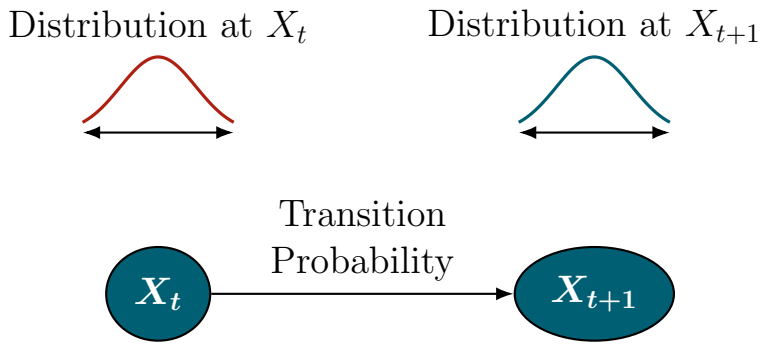


Figure 5.1: Randomly moving from state to state.

Figure 5.1 illustrates the transitions between two arbitrary states in a Markov chain, a type of stochastic process that adheres to the Markov property. This property ensures that the probability of transitioning to any future state depends only on the current state and not on the path taken to reach it. Each state in the diagram is associated with a unique probability distribution that dictates the likelihood of moving to subsequent states.

In figure 5.1, we see a crucial concept: "transition probability". Though it might seem like just another set of probabilities, it's essential for understanding how the process evolves from one state to another. For instance, if the current state is the digit '1', the transition probabilities define the likelihood of moving to any other digit in the state space. These probabilities help us map out the dynamics of state changes in the chain. So, let's define our process and the Markov property so than we can start cracking codes!

5.1.1 Have You Heard the Latest? Markov Broke the Chains and Escaped.

Markov chains are a type of stochastic process where the probability of transitioning to the next state depends only on the current state and not on the sequence of events that preceded it. This property is known as the Markov property or memorylessness. Markov chains are characterized by:

- **States**: A finite or infinite set of states that the system can be in.
- **Transition probabilities**: Each pair of states is associated with a probability that defines the likelihood of moving from one state to another.

Now that the introductions are out of the way, let's mathematically define the Markov property. If we define X_t to be a stochastic process satisfying the Markov property, it means that the probability of the distribution of the future state X_{t+1} depends only on the current state X_t and is conditionally independent of the past states $X_0, X_1, ..., X_{t-1}$. If we wish to translate this to an equation, we get:

$$P(X_{t+1} = x \mid X_t = x_t, X_{t-1} = x_{t-1}, \ldots, X_0 = x_0) = P(X_{t+1} = x \mid X_t = x_t) \tag{5.2}$$

Equation 5.2 represents our transition probabilities; if we were to apply it to our example:

$$P(X_3 = 3 \mid (X_1 = \text{Start}, X_2 = 1)) = P(X_3 = 3 \mid X_2 = 1)) \tag{5.3}$$

So, to have all our transition probabilities, we need calculations like 5.3, but for all the pairs we can create with our state space. Well, we can construct quite a few pairs, which could get messy, but since we have a beautiful way to store pairs, I foresee we can avoid getting dirty; yes, I am talking about a matrix! But how will we build such a matrix? I mean, it has to have probabilities somehow. We always used frequencies to compute probabilities, so what about if we start there? As we explore new territory, let's decrease the complexity by creating a new state space with just the digits; once we are comfortable with it, we will scale for all the states and the whole dataset. So our new space is:

$$0, 1, 2, 3, 4, 9 \tag{5.4}$$

With this new state space 5.4 defined, we can create a 6×6 matrix for which the indices are the elements of this exact state space:

	0	1	2	3	4	9
0	0	0	0	0	1	1
1	0	0	0	1	0	0
2	0	1	0	0	0	0
3	0	0	1	0	0	0
4	0	1	0	0	0	0
9	0	0	0	0	0	0

(5.5)

In our analysis, matrix 5.5 represents the frequencies of observed transitions between states. Each entry, such as the entry for $1 \to 3$, which is populated with a 1 because we observed this transition once, indicates the number of times we've observed transitions from one state to another. However, for our purposes in studying a Markov chain, we need to convert these frequencies into transition probabilities. Well, we can go on about this by dividing frequencies by totals as per the probabilities formula, but we have two options for how we do this: row-wise or column-wise.

The correct interpretation of this matrix should be row-wise because each row represents a unique starting state, and the entries in that row show all possible transitions from that state to others. This setup directly supports the calculation of the conditional prob-

abilities essential for Markov chains, specifically, the likelihood of transitioning to any other state from a given current state.

To convert the frequencies in 5.5 to probabilities, we must normalize each row so that the sum of all entries in a row equals 1. This normalization is achieved by dividing each frequency in a row by the total number of transitions observed from that row's state. For example, if state 0, which corresponds to the first row of the matrix, appears twice in our dataset, the entries for $P(4|0)$ and $P(9|0)$ are calculated by dividing the frequencies at these positions by 2, thereby converting the observed frequencies into probabilities:

$$P(4 \mid 0) = \frac{1}{2}, \quad P(9 \mid 0) = \frac{1}{2}$$

If we apply this same logic to all entries of the matrix 5.5:

$$P = \begin{array}{|c|c|c|c|c|c|c|} \hline & 0 & 1 & 2 & 3 & 4 & 9 \\ \hline 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ \hline 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ \hline 2 & 0 & 1 & 0 & 0 & 0 & 0 \\ \hline 3 & 0 & 0 & 1 & 0 & 0 & 0 \\ \hline 4 & 0 & 1 & 0 & 0 & 0 & 0 \\ \hline 9 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline \end{array} \tag{5.6}$$

The matrix 5.6 is often called a transition matrix for a Markov process. The matrix P, is a square matrix that describes the probabilities of transitioning from one state to another in a discrete-time stochastic process. The matrix has the following properties:

- **Size**: The transition matrix P has a size $n \times n$, where n is the number of possible states in the Markov process.
- **Entries**: Each entry p_{ij} represents the probability of transitioning from state i to state j in one time step.
- **Row sums**: The sum of the entries in each row equals 1. This is because the probabilities of moving from a given state i to all possible states j must add up to 1, ensuring that the system accounts for all potential transitions.

- **Non-negative entries**: Each entry p_{ij} is non-negative, as probabilities cannot be negative. Thus, $0 \leq p_{ij} \leq 1$.

Formally, the transition matrix can be represented as:

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{bmatrix}$$

Each row i in the transition matrix P represents the probabilities of transitioning from state i to other states j. Having formally defined the transition matrix P, now is an ideal time to create one for the entire dataset, focusing specifically on the state space as previously defined, which includes only the digits. While the full dataset is extensive and cannot be completely detailed here, I have utilized computational tools to generate the transition matrix based on the sample dataset we examined earlier. This approach allows us to effectively set up and analyze the transitions within our defined state space:

$P =$

	0	1	2	3	4	9
0	0	0.2	0	0	0.4	0.4
1	0.2	0	0.3	0.5	0	0
2	0.1	0.5	0	0.4	0	0
3	0.1	0.3	0.3	0	0.3	0
4	0.4	0	0	0.2	0	0.4
9	0.1	0	0	0	0	0.9

(5.7)

We build the transaction matrix 5.7 by counting the frequencies of transitions between states, similar to the methodology we used to create 5.6, then normalize them by the total transitions observed from the corresponding initial state, ensuring that each row equals 1. It is true that, in this specific case, our data were already in pairs, but usually, the data we receive is a sequence, for example:

$$\text{B}, \text{C}, \text{B}, \text{C}, \text{A}, \text{B}, \text{F}$$

This is an arbitrary example; those letters don't have any special meaning. In this situation, we need an extra step to calculate our

transition matrix, that is, first, to transform that sequence into pairs:

$$\mathrm{B} \to \mathrm{C}, \quad \mathrm{C} \to \mathrm{B}$$

$$\mathrm{B} \to \mathrm{C}, \quad \mathrm{C} \to \mathrm{A}$$

$$\mathrm{A} \to \mathrm{B}, \quad \mathrm{B} \to \mathrm{F}$$

Having a matrix allows us to explore different avenues as we now have access to the whole suite of tools that linear algebra offers. For example, we know that a matrix can be the vehicle for linear transformations, so when we multiply a vector by a matrix, we transform that vector from one space into another. It will be interesting to understand what happens if we multiply a vector by matrix P. The question is, what will the entries of this vector represent? If the matrix entries are transitions probabilities, what if we define a vector whose entries contain the state of the system at time t? We will be understanding what happens to a state of the system as we transform it by the transition matrix P. Let's explore this by defining a vector v_0 representing the state of the system at $t = 0$:

$$\vec{v_0} = [0.16, 0.16, 0.16, 0.16, 0.16, 0.16] \tag{5.8}$$

I feel it is time to clarify the lingo a bit, so we don't get lost. We decided to represent our transition probabilities with a matrix, bringing us the possibility of linear transformations. For that, we define $\vec{v_0}$ as a state probability vector. Think of it this way; our t is the index of the digit of our code, so $t = 0$ represents the first digit. So because we did not consider the Start and End states, $\vec{v_0}$ says that for the first digit, we have an equal probability of it being any number in your state space, meaning a 16.66% chance of it being a 0, 16.66% of it being a 1 and so on; because it is our first vector we call it the initial state vector. There is no direct meaning for the 16.66%'s; they were selected randomly for the purpose of the experiment.

Having a vector like 5.8 provides us with the mathematical framework necessary to describe the state of our system at any given time t. The indices of our state vector correspond to the columns of the matrix 5.7; for example, position 0 aligns with state 0 (representing the digit 0), position 1 with state 1 (representing the digit 1), and so

forth, up to position five which aligns with state 9. By modeling our sequence of numbers that form a code as a Markov stochastic process, we effectively commit to representing our entire system through the probabilities of transitioning between states. Each entry in the state vector then quantifies the likelihood of the system being in each respective state at the subsequent time step. Thus, each vector encapsulates a state distribution, and specifically for $t = 0$, it defines our initial state distribution as follows:

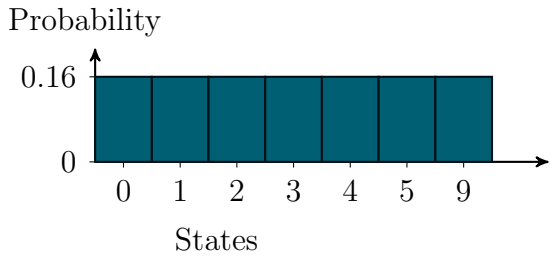


Figure 5.2: A Kit Kat bar? Ah no, just the state distribution of $\vec{v_0}$.

Figure 5.2 showcases the probability distribution for the initial state of our system at time $t = 0$, representing the potential first digits of our codes. The histogram uniformly allocates a probability of 0.16 to each state, from 0 through 9, indicating an equal likelihood of the system beginning with any of these digits at state $t = 0$. Next, we will investigate the dynamics of this system when the initial state vector, denoted as $\vec{v_0}$, is multiplied by the transition matrix P:

$$\vec{v_1} = \vec{v_0} \cdot P \tag{5.9}$$

Let's expand 5.9 using the dot product formula:

$$\vec{v_1}[j] = \sum_{i=0}^{n} \vec{v_0}[i] \cdot P_{ij}$$

Where j represents the index of the entries of our new vector $\vec{v_1}$. Time to input some numbers? Let's go! We will start with $\vec{v_1}[0]$, the first position of our vector representing the index of the state 0 in our current example:

$$\vec{v_1}[0] = \sum_{i=0}^{n} \vec{v_0}[i] \cdot P_{i0}$$

That is a weighted sum of $\vec{v_0}$ with the first column of P:

$$\vec{v}_1[0] = \begin{bmatrix} 0.16 & 0.16 & 0.16 & 0.16 & 0.16 & 0.16 \end{bmatrix} \cdot \begin{bmatrix} 0 \\ 0.2 \\ 0 \\ 0 \\ 0.4 \\ 0.4 \end{bmatrix}$$

That results in:

$$\vec{v}_1[0] = 0.16 \cdot 0 + 0.16 \cdot 0.02 + 0.16 \cdot 0 + 0.16 \cdot 0.4 + 0.16 \cdot 0.4$$

$$\vec{v}_1[0] = 0.144$$

If we do the same calculations for the remaining positions i of $\vec{v}_1$ we end up with:

$$\vec{v}_1 = [0.144, 0.16, 0.096, 0.176, 0.112, 0.272]$$

We first notice that the vector entries add up to 1, suggesting that this is another state probability, but now, for $t = 1$. Can this be where we land from $t = 0$? Remember that each vector represents a state, and each state corresponds to a random variable; this is why we have several entries with likelihoods. Also, the design of P ensures that the sum of probabilities of transitioning from any state to all possible next states equals 1. Thus, when you sum all entries of $\vec{v}_1$, you get 1, validating that $\vec{v}_1$ maintains the definition of a probability distribution:

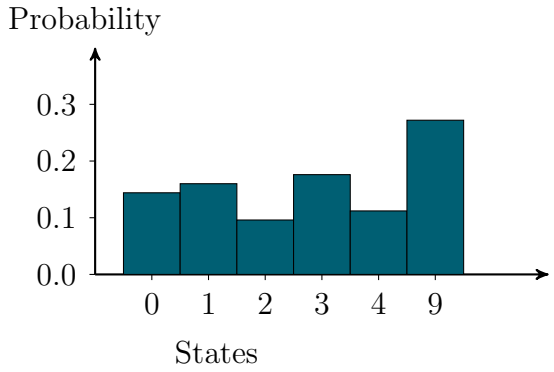


Figure 5.3: Yeah I know, just another state distribution and still no chocolate.

The histogram in figure 5.3 represents our state probabilities at $t = 1$, indicating where the system is likely to be after starting with the initial state vector $\vec{v}_0$. Specifically, we initiate our process with the vector $\vec{v}_0$ and multiplied it by our transition matrix P to arrive at $\vec{v}_1$. This operation provides an insight into the probable state distribution at $t = 1$. It is important to note that the transition matrix models the likelihoods of transitioning between pairs of states, which we have derived from our video feed analysis. Therefore, after starting from $\vec{v}_0$, there is a 14% probability that the next state will be 0, another 16% that it will be 1, 10% that it will be 2, 18% that it will be 3, 11% that it will be 4 and finally 27% that the digit is an 9. It makes sense for us to carry on computing state distributions:

$$\vec{v}_2 = \vec{v}_1 \cdot P \tag{5.10}$$

Continuing with this methodology would mean that for every next state, we must have the previous one, and that seems like a lot of notation and calculations; come on, mathematicians do not like messy equations. So instead of calculating $\vec{v}_3$ by multiplying $\vec{v}_2$ by P, let's try to work that formula 5.10 a bit. An obvious first step will be to replace $\vec{v}_1$ with $\vec{v}_0 \cdot P$:

$$\vec{v}_1 = \vec{v}_0 \cdot P$$

$$\vec{v}_2 = \vec{v}_1 \cdot P = \vec{v}_0 \cdot P \cdot P = \vec{v}_0 \cdot P^2$$

I like that a lot! The index of the vector and the exponent of P are equal. We can't just ignore this; maybe there is a relationship. Let's explore a $t = 3$:

$$\vec{v}_3 = \vec{v}_2 \cdot P = \vec{v}_0 \cdot P^2 \cdot P = \vec{v}_0 \cdot P^3$$

What about an index n ?

$$\vec{v}_n = \vec{v}_{n-1} \cdot P = \vec{v}_0 \cdot P^{n-1} \cdot P = \vec{v}_0 \cdot P^n \tag{5.11}$$

Wait; so with just $\vec{v}_0$ and P, we can get to any state?!?! Now we are talking! Well, what about, for example, if we wish to know what are the likelihoods of a given digit in position 7 of the code are:

$$\vec{v}_7 = \vec{v}_0 \cdot P^7, \quad \vec{v}_7 = [0.116, 0.079, 0.052, 0.082, 0.074, 0.554]$$

Probably the 7^{th} digit will be a 9; we do have a 55% chance of that. Although we haven't finalized the code yet, we can already understand where this system is likely to be at any given state n by having v_0 and the transition matrix P. This n is crucial for us, as it will indicate the length of our code. This piques my interest in further exploring what's happening at state n, and perhaps from there, we can estimate the size of the sequence we need. For instance, if we anticipate the code will be composed of 8 digits, then $n = 8$. Clearly, what influences the state distribution at state n is the matrix P, so an interesting aspect to analyze is how P evolves as we increase the exponent n. Here's what we can do: we will compute numerous iterations of P^i where $i = 1, 2, 4, 5, \ldots, n$, with n being a large number. Then we will calculate the Frobenius norm between consecutive pairs of P:

$$\|A - B\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |(a_{ij} - b_{ij})|^2} \tag{5.12}$$

If you read my first book on linear algebra, this formula 5.12 for the Frobenius norm will be no stranger to you. Suppose you haven't. Well, bad luck ... Ah! Just joking! Equation 5.12 represents the Frobenius norm, $\|A - B\|_F$ that quantifies the difference between two matrices A and B by taking the square root of the sum of the squares of the differences between their corresponding elements. This norm effectively measures the "distance" between A and B, where a higher Frobenius norm indicates a greater difference between the two matrices. So I will use a computer to create several powers of the transition matrix P, and will then compute the Frobenius norm between consecutive pairs; let's go over the example between P and P^2:

$$P = \begin{bmatrix} 0 & 0.2 & 0 & 0 & 0.4 & 0.4 \\ 0.2 & 0 & 0.3 & 0.5 & 0 & 0 \\ 0.1 & 0.5 & 0 & 0.4 & 0 & 0 \\ 0.1 & 0.3 & 0.3 & 0 & 0.3 & 0 \\ 0.4 & 0 & 0 & 0.2 & 0 & 0.4 \\ 0.1 & 0 & 0 & 0 & 0 & 0.9 \end{bmatrix}, P^2 = \begin{bmatrix} 0.24 & 0 & 0.06 & 0.18 & 0 & 0.52 \\ 0.08 & 0.34 & 0.15 & 0.12 & 0.23 & 0.08 \\ 0.14 & 0.14 & 0.27 & 0.25 & 0.16 & 0.04 \\ 0.21 & 0.17 & 0.09 & 0.33 & 0.04 & 0.16 \\ 0.06 & 0.14 & 0.06 & 0 & 0.22 & 0.52 \\ 0.09 & 0.02 & 0 & 0 & 0.04 & 0.85 \end{bmatrix}$$

Now if we apply 5.12 to P and P^2:

$$\|P - P^2\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} |(p_{ij} - p_{ij}^2)|^2} = 1.439$$

Where p_{ij} are the elements of matrix P and p_{ij}^2 are their counterparts but for matrix P^2. Now we will do the same thing but for a large n and compute the Frobenius norm between consecutive pairs. For example, the next one will be P^2 and P^3 and then P^3 and P^4 and so on:

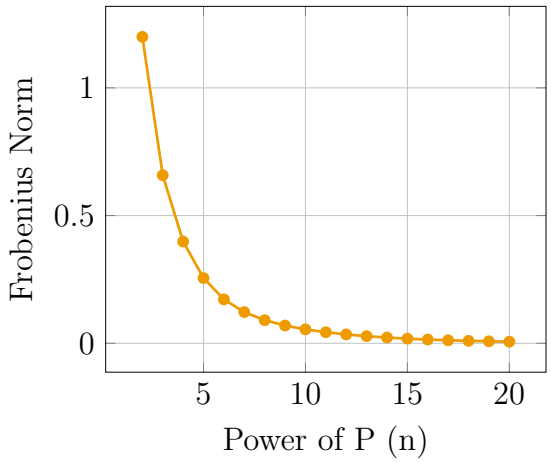


Figure 5.4: Nothing changes after a certain n - steady!

Figure 5.4 highlights the Frobenius norms between consecutive powers of P, showcasing a noteworthy pattern. As the matrix is raised to the 15^{th} power ($n = 15$), these norms approach zero, signalling that the transition probabilities between digits begin to stabilize and cease to change significantly. This stabilization suggests that our system of digits reaches a state of equilibrium at this exponent, meaning the behavior of transitioning between digits becomes predictable and constant.

At equilibrium, the probabilities that define transitions between digits (the state probabilities) stabilize to a fixed distribution. This fixed distribution, often called the "steady state" in our digit system, is vital because it indicates a long-term behavior where the likelihood of transitioning from one digit to another remains constant, regardless of the initial conditions of the sequence. This steady state

is crucial for predicting future states of our digit sequence system, assuming the system's dynamics, the ways in which digits transition, remain consistent. Thus, understanding this steady state not only provides an insight into the sequence's inherent patterns, but it also ensures that our predictions about future digit transitions are reliable and independent of how the sequence started.

What about if we calculate ours to take more tangible conclusions? We just saw that we get to equilibrium when the system's behavior becomes constant over time. Mathematically, we represent this by the condition where repeatedly multiplying the transition matrix P by itself (raising it to the power of n) in a matrix that, when multiplied by the state vector, yields the same vector, signifying no change in state probabilities.

Doesn't this remind you something? No, it is not that people should speak with each other rather than look at their phones. It is a linear algebra concept representing a linear transformation in which the original vector gets transformed on itself or a vector colinear with the original vector, the eigenvectors. Let's bring back this definition:

$$\vec{v}P = \lambda\vec{v}$$

Well, for our particular case, we want the vector to be exactly the same; that is the whole point behind a steady distribution. That is easy, we can set $\lambda = 1$:

$$\vec{v}P = \vec{v}$$

Usually the steady state distribution is represented by π, so:

$$\pi P = \pi \tag{5.13}$$

If we solve that linear system, we are off to the races; I got a tip that the horse named "Step Aside" is running hot, so perhaps we can place a bet, shit... forget those races, we're talking about the steady state distribution, of course, yes, if we solve the linear system 5.13 we will get our steady-state distribution. The methodology to solve such systems is out of the book's scope; however, if you would like to know about linear algebra, consider the volume 1 of this same series. Now, after asking the computer for help, this is our steady

state distribution:

$$\pi = [0.112, 0.053, 0.031, 0.051, 0.06, 0.69]$$

Let's plot this:

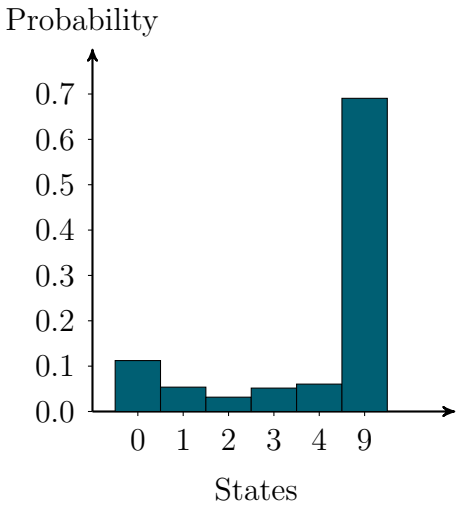


Figure 5.5: I think at some point state 9 will be present.

