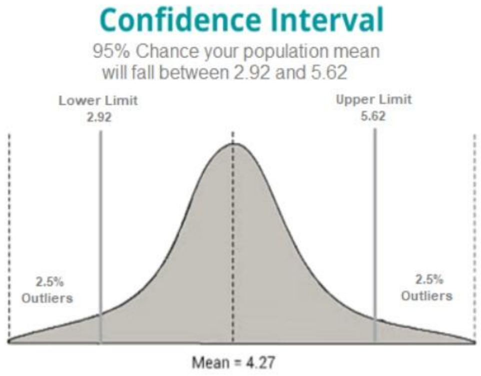
Hypothesis Testing

THE KNOWLEDGE HOUSE
Celebrating 10 Years of Diversifying Tech

Agenda - Schedule

1. Distributions Review & Z-score
2. Introduction to Hypothesis Testing
3. Hypothesis Testing Thought Experiment
4. Break
5. TLAB



Where do we expect most of our data to fall?

Agenda - Announcements

- TLAB #2 due 4/21

Agenda - Goals

- Review the concept of probability distributions
- Understand what a hypothesis test is using a real example
- Identify the steps to hypothesis testing

Distributions Review

Distributions Review

We've spoken about the different **distributions** that you might encounter during your data analysis.

Next week, we will formally begin the process of doing **effective data analysis** using the **pandas** package.

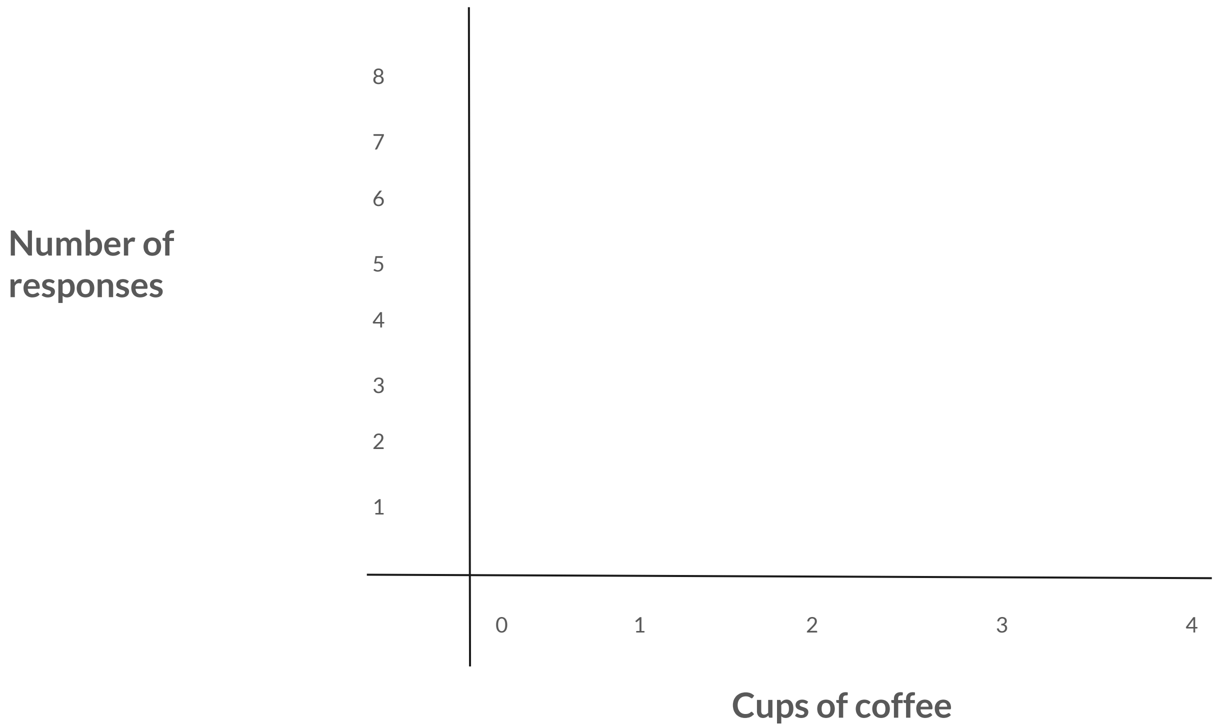
However, before we open this new chapter, let's dive back into the world of distributions and explore **which shapes we might encounter when looking for changes in our dataset.**