The steady state distribution describes the likelihood that a system will be in a particular state after a large number of transitions (steps) have occurred, reaching a constant probability. Therefore, the digit 9 is the one that will most likely appear at the end of a code sequence. Just look at that big bar; it's higher than Snoop Dogg at 70%! It is important to remember that the steady state distribution fundamentally differs from a histogram generated from the data. With the steady-state distribution, we are asking, "Where will our system be in n steps?" By contrast, a histogram answers the question, "What has happened in the past?" This distinction, though subtle, is quite powerful.

Let me illustrate a use case where I applied this concept a few times. Imagine you are in a marketing department analyzing the efficiency of your marketing channels. For simplicity, let's consider four channels: Email, Pay-Per-Click (PPC)-these are ads placed on platforms like Google where people get charged every time someone clicks the ad, Direct—when a user lands directly on your website, and Display-those persistent banners that seem to follow us everywhere, the digital version of stalkers. Suppose we decide to create

a model for a customer's journey to a sale via a sequence of digital touchpoints, for example:

$$\text{PPC} \rightarrow \text{Email} \rightarrow \text{Display} \rightarrow \text{Direct} \rightarrow \text{PPC}$$

This sequence could represent a customer's path to purchase, but it also shows that we are potentially paying three times to acquire the same customer due to the two PPC touchpoints and the Display one. Understanding that the steady state distribution represents the long-term view of the system allows us to analyze the likelihood that customers are concluding their journeys with PPC. We might experiment with strategies to alter this probability if we observe a high likelihood, similar to the prominent bar in the plot 5.5. For instance, we could introduce an email follow-up after the third touchpoint. Furthermore, adjusting the transition probabilities in P allows us to simulate how changes affect the steady-state distribution. This enables us to set targets and predict outcomes, such as how "reducing the transition from Direct to Email by $X\%$" might decrease the PPC likelihood in the steady state distribution by $Y\%$.

Not all Markov chains converge to equilibrium and have a unique steady-state probability distribution. To guarantee such features, a chain must have two properties: it has to be Irreducible and Aperiodic:

- **Irreducibility** - A Markov chain is irreducible if it is possible to get from any state to any other state in a finite number of steps, although the number of steps may vary depending on the pair of states. This property ensures that no state is isolated and all parts of the state space are connected.

- **Aperiodicity** - Aperiodicity in a Markov chain refers to how frequently a state can revisit itself over time. A state is called aperiodic if it does not follow a fixed pattern of returning to itself only at regular intervals. In more straightforward terms, there isn't a set number of steps after which the state is certain to return; it can return at any time, unpredictably.

If the chain has these two features, it is called 'Ergodic'; what do you think about that name? Don't you think it would also be an excellent name for a small dog? Maybe those terms are still a bit abstract, so let's visualize them with some small chains:

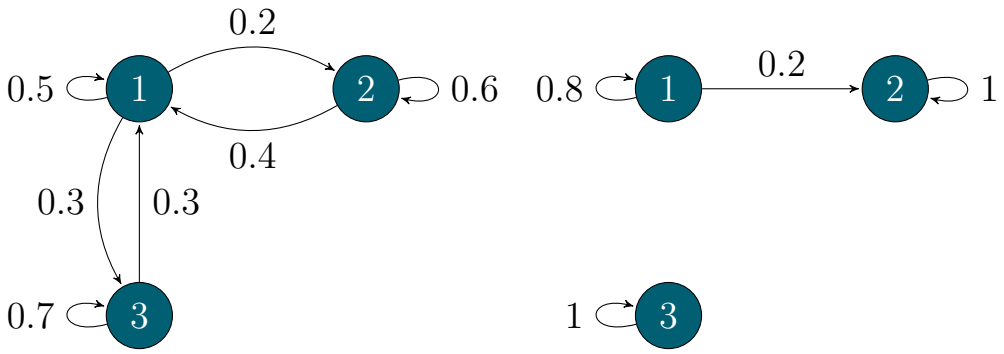


Figure 5.6: An Irreducible Markov Chain vs a Non-Irreducible one.

The left panel of figure 5.6 illustrates an irreducible Markov chain, where all states are interconnected. This interconnection ensures that it is possible to transition from any state to any other state, facilitating a predictable and unified long-term behavior across the system. As a result, the system converges to a single steady-state distribution, regardless of the initial state. By contrast, the right panel shows a non-irreducible Markov chain, where certain states or clusters are isolated, making transitions between some groups of states impossible. Such isolation can lead the system to have multiple steady states or potentially none, depending on where it starts. For example, starting in a state labeled 3 in a non-irreducible part of the chain might mean being stuck there, unable to transition to other states. Now, let's check "Aperiodicity":

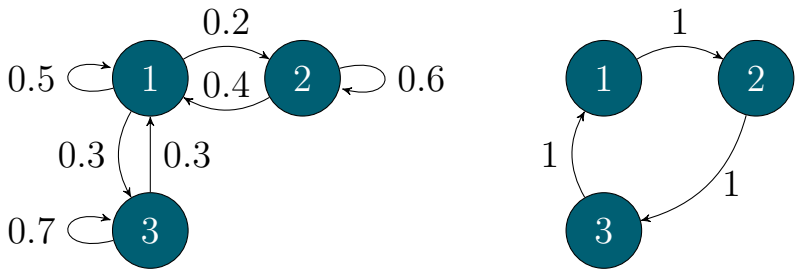


Figure 5.7: To period or not to period, is that really the question?

In Figure 5.7, we again contrast an aperiodic chain with a periodic one across two panels. The left panel displays a state diagram

for an aperiodic chain, where each state can transition to itself and others without adhering to a fixed interval. When considering the steady state, defined as a condition where state probabilities are independent of initial conditions, the flexibility in transition intervals significantly contributes to achieving equilibrium. In aperiodic chains, as the number of transitions increases, the absence of a fixed return pattern to any state enhances the likelihood of the system reaching an equilibrium point.

Looking to the right panel, we observe the opposite behavior; there is an explicit periodic behavior. In periodic Markov chains, the probability of being in a particular state at a given time depends on the starting state and the number of transitions modulated by the cycle period. This dependency means state probabilities oscillate according to the cycle's length without settling. Each state revisits itself predictably after a fixed number of steps, creating a dependency cycle that does not diminish over time. If you pick any starting state, we will end up on the same step in 3 time steps. This persistent memory of the starting point precludes the achievement of a steady state, where the memory of the initial state should ideally fade entirely.

I know they are all excellent properties, definitions, and, more importantly, tools we can use to tackle sequential problems. Still, so far, we don't know what we are going to put on the keypad to enter that Diamond Center; we only know that, the code probably ends on a nine. I promise we will crack that code, but you know how we roll; if we are working towards a goal and can learn about other stuff while doing so, we will take it. Yes, the steady-state distribution might not be the best framework in which to decipher a code, but there are a lot of use cases for that, just like that marketing scenario.

If we wish to start building sequences of digits, I suggest we go back and look at our transition matrix P. This guy represents the likelihood of where we will go given where we are; wow, this sentence would provide a good post for some internet guru. In any case, we can use that information and start a simulation process, hoping that some of the sequences will repeatedly come out of our simulation adventure. The problem is, with the current transition

matrix, where do we stop? We have no indication of how many digits the code has, but there is something in our dataset that might save our asses. We decided to leave out two states: Start and Enter. I do have a confession; I only left them out so we could study the steady-state distribution, as every time I worked with a marketing department, I used that bad boy, so I hope you forgive my detour.

With the inclusion of Start and Enter states in our model, we define clear beginning and end points for the code's execution. However, this structure means we sacrifice Ergodicity; for instance, once we reach the Enter state, transitions to other states are not possible, which deviates from the typical properties of a fundamental chain.

Despite this, the lack of a steady-state distribution is acceptable in our specific use case, as it does not significantly impact our objectives. Now, let's move forward by defining our new state space, followed by an updated transition matrix:

$$0, 1, 2, 3, 4, 9, \text{Start}, \text{Enter}$$

We will do the same as we did before to get the transition matrix 5.7 but now we will consider every pair in our dataset:

$P =$

	Start	0	1	2	3	4	9	Enter
Start	0	0.3	0.25	0.2	0.25	0.0	0.0	0.0
0	0	0	0.2	0.3	0.2	0.28	0.0	0.01
1	0	0.3	0	0.3	0.2	0.2	0.0	0.0
2	0	0.25	0.23	0	0.3	0.2	0.0	0.02
3	0	0.2	0.3	0.25	0	0.25	0.0	0.0
4	0	0.15	0.2	0.2	0.25	0	0.2	0.0
9	0	0.1	0.0	0.0	0.0	0.1	0	0.8
Enter	0	0	0	0	0	0	0	1

(5.14)

A transition matrix such as 5.14, in essence, works like a map for our system given a current state; it tells us where we are more likely to go, a phenomenon that opens up the opportunity to generate some path, as we will have a probabilistic framework for each step. We know that all our codes begin with the key Start, so let's see where the system is likely to be after that:

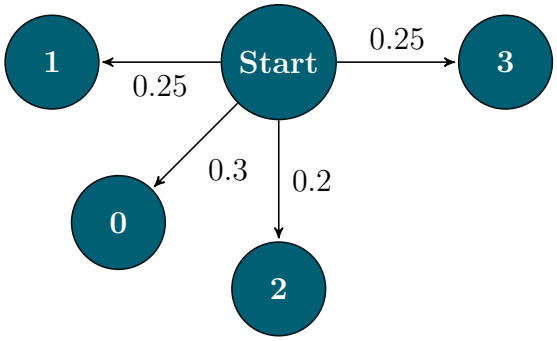


Figure 5.8: State transition diagram from the Start state.

Figure 5.8 represents the state distribution of the Start state; in other words, all the possibilities of where we will end up after Start and how likely we are to get there. For example, 25% of the time, from Start, we go to 3. Let's assume this happened, and now we are at state 3:

$$\text{Start} \rightarrow 3$$

Now we moved from 'Start' to '3' and that means we have a different state distribution:

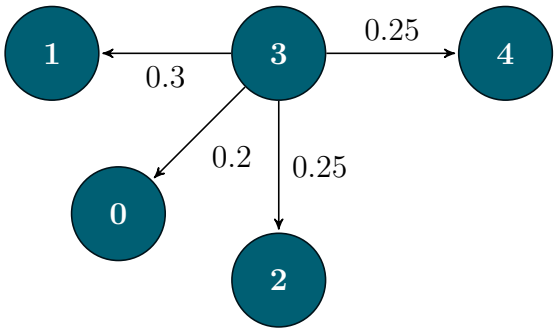


Figure 5.9: Now the transitions but from 3.

Figure 5.9 illustrates the potential state transitions from state 3. Continuing from here, we can select another state and iterate this process until we reach the Enter state, which represents completing one possible code sequence. This outcome, however, is just one among many possibilities. For instance, if the transition from Start had led to state 2 instead of state 3, the resulting sequence would have differed significantly.

The fact that we can explore different paths with our transitions matrix suggests a wide range of possibilities for the code that will open the garage, and that means one thing for us: we will have to explore them thoroughly.

To explore the full range of possibilities, what if we simulate thousands of sequences initiating from Start and terminating at Enter? By doing this, we can identify the most frequently occurring sequences, which could serve as our candidate codes. Yes, we are going to do some simulations! However, earlier, we had a nice equation from which we could draw numbers, now we have a discrete set-up with the state distributions that represent our randomness, so we must find a way to generate these sequences.

How do we go from here, where everything is "static"'? I mean, we have a transition matrix with probabilities but no equation to draw any numbers from it, as we had before with, for example, the PDFs. Now what do we do? Magic! Not really, but if we need something random, what about using a random number generator?

A random number generator starts with what is known as a "seed"—a starting number from which the generator begins its process. This seed undergoes a series of mathematical operations, including multiplication, addition, and modular functions, to ensure the output is sufficiently "random." Each result becomes the seed for the next round, creating a chain of random numbers. It's important to note that the sequence of operations used in generating these numbers isn't arbitrary; they are carefully selected and tested for statistical robustness to ensure that the randomness meets certain standards, one being that the numbers are between 0 and 1.

The question we are addressing here involves the use of random number generators in conjunction with our transition matrix to create sequences of digits, which will ultimately serve as our codes. Our goal, given that we are in a state t (representing a digit), is to determine the next number based on the state distribution at that particular time. By successfully linking a random number to a state distribution, we can then generate digits randomly according to this distribution.

In the previous paragraph, we observed that the process of generating random numbers ensures these values are always between 0 and 1, matching the range needed for our state distribution. However, consider the following scenario: if we draw a random number, say, 0.34, how do we relate this to a specific digit within our state

distribution? For example, referring to figure 5.9, what happens if we draw 0.34? Do we select 1 as the next digit because its likelihood is 0.3 and this is the closest to 0.34 among all digit probabilities? This seems problematic because any number we generate above 0.3 would cause us to select digit 1 as the next state, thereby creating a significant bias. Clearly, determining the next state by the proximity of state probabilities to the random number is not an effective strategy.

Instead, what if we assign a range for each digit in the state distribution? For instance, if the random number is between 0 and 0.2, we select digit 0; whereas if the random number falls within the interval of 0.2 to 0.45, we opt for digit 2. By defining such intervals for each digit in the state, we can avoid the biases introduced by merely calculating distances between random numbers and state probabilities.

However, we need to be careful in how we define these ranges or intervals; they must accurately reflect the probabilities of each digit at the given state. To achieve this, we can utilize frameworks that operate within the 0 to 1 range, such as the CDF and the PDF if we integrate it. Well, we don't have the PDF, so let's shift our attention to the CDF to define our intervals. But how can we build such thing for a state distribution?

Traditionally, when dealing with CDFs, the data follows a specific order that reflects a natural or logical sequence of events, as seen in typical statistical applications where the x-axis represents a progression through quantitative outcomes. This ordered representation ensures that each subsequent value on the x-axis includes the cumulative probability of previous values, which is central to the nature of a CDF. However, in the context of our state distributions for digit generation, the sequence or order of states on the x-axis can differ without affecting the functionality of the CDF. This is because each state's position is merely a label rather than a quantitative measure, and the probabilities can still accumulate correctly. Therefore, while in many cases the CDF is built upon a naturally ordered dataset, for our state distributions, the specific order of states is not inherently relevant. What about if we pick a state, and create

the CDF for it? Say state 3, the same one represented in diagram 5.9:

$$\begin{aligned} 0 &\to 0.2 \\ 1 &\to 0.2 + 0.3 = 0.5 \\ 2 &\to 0.5 + 0.25 = 0.75 \\ 4 &\to 0.25 + 0.75 = 1 \end{aligned} \tag{5.15}$$

In 5.15, we have outlined a cumulative distribution function for the state distribution following digit 3. As we established in the previous paragraph, the order of the states on the x-axis is not critical to the construction of the CDF; instead, it is the cumulative probabilities that matter. To construct this CDF, we simply sum the probabilities of transitioning to each subsequent state. For instance, there is a 20% probability of transitioning from 3 to state 0, which is represented by the interval $0 \to 0.2$. Following this, there is a 30% chance of moving from 3 to 1, so we extend the CDF to 0.5, adding the 0.3 probability of transitioning to 1 to the previous 0.2. This process is repeated for the remaining states, cumulatively adding each state's transition probability to define the complete CDF. This method ensures that each segment of the CDF correctly represents the probability of transitioning from 3 to each respective state, illustrating how each step in the function accounts for the accumulated likelihoods of moving to or past each state. So we have a CDF, now it is time to define some intervals, for example for the state 0:

$$\text{State 0:}[0, 0.2] \tag{5.16}$$

In 5.16, we are saying that, if we generate a random number between 0 and 0.2, we select the state 0, perhaps a bold statement, but let me justify it. Because the random numbers are uniformly generated and always between 0 and 1, this means that with a big enough sample, 20% of them will be between 0 and 0.2. So, if our system moves to 0 from 3 20% of the time, this interval will reflect that! We can use the same tecnique to create the other intervals:

$$\text{State 1:}]0.2, 0.5], \quad \text{State 3:}]0.5, 0.75], \quad \text{State 4:}]0.75, 1]$$

Each interval's width precisely matches the transition probability from 3 to the state represented by that interval. With this strategy

in place to generate sequences, if, for instance, a randomly generated number is 0.65, we select state 3 as our next digit. We then continue to apply this same process, systematically generating and mapping new random numbers to their corresponding states:

1. Draw another random number.
2. Create the CDF for the current state distribution.
3. Now, state 2.
4. Select the next digit.

This process continues until we reach the Enter state, our stopping criterion. Once a sequence completes, we start again to generate another sequence, and we repeat this procedure until we have enough sequences to analyze. For this experiment, let's say we aim to generate 10000 codes using the methodology described above:

$$\begin{array}{ll} \text{Code: Start} \to 0 \to 4 \to 9 \to \text{Enter} & \text{Count: } 478 \\ \text{Code: Start} \to 3 \to 4 \to 9 \to \text{Enter} & \text{Count: } 373 \\ \text{Code: Start} \to 1 \to 4 \to 9 \to \text{Enter} & \text{Count: } 321 \\ \text{Code: Start} \to 2 \to 4 \to 9 \to \text{Enter} & \text{Count: } 241 \\ \text{Code: Start} \to 0 \to 3 \to 4 \to 9 \to \text{Enter} & \text{Count: } 120 \end{array} \tag{5.17}$$

And there we have it, after 10000 simulations, 5.17 represents the most observed frequencies and our best candidates with which to try and enter the garage. To obtain such sequences, we used the methodology we described above and then just counted frequencies amongst the 10000 samples of codes we got from the simulation process.

What a journey just to get a code! It all started with a sample of guards' hourly visits to the keypad and the need to understand how many code inputs we could capture on a given day. With that data we created two different methodologies to understand the success of installing a camera unnoticed (the binomial distribution), and then the number of code inputs we could capture on a given day (the Poisson distribution).

With the data we had access to, our chances of installing the camera without bumping into a guard were meager. However, we met Rita, who provided us with great intel: "When my flag is out, our team is playing." This became a new, critical piece of evidence that we incorporated into the beliefs provided by the binomial distribution. This revelation helped us see a window in which to install the camera, and we used Rita's flag as a proxy to reduce the likelihood of encountering guards.

That led us to Bayes' Law, a phenomenal mathematical framework that is so simple yet powerful. When we decided to opt for the Poisson distribution to model the rate of visits to the keypad as rigorous thieves... DAMN IT, not again, I meant mathematicians; we decided to test our data with some new methodology we deduced. The Chi-Square test led us to understand that although the evidence wasn't sufficient to reject H_0 and disregard the Poisson distribution, the power of the test could have been more encouraging.

However, we made good use of Bayes' Law: The prior reflects our initial beliefs, and the likelihood will update us with new evidence. With that came the concept of the conjugate prior, nearly a mathematical trick to facilitate calculations. With this handyman came the Beta, the Gamma, and the Uniform distributions, which helped us understand how the rate of visits λ behaved.

When everything seemed ready for us to check some videos and obtain some codes, things were revealed to be more complex than we expected. Beatriz was nowhere to be seen. Luckily, we gained access to some tables, one of them containing features from numbers, such as symmetry and the number of loops in numbers. With that, we explored a machine learning technique called Naive Bayes, which allowed us to classify numbers using Bayes' Law to understand the likelihood of those same features against a labeled number.

Finally, we had to create a framework to simulate sequences because the information we received came in pairs of numbers. Who is better at sequences than Markov? A new definition of a stochastic process and how to model them resulted in us learning about the Markovian property; "Today, we only care about yesterday." With this, we created a list of sequences of digits based on simulations

that were the candidates for codes to enter the garage. And this means we can now enter the building.

Notarbartolo understood that the devil was in the details of the intricate world of security. The supposedly state-of-the-art system had its technological Achilles' heel. For instance, the outdated VCRs, which relied on physical tapes, presented a significant vulnerability: tapes could be stolen or swapped, effectively erasing any evidence of the heist. These days, they no longer have tapes; a live feed goes directly to a security company, so that there is no way to deactivate it without the police arriving promptly.

From the garage door to the vault, we must enter another door and navigate a L-shaped path consisting of two corridors with four cameras. Each leg of the L has two cameras positioned at its ends. We can observe the first two cameras from the second door, but the last two are unobservable due to the angle of the walls:

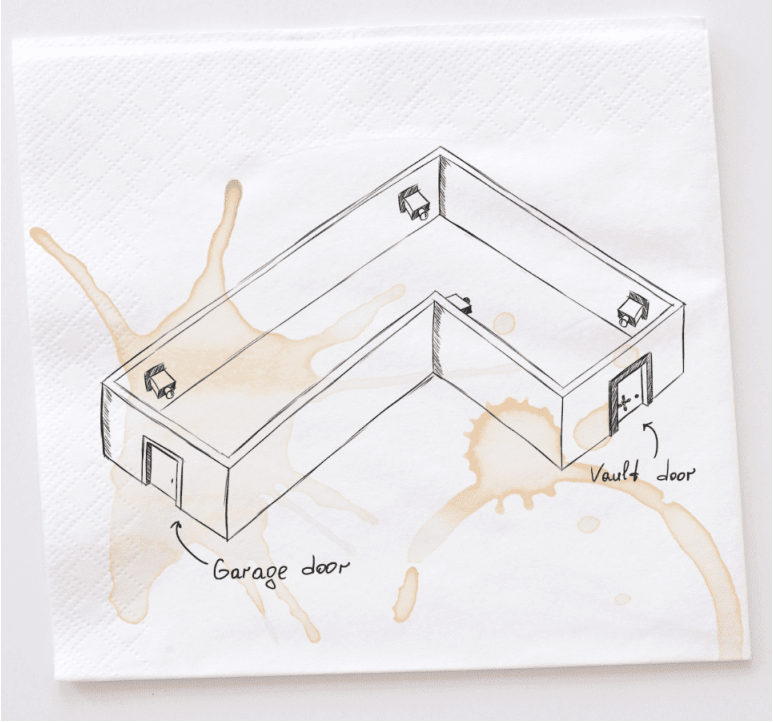


Figure 5.10: The L that leads to the gems.

Yeah, things are trickier for us, but let's not give up just yet. I have a dataset that we might be able to do something with:

TimeSlot	Camera1	Camera2	Camera3	Camera4
00:00	160	150	140	45
00:10	150	140	130	60
00:20	140	130	120	0
00:30	330	120	110	10
00:40	120	110	100	20
00:50	110	100	90	30
01:00	100	90	80	40
01:10	90	80	70	50
01:20	80	70	60	60
01:30	70	60	50	70
01:40	60	50	200	80
01:50	50	40	30	90
02:00	40	30	20	100
02:10	30	20	10	110
02:20	20	10	5	120
02:30	10	5	0	130
02:40	5	0	181	140
02:50	0	45	120	150
03:00	33	57	197	160

Table 5.1: The Diamond Centre camera angles.

In table 5.1, we have a dataset that captures the directional angles of four ceiling-mounted cameras. The angles range from 0 to 360 degrees; we have data measurements for all four cameras in 10-minute intervals.

An angle of 180 degrees indicates that a camera is pointing directly to the right, parallel to the ground, and an angle of 0 degrees means it is pointing directly to the left. These specific angles are noteworthy because they represent potential blind spots. Importantly, each camera maintains its set angle for the entire 10-minute interval, which can lead to detection issues in the corridor due to suboptimal positioning.

If we find instances where all the cameras are positioned at dead angles, this might just have save us, as I was starting to get concerned about how the hell we were going to bypass all these cameras. Let's further explore this data and see if we can capture more surveillance blind spots.

Ideally, we need to find a moment when all the cameras are in a dead angle, creating a complete blind spot over both corridors, allowing us to go from the door in the garage to the vault door unnoticed. It seems like a complex task for such a small sample of data, but let's see what we can do with it. How about we start simple; say we pick camera one first, check if we can do something with it, and then scale to all cameras?

One strategy could involve creating a histogram, fitting a distribution, and then sampling from this distribution to augment our dataset. However, this approach might lead us towards potential pitfalls. Allow me to explain. First, this method assumes that each sample from camera 1's angle is independent, an assumption that stretches credibility in this context. Such independence implies that the camera might remain stopped at a given position for multiple 10-minute intervals or oscillate back and forth randomly. Neither scenarios seems very realistic because I imagine the camera moving sequentially through different angles and values. Therefore, we need a process that acknowledges some level of dependency, for example, something that allows us to compute the likelihood of a position given we came from a specific one:

$$P(\text{Camera 1 is at a } 140^\circ \mid \text{Camera 1 was at a } 150^\circ) \tag{5.18}$$

The angles in 5.18 are arbitrary and just to show what we are aiming to build, some type of model that allows us to understand dependencies. Moreover, the assumption that the cameras operate independently of each other might lead us straight into a nice cell. If the security cameras are designed to prevent unauthorized people from getting from the garage door to the vault, there is no way that these devices don't move in synchrony. Otherwise, this will be a fragile system.

Thus, we find ourselves with a new paradigm, not only because we need to account for dependencies within a single variable, such as sequences of angles of a single camera, but also because we must manage dependencies across multiple cameras and, therefore, multiple variables.

5.2 Do We Really Need More Variables? Multivariate Distributions.

Welcome to the exciting world of multidimensionality! Right, but before starting to deduce methodology to accommodate four

different variables related to all the cameras, let's ease ourselves into this topic. Do you miss the bell curve? Me too! Why don't we explore a multidimensional normal distribution if that is the case? If we have a random variable X that follows a normal distribution, we can represent it by:

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

This is nothing new for us, a normal distribution defined by its mean μ and variance σ^2. Now we don't just have one random variable; our example comes with four of them; but you know the saying, if it is good for n, it must work for four, so let's consider n multiple random variables, $X = (X_1, X_2, ..., X_n)$ where $X_1, X_2, .., X_n$ have normal distributions.

For our explorations of the multivariate Gaussian distribution, I propose two dimensions, as we can create some visuals to better understand the dynamics of their joint probability distribution, which is a function that simultaneously reflects the behavior of these two variables. So, if we go with X_1 and X_2, we can defined them like:

$$X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2), \quad X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$$

Two vanilla ice creams... Two vanilla random variables X_1, X_2 normally distributed defined by their respective population means μ_1, μ_2 and population variances σ_1^2, σ_2^2. Cool, now we can create one of my favourite mathematical elements, the equation! Let's start by defining the PDFs of X_1 and X_2 as:

$$f_{X_1}(x), \quad f_{X_2}(x) \tag{5.19}$$

Equation 5.19 represents the PDFs for the variables X_1 and X_2, the two individuals we need to find a way to combine to have a joint PDF equation. With regard to combining $f_{X_1}(x)$ and $f_{X_2}(x)$, we have four options: add, subtract, multiply, or divide. Since the PDFs in the continuous case return densities, subtracting them is a bad idea because we might end up with negative values, which is a no-no. In terms of dividing them, the result would be hard to interpret, and there is no meaning to a ratio of PDFs in probability theory. Well, with those two excluded, we are left with the multiplication

and the addition. Adding them will mean that we are looking for what happens when we observe X_1 or X_2, and that is not what we are looking for; we want to know what happens when both X_1 and X_2 occur, so:

$$f_{X_1,X_2}(x_1, x_2) = f_{X_1}(x_1) \cdot f_{X_2}(x_2) \tag{5.20}$$

That is not bad; we just have to multiply the PDFs, and we are on our way; it seems too good to be true. And that is because it is. Doing this, we consider X_1 and X_2 independent; therefore, 5.20 does not capture any dependency between variables. One of the motivations for creating a joint probability distribution was to understand the possible dependency between cameras to explore the possibility of a blind spot. A way to introduce such a concept into 5.20 is by including conditional probabilities, for example:

$$f_{X_1,X_2}(x_1, x_2) = f_{X_1}(x_1) \cdot f_{X_2|X_1}(x_2|x_1) \tag{5.21}$$

With 5.21, the conditional term on the right side of the multiplication incorporates the dependencies that might occur between X_1 and X_2. Next, we integrate equation 5.21 between plus and minus infinity. Would you care to do this integration? No? As we stand, me neither. I don't even have to expand that integral to know if it will be nasty, so before jumping into a maze of madness and calculations, let's see if there is another way we can take it; I mean, we are dealing with normal distributions, which contain properties that can simplify calculations. Let's stay on the fact that we must capture the variance between random variables. Well, we have a tool for that, remember? That's the covariance matrix!

$$\Sigma = \begin{bmatrix} \sigma_{X_1}^2 & \sigma_{X_1X_2} \\ \sigma_{X_1X_2} & \sigma_{X_2}^2 \end{bmatrix} \tag{5.22}$$

Using a covariance matrix is particularly effective when handling variables like X_1 and X_2, which share the same distribution and the same scale. This uniformity ensures that variance, one of the critical parameters that characterize these variables, is consistently applied across the analysis. In the covariance matrix, diagonal entries like $\sigma_{X_1}^2$ and $\sigma_{X_2}^2$ directly represent the variance of each variable, indicating their variability. Off-diagonal entries, σ_{X_1,X_2} and σ_{X_2,X_1},

measure the covariance, revealing how these variables change about each other. Having a matrix requires us to change our notation a bit because, if we want to leverage the power of linear algebra, we must operate with scalars, vectors, and matrices. As we have two means, let's start by defining a couple of vectors:

$$\vec{X} = [X_1, X_2], \quad \vec{\mu} = [\mu_1, \mu_2] \tag{5.23}$$

In 5.23, we have two vectors: $\vec{X}$ represents our random variables, whereas $\vec{\mu}$ contains our means. The next step is to understand how we can incorporate all this algebra into a Gaussian PDF, so let's bring one of those back:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{5.24}$$

We can see that, in equation 5.24, the mean and the variance are both in the exponent and the normalization factor, so we must find a way to scale these univariate terms to multidimensional elements. Or, say the same but in street maths, where are we going to squeeze those vectors and matrices in. Let's start high up on the second floor with the exponent:

$$-\frac{(x-\mu)^2}{2\sigma^2} \tag{5.25}$$

Now, instead of σ^2 representing the variance, we have a covariance matrix Σ. Because μ is playing very badly, it will be substituted by a vector $\vec{\mu} = [\mu_1, \mu_2]$ and x will follow along and be defined as vector $\vec{x} = [x_1, x_2]$. We see a σ^2 in the denominator of 5.25, and that bad boy, in our multivariate case, must include the dependencies between X_1 and X_2. Therefore, it has to be replaced with Σ; the problem is that we don't have the division operation in linear algebra, so we can't just replace it on the equation. Instead, we multiply by the inverse of the same matrix to achieve a similar operation. [REMEMBER TO USE THIS AS AN UPSELL!] We cover such topics in my first book on linear algebra; maybe consider checking it out if linear algebra interests you. Anyhow, our first candidate for the exponent is:

$$\Sigma^{-1}$$

Next, we have that gentleman, $(x - \mu)^2$. If you recall, when we deduced the expression for the PDF of the univariate normal distribution, we said that the term was there to calculate the distance of x to the mean, and because a distance has to be positive, we squared it. We must find a way to achieve the same goal but with vectors and matrices. Remember that our new x is now a vector $\vec{x} = [x_1, x_2]$ and μ is equal to $\vec{\mu} = [\mu_1, \mu_2]$. We can try multiplying the square of the difference of those vectors:

$$[\vec{\mu} - \vec{x}]^2 \cdot \Sigma^{-1} \tag{5.26}$$

While it's true that the multiplication in 5.26 is mathematically sound; when we perform element-wise squaring on the vector we focus solely on the magnitudes of deviations from the means of each variable, ignoring the way these variables might be influencing each other. Each squared term $(x_i - \mu_i)^2$ considers only the variance component for a single dimension and loses any information about the relationship between different dimensions. We can avoid this by making use of the quadratic form:

$$[\vec{x} - \vec{\mu}]^T \Sigma^{-1} [\vec{x} - \vec{\mu}] \tag{5.27}$$

The quadratic form 5.27 in the multivariate Gaussian distributions effectively measures the "distance" of a data point from the mean, accounting for both variances and covariances. This scalar value is crucial because it generalizes the concept of squared distance from univariate to multivariate contexts, where simply summing squared deviations fails to consider inter-variable relationships. Now, if we manage to address the normalization factor, we will have everything we need to finalize our equation:

$$\sqrt{2\pi\sigma^2} \tag{5.28}$$

When dealing with the scaling of the exponent, the transition from σ^2 to the covariance matrix was relatively smooth. We could do that due to the possibility of representing the other elements, namely, x and μ, in vector form. Now, we have a different paradigm to solve. We know that the sole purpose of that 5.28 is to make the PDFs to integrate to one. Simply replacing σ^2 with Σ won't cut it

because, as we covered before, we don't use divisions with matrices . Also, the result of the numerator must be a number and not an algebraic structure. So, we need to find a way to incorporate all the information present in the covariance matrix but in the form of a scalar. What about the determinant? We know that this guy represents an area or a volume and, therefore, a scalar, but will it provide us with the information we need?

To understand if the determinant will provide us with what we need, we first need to understand what shape are we computing an area, or its volume. In one dimension, a Gaussian distribution is represented by a simple bell curve, centered at the mean and spreading out based on its variance. When we extend this concept to two dimensions, the distribution doesn't just spread out in one direction but in two, leading to a shape that can tilt and stretch based on how the two variables relate to each other. Well, this results in a ellipse, whose orientation and width are dictated by how much one variable affects the other, capturing not just individual variances but also their covariance, or mutual variation:

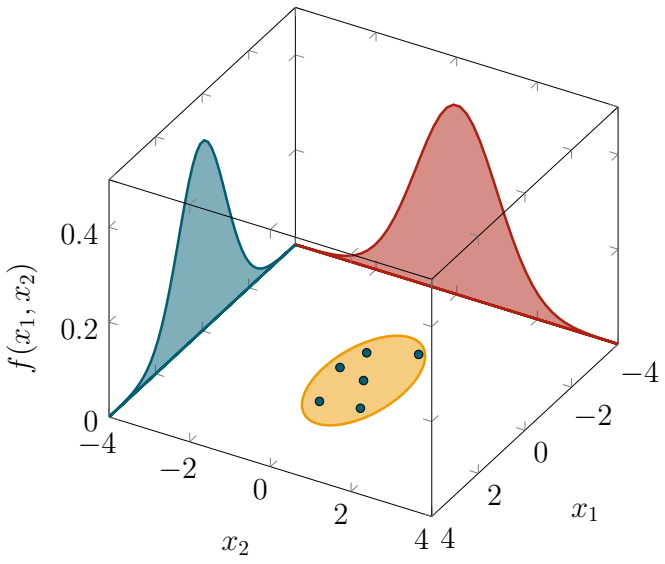


Figure 5.11: Two mountains and a lake.

In figure 5.11, we depict two Gaussian distributions using blue and red colors. At the center, a yellow ellipse represents the area where observations from the joint distribution are most likely to occur. This visual representation ties directly into the role of the

covariance matrix, which is fundamental in shaping this ellipse.

The determinant of the covariance matrix is a scalar value that indicates the volume of the multi-dimensional ellipsoid shaped by these covariances. This volume reflects how much space the data occupies when considering all dimensions together, akin to measuring the "size" of the statistical space, the yellow shape in our plot. The determinant is crucial for normalization, as it adjusts the scale of the multivariate Gaussian distribution to ensure that the probability density function integrates into one. This integration is a fundamental requirement for all probability distributions.

Furthermore, the normalization factor in a multivariate normal distribution, derived from the determinant of the covariance matrix, adjusts for changes in the 'size' or 'volume' of the statistical space. This volume is influenced by the spread and correlations among the variables. The determinant itself is a complex function of all the elements of the covariance matrix, not just the individual variances or pairwise covariances. For instance, if the covariances between variables increase, this generally suggests a broader spread across these dimensions, potentially increasing the determinant. However, the exact effect on the determinant depends on the intricate balance of all covariance matrix elements. As the determinant increases, the normalization factor — which includes the term $\sqrt{(2\pi)^n \det(\Sigma)}$ where n is the number of variables — correspondingly decreases. This scaling down of the normalization factor is essential to maintain the total probability at one, adjusting not just the height but the overall shape of the distribution to reflect increased variability and maintain proper probabilistic calibration. The factor $(2\pi)^n$ ensures that this adjustment is appropriately scaled across all dimensions of the distribution:

$$\frac{1}{(2\pi)^{\frac{n}{2}} \cdot |\Sigma|^{\frac{1}{2}}} \tag{5.29}$$

5.2.1 All Those Bells Would Make Any Fire Station Envious: The Multivariate Gaussian Distribution.

With 5.29 we are now finally ready to have our equation for the PDF of the multivariate Gaussian distribution:

$$f(\vec{x}) = \frac{1}{(2\pi)^{\frac{n}{2}} \cdot |\Sigma|^{\frac{1}{2}}} \cdot e^{\left(-\frac{1}{2}(\vec{x}-\vec{\mu})^T \Sigma^{-1}(\vec{x}-\vec{\mu})\right)} \tag{5.30}$$

Where:

- $\vec{x}$ and $\vec{\mu}$ are vectors,
- Σ is the covariance matrix,
- $|\Sigma|$ is the determinant of the covariance matrix,
- n is the number of dimensions.

And there we have it, our first multivariate PDF! Equation 5.29 depicts the density function for n random variables X provided that all of them are Gaussian. This distribution is widely used in machine learning; a good example is the Gaussian Mixture Models (GMM), a topic outside the scope of this book, but you will have no issue learning it if you wish. Now, what comes after a series of deductions? Ice cream! We can have some, but only under the condition that it comes accompanied by an example of a multivariate Gaussian distribution. So, let's consider two random variables X_1 and X_2 with the following distributions:

$$X_1 \sim \mathcal{N}(2, 1), \quad X_2 \sim \mathcal{N}(4, 2)$$

Cool; now, to kick things off, we need a vector with the means as well as a covariance matrix:

$$\vec{\mu} = \begin{bmatrix} 2 \\ 4 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 2 \end{bmatrix}$$

The vector $\vec{\mu}$ contains the means of our variables X_1 and X_2. And Σ is our covariance matrix where the entries that populate the

diagonal are the variance of the same random variables. The other entries represent the dependency between variables, and I calculated them for us using the same formulae we studied when covariance was introduced. I won't be providing the dataset as the goal of the exercises is to understand the multivariate Gaussian distribution further, so let's assume both those parameters $\vec{x}$ and Σ to be true. Our focus now shifts to Σ, as we must invert it and also compute its determinant:

$$|\Sigma| = (1 \cdot 2) - (0.5 \cdot 0.5) = 2 - 0.25 = 1.75$$

That gives us our determinant. Next, we need the inverse:

$$\Sigma^{-1} = \frac{1}{1.75}\begin{bmatrix} 2 & -0.5 \\ -0.5 & 1 \end{bmatrix} = \begin{bmatrix} \frac{8}{7} & -\frac{2}{7} \\ -\frac{2}{7} & \frac{4}{7} \end{bmatrix}$$

Now we just need to substitute the values of μ and Σ into the equation 5.29:

$$f(x_1, x_2) = \frac{1}{(2\pi)\sqrt{1.75}} e^{\left(-\frac{1}{2}\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} 2 \\ 4 \end{bmatrix}\right)^T \begin{bmatrix} \frac{8}{7} & -\frac{2}{7} \\ -\frac{2}{7} & \frac{4}{7} \end{bmatrix} \left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} 2 \\ 4 \end{bmatrix}\right)\right)}$$

Impressive expression, but what can we do with it? We can do everything we did with the univariate case, but for n variables. For example we can compute densities; say we have the point $(1, 2)$:

$$\begin{bmatrix} 1 \\ 2 \end{bmatrix} - \begin{bmatrix} 2 \\ 4 \end{bmatrix} = \begin{bmatrix} -1 \\ -2 \end{bmatrix}$$

$$f\left(\begin{bmatrix} 1 \\ 2 \end{bmatrix}\right) = \frac{1}{2\pi\sqrt{1.75}} \cdot e^{\left(-\frac{1}{2}\begin{bmatrix} -1 & -2 \end{bmatrix}\begin{bmatrix} \frac{8}{7} & -\frac{2}{7} \\ -\frac{2}{7} & \frac{4}{7} \end{bmatrix}\begin{bmatrix} -1 \\ -2 \end{bmatrix}\right)}$$

$$= \frac{1}{2\pi\sqrt{1.75}} \cdot e^{\left(-\frac{1}{2}\left(\frac{8}{7} + \frac{16}{7} - \frac{8}{7}\right)\right)}$$

$$= \frac{1}{2\pi\sqrt{1.75}} \cdot e^{\left(-\frac{16}{14}\right)} \approx 0.038$$

That means that the point $(1, 2)$ has a density of approximately 0.038. If we wish to understand probabilities, because we are in a continuous scenario, we know that we need a range and integrals. On

top of this, we have two dimensions, so we must employ a multivariate integral. For example, say you want to calculate the likelihood of x_1 and x_2 being between (1,2) and (2,4):

$$\int_1^2 \int_2^4 \frac{1}{2\pi\sqrt{1.75}} e\left(-\frac{1}{2}\left(\begin{pmatrix}x_1\\x_2\end{pmatrix}-\begin{pmatrix}2\\4\end{pmatrix}\right)^T \begin{pmatrix}\frac{8}{7} & -\frac{2}{7}\\ -\frac{2}{7} & \frac{4}{7}\end{pmatrix}\left(\begin{pmatrix}x_1\\x_2\end{pmatrix}-\begin{pmatrix}2\\4\end{pmatrix}\right)\right) dx_2\, dx_1 \approx 0.141$$

We come up with a probability of 0.141. In practical terms, this distribution is adept at modeling events that exhibit both Gaussian properties and covariances. For example, in a retail setting, a company might use a multivariate Gaussian distribution to analyze customer data on spending amounts and shopping frequency. This statistical model is particularly useful because it incorporates the covariances between these behaviors, providing insights into how they tend to vary together across the customer base. Perhaps, the covariance matrix from this model could reveal that customers who frequently shop also tend to spend more. This isn't just about noting that these behaviors occur together; it's about understanding the extent to which one behavior might predict another.

By mapping out these relationships, the company can better understand customer patterns. Such insights could help tailor marketing strategies, optimize store layouts, and even predict future sales trends. Essentially, the multivariate Gaussian distribution offers a powerful tool for grasping the complex dynamics of customer behavior in a nuanced and quantitatively robust way, enabling businesses to make informed decisions that are directly influenced by observed customer trends.

We have officially ventured into the realm of multidimensional probabilities. So far, everything has been smooth sailing, because we assumed that the distributions for all the random variables were normal, which facilitated our calculations. However, an important question arises: what if, for example, we have three random variables, of which one is normally distributed, another follows a Beta distribution, and a third one adheres to a Gamma distribution? Attempting to multiply these PDF's and capture their dependencies is feasible but is equally a nightmare.

This brings us to our task at hand: to capture blind spots on

camera angles so we can get from the door in the garage to the vault. This exploration of multidimensional probability density functions began because we hypothesized that the movements of the four cameras might be dependent. However, we might have a problem because, although we have a framework to create joint distributions, it still requires several things: We need to know the underlying distribution of the variables, such as in our example where X_1,X_2 followed a normal distribution, to have the ability to multiply the PDFs of the variables, and, lastly, a lot of will power to do those multiplications, which means that there has got to be a better way to do this. Well, let's see if there is!