Distributions Review

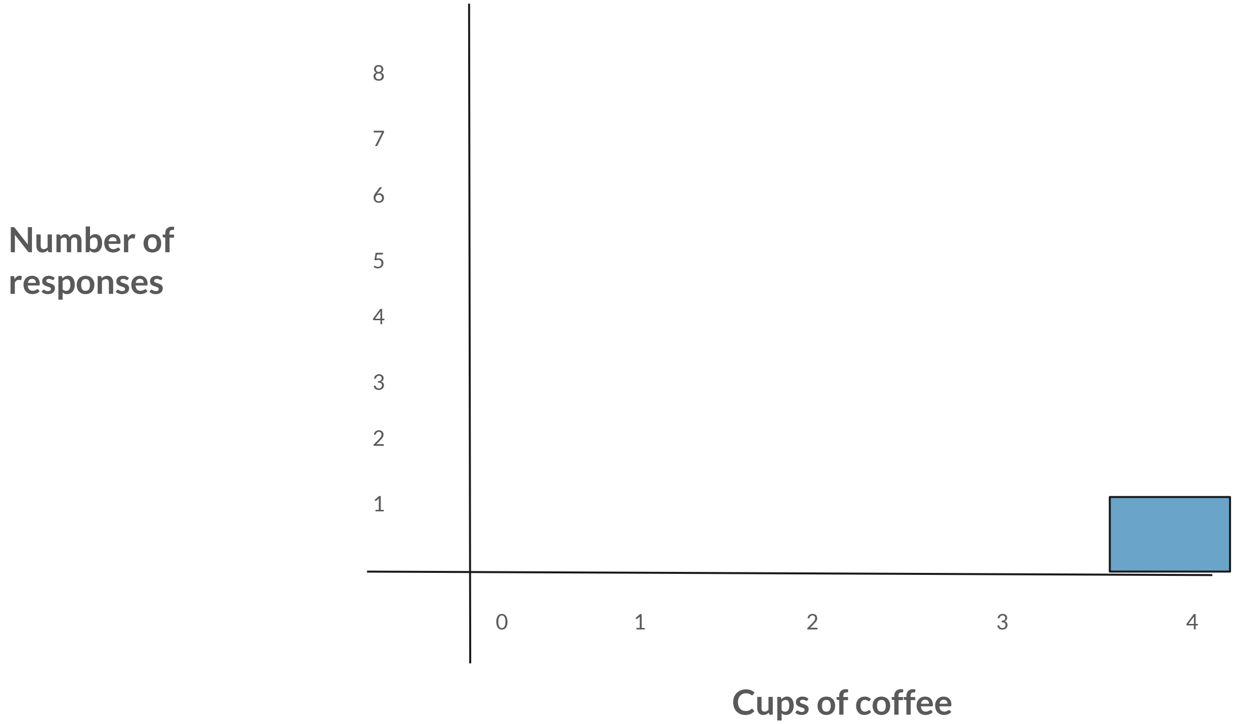
To ensure that we're operating off of the same vocabulary for the next couple of slides, let's review how a distribution visualization is formed once more.

Let's say we are measuring the caffeine consumption of a typical American.