In table 5.3 we have four cameras C_1, C_2, C_3, C_4 and their joint distribution can be define as:

$$P(C_1, C_2, C_3, C_4) \tag{5.31}$$

With 5.31, we are in a 4-dimensional space, and one way to create that joint distribution is to assume that all the variables are Gaussian. Maybe that is true, but if we think about it, are the cameras all likely to be at the same angle the majority of the time? Probably not; plus, with our data sample, we can't be sure of that, so let's try not to make any assumptions regarding the underlying distributions. If we make no assumption, then we are stuck, and whenever this happens we need to deduce some maths. Firstly let's try to understand the dependencies between cameras, for example:

$$P(C_1|C_2 = c_2, C_3 = c_3, C_4 = c_4) \tag{5.32}$$

Where c_1, c_2, c_3, c_4 represent arbitrary values for camera angles. Equation 5.32 tells us about the positional behavior of C_1 given some arbitrary positions of the other cameras. Not only are we capturing dependencies, but we've also reduced the dimensionality from the four dimensions in 5.31 to just one! In 5.31, we have four random variables, whereas in 5.32, we are just studying what happens to C_1 given the angles of the other cameras. Thus, we managed to decrease the complexity of our task. Let's follow the same logic and create conditional probabilities for all the remaining cameras to see if we are really on to something:

$$P(C_2|C_1 = c_1, C_3 = c_3, C_4 = c_4)$$

$$P(C_3|C_1 = c_1, C_2 = c_2, C_4 = c_4)$$

$$P(C_4|C_1 = c_1, C_2 = c_2, C_3 = c_3)$$

Still skeptical? I get it—how can we link all these conditional probabilities together? Well, I've got an idea! Each conditional probability gives us insights into the behavior of one camera given the positions of the others, meaning that we're effectively capturing their interdependencies. So, what if we use these conditional equations to generate new points from the joint distribution 5.31? Let's picture this: we could start with a randomly selected point, say $C^0 = (c_1^{(0)}, c_2^{(0)}, c_3^{(0)}, c_4^{(0)})$. Here, C represents a point with four coordinates, each representing a camera position; for example, $c_1^{(0)}$ is position 0 of camera one. In the superscript in C, the zero represents the first point. Now, we can go to our conditional probability PDF and draw a new point $C^1 = (c_1^{(1)}, c_2^{(1)}, c_3^{(1)}, c_4^{(1)})$:

$$c_1^{(1)} \sim P(C_1|C_2 = c_2^{(0)}, C_3 = c_3^{(0)}, C_4 = c_4^{(0)})$$

After obtaining a new angle value for camera one, denoted as $c_1^{(1)}$, we need to consider its impact on the positioning of the other cameras, starting with camera 2. This approach reflects a cascading update, where each new camera position potentially influences the next:

$$c_2^{(1)} \sim P(C_2|C_1 = c_1^{(1)}, C_3 = c_3^{(0)}, C_4 = c_4^{(0)})$$

With the new position of camera one set, we now assess where camera 2 might be located. This position, $c_2^{(1)}$, is drawn based on the updated position of camera one and the unchanged positions of cameras 3 and 4. This step ensures that each camera's position is updated sequentially, maintaining a logical order of dependency and influence:

$$c_3^{(1)} \sim P(C_3|C_1 = c_1^{(1)}, C_2 = c_2^{(1)}, C_4 = c_4^{(0)})$$

Next, we determine the position of camera 3, denoted as $c_3^{(1)}$, considering the new positions of cameras 1 and 2. This step illustrates the adaptive nature of our model, where each camera's new position takes into account not just the initial settings but also the

dynamic adjustments of related cameras. The position of camera 4 remains as the only unchanged reference at this point:

$$c_4^{(1)} \sim P(c_4 | C_1 = c_1^{(1)}, C_2 = c_2^{(1)}, C_3 = c_3^{(1)})$$

Finally, we update the position of camera 4, $c_4^{(1)}$, using the newly determined positions of all other cameras. This final step completes one full cycle of updates across all cameras, setting the stage for further iterations if needed. This process emphasizes the interconnected nature of camera positions and showcases how each update depends on a collaborative understanding of the entire system's state.

With this, we have our sample for camera position:

$$C^1 = (c_1^{(1)}, c_2^{(1)}, c_3^{(1)}, c_4^{(1)})$$

For more points, we can keep the momentum going, we can use C^1 to generate C^2, and so on. Each step uses the latest sample as the basis for the next, ensuring that our sampling reflects how these cameras influence each other in real time. By repeatedly sampling in this manner, we not only generate more data points, but also effectively weave together these conditional insights to reconstruct the broader, more complex pattern of the joint distribution. This iterative approach allows us to explore and map out the blind spots among the cameras without ever explicitly defining the entire four-dimensional probability landscape:

$$C^0 \to C^1 \to C^2 \to ... \to C^n \tag{5.33}$$

Diagram 5.33 represents what we will do when applying the process we just described, where we will get n points, where the next one depends on the current point... wait, Markov again?

5.2.2 Wait, Did Markov Brake the Chains and Run Away To Monte Carlo?

Indeed, that is the Markovian property, and because we combine the Markov chain with Monte Carlo simulation, this technique is called Markov Chain Monte Carlo. MCMCs, the acronym for this

method, is the umbrella term for the technique of sampling from a multivariate distribution while capturing dependencies. Within MCMC, we have a selection of different algorithms with which to sample a distribution; the particular one we just discussed is called Gibbs sampling.

5.2.2.1 I Can Never Go Where I Wish! Gibbs Sampler.

Gibbs sampling is a MCMC method used to generate samples from a multivariate probability distribution, particularly when direct sampling is difficult. This method is employed when the joint distribution is complex, but the conditional distributions are easier to handle. The Gibbs sampling algorithm involves the following steps:

1. **Initialization:** Begin with an arbitrary starting point $X^{(0)} = (x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)})$, where n is the number of variables in the distribution.

2. **Iterative Sampling:** For each iteration t, update each variable sequentially from its conditional distribution, conditioned on the current values of all other variables:

$$\begin{aligned} x_1^{(t+1)} &\sim p(x_1 | x_2^{(t)}, x_3^{(t)}, \ldots, x_n^{(t)}) \\ x_2^{(t+1)} &\sim p(x_2 | x_1^{(t+1)}, x_3^{(t)}, \ldots, x_n^{(t)}) \\ &\vdots \\ x_n^{(t+1)} &\sim p(x_n | x_1^{(t+1)}, x_2^{(t+1)}, \ldots, x_{n-1}^{(t+1)}) \end{aligned}$$

3. **Convergence:** Repeat the above step until the samples converge to the target distribution. The convergence is typically assessed using statistical diagnostics for MCMC.

Gibbs sampling is particularly effective, because it constructs a Markov chain whose stationary distribution is the target distribution, but alongside the concept of convergence, it also brings some ugly equations. Conversion of a MCMC is a new concept for us, so let's try to understand it a bit better.

Because we start the method with a random point, the initial samples generated by the Gibbs sampling algorithm may not accurately represent the target distribution we are trying to approximate. Convergence in this context means that as we continue to iterate—updating each variable in sequence based on its conditional probability given the current values of all other variables—the distribution of the sampled values starts to stabilize and align closely with the true underlying distribution we aim to model.

As the iterations progress, the influence of the initial, arbitrary starting point diminishes. This stabilization indicates that the samples are becoming independent of the initial conditions and are now being drawn from the probability distribution of interest.

Imagine you've just moved to a new city without a map. Initially, your explorations start from a random point—your new home. On the first few outings, your understanding of the city is very limited, biased towards the routes and areas immediately surrounding your home. This initial phase is much like the early iterations in Gibbs sampling, where samples are highly influenced by the arbitrary starting point.

As you continue to explore more, venturing further from your starting point on different days, taking various turns and visiting new neighborhoods, your mental map of the city starts to form more comprehensively. This process resembles the Gibbs sampling iterations where each variable (or landmark and route in this analogy) is updated based on your latest explorations (or the latest sampled values).

Eventually, your understanding of the city's layout stabilizes. You gain confidence in navigating between any two points, understanding the overall structure and how different parts connect. This level of familiarity and ease in navigation indicates that you have "converged" in your knowledge of the city layout. Just as in Gibbs sampling, once convergence is achieved, your explorations (or samples) truly reflect the city (or the target distribution), and are not just biased by your initial, random starting point.

Here is a crazy idea, because this is all new to us, what about if

we go back to our multivariate Gaussian and do a couple of iterations of the method? This strategy will be advantageous because we know the distribution well; we can thus validate the technique, apply the equations, and understand when to reach conversion. Good idea? I think so too:

$$\begin{aligned} X_1 \sim \mathcal{N}(2,1), \quad X_2 \sim \mathcal{N}(4,2) \\ \vec{\mu} = \begin{bmatrix} 2 \\ 4 \end{bmatrix}, \qquad \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 2 \end{bmatrix} \end{aligned} \tag{5.34}$$

We know the Gaussian distribution defined by 5.34, as it is the same as the one we used in the multivariate Gauss PDF deduction. Our goal now is to ignore that joint PDF and use the Gibbs sampler to see if, upon conversion, our sampled dataset resembles the same behavior as the one reflected by the defined multivariate Gaussian distribution. So the first thing we need to start this party is a point. For the point selection, we have the same narrative as always; we can select it randomly unless we have some prior knowledge of the domain. As always though, randomly selecting things comes with some setbacks; it can impact the burnout period.

In Gibbs sampling and other Markov Chain Monte Carlo methods, the burn-in period is a critical concept. It refers to the initial phase of the algorithm during which the generated samples are not considered representative of the target distribution. This phase is crucial because the Markov chain needs time to reach a state of equilibrium where the samples are statistically reliable and reflective of the distribution being explored. The length of the burn-in period can be influenced by the initial point selection; if the initial point is far from regions of high probability within the distribution, the chain may take longer to "forget" this start and converge to the equilibrium. Consequently, the samples collected during the burn-in period are typically discarded, as they do not accurately represent the desired characteristics of the target distribution.

OK, now we need to define a method to progress along the chain, moving from our initial state to subsequent states. Returning to our description of the Gibbs sampler, we utilize the conditional probability density functions to determine the likely next state given a prior state. For example, the probability density function of transitioning

to X_1 given that X_2 is at a specific state x_2 is expressed as:

$$f(X_1 \mid X_2 = x_2) \tag{5.35}$$

This equation provides information on how the previous state X_2 influences or determines the behavior of X_1, guiding us in simulating the next step in our Markov chain. The idea behind the sampler is that working with distributions such as 5.35 is easier than deriving the full joint distributions, so we must find a PDF expression for 5.35. Now, we will "cheat" a little bit, and here's how. We will use the multivariate Gaussian PDF to derive the posterior distribution for 5.35 as we said we would ignore that same PDF. Because this is a rudimentary example and the multivariable Gauss is a well-defined joint probability distribution, there is no reason to use the Gibbs sampler. Still, we decided to do so in order to fully grasp the technique. However, the work we are about to do, deducing the PDF for 5.35, will not be in vain, as, when a time comes when we are faced with a complicated joint distribution, we might choose to use the Gaussian distribution to define the conditional PDF required by the Gibbs sampler. So, let's bring back the equation of the multivariate Gaussian distribution:

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_{X_1} \\ \mu_{X_2} \end{bmatrix}, \begin{bmatrix} \sigma^2_{X_1} & \sigma_{X_1X_2} \\ \sigma_{X_1X_2} & \sigma^2_{X_2} \end{bmatrix} \right)$$

The joint density function of X_1 and X_2 is:

$$f_{X_1,X_2}(x_1,x_2) = \frac{1}{2\pi\sqrt{\sigma^2_{X_1}\sigma^2_{X_2} - \sigma^2_{X_1X_2}}} \cdot e^{\left(-\frac{1}{2}\begin{bmatrix} x_1-\mu_{X_1} \\ x_2-\mu_{X_2} \end{bmatrix}^\top \begin{bmatrix} \sigma^2_{X_1} & \sigma_{X_1X_2} \\ \sigma_{X_1X_2} & \sigma^2_{X_2} \end{bmatrix}^{-1} \begin{bmatrix} x_1-\mu_{X_1} \\ x_2-\mu_{X_2} \end{bmatrix}\right)} \tag{5.36}$$

Equation 5.36 represents the equation for the probability density function that characterizes the joint distribution of X_1 and X_2. This setup is essential but not the complete answer to our inquiry; however, it certainly points us in the right direction.

When introducing the Gaussian PDF for a single variable x, we often visualize it with a histogram resembling a bell shape, emphasizing the central role of the mean and variance. One of the biggest fasctors responsible for this shape is the exponential decay function,

whose exponent centers the data around the mean and controls the spread of the data. This characteristic is crucial because it directly impacts how probabilities are assigned based on deviation from the mean. In a joint Gaussian framework for X_1 and X_2, the exponent in the joint density function serves a similar purpose:

$$e^{\left(-\frac{1}{2}\begin{bmatrix} x_1 - \mu_{X_1} \\ x_2 - \mu_{X_2} \end{bmatrix}^\top \Sigma^{-1} \begin{bmatrix} x_1 - \mu_{X_1} \\ x_2 - \mu_{X_2} \end{bmatrix}\right)} \tag{5.37}$$

That guy 6.7 shapes how X_1 and X_2 are jointly distributed around their means, and it reveals how knowledge of one influences our understanding of the other. By focusing on the exponent's structure, particularly the terms involving Σ^{-1} (the inverse of the covariance matrix), we can dissect the dependencies between X_1 and X_2. If we are to understand how X_1 impacts X_1 through the conditional probability $P(X_1 \mid X_2)$, we must isolate X_1 in this exponent. Our first target will be that Σ^{-1}:

$$\Sigma^{-1} = \frac{1}{\sigma^2_{X_1}\sigma^2_{X_2} - \sigma^2_{X_1X_2}} \begin{bmatrix} \sigma^2_{X_2} & -\sigma_{X_1X_2} \\ -\sigma_{X_1X_2} & \sigma^2_{X_1} \end{bmatrix} \tag{5.38}$$

Equation 5.38 is just the formula to invert Σ, but now if we replace Σ^{-1} in 6.7 with it, we will have more common terms to work with:

$$\frac{1}{\sigma^2_{X_1}\sigma^2_{X_2} - \sigma^2_{X_1X_2}} \left(\sigma^2_{X_2}(x_1 - \mu_{X_1})^2 - 2\sigma_{X_1X_2}(x_1 - \mu_{X_1})(x_2 - \mu_{X_2}) + \sigma^2_{X_1}(x_2 - \mu_{X_2})^2\right)$$

We are interested in understanding how the behavior of X_1 depends on different values of X_2. To analyze this, we fix X_2 at a specific, arbitrary value x_2, effectively analyzing X_1 under the condition $X_1|X_2 = x_2$. In this scenario, all the terms in our equation that involve X_2 are treated as constants. This simplification allows us to focus solely on the variability and behavior of X_1 without the influence of changing X_2 values, thereby simplifying our equation and making the analysis more straightforward:

$$\frac{\sigma^2_{X_2}}{\sigma^2_{X_1}\sigma^2_{X_2} - \sigma^2_{X_1X_2}} \left(x_1 - \left(\mu_{X_1} + \frac{\sigma_{X_1X_2}}{\sigma^2_{X_2}}(x_2 - \mu_{X_2})\right)\right)^2 \tag{5.39}$$

That expression 5.39 still seems threatening, but let's not forget that it represents the exponent of a Gaussian distribution, meaning that are dealing with a square difference between a variable and a constant divided by a constant:

$$\left(\frac{x-\mu}{\sigma}\right)^2 \tag{5.40}$$

Equations 5.39 and 5.40 are not that different, right? Check this mapping:

$$\mu \rightarrow \mu_{X_1} + \frac{\sigma_{X_1X_2}}{\sigma^2_{X_2}}(x_2 - \mu_{X_2}) \tag{5.41}$$

$$\sigma \rightarrow \frac{\sigma^2_{X_1}\sigma^2_{X_2} - \sigma^2_{X_1X_2}}{\sigma^2_{X_2}} \tag{5.42}$$

In equation 5.41 the term $\frac{\sigma_{X_1X_2}}{\sigma^2_{X_2}}(x_2 - \mu_{X_2})$ adjusts the mean of X_1 based on the observed value of X_2, scaled by the strength and direction of the linear relationship (covariance) between X_1 and X_2. This adjustment shifts the expectation for X_1.

The other guy represented by 5.42 calculates the variance of X_1 conditional on knowing $X_2 = x_2$. It quantifies the remaining uncertainty in X_1 after accounting for the information provided by X_2. The reduction in variance compared to $\sigma^2_{X_1}$ alone is due to the removal of uncertainty in X_1 that can be explained by the known value of X_2.

The adjustments to the mean and variance of X_1 effectively define the conditional probability $P(X_1|X_2 = x_2)$. By recalibrating X_1's expected value and reducing its variance based on the known value of X_2, these modifications capture how X_2's information directly influences and sharpens our predictions for X_1, embodying the essence of conditional probability. Yeah baby equation 5.39 represents the exponent of the PDF of conditional distribution $P(X_1|X_2 = x_2)$ and because of those similarities described above we can conclude that:

$$\mu_{X_1|X_2} = \mu_{X_1} + \frac{\sigma_{X_1X_2}}{\sigma^2_{X_2}}(x_2 - \mu_{X_2}) \tag{5.43}$$

$$\sigma^2_{X_1|X_2} = \frac{\sigma^2_{X_1}\sigma^2_{X_2} - \sigma^2_{X_1X_2}}{\sigma^2_{X_2}} \tag{5.44}$$

There we have it, the parameters that define our conditional probabilities. Now, we are ready to put this sampler to work. Oh, the initial point! What if we say $X^{(0)} = (0, 0)$? Yeah, I like that too:

$$\vec{\mu} = \begin{bmatrix} 2 \\ 4 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 2 \end{bmatrix}$$

Looking at equations 5.43 and 5.44 we need the following parameters:

$$\mu_{X_1} = 2, \quad \mu_{X_2} = 4$$
$$\sigma_{X_1 X_2} = 0.5, \quad \sigma^2_{X_1} = 1, \quad \sigma^2_{X_2} = 2$$

Now we can proceed to compute both the means and the variances. Let's start with the conditional mean, where $X_2 = 0$:

$$\mu_{X_1|X_2=0} = 2 + \frac{0.5}{2}(0 - 4) = 2 - 1 = 1$$

For the variance:

$$\sigma^2_{X_1|X_2} = \frac{1 \cdot 2 - 0.5^2}{2} = \frac{2 - 0.25}{2} = 0.875$$

Cool, with those results we can define the probability distribution for $X_1^{(1)}$:

$$X_1^{(1)} \sim P(X_1|X_2 = X_2^{(0)})$$
$$X_1^{(1)} \sim \mathcal{N}(1, 0.875) \tag{5.45}$$

The next step is to, generate a sample from the distribution 5.45 which once again, we will use software to get it. Drum roll ... $X_1^{(1)} = 0.995$. Now to get $X_2^{(1)}$ we must move one step further in the chain and calculate our conditional mean and variance by setting $X_1^{(1)} = 0.995$:

$$\mu_{X_2|X_1=0.995} \approx 4 + 0.5 \cdot (0.995 - 2) = 3.5025$$

$$\sigma^2_{X_2|X_1} = 1.75$$

So:

$$X_2^{(1)} \sim P(X_2|X_1 = X_1^{(1)})$$
$$X_2^{(1)} \sim \mathcal{N}(3.50, 1.75)$$

Mister computer, please spit a number from a normal distribution with a mean of 3.5 and a variance of 1.75 ...(weird computer noises)... 3.632. And that completes the steps necessary to get to our new point $X^{(1)} = (0.995, 3.632)$. If we needed another point, the process is the same, but now we start from $X^{(1)}$, to get $X^{(2)}$. For $X^{(3)}$ we start from $X^{(2)}$ and so on for all the extra samples we require. To do a quick sense check on the convergence of the method we will sample a few more points, plot them, and check if the graph will look anything like an ellipse:

$\boldsymbol{X^{(i)}}$	$\boldsymbol{X_1^{(i)}}$	$\boldsymbol{X_2^{(i)}}$
$X^{(1)}$	0.995	3.632
$X^{(2)}$	1.5636	3.6186
$X^{(3)}$	2.6329	1.5957
$X^{(4)}$	0.6031	4.7263
$X^{(5)}$	1.3832	2.5450
$X^{(6)}$	1.9123	1.3297
$X^{(7)}$	0.3485	2.8305
$X^{(8)}$	1.8906	1.9059
$X^{(9)}$	2.7498	1.5757
$X^{(10)}$	2.0366	2.7606

Table 5.2: The results of Gibbs sampler.

In table 5.2, we have represented ten different points, all sampled with the Gibbs sampler. Oh don't feel sorry for me, the calculations where not done by hand, I used a computer! Now let's plot these guys and check how they are displayed:

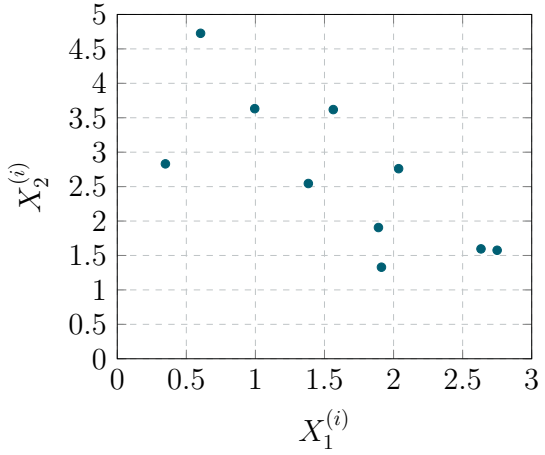


Figure 5.12: The beginning of what can be a great ellipse.

Well, Well, Well, the way the points in 5.12 are displayed seems like the beginning of a nicely shaped ellipse, which is the figure

we are looking for. Another sense check we can do is to compute the means of both samples $X_1^{(i)}$ and $X_2^{(i)}$ and compare them with the theoretical distributions we defined before exploiting the Gibbs sampler, so:

$$\bar{X_1} = 1.611, \quad \bar{X_2} = 2.652 \tag{5.46}$$

If you recall, the means for the theoretical distributions were 2 and 4, and they align with the sample means represented in 5.46, suggesting that with continued sampling, we should expect to approximate these vector mean values even more closely. However, even if this preliminary comparison indicates that the technique is functioning as intended, assessing convergence and the overall success of this method involves additional criteria

To truly explore the convergence of the Gibbs sampler, let's return to our main task: identifying a blind spot in the surveillance system. This robust framework allows us to sample while considering dependencies, drawing observations from the four camera angles at any time. The setup requires an initial point and a conditional likelihood. When selecting this likelihood, we need to consider how well it reflects the behavior of our data and whether we can perform calculations with it efficiently. Now, which distribution do we select? There are so many to pick from. Firstly, let's decide on the nature of our variable, the camera angles. Is it discrete or continuous? In this instance, the continuous case is more adequate as we are dealing with a camera motion and an angle. Our pool of distributions to select from just got reduced.

Considering all the continuous distributions we know: Beta, Normal, and Gamma, which will suit us best? The Gamma will not serve us well here; this is due to the skewness present in their curve. Picking up a distribution with such a feature can create bias toward a set of extreme angles and lead us to some biased conclusions on the blind spots. If skewness is a problem, do we need a symmetrical distribution, such as the normal? It might be our best option, as the tails representing the blind spot will be very similar.

Because the conjugated prior of the normal distribution is a normal distribution that results in our posterior, the distribution we are after is also normal. This simplifies our calculation, but are we over-

simplifying our task with this selection? We also know how to define a PDF for a multivariate Gauss, so why bother with a sampler? Another thing to consider is whether it makes sense to have a model such as the Gaussian one to describe the behavior of surveillance cameras. The answer is yes and yes! Let me justify this.

Granted, we just have four dimensions, so in this particular situation, it might seem like overkill to use the sampler, but the advantage is that we can do some hand calculations to deeply understand the method. If you think about the PDF of the multivariate Gaussian distribution, it is an inverse of a covariance matrix. Well, if we are working in a space with a decent number of dimensions, inverting that bad boy is a complicated and expensive task.

Moreover, Gibbs sampling is beneficial for iteratively refining estimates and exploring the parameter space more thoroughly. In scenarios where parameter estimates must be updated frequently, or the model needs to adapt dynamically to new information, Gibbs sampling provides a robust mechanism for convergence monitoring. Each iteration offers a snapshot of the system's state, allowing for continuous assessment and adjustment.

Now, about our specific problem. Assuming a joint Gaussian distribution for the angles of surveillance cameras suggests that these angles typically cluster around the mean value. This mean represents the usual or most probable orientation of the cameras, which is practical because, in a controlled environment like a surveillance setup, cameras are generally directed towards specific areas of interest and only deviate slightly from these positions under normal conditions. This assumption of slight deviation is reasonable as camera movements are typically subtle, resulting from minor adjustments or environmental factors.

Furthermore, the utility of a joint Gaussian distribution extends to capturing the relationship between the movements of different cameras through its covariance structure. This is particularly important in a surveillance context where cameras may need coordination to cover the maximum area effectively. The covariance in the Gaussian distribution quantifies how changes in the angle of one camera might be related to changes in another, helping in the

strategic placement and synchronization of camera movements to ensure optimal coverage and minimal blind spots. Therefore, this modeling approach not only simplifies the mathematical handling of the system's dynamics, but also enhances operational efficiency by providing insights into how different units influence each other.

Right, let's get this show started... oh I just remembered something: we have formulas to compute the parameters of a conditional Gaussian, but just for two dimensions ... AHHHHHH, there is always something else to do! Less complaining and more writing:

$$\mu_{X_1|X_2} = \mu_{X_1} + \frac{\sigma_{X_1X_2}}{\sigma^2_{X_2}}(x_2 - \mu_{X_2}) \tag{5.47}$$

$$\sigma^2_{X_1|X_2} = \frac{\sigma^2_{X_1}\sigma^2_{X_2} - \sigma^2_{X_1X_2}}{\sigma^2_{X_2}} \tag{5.48}$$

These formulae 5.47 and 5.48 were the ones arrived at when we just had X_1 and X_2. Now we have four cameras C_1, C_2, C_3 and C_4, so we go from:

$$X_1|X_2$$

To:

$$C_1|(C_2, C_3, C_4)$$

Hmmm, I have a cheeky idea... what about if we cooked chicken in Coca Cola? That is actually a real recipe and trust me, it is tasty! But the real idea is, what if we define this cat as:

$$C_B = \begin{pmatrix} C_2 \\ C_3 \\ C_4 \end{pmatrix}$$

With that we can use our equations 5.47 and 5.48; we just need to defined them as:

$$\mu_{C_1|C_B} \quad \sigma^2_{C_1|C_B}$$

OK, because we define C_B as a vector, we have no escape from linear algebra, so let's replace the terms in the formulae for the mean and variance with the corresponding algebraic structures (what do you think about that vernacular? Spectacular, no? In reality, these

structures are simply vectors and matrices, I just felt like showing off:

$$\mu_{C_1|C_B} = \mu_{C_1} + \Sigma_{C_1C_B}\Sigma_{C_B}^{-1}\left(C_B - \vec{\mu}_{C_B}\right) \tag{5.49}$$

In transitioning from 5.47 to 5.49, we shift our focus from a single variable, X_B, to a vector of variables, C_B. Consequently, the scalar variances (σ) are replaced by the covariance matrix (Σ), and the mean μ_{X_2} is expanded to a vector of means μ_{X_B}. In linear algebra, direct division by matrices is not defined, which necessitates the use of Σ^{-1}, the inverse of the covariance matrix. This operation is analogous to division in scalar arithmetic and allows for the proper handling of linear transformations involving multiple variables. By incorporating these adjustments, we establish a robust framework for calculating conditional distributions, enhancing our capacity to analyze multiple variables concurrently. We apply a similar approach to 5.48, adapting our methods to accommodate the complexities introduced by vector-based operations:

$$\sigma^2_{C_1|C_B} = \sigma^2_{C_1} - \Sigma_{C_1C_B}\Sigma^{-1}_{C_B}\Sigma_{C_BC_1} \tag{5.50}$$

The intuition to get from 5.48 to 5.50 is the same as the one we needed to go from 5.47 to 5.49. But there are new players whose shapes and forms we need to understand; these are the fellows $\Sigma_{C_1C_B}$ and $\Sigma_{C_BC_1}$. For that, I propose the path of the blessing of ignorance what I mean by this is, let's pretend C_1 and C_2 are just two random variables; in this case, C_2 tricked us into thinking it was alone, and the bastard sneaked three elements in instead of just himself, but so far we are still not aware of them. Say that we now want to define a multivariate normal distribution with C_1 and C_2; for that, we need:

$$\vec{\mu}, \quad \Sigma \rightarrow \vec{\mu} = \begin{pmatrix} \mu_{C_1} \\ \mu_{C_B} \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma^2_{11} & \sigma_{12} \\ \sigma_{21} & \sigma^2_{22} \end{pmatrix}$$

Wait what? C_B now you tell me that you are three values instead of one, and all of you want to eat chocolate mousse?!?!? Come on, man! FINE! We will accommodate all of you:

$$\vec{\mu} = \begin{pmatrix} \mu_{C_1} \\ \vec{\mu}_{C_B} \end{pmatrix} \rightarrow \begin{pmatrix} \mu_{C_1} \\ \mu_{C_2} \\ \mu_{C_3} \\ \mu_{C_4} \end{pmatrix}$$

Actually it was not that bad to have all the C guys in; we just needed a bigger vector. Now for the covariance:

$$\Sigma = \begin{pmatrix} \sigma_{C_1C_1} & \Sigma_{C_BC_1} \\ \Sigma_{C_1C_B} & \Sigma_{C_BC_B} \end{pmatrix} \rightarrow \begin{pmatrix} \sigma_{11}^2 & \begin{pmatrix} \sigma_{12} & \sigma_{13} & \sigma_{14} \end{pmatrix} \\ \begin{pmatrix} \sigma_{21} \\ \sigma_{31} \\ \sigma_{41} \end{pmatrix} & \begin{pmatrix} \sigma_{22}^2 & \sigma_{23} & \sigma_{24} \\ \sigma_{32} & \sigma_{33}^2 & \sigma_{34} \\ \sigma_{42} & \sigma_{43} & \sigma_{44}^2 \end{pmatrix} \end{pmatrix} \tag{5.51}$$

The partitioned covariance matrix Σ in equation 5.51 is structured to separate the variance and covariance components of multiple variables. The matrix is divided into four blocks. The top left element, σ_{11}^2, represents the variance of the first variable. The top right and bottom left blocks are vectors containing the covariances between the first variable and the remaining variables, denoted by σ_{12}, σ_{13}, and σ_{14} for the top right, and σ_{21}, σ_{31}, and σ_{41} for the bottom left. The bottom right block is a submatrix representing the variances and covariances among the remaining variables, with diagonal elements (σ_{22}^2, σ_{33}^2, σ_{44}^2) indicating their variances, and off-diagonal elements (σ_{23}, σ_{24}, σ_{32}, σ_{34}, σ_{42}, σ_{43}) representing their covariances. This partitioning simplifies the handling of multivariate distributions in computations.

Nice! We have everything we need to apply the Gibbs sampler in our dataset with the camera angles for four distinct cameras. First things first, we need a starting point, so why not use the time 00:00? There is no special reason behind this particular choice; it is any random guess; you can even pick a point that it is not in the table:

$$C^{(0)} = \begin{pmatrix} 160 \\ 150 \\ 140 \\ 45 \end{pmatrix} \tag{5.52}$$

The vector 5.52 represents our initial guess, the starting point for our sampling adventure. We define each of its coordinates with a superscript and an subscript corresponding to the point we are sampling and the camera to which the value belongs. For example, $C_1^{(0)}$ is the value of the first camera in point 0 of our simulation, and in this particular case, it is equal to 160. So, to move forward, we need to calculate the conditional distribution for $C_1^{(1)}$ given $C_2^{(0)} =$

150, $C_3^{(0)} = 140$, and $C_3^{(0)} = 45$. Let's start with the mean vector and the covariance matrix:

$$\vec{\mu} = \begin{pmatrix} 80.95 \\ 66.32 \\ 90.16 \\ 68.95 \end{pmatrix} \tag{5.53}$$

Each row of the vector 5.53 corresponds to the sample mean of a given camera, and they are ordered just like $C^{(0)}$ so the first entry of 5.53 is the sample mean of the first camera and so on. Now, moving to the covariance matrix:

$$\Sigma = \begin{pmatrix} 100 & 80 & 60 & 40 \\ 80 & 100 & 60 & 40 \\ 60 & 60 & 100 & 20 \\ 40 & 40 & 20 & 100 \end{pmatrix}$$

There we have it, our covariance matrix Σ, whose entries we also calculated with the values of our dataset 5.3. We need a few more individuals to have everything necessary to compute both 5.49 and 5.50. These equations will respectively give us the mean and the variance of our normal distribution that defines $C_1^{(0)}$, but for that, we need to compute some other terms, for example, $\Sigma_{X_1 X_B}$. As we are no longer working with X's, let's adapt that formula to our current notation:

$$X_1 \rightarrow C_1^{(0)}$$

Previously, X_1 was the value whose conditional normal distribution we wanted to define; however, now we want to sample a value from the first camera, and that is defined by $C_1^{(0)}$. In regards to X_B, this is a vector that has for entries for all the other points so that we can define a C_B instead:

$$C_B = \begin{pmatrix} C_2^{(0)} \\ C_3^{(0)} \\ C_4^{(0)} \end{pmatrix} = \begin{pmatrix} 150 \\ 140 \\ 45 \end{pmatrix}$$

Nice! Now we just need to follow the structure of 5.51 to get $\Sigma_{C_1^{(1)} C_B}$:

$$\Sigma_{C_1^{(1)} C_B} = \begin{pmatrix} 80 & 60 & 40 \end{pmatrix}$$

Instead of Σ_{X_B} we have Σ_{C_B}; to define it we will again use the structure of 5.51:

$$\Sigma_{C_B} = \begin{pmatrix} 100 & 60 & 40 \\ 60 & 100 & 20 \\ 40 & 20 & 100 \end{pmatrix}$$

Let's keep it going! The next one is $X_B - \vec{\mu_{X_B}}$, which in our current universe is $C_B - \vec{\mu_{C_B}}$:

$$C_B - \vec{\mu_{C_B}} = \begin{pmatrix} 150 \\ 140 \\ 45 \end{pmatrix} - \begin{pmatrix} 66.32 \\ 90.16 \\ 68.95 \end{pmatrix} = \begin{pmatrix} 83.68 \\ 49.84 \\ -23.95 \end{pmatrix}$$

Since we are on a roll, what about if we invert something? For example, this guy:

$$\Sigma_{C_B}^{-1} = \begin{pmatrix} 0.018 & -0.010 & -0.005 \\ -0.010 & 0.016 & 0.001 \\ -0.005 & 0.001 & 0.012 \end{pmatrix}$$

And just like that, we are ready to do some calculations:

$$\begin{aligned} \Sigma_{C_1^1 C_B} \Sigma_{C_B}^{-1} &= \begin{pmatrix} 80 & 60 & 40 \end{pmatrix} \begin{pmatrix} 0.018 & -0.010 & -0.005 \\ -0.010 & 0.016 & 0.001 \\ -0.005 & 0.001 & 0.012 \end{pmatrix} \\ &\approx \begin{pmatrix} 0.64 & 0.2 & 0.14 \end{pmatrix} \end{aligned}$$

For our mean:

$$\begin{aligned} \mu_{C_1^{(1)}|C_B} &= 80.95 + \begin{pmatrix} 0.64 & 0.2 & 0.14 \end{pmatrix} \begin{pmatrix} 83.68 \\ 49.84 \\ -23.95 \end{pmatrix} \\ &\approx 141 \end{aligned}$$

Next, we have our variance:

$$\begin{aligned} \sigma^2_{C_1^{(1)}|C_B} &= 100 - \begin{pmatrix} 80 & 60 & 40 \end{pmatrix} \begin{pmatrix} 0.018 & -0.010 & -0.005 \\ -0.010 & 0.016 & 0.001 \\ -0.005 & 0.001 & 0.012 \end{pmatrix} \begin{pmatrix} 80 \\ 60 \\ 40 \end{pmatrix} \\ &\approx 31.20 \end{aligned}$$

Good work, now we have all the parameters to defined our conditional normal distribution for $C_1^{(1)}$:

$$C_1^{(1)} \sim \mathcal{N}(141, 31.20) \tag{5.54}$$

The only thing stopping us getting our first value for our camera angle is to draw a random number from 5.54 and that will be done by the computer. I have got news, the number has just arrived:

$$C_1^{(1)} = 135.88$$

Next we need to sample a value $C_2^{(1)}$; the process is very similar to what we just did to obtain $C_1^{(1)}$, the only difference is on how we set up C_B:

$$C_B = \begin{pmatrix} C_1^{(1)} \\ C_3^{(0)} \\ C_4^{(0)} \end{pmatrix} = \begin{pmatrix} 135.88 \\ 140 \\ 45 \end{pmatrix} \tag{5.55}$$

The new vector C_B contains the updated value of our new angle for camera 1. From this point one, the calculations are very similiar to the ones shown above, therefore they will be omitted and I will present the final values for the mean and variance:

$$\mu_{C_2^{(1)}|C_B} = 105.97, \quad \sigma^2_{C_2^{(1)}|C_B} = 466.88$$

Which means that:

$$C_2^{(1)} \sim \mathcal{N}(105.97, 466.88)$$

Another sample coming our way:

$$C_2^{(1)} = 115.6$$

For the next two points we do the same thing, update C_B with the new point sampled from the iteration before and compute the correspondent mean and standard deviation meaning that for $C_3^{(1)}$ and $C_4^{(1)}$ we have that:

$$C_B = \begin{pmatrix} C_1^{(1)} \\ C_2^{(1)} \\ C_4^{(0)} \end{pmatrix} = \begin{pmatrix} 135.88 \\ 115.6 \\ 45 \end{pmatrix}$$

$$\mu_{C_3^{(1)}|C_B} = 118.20, \quad \sigma^2_{C_3^{(1)}|C_B} = 2681.74$$
$$C_3^{(1)} \sim \mathcal{N}(118.20, 2681.74) \rightarrow C_3^{(1)} = 127.25$$

$$C_B = \begin{pmatrix} C_1^1 \\ C_2^1 \\ C_3^1 \end{pmatrix} = \begin{pmatrix} 135.88 \\ 115.6 \\ 127.25 \end{pmatrix}$$
$$\mu_{C_4^{(1)}|C_B} = 43.72, \quad \sigma^2_{C_4^{(1)}|C_B} = 544.67$$
$$C_4^{(1)} \sim \mathcal{N}(43.72, 544.67) \rightarrow C_4^{(1)} = 52.13$$

And there it is, our new freshly sampled $C^{(1)}$!

$$C^{(1)} = (135.88, 115.6, 127.25, 52.13)$$

The process to sample another point $C^{(2)}$ is exactly the same as what we did to get $C^{(1)}$, but instead of starting from $C^{(0)}$, we start from $C^{(1)}$. For our analysis, we need to draw more points to understand our joint distribution $P(C_1, C_2, C_3, C_4)$, aiming to reflect the dependencies and probabilities of the camera angles. To do so, we will use computer software and generate 10000 samples of 4-dimensional points. The code for the method will be available, and you can play around with it.

With this simulation, we will focus on two things: first, we need to learn how to verify it the method converged. Secondly, we will use our posterior distribution to check our chances of finding a blind spot. Let's tackle the convergence of the method first, because if this is not verified, we can't use our sample to make any inferences. Practically, we achieve convergence when:

- **Trace Plots**: The sampled values of each variable (e.g., camera angles) should display a random scattering around a constant value without drifting away or showing patterns of increase or decrease over iterations.
- **Autocorrelation**: Autocorrelation, in the context of Gibbs sampling, refers to the correlation of a sample with itself over successive iterations or time lags. High autocorrelation at

lower lags indicates that consecutive samples are similar, meaning the sampler has not moved far in the sample space. This can lead to inefficiencies, as it takes longer for the sampler to explore the entire distribution. However, minimal autocorrelation at higher lags suggests that successive samples become more independent of each other as the chain progresses. This independence is crucial for the convergence of the Gibbs sampler, as it indicates that the sampler is effectively exploring the target distribution and providing a representative sample. Reducing autocorrelation helps ensure that the sampler does not get "stuck" in one region of the distribution, thereby improving the quality and reliability of the estimates generated. In essence, low autocorrelation in the Gibbs sampling process is a sign of good mixing and efficient exploration of the parameter space, leading to faster and more accurate convergence to the target distribution.

I generated 10000 samples of points with four angle values, one per camera. When I generated the samples, I also created some trace plots:

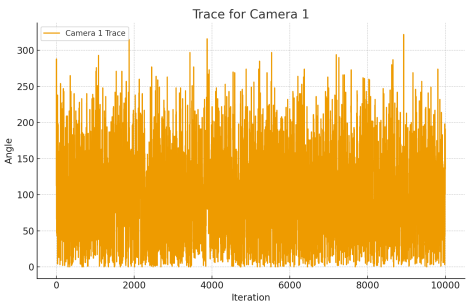


Figure 5.13: Trace for Camera 1.

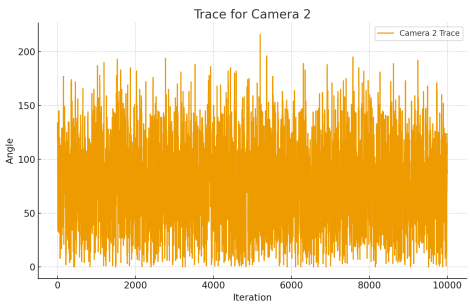


Figure 5.14: Trace for Camera 2.

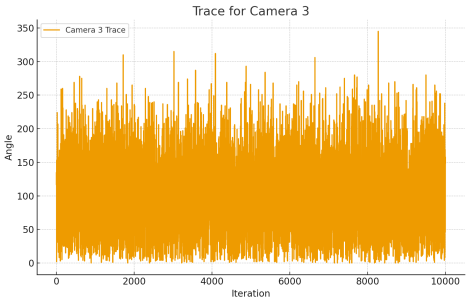


Figure 5.15: Trace for Camera 3.

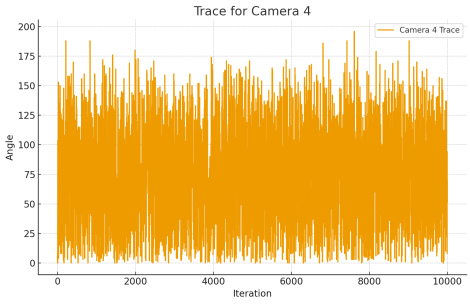


Figure 5.16: Trace for Camera 4.

To create the trace graphs above, we must plot all the individual sampled angle values per camera, showcasing the evolution of sampled values for each camera's angle across 10000 iterations of Gibbs sampling. These plots are essential for understanding how the algorithm samples from the conditional distributions and adjusts with each iteration. As you look at the plots, you'll notice that the values for each camera begin to fluctuate less and less, eventually stabilizing around a particular range. This behavior suggests that Gibbs sampling effectively finds a consistent and stable average for the angles, which indicates the algorithm's convergence.

One thing to mention is that to ensure the practicality of our model, we applied constraints to the sampled angles, restricting them to realistic values between 0 and 360 degrees. In simple terms, every time we sampled a number outside that range, we ignored it and sampled again. This truncation was necessary because mathematical sampling might occasionally propose unrealistic values outside this range. By truncating the data, we guarantee that all sampled values fall within a physically feasible range, maintaining the integrity and applicability of the model to real-world scenarios where camera angles must reflect possible physical positions.

So, we did not create an equation for a joint distribution, but we have 10000 points of this bad boy, and with that, we can do some inference. But how can we use this data to help us bypass the cameras?

We know that the two corridors form an L-shape and that we can observe the first two cameras, but knowing the positions of cameras three and four installed on the second corridor will be very tricky, if not impossible.

During his regular visits to the vault, Notabartolo noticed that each safety deposit box had a plastic layer behind the lock instead of metal. This crucial observation led him to devise a specialized tool for breaking through the plastic and prying open the metal door. He and his crew later created this tool in Turin and successfully used it during the heist.