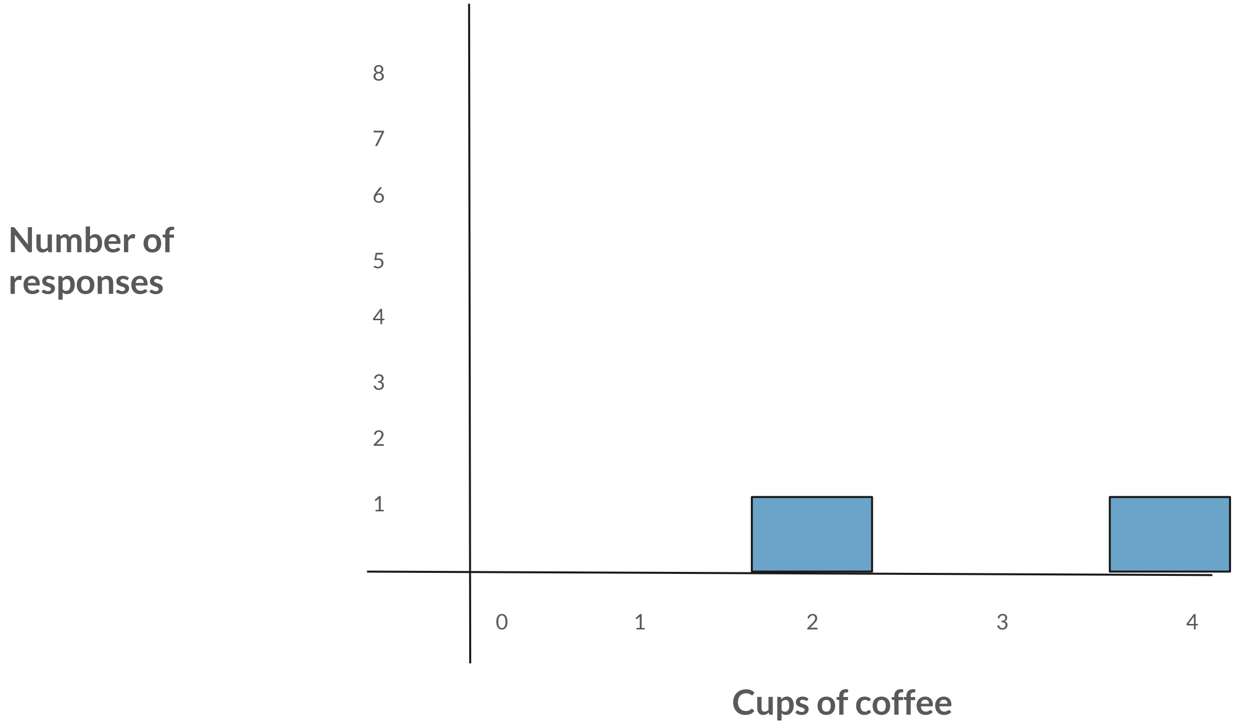
We get the funds to randomly call 30 phone numbers across America (**anyone remember why 30?**).



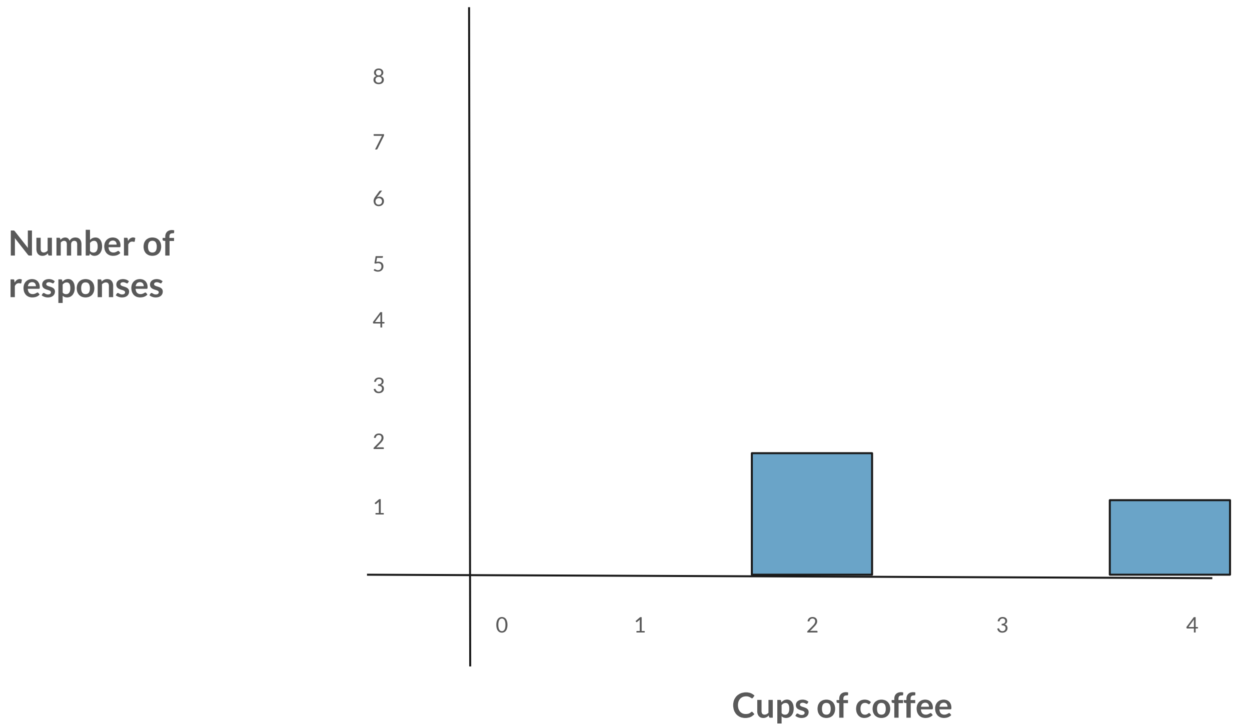
We will slowly build our histogram as we **collect survey responses**.