Inspired by Notabartolo's ingenuity, I designed a compact periscope

capable of measuring the angles of security cameras. We can utilize this device to analyze the positions of the first two cameras. If these angles create a blind spot, we can then use our data to estimate the likelihood that cameras 3 and 4 will be positioned similarly, potentially allowing us to evade detection completely along the entire corridor.

In essence, we need the probability of cameras 3 and 4 allowing a blind spot, given what we know about cameras 1 and 2 are. Firstly, let's define this blind spot:

$$\theta < 35^\circ \quad \text{or} \quad \theta > 115^\circ$$

We will consider a camera unable to detect us if the angle θ sits at values lower than 35 degrees or higher than 115. Cool, let's create a function I that represents this:

$$I(x) = \begin{cases} 1 & \text{if } x < 35^\circ \text{ or } x > 115^\circ \\ 0 & \text{otherwise} \end{cases} \tag{5.56}$$

Equation 5.56 defines our binary function I, which attributes a value of 1 if the camera angle is inside our blind spot range or zero otherwise. This same function will help us calculate the probability of cameras 3 and 4 being in blind spot, given the other two cameras also are:

$$P(I(C_3) = 1 \cap I(C_4) = 1 \mid I(C_1) = 1 \cap I(C_2) = 1) \tag{5.57}$$

The function I will return a 0 if we are not in the angle range defined as blind spot or 1 otherwise. This makes 5.57 the likelihood that we are looking for and can further expand it:

$$\frac{P(I(C_1) = 1 \cap I(C_2) = 1 \cap I(C_3) = 1 \cap I(C_4) = 1)}{P(I(C_1) = 1 \cap I(C_2) = 1)}$$

Well, now we need to count numbers; it is that simple. Suppose we apply that I function to our 10000 points. What will come of this is a new dataset where instead of camera angles, we have a collection of zeros and ones, for example:

C_1	C_2	C_3	C_4
160	150	140	25
150	140	130	60

$\xrightarrow{I(.)}$

$I(C_1)$	$I(C_2)$	$I(C_3)$	$I(C_4)$
1	1	1	1
1	1	1	0

Table 5.3: From angles to blind spot with I.

Diagram 5.3 showcases how the function I is applied to our dataset. On the left hand side of the diagram, a table lists two data points, each comprising four measurements from cameras, indicating different camera angles. The right hand side of the diagram presents these same data points after transformation by our blind spot function I, which is essential for our subsequent analysis.

To compute the likelihood that cameras 3 and 4 are in a blind spot, we specifically consider cases where both cameras 1 and 2 are confirmed to have blind spots. We start by extracting all such instances from our dataset. Then, we focus on these selected instances to determine how often cameras 3 and 4 also have blind spots. The probability is calculated by dividing the number of instances where all four cameras have blind spots by the total instances where cameras 1 and 2 both have blind spots. After performing this calculation, it turns out the odds are comparable to those of a coin flip:

$$P(I(C_3) = 1 \cap I(C_4) = 1 | I(C_1) = 1 \cap I(C_2) = 1) \approx 50\%$$

Firstly, we needed the word coin somewhere in a book about statistics. Secondly, we have a 50% chance that if cameras one and two are in a blind spot, the remaining cameras will also be, giving us a 10-minute interval to run from the door into the safe door. Well good luck us!

An application of the Gibbs sample to the world of machine learning can be seen in Natural Language Processing (NLP). In NLP, it is often necessary to group documents by topic to enhance information retrieval, summarization, and understanding of large bodies of text. For example, a digital library might need to classify thousands of articles into thematic categories such as “technology,” “healthcare,” or “politics” to improve search functionality and user navigation.

We can use the Latent Dirichlet Allocation (LDA), a probabilistic model that assumes documents are mixtures of various topics, and

characterizes each topic through a distribution over words. This model is particularly adept for topic modeling because it accommodates the intuitive notion that documents cover multiple topics to varying degrees. For instance, a news article about health technology might predominantly cover "technology" and include aspects of "healthcare".

Gibbs sampling fits naturally into the framework of LDA due to its ability to handle the complex joint distributions involved. Since direct sampling from these distributions is not feasible, Gibbs sampling allows us to efficiently sample from the conditional distributions of each variable iteratively.

With an introduction to Markov Chain Monte Carlo methods, we now have a powerful numerical integration technique at our disposal, especially useful when dealing with posteriors that are unknown or difficult to compute. This method serves as a key, bringing us to the door of a vault filled with valuable insights. The final step is to unlock that door and access the diamonds within. We're on the brink of our plan's grand finale, with only the main vault standing between us and the glittering treasures that await. However, we have a gap in our expertise that cannot be ignored. We can't crack that safe, and we need to recruit one more team member to help with our final push. Below, there is a diagram of the 2003 vault and its security features:

1 – Combination dial (0-99), 2 – Keyed lock, 3 – Seismic sensor (built-in)

4 – Locked steel grate, 5 – Magnetic sensor, 6 – External security camera

7–Keypad for disarming sensors, 8–Light sensor, 9–Internal security camera

10 – Heat/motion sensor (approximate location)

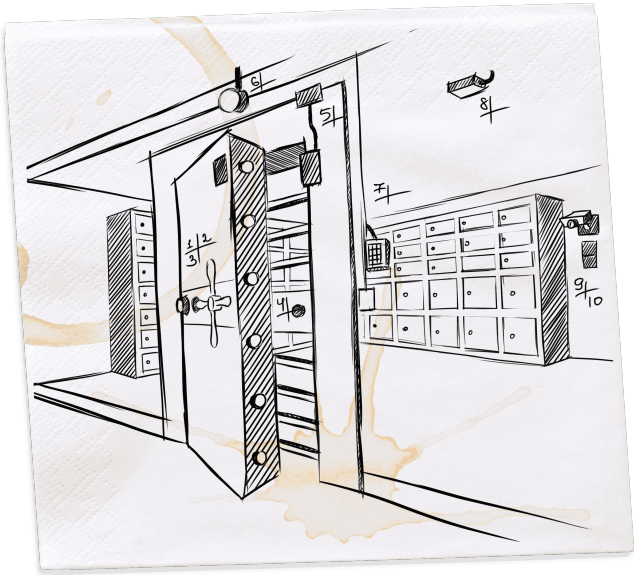


Figure 5.17: A possible configuration of the vault.

The robbers behind the Antwerp Diamond Heist demonstrated meticulous planning and execution to bypass these formidable security measures. They began by constructing a detailed replica of the vault, which allowed them to practice and perfect their safe-cracking techniques without risk. They managed to capture the vault's key and combination by covertly placing a hidden camera across from the vault door during a routine visit, enabling them to replicate the key and memorize the lock combination.

On the night of the heist, they used hairspray to temporarily disable the heat and motion sensors, preventing these systems from detecting their presence. To deceive the magnetic sensors, they crafted a custom-made aluminum plate, allowing them to manipulate the sensors without triggering an alarm. They also covered internal security cameras with black plastic bags to avoid video detection.

In addition, they strategically avoided triggering the light sensor by maintaining minimal light usage as they worked inside the vault. The final step in their elaborate bypass strategy involved manipulating the vault's wiring. They connected bypass devices to the security system's wiring, fooling it into registering normal operational status while they emptied the vault of its contents.

Unfortunately, I don't know if the safe is the same as the one depicted in 5.17. Additionally, our contact inside the safe company

responsible for the installation at the Diamond Center is getting cold feet and is currently unwilling to cooperate. Although I'm confident I can persuade him in time, we must act quickly. Since neither of us is a specialist in safe cracking, we need to recruit additional team members without further delay.

Given our lack of specific information about the exact model installed at the Diamond Center, here's our recruitment strategy: After extensive research, I obtained the blueprints of the top five most commonly installed vaults worldwide. I have created replicas of each safe and invited eight different thieves to test their skills by attempting to crack them open:

Vault ID	Thief	Trials	Success	Success Rate
V001	Alice	20	15	0.75
V001	Bob	20	8	0.40
V001	Charlie	20	12	0.60
V002	Alice	20	10	0.50
V002	Bob	20	13	0.65
V002	Charlie	20	11	0.55
V003	Alice	20	6	0.30
V003	Bob	20	14	0.70
V003	Charlie	20	9	0.45

Table 5.4: A sample of our safe breaking experiment.

Table 5.4 outlines a dataset sample collected during a simulated evaluation of 5 skilled thieves attempting to breach 8 unique vaults. Each thief was allowed a maximum of 8 minutes per attempt, and we were closely monitoring their efficiency and skill under strict constraints. Here is a revised description of the columns in table 5.4:

- **Vault ID**: Identifier for the specific vault targeted, each representing a unique security configuration.
- **Thief**: The name of the thief attempting the breach, useful for assessing individual performance against different vaults.
- **Trials**: The total number of attempts each thief made on a specific vault.
- **Successes**: The number of successful breaches made within the allowed time.

- **Success Rate**: Calculated by dividing the number of successful attempts by the total number of trials, represented as a probability between 0 and 1, indicating the efficiency of the thief.

Expanding our team brings two main risks: lower earnings per person and a higher chance of complications. Given these stakes, it's vital to choose the right candidate for our safe-cracking operation. Ideally, we should add only one person, emphasizing the need for someone with a broad range of skills to streamline our decision-making. Since the safes are fitted with diverse alarm systems and configurations, the performance of each thief varies depending on the model of the safe. The data in table 5.4 is essential as it helps us customize our strategy for effectively cracking any safe we encounter.

Before going deep into equations and probability distributions, let's consider the unique setup of our recruitment task. Since we are unsure of the specific vault currently at the Diamond Center and whether our contact will provide any information, we must be prepared for both scenarios. If we lack specific information, our safest bet is to recruit the thief with the best overall skill. However, if we know the exact vault model, we can select the thief who demonstrated the best performance with that particular vault.

To achieve this, we need to understand the individual skill level of each thief per vault and also establish a metric indicating a global level of aptitude for safe cracking among the thieves. This dual approach will ensure we can adapt our strategy based on the information available, optimizing our chances of success.

It will be a big stretch to think that the ability a given thief has to crack any safe is entirely independent of its global skills. Well, we must capture this relationship, and because we have a set of data, what about if we initiate a model for this new paradigm by defining a likelihood? The Bayes law never failed us and has tremendous flexibility, so this seems like our best shot.

If one of our goals is to understand the success and failures of a given thief when they are operating any vault, we have a beautiful

distribution for these setups:

$$P(y_{ij} \mid p) \sim \text{Binomial}(n_{ij}, p) \tag{5.58}$$

Equation 5.58 models the count of successful attempts by thief i on vault j using a binomial distribution. Here, y_{ij} represents the number of successful cracks by thief i on vault j, and n_{ij} denotes the total number of attempts thief i makes on vault j. The success parameter p is a fixed probability of success that is assumed to be constant across all attempts and all vaults, which simplifies the model but does not account for variability between different thieves or vaults. This assumption can be overly simplistic as it does not reflect the potential differences in skill levels among thieves or the varying difficulties of different vaults. To address this, let's introduce a parameter p_{ij}, which represents the specific probability of success for a given thief on a particular vault:

$$P(y_{ij} \mid p_{ij}) \sim \text{Binomial}(n_{ij}, p_{ij}) \tag{5.59}$$

With equation 5.59, we establish a more robust model as the success of each thief depends on a personalized parameter p_{ij}. This modification allows us to more accurately model the variability in success probabilities among different thieves and vaults. So far, equation 5.59 captures the overall success of each thief cracking a specific vault. However, we still lack a measure of each thief's overall skill, which is crucial because we might be inclined to recruit those who demonstrate the best overall success rate, showcasing superior skills. The proficiency of each thief will impact how well they crack a given safe, or in other words, our parameter p_{ij} is influenced by:

$$P(p_{ij} \mid \theta_i) \tag{5.60}$$

Where θ_i represents the overall success rate per each thief i. Now, we need to define a probability distribution for 5.60. However, if it depends on θ_i, let's first focus on this particular distribution. We defined this parameter to be the rate of success of each thief globally, something that could provide us with a proxy for their skill set. One distribution we study that can be bounded between 0 and 1 is the Beta distribution. Therefore, this is a good candidate for our θ_i distribution:

$$P(\theta_i) \sim \text{Beta}(\alpha, \beta) \tag{5.61}$$

The next step is to find a way to define a distribution for 5.60 that is somewhat related to 5.61. Let's take a step back and look at this from a birds-eye view. In essence, we have a likelihood defined by equation 5.58; this comes from what we observed in our data. We then defined p_{ij}, which is the probability of success of thief i cracking vault j; in essence, that is our prior. So if our likelihood is a binomial, our conjugated prior is a Beta distribution, which is again adequate to module rates of success. Cool, so now we know that:

$$P(p_{ij} \mid \theta_i) \sim \text{Beta}(\alpha_p, \beta_p) \tag{5.62}$$

Now we must find a way to relate 5.60 to 5.61, and the reality is that we don't have much wiggle room here; it has to be through α_p and β_p. Let's think about these two guys in the context of a general Beta distribution. When we studied the impact of the parameters on the shape of this same curve, we concluded that increasing α_p relative to β_p increases the mean of the distribution, shifting it closer to 1, which will indicate a higher probability of success p_{ij}. Conversely, a higher increase of β relative to α shifts the distribution mean closer to 0, indicating a lower probability of success. Another way to verify this is by analyzing the expected value formula:

$$E[X] = \frac{\alpha}{\alpha + \beta} \tag{5.63}$$

Suppose we apply the above mentioned dynamics to α_p and β_p. In that case, we can infer that 5.63 increases in expected values when α_p increases relative to β_p and the opposite when β_p increases relative to α_p. So, if θ_i represents the overall likelihood of success of a thief i and α_p is the parameter that will "control" the success in 5.62, we can define it by θ_i which means that the failures or β_p is equal to $1 - \theta_i$:

$$P(p_{ij} \mid \theta_i) \sim \text{Beta}(\theta_i, 1 - \theta_i) \tag{5.64}$$

When modeling probabilities such as p_{ij} using a Beta distribution parameterized directly as $\text{Beta}(\theta_i, 1 - \theta_i)$, the model might fail to represent the true variability and confidence in these probabilities adequately. This formulation tightly binds the distribution around the mean value θ_i, with very limited variance, implying an unrealistic assumption that we are almost certain about θ_i's value, even before

observing substantial data. This can lead to overfitting, where the model reacts too strongly to small samples or outliers and may not generalize well to new or unseen data. To address these issues, introducing a scaling factor K in the formula, $\text{Beta}(\theta_i \cdot K, (1-\theta_i) \cdot K)$, provides a way to adjust the influence of prior knowledge by controlling the spread of the distribution:

$$P(p_{ij} \mid \theta_i) \sim \text{Beta}(K \cdot \theta_i, K \cdot (1-\theta_i)) \tag{5.65}$$

With this step, we have created a hierarchical framework where we have three distinct levels:

- **Level One - Observed Data (Likelihood y_{ij})**: Is the observed number of successes (e.g., successful vault cracks) for thief i on vault j. It is modeled using a Binomial distribution because y_{ij} represents the count of successes out of a given number of attempts n_{ij}.

  $$P(y_{ij} \mid p_{ij}) \sim \text{Binomial}(n_{ij}, p_{ij})$$

  Where p_{ij} is the probability of success by thief i on vault j.

- **Level Two - Probability of Success p_{ij} (Conditional on θ_i)**: p_{ij} is the vault-specific success probability for each thief, influenced by the thief's overall ability θ_i and potentially by other vault-specific factors. Instead of being directly observed, p_{ij} is estimated and follows a Beta distribution as influenced by θ_i:

  $$P(p_{ij} \mid \theta_i) \sim \text{Beta}(\theta_i \cdot K, (1-\theta_i) \cdot K)$$

- **Level Three - Overall Skill Level of Each Thief θ_i** : θ_i represents the underlying or overall success probability that affects all vault attempts by thief i. This parameter captures the general capability or skill level of the thief.

  $$P(\theta_i) \sim \text{Beta}(\alpha, \beta)$$

We can draw a little diagram to help us understand the hierarchy present in our model:

$$\theta_i \sim \text{Beta}(\alpha, \beta)$$
$$\Big\downarrow \text{depends on}$$
$$p_{ij} \sim \text{Beta}(\theta_i K, (1-\theta_i)K)$$
$$\Big\downarrow \text{informs}$$
$$y_{ij} \sim \text{Binomial}(n_{ij}, p_{ij})$$

So, for each thief-vault combination (i, j), there is a specific probability p_{ij} that represents the thief's chance of successfully cracking that particular vault. This probability depends on θ_i and is modeled as a Beta distribution influenced by the thief's overall skill level, scaled by a factor K to adjust the variance and better capture the confidence in these probabilities. Models defined with a hierarchy such as this are called "Hierarchical Bayesian models".

Chapter 6

Structured Uncertainty: Hierarchical Bayesian Models.

Hierarchical Bayesian models (HBMs) provide a flexible framework for modeling complex data structures with multiple levels of variability. These models are particularly useful for incorporating uncertainty at various levels of the data hierarchy, allowing for more accurate predictions and inferences.

An HBM typically involves multiple layers of parameters, each defined by prior distributions that are conditioned on parameters from higher levels in the hierarchy. This approach allows for the integration of both prior knowledge and observed data:

Model Levels

- **Level 1: Data Layer** - This level involves the observed data, which are often conditioned on parameters from the next level. The data model might be expressed as:

  $$P(y_{ij} \mid \theta_{ij}) \sim F(\theta_{ij}, \phi),$$

  Where y_{ij} represents the observed data, θ_{ij} are the parameters directly affecting the data, and ϕ are known hyperparameters or additional data-specific parameters.

- **Level 2: Parameter Layer** - The parameters that directly affect the data are themselves modeled, typically conditioned on higher-level parameters or hyperparameters:

  $$P(\theta_{ij} \mid \psi_i) \sim G(\psi_i, \tau),$$

The parameter's ψ_i are higher-level parameters or hyperparameters, and τ represents the hyperparameters governing this distribution.

- **Level 3: Hyperparameter Layer** - The higher-level parameters are often modeled using prior distributions, which can be informed by previous studies, expert knowledge, or hierarchical structuring:

$$P(\psi_i) \sim H(\lambda), \tag{6.1}$$

Here λ includes any hyperparameters or fixed parameters needed to define the prior distribution.

HBMs are advantageous in scenarios where data are nested or grouped in complex ways. They allow for:

- Accurate reflection of the uncertainty and variability at each level of the model.

- Sharing of information across groups or levels, which can enhance estimates, especially in data-sparse regions.

- Flexible incorporation of varying levels of prior knowledge through the hierarchical structure.

How are we going to estimate all these parameters? The solution is to use Markov Chain Monte Carlo methods. Our model doesn't rely on a single multivariate distribution but involves numerous parameters that depend on each other. This interdependence creates complexity that traditional statistical methods struggle to handle effectively. MCMC's methods are well-suited for this task as they can sample from complex and high-dimensional posterior distributions. No, we aren't going to use the Gibbs sampler again; instead, why not learn a new technique?

6.1 Don't Celebrate Yet! Metropolis-Hastings Still Has to Accept Us.

The "Metropolis-Hastings" (MH) algorithm, like the Gibbs sampler, is designed to generate a sequence of samples from a probability distribution by constructing a Markov chain with the desired distribution as its equilibrium state. Both methods utilize the properties of Markov chains to ensure that this sequence of samples converges to the target distribution. Sampling in each algorithm is performed iteratively, where parameter estimates are updated based on the current state to generate the next state. This iterative process adheres strictly to the Markov property, meaning that each new sample depends only on the immediate previous sample.

Unlike the Gibbs sampler, which deterministically cycles through each variable to sample from its conditional distribution given all other variables, MH uses a proposal distribution to generate potential future states. This proposal can be any distribution $Q(x' \mid x)$ from which it is easy to sample, where x and x' represent the current and proposed states, respectively. Fancy wording, I know. What this means is that the algorithm will propose a new point drawn from the distribution $Q(x' \mid x)$. Following this, there is an acceptance/rejection criterion applied to this point. It is similar to the questioning we would undergo if we get caught, but the outcome for us is a bit different than for the point. In the case of acceptance, the point becomes a new sample, whereas we will go to jail... So, we have another MCMC that works on samples from a proposal distribution that are either accepted or rejected based on a certain criterion, which means we need to define two things: the proposal distribution $Q(x' \mid x)$ and the acceptance criterion. Let's start with the proposal distribution.

The proposal distribution is a conditional probability distribution from which a new candidate state x' is generated, given the current state x. This distribution is not the joint distribution we are sampling from; instead, it's an auxiliary distribution used to

generate new points. Imagine a rectangle representing the probability space of our target distribution, which contains all possible points. The challenge is that we only know this space partially. By proposing new points that are tested against the acceptance criteria, this distribution is helping us explore this rectangular space and its variance, because accepting a point means we are inside the rectangle, whereas a rejection means otherwise. We can use two types of proposal distributions:

- **Symmetrical around the current state**: Use a symmetric proposal distribution when the target distribution appears balanced around a central point. This type of distribution proposes new states equally in all directions from the current state, making it ideal when the target has no directional bias. It simplifies calculations since the probability of moving from any state x to x' is the same as moving back from x' to x.
- **Asymmetric distributions**: Use an asymmetrical proposal distribution when the target distribution is skewed or has an irregular shape. This allows you to propose moves that are more likely to explore less dense areas or navigate the structure of the distribution more effectively, accommodating its uneven characteristics.

But wait, isn't the target distribution what we are looking for? How can we know if it is symmetrical or not? Indeed, a good question! To determine whether a distribution is symmetrical when its form is unknown, we can employ several practical approaches. For example, we can start with exploratory data analysis using visual tools like histograms to get an intuitive sense of symmetry. Or, since we have computers, we can simply simulate both symmetrical and asymmetrical scenarios and check which one converges more effectively.

Now that we have a mechanism to draw points, we must determine if they are good candidates for our samples, which brings us to our acceptance criteria. Well, the first thing we must consider in order to define the way we say "yes please" or "take a walk" to a point is, in fact, the points: the original x and x'. What measure

can we use to check if a point is probably part of a distribution? What about the density? Density is a measure of the concentration of mass around a point, so, if we find a bigger density around x' than x, this will indicate that the new proposed point is indeed a good candidate. And in order to have densities, we need a posterior:

$$P(x \mid \text{data}) = L(\text{data} \mid x) \cdot P(x) \tag{6.2}$$

Where $P(x \mid \text{data})$ is the posterior, $L(\text{data} \mid x)$ is the likelihood, and $P(x)$ is the prior. The big advantage of the MCMC's is that they save us from doing multiplications such as the one in 6.2 when the probabilities and priors are complex; however, in this particular case, because we are evaluating the posterior "locally," we just need the value of it at x and x'. So, we can compute the probability and prior values for them and then divide the results:

$$\alpha = \frac{P(x' \mid \text{data})}{P(x \mid \text{data})} \tag{6.3}$$

So, if the result of 6.3 is bigger than one, it means that $P(x')$ is higher, and therefore, we observe more density around the state x', making it a good candidate. Conversely, if that density is lower than one, x has more density around it, which is not ideal.

The formula 6.3 alone is inadequate as it ignores the proposal distribution, which is essential for correctly assessing the likelihood of transitioning between states. Without considering the proposal probabilities $Q(x' \mid x)$ and $Q(x \mid x')$, the acceptance ratio, we might introduce biases, favoring certain transitions disproportionately. For instance, if $Q(x' \mid x)$ is high while $Q(x \mid x')$ is low, it implies an easier transition to x' from x than vice versa. Relying solely on $\frac{P(x'|\text{data})}{P(x|\text{data})}$ would result in overrepresenting x' in the sample set because x' would be accepted more frequently than is justified, potentially leading to incorrect sampling and failure of the Markov chain to converge to the target distribution $P(x \mid \text{data})$. To address this issue, we refine our methodology to include the proposal distribution in the acceptance ratio:

$$\alpha = \min\left(\frac{P(x' \mid \text{data}) \cdot Q(x \mid x')}{P(x \mid \text{data}) \cdot Q(x' \mid x)}\right) \tag{6.4}$$

This adjustment guarantees that the acceptance probability always falls within the [0, 1] range, making it a suitable criterion for accepting or rejecting the proposed state. A significant simplification occurs when the proposal distribution Q is symmetrical, so that $Q(x' \mid x) = Q(x \mid x')$. In this scenario, the term $Q(x \mid x')/Q(x' \mid x)$ equals 1, thereby simplifying the acceptance ratio to:

$$\alpha = \min\left(1, \frac{P(x' \mid \text{data})}{P(x \mid \text{data})}\right)$$

This simplification is beneficial because it reduces computational overhead and enhances algorithm efficiency; this is particularly useful when the target distribution is symmetrically centered around the current state, allowing equal exploration in all directions without bias. This symmetrical setting facilitates more straightforward calculations and can expedite the convergence of the Markov chain. Having computed α, we need a basis for comparison to decide whether the proposed point should be retained. A deterministic threshold could introduce bias, disproportionately favoring either high-density or low-density points. Therefore, a static threshold seems unsuitable. So, if deterministic is no bueno, let's go random baby!

The uniform distribution is an excellent choice because it assigns equal probability to all outcomes within its range. By setting its bounds between 0 and 1, we can ensure controlled parameter space exploration. Thus, the proposed state x' is accepted if:

$$u \leq \alpha$$

Where u is a random number sampled from a uniform distribution $U[0, 1]$. This completes the deduction of the Metropolis-Hastings algorithm. Because we have a Bayesian model with quite a few parameters and distributions, we will first apply the MH sampling method to a simple two-dimensional example to solidify the methodology. Then, we will draw from our more complex and hierarchical model.

If we can engage with a simpler example, why not learn a new distribution? I was thinking the same, and there is a useful one for us to learn. A two-in-one! The best deal we can get, except when speaking about kicks in the butt; in that case, one is better than two.

6.1.1 Against the Clock: The Exponential Distribution.

Meet our friend, the exponential distribution, a continuous probability distribution controlled by the elites, oh... that is the government; what controls the exponential is a λ parameter, my bad! This distribution is widely utilized in various fields, including machine learning, where it serves to model the time between events in scenarios where events occur continuously and independently at a constant average rate. This distribution is especially useful in machine learning for survival analysis, which involves modeling time until an event, such as system failure or churn prediction. Additionally, the exponential distribution finds an application in anomaly detection, where the time between rare events can be crucial for identifying unusual patterns. The probability density function of the exponential distribution for a random variable X is given by:

$$f(x;\lambda) = \lambda e^{-\lambda x} \quad \text{for } x \geq 0,$$

Where λ is the rate parameter, representing the reciprocal of the mean. This function describes how likely it is for x, a particular time interval, to occur at a specific rate λ. The cumulative distribution function is:

$$F(x;\lambda) = 1 - e^{-\lambda x} \quad \text{for } x \geq 0.$$

For the expected value of an exponential distribution, we have:

$$E(X) = \frac{1}{\lambda}.$$

Finally the variance of the exponential distribution is:

$$\text{Var}(X) = \frac{1}{\lambda^2}.$$

We will skip all the deductions as they have been done multiple times throughout the book. However, we need to see this distribution's face:

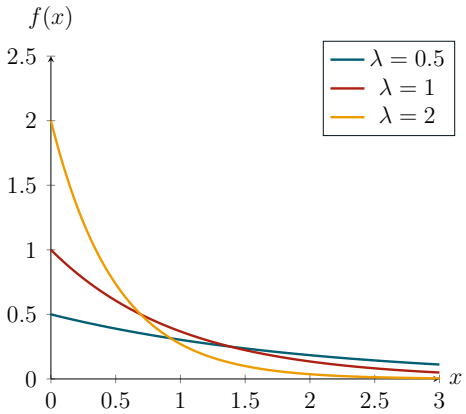


Figure 6.1: The higher the λ, the cooler the slide.

In the exponential distribution, the parameter λ is known as the rate parameter, and it significantly influences the behavior of the distribution. The PDF of the exponential distribution is defined as $f(x;\lambda) = \lambda e^{-\lambda x}$, where $x \geq 0$. From this formulation, several impacts of λ can be observed:

1. Higher λ: As λ increases, the probability density function becomes steeper. This is evident from the plots where the yellow line ($\lambda = 2$) falls off more rapidly than the red ($\lambda = 1$) and blue ($\lambda = 0.5$) lines. A higher λ indicates a higher rate of events per unit time, resulting in shorter intervals on average between events.

2. Shift in Peak: The peak of the distribution (its mode) is always at $x = 0$ for any positive λ, but the height of the peak increases as λ increases. This reflects a higher probability of very short intervals between events when events occur more frequently (higher λ).

3. Decay Rate: The rate at which the probability density decreases as x increases is also controlled by λ. A higher λ leads to a faster decay, meaning the probability of longer intervals occurring becomes significantly lower compared to distributions with a smaller λ.

The rate parameter λ in the exponential distribution must always be positive to ensure that the probability density function (PDF)

$f(x;\lambda) = \lambda e^{-\lambda x}$ for $x \geq 0$ remains a valid probability distribution. A positive λ guarantees that $f(x;\lambda)$ is non-negative across its domain and that its integral from zero to infinity sums to 1, fulfilling the basic criteria for a probability distribution. When $\lambda > 0$, the distribution effectively models real-world stochastic processes where the probability of an event occurring decreases exponentially with time, such as the decay of a radioactive particle or the arrival of a customer. Conversely, if $\lambda \leq 0$, the distribution fails both mathematically and conceptually: $\lambda = 0$ results in a PDF of zero across all x, contributing nothing to the total probability, and $\lambda < 0$ leads to an unbounded exponential increase in the function $e^{-\lambda x}$, violating the criteria that the integral of the PDF must equal 1. This makes no practical sense for describing positive durations or waiting times.

Overall, λ serves as a scaling factor on the x-axis and inversely affects the expected value of the distribution. The mean (or expected value) of the exponential distribution is $\frac{1}{\lambda}$, so a higher rate parameter results in a lower average time between events, illustrating a faster-paced underlying process.

Now that the introductions are out of the way, let's consider a dataset with two variables x_1 and x_2 that we believe are exponentially distributed:

x_1	x_2
0.939	0.021
6.020	3.504
2.633	1.786
1.826	0.239
0.339	0.201
0.339	0.203
0.120	0.363
4.022	0.744
1.838	0.566
2.463	0.344

Table 6.1: Just a sample for two variables each drawn from an exponential distribution.

Say that we believe that dependencies exist between x_1 and x_2 for example, x_1 can be the time until a customer makes their first purchase and x_2 the time until the same customer makes their second purchase. We can see a scenario where x_1 impacts x_2. If a customer takes a long time to make their first purchase, they will either be

more cautious and take longer for subsequent purchases or become more comfortable and make the second purchase more quickly.

To model this, we will assume that both x_1 and x_2 are exponentially distributed, and our goal is to create a joint posterior distribution that captures these dependencies. Let's consider just x_1 and create a model for the parameter λ. We will be pretty much in the same scenario we were in when we introduced Bayes' law and estimated a posterior for λ for the Poisson distribution; however, now, we have two λ's, representing two exponential distributions, which we assumed are linked. We know that multivariate distributions are a pain in the ass, so let's create our equations as a single exponential and then scale them to multi-dimensions, and then outer space. OK, maybe not outer space...

Univariate Likelihood:

Because we assumed that the data is exponentially distributed, we will have the following equation for probability:

$$P(\lambda_1|x_{1i}) = \prod_{i=1}^{n} f(x_{1i}|\lambda_1)$$

$$P(\lambda_1|x_{1i}) = \prod_{i=1}^{n} \lambda_1 e^{-\lambda_1 x_{1i}}$$

$$P(\lambda_1|x_{1i}) = \lambda_1^n \cdot e^{-\lambda_1 \sum_{i=1}^{n} x_{1i}}$$

Where x_{1i} are the observations of our variable x_1. Now for the case where we consider both x_1 and x_2 because we have the individual observations to be independent, the likelihood function for each set of observations can be constructed by multiplying the probability density functions (PDFs) of individual observations:

Multivariate Case:

$$L(\lambda_1|x_{1i}) = \prod_{i=1}^{n} f(x_{1i}|\lambda_1) = \prod_{i=1}^{n} \lambda_1 e^{-\lambda_1 x_{1i}} = \lambda_1^n e^{-\lambda_1 \sum_{i=1}^{n} x_{1i}}$$

$$L(\lambda_2|x_{2i}) = \prod_{i=1}^{n} f(x_{2i}|\lambda_2) = \prod_{i=1}^{n} \lambda_2 e^{-\lambda_2 x_{2i}} = \lambda_2^n e^{-\lambda_2 \sum_{i=1}^{n} x_{2i}}$$

Combining the Likelihood Functions:

$$\begin{aligned} L(\lambda_1, \lambda_2 | x_{1i}, x_{2i}) &= L(\lambda_1 | x_{1i}) \cdot L(\lambda_2 | x_{2i}) \\ &= \lambda_1^n e^{-\lambda_1 \sum_{i=1}^n x_{1i}} \cdot \lambda_2^n e^{-\lambda_2 \sum_{i=1}^n x_{2i}} \end{aligned}$$

Where x_{1i} and x_{2i} are the observations of both x_1 and x_2 respectively. Having defined the likelihoods, we are now ready to move to the priors. Once again, let's start with the univariate case:

Univariate Prior:

$$\lambda_1 \sim \text{Gamma}(\alpha, \beta)$$

Because our likelihood is an exponential, we choose the conjugated prior, which comes as a Gamma function. The next step is the multivariate prior:

Multivariate Prior:

$$\log(\lambda_1), \log(\lambda_2) \sim \mathcal{N}(\mu, \Sigma)$$

$$\mu = \begin{pmatrix} \log(\mu_1) \\ \log(\mu_2) \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2 \\ \sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

Assuming a relationship between λ_1 and λ_2, we chose a multivariate Gaussian distribution as our prior because the covariance matrix can capture the dependencies between the exponential rates. Additionally, we introduced the logarithm transformation. This step is crucial because λ values must be positive, and the normal distribution, defined only by its mean and standard deviation, can produce negative values. By applying the logarithm, we ensure that all values of λ remain positive, thus avoiding any negative rates. We are nearly there, the only thing missing are the so desired posteriors:

Univariate Posterior:

$$\lambda_1 | x_{1i} \sim \text{Gamma}\left(\alpha + n, \beta + \sum_{i=1}^{n} x_{1i}\right)$$

The result of multiplying an exponential function by a gamma is another gamma, it is a direct result of multiplying the likelihood

by the conjugate prior. Moving on to the multivariate case:

Multivariate Posterior:

$$P(\lambda_1, \lambda_2 | x_{1i}, x_{2i}) \propto (\lambda_1 \lambda_2)^{\alpha-1} e^{-\beta(\lambda_1+\lambda_2)} \cdot \lambda_1^n e^{-\lambda_1 \sum_{i=1}^n x_{1i}} \cdot \lambda_2^n e^{-\lambda_2 \sum_{i=1}^n x_{2i}}$$

Since we have a new toy, we should play with it. The time has come for us to apply the Metropolis-Hastings algorithm; for that, we have nearly everything we need, so let's go over what is missing, and then we will proceed with the algorithm and draw one pair of λ's. To start with, we need values to complete our prior, namely a covariance matrix and our vector with the means:

$$\vec{\mu} = (-1, -0.5), \quad \Sigma = \begin{pmatrix} 0.75 & 0.3 \\ 0.3 & 0.5 \end{pmatrix}$$

Those values are our prior knowledge; as we know by now, they can be estimated in various ways, for example, from past experiments. Remember that the exponential distribution uses a negative exponent, such as $e^{-1} \approx 0.3679$, which is indeed positive. Moving on, we need a suitable proposal distribution. I recommend the normal distribution, and here's why: it allows precise control over parameter space exploration due to its definable mean and variance. Moreover, its symmetrical shape simplifies the computation of acceptance ratios, essential for maintaining a detailed balance in the sampling process. Additionally, the normal distribution's tail characteristics facilitate fine local adjustments and, occasionally, more significant leaps, enabling thorough exploration of the posterior landscape across various scenarios. This makes it an optimal choice for our proposal distribution:

$$\mathcal{N}\left(\begin{pmatrix} \log(\lambda_1) \\ \log(\lambda_2) \end{pmatrix}, 0.1 \times I\right) \tag{6.5}$$

Equation 6.5 defines our proposal distribution as a multivariate Gaussian distribution, characterized by a distinctive covariance matrix scaled by 0.1 alongside the identity matrix. This scaling factor effectively controls the step size, thus influencing how broadly we explore the parameter space. The choice of 0.1 for the covariance matrix is crucial; it tempers the extent of distribution spread, ensuring that the exploration is conservative. This strategy avoids huge

steps that could result in frequent rejections, promoting a steadier and more reliable convergence toward the posterior distribution.

Now, with a pair of values for λ_1 and λ_2 we will finally be ready to kick things of:

$$\lambda_1^{(0)} = 1, \lambda_2^{(0)} = 1 \rightarrow \vec{\lambda}^{(0)} = (1, 1)$$

The next step will be to go to our proposal distribution 6.5 and draw a new vector $\vec{\lambda}^{(1)}$:

$$Q(\vec{\lambda}^{(1)} \mid \vec{\lambda}^{(0)}) = \mathcal{N}(\log(\vec{\lambda}^{(0)}), 0.1 \times I)$$

We define our proposal distribution with the letter Q. To sample $\vec{\lambda}^{(1)}$ we centre Q around $\vec{\lambda}^{(0)}$ and explore what is around it:

$$\begin{pmatrix} \log(\lambda_1^{(1)}) \\ \log(\lambda_2^{(1)}) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \log(\lambda_1^{(0)}) \\ \log(\lambda_2^{(0)}) \end{pmatrix}, 0.1 \times I \right)$$

Because $\log(\lambda_1^(0)) = 0$ and $\log(\lambda_2^(0)) = 0$, our proposal distribution becomes:

$$\begin{pmatrix} \log(\lambda_1^{(1)}) \\ \log(\lambda_2^{(1)}) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, 0.1 \times I \right) \tag{6.6}$$

We will ask our friend the computer to provide us with a proposal vector $\vec{\lambda}^{(1)}$ by sampling 6.6:

$$\begin{pmatrix} \log(\lambda_1^{(1)}) \\ \log(\lambda_2^{(1)}) \end{pmatrix} = \begin{pmatrix} -0.3195 \\ -0.1590 \end{pmatrix} \tag{6.7}$$

Now to transform the proposed values back from the log-scale:

$$\lambda_1^{(0)} = e^{\log(\lambda_1^{(0)})}, \quad \lambda_2^{(1)} = e^{\log(\lambda_2^{(1)})}$$

$$\begin{pmatrix} \lambda_1^{(1)} \\ \lambda_2^{(1)} \end{pmatrix} = \begin{pmatrix} 0.7265 \\ 0.8530 \end{pmatrix}$$

Next, we need to compute our accepting criteria α:

$$\alpha = \min\left(1, \frac{P(\vec{\lambda}^{(1)} \mid \text{data})}{P(\vec{\lambda}^{(0)} \mid \text{data})} \cdot \frac{Q(\vec{\lambda}^{(0)} \mid \vec{\lambda}^{(1)})}{Q(\vec{\lambda}^{(1)} \mid \vec{\lambda}^{(0)})}\right) \tag{6.8}$$

Because we know that our posterior has the following expression:

$$P(\lambda \mid \text{data}) \propto L(\text{data} \mid \lambda)P(\lambda) \tag{6.9}$$

If we use the result 6.9 in 6.8

$$\alpha = \min\left(1, \frac{L(\text{data} \mid \vec{\lambda}^{(1)})P(\vec{\lambda}^{(1)})}{L(\text{data} \mid \vec{\lambda}^{(0)})P(\vec{\lambda}^{(0)})} \cdot \frac{Q(\vec{\lambda}^{(0)} \mid \vec{\lambda}^{(1)})}{Q(\vec{\lambda}^{(1)} \mid \vec{\lambda}^{(0)})}\right) \tag{6.10}$$

Because our proposal is a normal distribution and therefore symmetrical:

$$Q(\vec{\lambda}^{(1)} \mid \vec{\lambda}^{(0)}) = Q(\vec{\lambda}^{(0)} \mid \vec{\lambda}^{(1)}) \rightarrow \frac{Q(\vec{\lambda}^{(0)} \mid \vec{\lambda}^{(1)})}{Q(\vec{\lambda}^{(1)} \mid \vec{\lambda}^{(0)})} = 1$$

Meaning that 6.10 simplifies to:

$$\alpha = \min\left(1, \frac{L(\text{data} \mid \vec{\lambda}^{(1)})P(\vec{\lambda}^{(1)})}{L(\text{data} \mid \vec{\lambda}^{(0)})P(\vec{\lambda}^{(0)})}\right) \tag{6.11}$$

Right, 6.11 defines our α, now we just need to input the entries of each of vectors λ's in our posterior... what posterior you say? That one we calculated above:

$$P(\lambda_1, \lambda_2|\{x_{1i}, x_{2i}\}) \propto (\lambda_1\lambda_2)^{\alpha-1} e^{-\beta(\lambda_1+\lambda_2)} \cdot \lambda_1^n e^{-\lambda_1 \sum_{i=1}^n x_{1i}} \cdot \lambda_2^n e^{-\lambda_2 \sum_{i=1}^n x_{2i}}$$

I am going to do these calculations with a computer, but in essence we just need to plug in the values of the λ's from each vector and the entries of our data to get both posteriors:

$$\alpha = \min(1, 1.440) = 1$$

Because our u comes from a uniform distribution between 0 and 1, $U[0, 1]$, we can skip this bit as u will never be higher than one and our acceptance criteria is $u \leq \alpha$. Therefore we accept the point that was proposed to us, and we say thank you very much for this:

$$\vec{\lambda}^{(1)} = (0.726, 0.853)$$

And that is it! Now, if we had rejected the proposed vector $\vec{\lambda}^{(1)}$ we would have to go back to our proposal and get a new pair of λ's,

and this processes would be repeated until some pair gets accepted. To go forward and draw a second sample of λ's, namely $\vec{\lambda}^{(2)}$, one just need to do the exact same thing as we did for $\vec{\lambda}^{(2)}$ but we centrer the distribution around that vector $\vec{\lambda}^{(2)}$.

It is time for us to finalize our plans and decide on how many people we will take on our team to rob the diamonds, which brings us back to our recruitment task. To recruit the best overall thief and the thief or thieves best equipped to crack a specific vault, we developed a framework based on Bayesian statistics, a Bayesian hierarchical model.

Now, we need to estimate all the parameters that make our model, so we decided to deduce and use the MH algorithm. Now, how will we do this? Let's take it step by step. Considering we have eight thieves and five vaults, we'll start with an example involving just two thieves and two vaults. This will help us grasp the underlying mechanics of the model. Once we understand this, we can extend our approach to include all thieves and vaults using a computer. Demonstrating the process with a smaller subset will be beneficial for a thorough understanding, ensuring that we can scale up if necessary, so let's remind ourselves of our model levels:

- **Level One - Observed Data (Likelihood y_{ij})**:

$$P(y_{ij} \mid p_{ij}) \sim \text{Binomial}(n_{ij}, p_{ij})$$

  Where p_{ij} is the probability of success by thief i on vault j.

- **Level Two - Probability of Success p_{ij} (Conditional on θ_i)**:

$$P(p_{ij} \mid \theta_i) \sim \text{Beta}(\theta_i \cdot K, (1 - \theta_i) \cdot K)$$

- **Level Three - Overall Skill Level of Each Thief θ_i** :

$$P(\theta_i) \sim \text{Beta}(\alpha, \beta)$$

We need three things: Cookies, milk, and a blanket... ahhh, just a little flashback! What we do need is a couple of vaults, a couple of thieves, and a strategy for navigating these three hierarchy levels

with the Metropolis-Hastings algorithm. Our assumption is that the overall skill of each thief impacts their success rate at opening a given vault, which logically leads us to begin at level three with our distribution, θ.