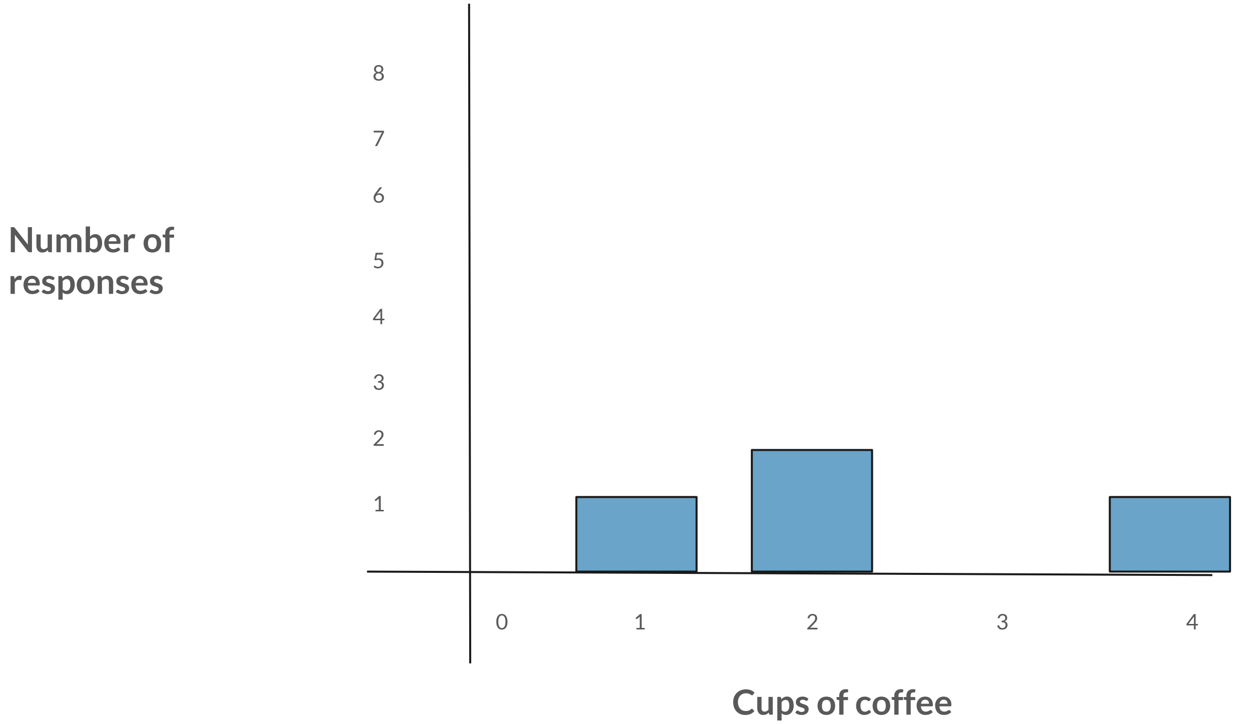
We call the first phone number. A participant named Farukh tells us he drank 4 cups of coffee in one day! We create our first bar for the “4” quantity on our histogram to **indicate 1 person drank 4 cups of coffee**.



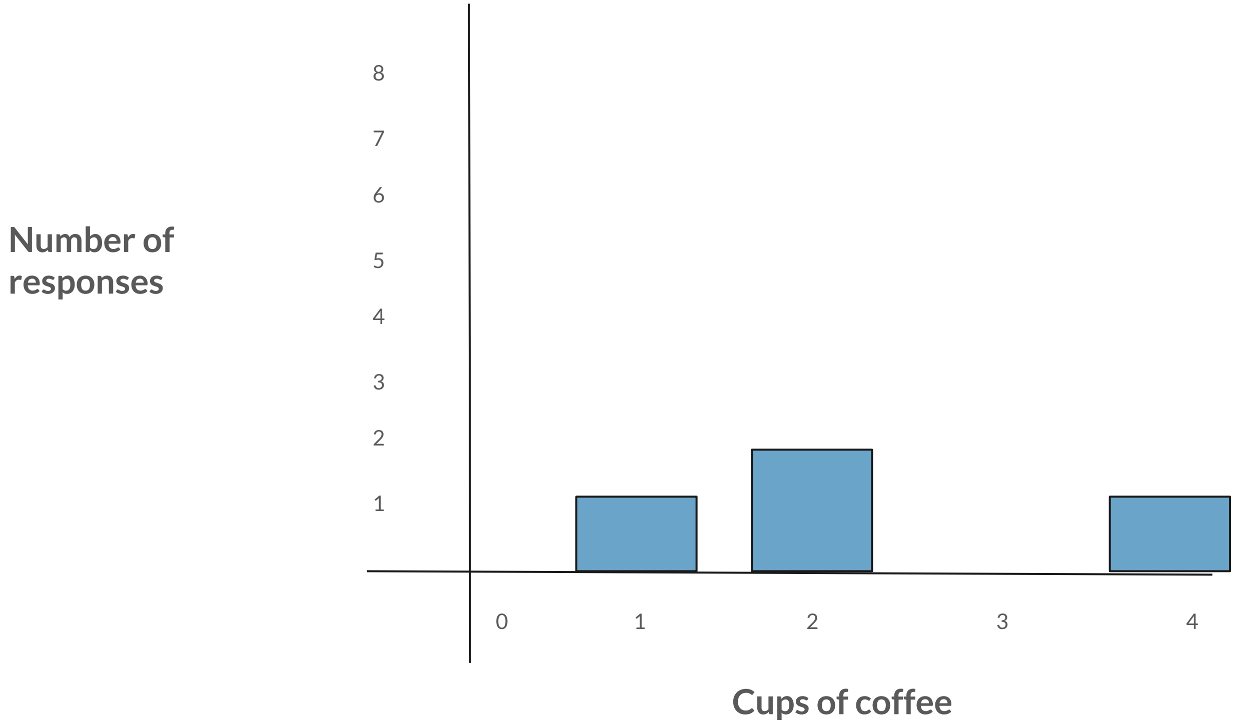
We call the next participant and they give us a more reasonable response of **2 cups of coffee**. We then indicate **1 person drank 2 cups of coffee**.