Our candidates are Alice and Bob, and we'll be examining their success at vaults V001 and V002. Having established the individuals and the vaults we'll be looking at; we now need to consider their overall success rates. These will form our beliefs that define a prior:

1. **Initialize the overall skill levels $\boldsymbol{\theta_i}$:**

   $$\theta^{(0)}_{\text{Alice}} = 0.7, \quad \theta^{(0)}_{\text{Bob}} = 0.5$$

   These correspond to our prior knowledge; they can be educated guesses, estimated with our dataset; in this care, we averaged out the success rate overall vaults per thief.

2. **Initialize the specific success probabilities $\boldsymbol{p_{ij}}$:**
   First, we need to assume a value for K; let's go with $K = 10$ for moderate confidence. Now we must calculate the parameters for the Beta distribution that defines p_{ij} and characterizes level:

   $$P(p_{ij} \mid \theta_i) \sim \text{Beta}(\theta_i \cdot K, (1 - \theta_i) \cdot K)$$

   So for example, when $i = \text{Alice}$ and $j = \text{V001}$ we have:

   $$p_{\text{Alice, V001}} \sim \text{Beta}(0.7 \cdot 10, (1 - 0.7) \cdot 10)$$

   Doing the calculations for both thieves and vaults:

   $$p^{(0)}_{\text{Alice, V001}} \sim \text{Beta}(7, 3), \quad p^{(0)}_{\text{Alice, V002}} \sim \text{Beta}(7, 3)$$

   $$p^{(0)}_{\text{Bob, V001}} \sim \text{Beta}(5, 5), \quad p^{(0)}_{\text{Bob, V002}} \sim \text{Beta}(5, 5)$$

   If we draw samples for all those Beta's:

   $$p^{(0)}_{\text{Alice, V001}} = 0.628, \quad p^{(0)}_{\text{Alice, V002}} = 0.493$$

   $$p^{(0)}_{\text{Bob, V001}} = 0.134, \quad p^{(0)}_{\text{Bob, V002}} = 0.580$$

   This marks the initialization part of our algorithm. Before, we did it by selecting two λ's; however, as we are now operating in a hierarchical framework and our likelihood y_{ij} depends on two other levels, we have to initiate both of them.

3. **Propose new values**: This step is similar to what we did before when we used a multivariate Gaussian. However, we are not considering any dependencies between thieves or vaults, so we can propose values independently. For this we will create a new value for θ per thief and with that update p_{ij}, creating than p'_{ij}:

$$\theta'_{\text{Alice}} = \theta_{\text{Alice}} + \epsilon \quad \text{Where} \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

This step is a method for exploring the probability distribution starting from the current state. If we are at θ_{Alice} and wish to move to a new state θ'_{Alice}, one approach is to introduce a small perturbation ϵ. To ensure that we do not stray too far from θ_{Alice}, ϵ is defined to follow a normal distribution centered at 0, with a small standard deviation. This approach allows for a controlled exploration of the current state. So, if we want to propose a new θ for both Alex and Bob:

$$\theta'_{\text{Alice}} = \theta_{\text{Alice}} + \epsilon \quad \text{Where} \quad \epsilon \sim \mathcal{N}(0, 0.02)$$

$$\theta'_{\text{Bob}} = \theta_{\text{Bob}} + \epsilon \quad \text{Where} \quad \epsilon \sim \mathcal{N}(0, 0.02)$$

Once we have the proposals for the new θ's, we need to verify that they are positive; otherwise, we have to generate a new ϵ. Don't forget θ represents a rate, therefore it must be between 0 and 1:

$$\epsilon_{\text{Alice}} = 0.015, \quad \epsilon_{\text{Bob}} = -0.010$$

Now let's get our new θ's:

$$\theta'_{\text{Alice}} = \theta_{\text{Alice}} + \epsilon_{\text{Alice}} = 0.7 + 0.015 = 0.715$$

$$\theta'_{\text{Bob}} = \theta_{\text{Bob}} + \epsilon_{\text{Bob}} = 0.5 - 0.010 = 0.490$$

The next step is to update the p_{ij} with the new θ':

$$p'_{\text{Alice, V001}} \sim \text{Beta}(\theta'_{\text{Alice}} \cdot K, (1 - \theta'_{\text{Alice}}) \cdot K)$$

$$p'_{\text{Alice, V002}} \sim \text{Beta}(\theta'_{\text{Alice}} \cdot K, (1 - \theta'_{\text{Alice}}) \cdot K)$$

$$p'_{\text{Bob, V001}} \sim \text{Beta}(\theta'_{\text{Bob}} \cdot K, (1 - \theta'_{\text{Bob}}) \cdot K)$$

$$p'_{\text{Bob, V002}} \sim \text{Beta}(\theta'_{\text{Bob}} \cdot K, (1 - \theta'_{\text{Bob}}) \cdot K)$$

Next, we can use the values of θ_{Alice} and θ_{Bob} and the ϵ's to compute our p_{ij}:

$$p'_{\text{Alice, V001}} \sim \text{Beta}(0.7223{\cdot}10, (1{-}0.7223){\cdot}10) = \text{Beta}(7.223, 2.777)$$

$$p'_{\text{Alice, V002}} \sim \text{Beta}(0.7223{\cdot}10, (1{-}0.7223){\cdot}10) = \text{Beta}(7.223, 2.777)$$

$$p'_{\text{Bob, V001}} \sim \text{Beta}(0.5037{\cdot}10, (1{-}0.5037){\cdot}10) = \text{Beta}(5.037, 4.963)$$

$$p'_{\text{Bob, V002}} \sim \text{Beta}(0.5037{\cdot}10, (1{-}0.5037){\cdot}10) = \text{Beta}(5.037, 4.963)$$

Sample, sample, sample!

$$p'_{\text{Alice, V001}} = 0.787, \quad p'_{\text{Alice, V002}} = 0.750$$

$$p'_{\text{Bob, V001}} = 0.581, \quad p'_{\text{Bob, V002}} = 0.389$$

4. **Calculate Acceptance Ratios $\boldsymbol{\alpha_{\theta_i}, \alpha_{p_{ij}}}$**: After the proposal, we get married... or we accept or reject our new suggested θ_i and $\alpha_{p_{ij}}$. The way we do it is similar to what we did in the example where we first applied the MH algorithm, but this time, we have to compute two α's:

$$\alpha_{\theta_i} = \min\left(1, \frac{L(\text{data} \mid \theta'_i)P(\theta'_i)}{L(\text{data} \mid \theta_i)P(\theta_i)}\right) \tag{6.12}$$

$$\alpha_{p_{ij}} = \min\left(1, \frac{L(\text{data} \mid p'_{ij})P(p'_{ij})}{L(\text{data} \mid p_{ij})P(p_{ij})}\right) \tag{6.13}$$

Because the proposal is based upon a perturbation that follows a normal distribution, we are in a case with symmetry $Q(\theta'_{\text{Alice}} \mid \theta_{\text{Alice}}) = Q(\theta_{\text{Alice}} \mid \theta'_{\text{Alice}})$, and that results in 1. With the expressions to compute the α's, we are multiplying a likelihood by a prior, which is approximately a posterior:

$$P(\theta_i \mid \text{data}) \propto L(\text{data} \mid \theta_i) \cdot P(\theta_i)$$

$$P(p_{ij} \mid \text{data}) \propto L(\text{data} \mid p_{ij}) \cdot P(p_{ij})$$

OK, this is the plan of attack, we flank θ, compute that likelihood and before he turns around we have already done the multiplication

with the prior. Because θ_i affects p_{ij}, $L(\theta_i)$ is calculated through the probability of all parameters associated with p_{ij}:

$$L(\text{data}|\theta_i) = \prod_j \text{Binomial}(y_{ij} \mid n_{ij}, p_{ij}(\theta_i)) \tag{6.14}$$

Now for the prior of θ we said it is a Beta distribution:

$$P(\theta_i) = \text{Beta}(\alpha, \beta) \tag{6.15}$$

By multiplying 6.14 and 6.15 we get the expression of our posterior:

$$P(\theta_i \mid \text{data}) \propto \left(\binom{n_{ij}}{y_{ij}} \theta_i^{y_{ij}} (1-\theta_i)^{n_{ij}-y_{ij}} \right) \cdot \left(\frac{\theta_i^{\alpha-1}(1-\theta_i)^{\beta-1}}{B(\alpha,\beta)} \right)$$

Given the multiplication of powers of θ_i and $(1-\theta_i)$, the posterior can be simplified to a new Beta distribution:

$$P(\theta_i \mid \text{data}) = \text{Beta}(\theta_i; \alpha + y_{ij}, \beta + n_{ij} - y_{ij}) \tag{6.16}$$

We are nearly done with α_{θ_i}, we just need to evaluate 6.16 at our proposed θ' and original θ. We need more information for that: α and β, two individuals we haven't yet defined. However, we will use one of the best techniques in statistics; we will assume them! We know that if $\alpha = \beta = 1$, we have a uniform distribution as we never met these thieves before. We do not know their skill, so this seems a good assumption. Now for y_{ij}, representing the number of successes, n_{ij} the total trials:

$$n_{\text{Alice, V001}} = 20, \quad y_{\text{Alice, V001}} = 15$$

$$n_{\text{Alice, V002}} = 20, \quad y_{\text{Alice, V002}} = 10$$

$$n_{\text{Bob, V001}} = 20, \quad y_{\text{Bob, V001}} = 8$$

$$n_{\text{Bob, V002}} = 20, \quad y_{\text{Bob, V002}} = 10$$

Expanding our formula 6.12 for both our thieves:

$$\alpha_{\theta_{\text{Alice}}} = \min\left(1, \frac{\text{Beta}(\theta'_{\text{Alice}}; \alpha + y_{ij}, \beta + n_{ij} - y_{ij})}{\text{Beta}(\theta_{\text{Alice}}; \alpha + y_{ij}, \beta + n_{ij} - y_{ij})}\right)$$

$$\alpha_{\theta_{\text{Bob}}} = \min\left(1, \frac{\text{Beta}(\theta'_{\text{Bob}}; \alpha + y_{ij}, \beta + n_{ij} - y_{ij})}{\text{Beta}(\theta_{\text{Bob}}; \alpha + y_{ij}, \beta + n_{ij} - y_{ij})}\right)$$

Now we are ready!

$$\alpha_{\theta_{\text{Alice}}} = \min\left(1, \frac{\text{Beta}(0.715, 16 + 11, 6 + 11)}{\text{Beta}(0.7; 16 + 11, 6 + 11)}\right)$$

$$\alpha_{\theta_{\text{Bob}}} = \min\left(1, \frac{\text{Beta}(0.490, 9 + 11, 13 + 11)}{\text{Beta}(0.5; 9 + 11, 13 + 11)}\right)$$

We will perform the divisions of the Betas densities at the specific points with a computer:

$$\alpha_{\theta_{\text{Alice}}} = \min(1, 1.043) = 1$$

$$\alpha_{\theta_{\text{Bob}}} = \min(1, 0.931) = 0.931$$

Next, we need to do something very similar, but for our other parameter $\alpha_{p_{ij}}$. Starting with the likelihood of p_{ij}, and this one directly relates to the observable data:

$$L(\text{data} \mid p_{ij}) = \text{Binomial}(y_{ij} \mid n_{ij}, p_{ij})$$

For the priors π, they are both straightforward and have already been defined:

$$P(p_{ij}) = \text{Beta}(\theta_i \cdot K, (1 - \theta_i) \cdot K)$$

The calculations of the posterior will be very similar to the previous example and will be omitted:

$$P(p_{ij} \mid \text{data}) = \text{Beta}\left(p_{ij}; \theta_i \cdot K + y_{ij}, (1 - \theta_i) \cdot K + (n_{ij} - y_{ij})\right)$$

We will have four different α's:

$$\alpha_{p_{\text{Alice, V001}}} = \min\left(1, \frac{\text{Beta}(p'_{\text{Alice, V001}}; \theta'_{\text{Alice}} \cdot K + 15, (1 - \theta'_{\text{Alice}}) \cdot K + 5)}{\text{Beta}(p_{\text{Alice, V001}}; \theta_{\text{Alice}} \cdot K + 15, (1 - \theta_{\text{Alice}}) \cdot K + 5)}\right)$$

$$\alpha_{p_{\text{Alice, V002}}} = \min\left(1, \frac{\text{Beta}(p'_{\text{Alice, V002}}; \theta'_{\text{Alice}} \cdot K + 10, (1 - \theta'_{\text{Alice}}) \cdot K + 10)}{\text{Beta}(p_{\text{Alice, V002}}; \theta_{\text{Alice}} \cdot K + 10, (1 - \theta_{\text{Alice}}) \cdot K + 10)}\right)$$

$$\alpha_{p_{\text{Bob, V001}}} = \min\left(1, \frac{\text{Beta}(p'_{\text{Bob, V001}}; \theta'_{\text{Bob}} \cdot K + 8, (1 - \theta'_{\text{Bob}}) \cdot K + 12)}{\text{Beta}(p_{\text{Bob, V001}}; \theta_{\text{Bob}} \cdot K + 8, (1 - \theta_{\text{Bob}}) \cdot K + 12)}\right)$$

$$\alpha_{p_{\text{Bob, V002}}} = \min\left(1, \frac{\text{Beta}(p'_{\text{Bob, V002}}; \theta'_{\text{Bob}} \cdot K + 10, (1 - \theta'_{\text{Bob}}) \cdot K + 10)}{\text{Beta}(p_{\text{Bob, V002}}; \theta_{\text{Bob}} \cdot K + 10, (1 - \theta_{\text{Bob}}) \cdot K + 10)}\right)$$

We just expanded the formula 6.13 for each thief and vault. Now we can input the corresponding values and compute those bad boys:

$$\alpha_{p_{\text{Alice, V001}}} = \min\left(1, \frac{\text{Beta}(0.787; 0.715 \cdot 10 + 15, (1 - 0.715) \cdot 10 + 5)}{\text{Beta}(0.628; 0.7 \cdot 10 + 15, (1 - 0.7) \cdot 10 + 5)}\right)$$

$$\alpha_{p_{\text{Alice, V002}}} = \min\left(1, \frac{\text{Beta}(0.750; 0.715 \cdot 10 + 10, (1 - 0.715) \cdot 10 + 10)}{\text{Beta}(0.493; 0.7 \cdot 10 + 10, (1 - 0.7) \cdot 10 + 10)}\right)$$

$$\alpha_{p_{\text{Bob, V001}}} = \min\left(1, \frac{\text{Beta}(0.581; 0.490 \cdot 10 + 8, (1 - 0.581) \cdot 10 + 12)}{\text{Beta}(0.134; 0.5 \cdot 10 + 8, (1 - 0.5) \cdot 10 + 12)}\right)$$

$$\alpha_{p_{\text{Bob, V002}}} = \min\left(1, \frac{\text{Beta}(0.389; 0.490 \cdot 10 + 10, (1 - 0.490) \cdot 10 + 10)}{\text{Beta}(0.389; 0.5 \cdot 10 + 10, (1 - 0.5) \cdot 10 + 10)}\right)$$

Computer please help!

$$\alpha_{p_{\text{Alice, V001}}} = \min(1, 2.392) = 1$$

$$\alpha_{p_{\text{Alice, V002}}} = \min(1, 0.192) = 0.192$$

$$\alpha_{p_{\text{Bob, V001}}} = \min(1, 374.490) = 1$$

$$\alpha_{p_{\text{Bob, V002}}} = \min(1, 1.04) = 1$$

Now that we have all six α's, we need to draw a value u from a normal distribution. We have that $u = 0.05$, and because our acceptance criteria is $u \leq \alpha$, we accept all our proposed new points. Let's organize our initial values and our new sampled ones:

Parameter	Initial Value	Proposed Value	$u \leq \alpha$
θ_{Alice}	0.7	0.715	$0.05 \leq 1 \rightarrow$ Accepted
θ_{Bob}	0.5	0.490	$0.05 \leq 0.931 \rightarrow$ Accepted
$p_{\text{Alice, V001}}$	0.628	0.787	$0.05 \leq 1 \rightarrow$ Accepted
$p_{\text{Alice, V002}}$	0.493	0.750	$0.05 \leq 0.192 \rightarrow$ Accepted
$p_{\text{Bob, V001}}$	0.134	0.581	$0.05 \leq 1 \rightarrow$ Accepted
$p_{\text{Bob, V002}}$	0.580	0.389	$0.05 \leq 1 \rightarrow$ Accepted

Table 6.2: Initial and proposed values for Metropolis-Hastings algorithm.

If we wish to carry on sampling values for both the overall rate of success θ of each thief and their ability to crack individual safes, our p_{ij}, we have to carry on with the MH algorithm, but now from the new proposed values. The process is the same as before; if any proposed value gets rejected, we must keep drawing values until they get accepted. We indeed have more vaults and more thieves; however, if we understand how to sample two thieves and two vaults, the process to add either vaults or thieves is analogous.

After running 1000 samples, I calculated the expected values for the overall skill level θ_i and the success rate of each thief on a given vault p_{ij}. For detailed plots of each distribution and convergence analysis, please refer to the code notebooks accompanying this book. These notebooks included the code, and all relevant distribution plots. Although we focus primarily on the recruitment task here, I encourage you to explore the notebooks for a deeper analysis and additional plots demonstrating the method's convergence. Now, let's proceed with the results and our recruitment task:

Thief	θ
Alice	0.625
Bob	0.508
Charlie	0.412
Dave	0.277
Eve	0.411
Frank	0.412
Grace	0.588
Helen	0.368

Table 6.3: Final θ values for each thief.

In table 6.3 we have the expected value of the rate of success of our θ_i distributions, which represents our overall skill per thief. We can see that Alice is the better all-round thief in terms of skill, with a success rate of around 63%. Next let's check how they performed with each individual safe:

Thief	V001	V002	V003	V004	V005
Alice	0.707	0.540	0.609	0.6413	0.674
Bob	0.434	0.603	0.536	0.501	0.471
Charlie	0.436	0.372	0.470	0.404	0.336
Dave	0.327	0.292	0.261	0.227	0.193
Eve	0.471	0.404	0.437	0.371	0.339
Frank	0.501	0.437	0.402	0.367	0.303
Grace	0.662	0.595	0.628	0.562	0.529
Helen	0.420	0.387	0.354	0.322	0.289

Table 6.4: Final p_{ij} values for each thief and vault

Table 6.4 provides a clear comparison of each thief's success rates across different vault configurations. Here's a succinct analysis based on the presented data:

- **Alice**: Exhibits the highest success rate in vault V001 at 70.7% and consistently performs well across the other vaults with her lowest being 54.0% in V002.

- **Bob**: Shows variable performance with his best in vault V002 at 60.3% and his lowest in vault V005 at 47.1%.
- **Charlie**: Has his highest success rate in vault V003 at 47.0%, with a significant drop to 33.6% in vault V005, indicating a struggle with this configuration.
- **Dave**: Consistently scores lower across all vaults, with a peak of 32.7% in V001 and decreasing to 19.3% in V005.
- **Eve**: Maintains a narrow range of success, peaking at 47.1% in vault V001 and dipping to 33.9% in vault V005.
- **Frank**: Achieves his best rate in vault V001 at 50.1%, with a gradual decline to 30.3% in vault V005.
- **Grace**: Strong performances across the board, peaking at 66.2% in vault V001 and showing her lowest at 52.9% in V005.
- **Helen**: Starts with 42.0% in vault V001 and sees a consistent decrease to 28.9% in vault V005.

If our contact provides information about the specific safe installed at the Diamond Centre, we can select the most skilled thief for that particular safe. However, if we do not receive this information, we will choose Alice, as she demonstrates the highest overall skill level.

Leonardo Notarbartolo and his team when inside the vault, they used a hand-cranked drill to open safe-deposit boxes, gathering gold bars, cash, and diamonds. Their duffel bags quickly filled with loot. By 5:30 am, they had completed the theft and transported the bags up the stairs, out of the building, and into Notarbartolo's car.

After retreating to an apartment, they split up the next day, escaping in two cars along separate routes. Notarbartolo was with Speedy, who panicked and scattered incriminating evidence in a forest instead of destroying it during their return to Italy. Later, a local discovered this evidence, leading the police to trace the heist back to Notarbartolo and his team. Key clues, such as a grocery receipt for salami, tied the thieves to the crime. The authorities arrested

Notarbartolo and his accomplices using phone records and DNA evidence. Despite their meticulous planning, the gang was captured, with the exception of the King of keys, though much of the stolen loot remains unrecovered.

Do you feel nervous knowing they went to jail? Well, you don't have to! We meticulously planned our getaway using measures of central tendency and spread. Deciding on the timing for our entry into the Diamond Centre proved challenging. We employed various techniques, including the probability density functions, the normal distribution, the central limit theorem, confidence intervals, and hypothesis testing. After all, we chose the early morning hours over the weekend, just like the mastermind Notarbartolo did.

Our next challenge was the garage access. We strategically installed a camera to capture the code from a video feed. To determine the best time for camera installation and the daily frequency of key code entries we turned to the binomial and Poisson distributions. Aware of the risk of detection, we sought guidance from Bayes. His theorem was instrumental in updating our beliefs with evidence, a crucial step in handling the video feed and deciphering the numbers.

Mr. Markov then assisted us in creating a sequence for the digits that likely made up a valid code, laying the groundwork for our use of MCMC techniques. The Gibbs sampler revealed blind spots from the door inside the garage to the vault. With everything in place to access the safe, we realized the need for additional expertise and launched a recruitment mission. Utilizing a hierarchical Bayesian model and a new MCMC algorithm, the Metropolis-Hastings, we identified two candidates to join our team and crack open the safe where our gems awaited.

We are now ready, but not to crack a safe and rob diamonds (don't do it; our plan will fail—It was just a theme to make the subject lighter). Instead, we're ready to confidently step into the world of machine learning. In volume 1 and 2 of this series we built a solid foundation in linear algebra and calculus and now, after adding probability and statistics, we're well-equipped to understand and creatively apply a vast array of machine learning algorithms.

This series aimed to demystify mathematics and inspire confidence and motivation to continue learning. More than just importing code libraries and analyzing results, I wanted us to understand the mechanics behind the algorithms. By grasping the mathematics that works "under the hood," we can utilize these algorithms more effectively, adapt and apply them to various applications. So why stop now? After all the hard work of going through plots and equations, it is a waste not to use them. I'll see you on the other side, where we and the machines learn.

Peace

Jorge