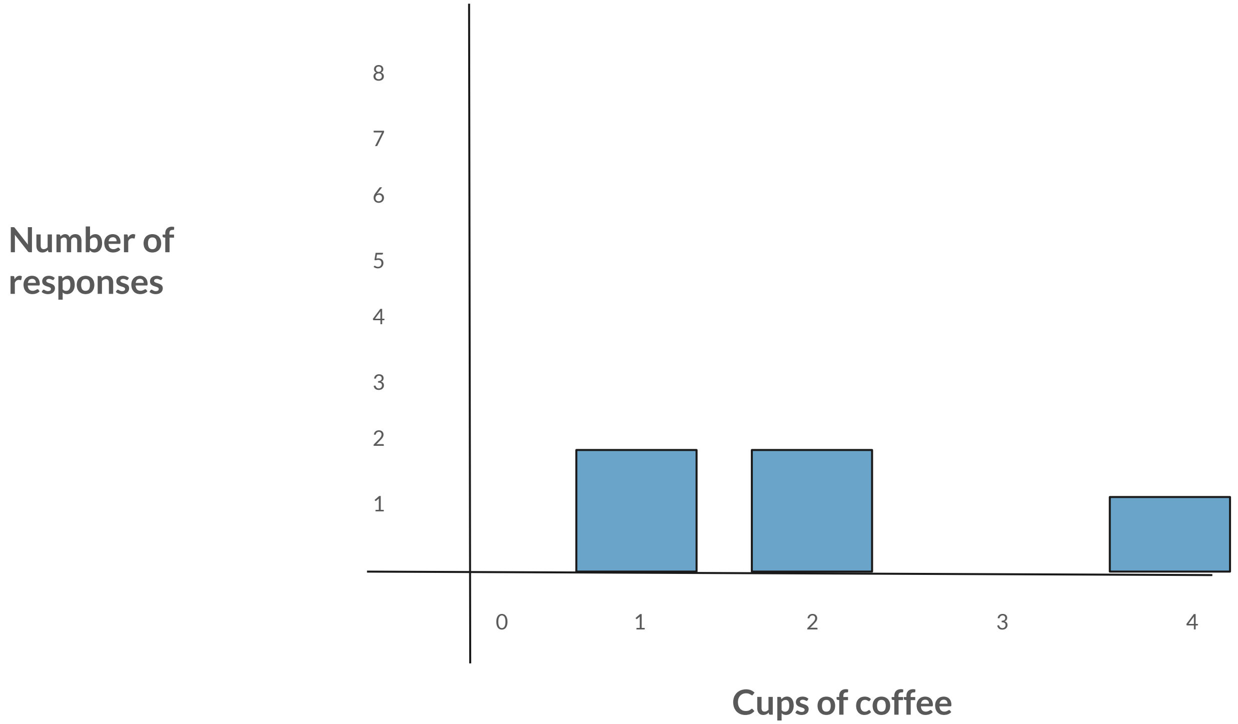
The next randomly selected participant also tells us they drank 2 cups of coffee, this bumps our “2” bar. We now see that **2 participants have drank 2 cups of coffee** and **1 participant has drank 4.**



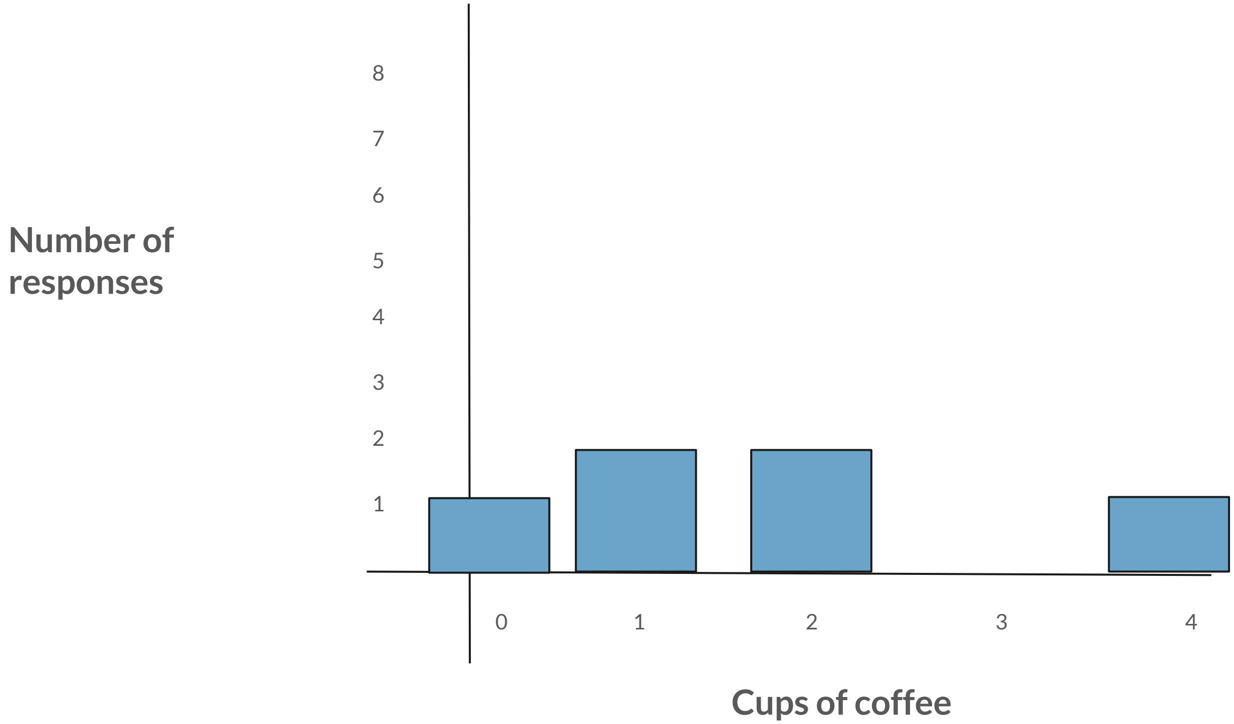
Our next participant tells us they drank **1 cup of coffee**



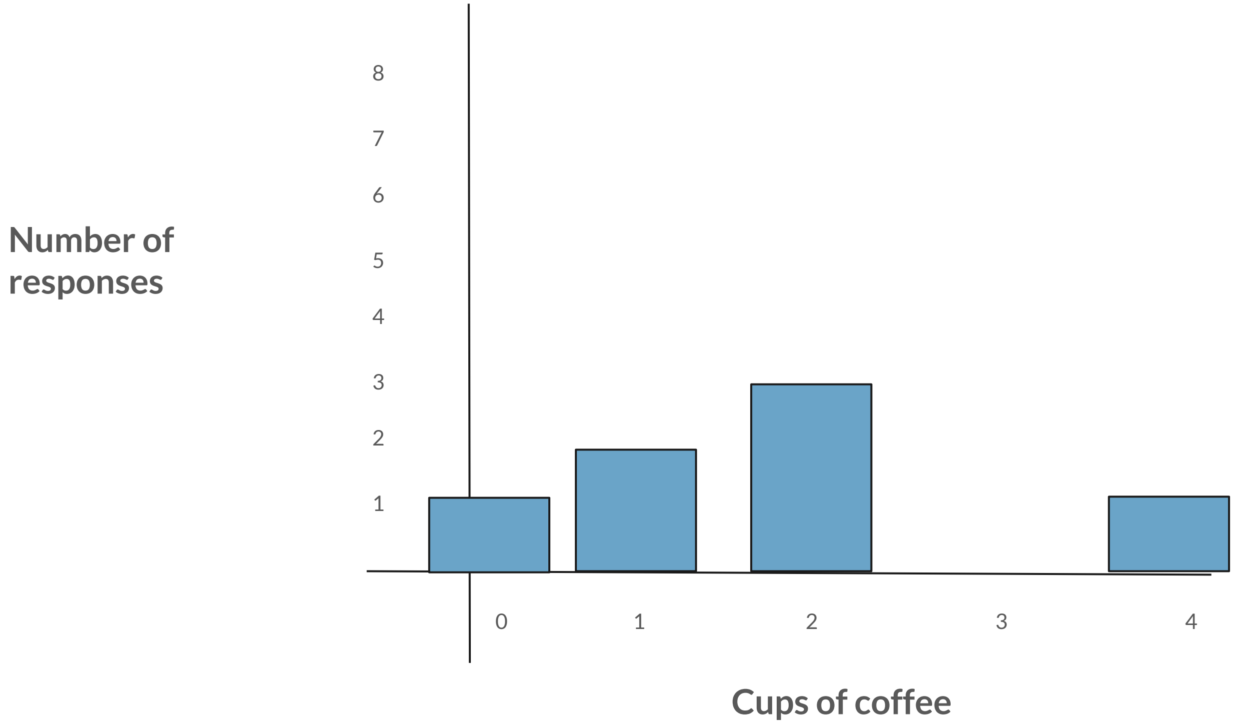
We then complete our data collection by surveying the rest of the 30 participants.



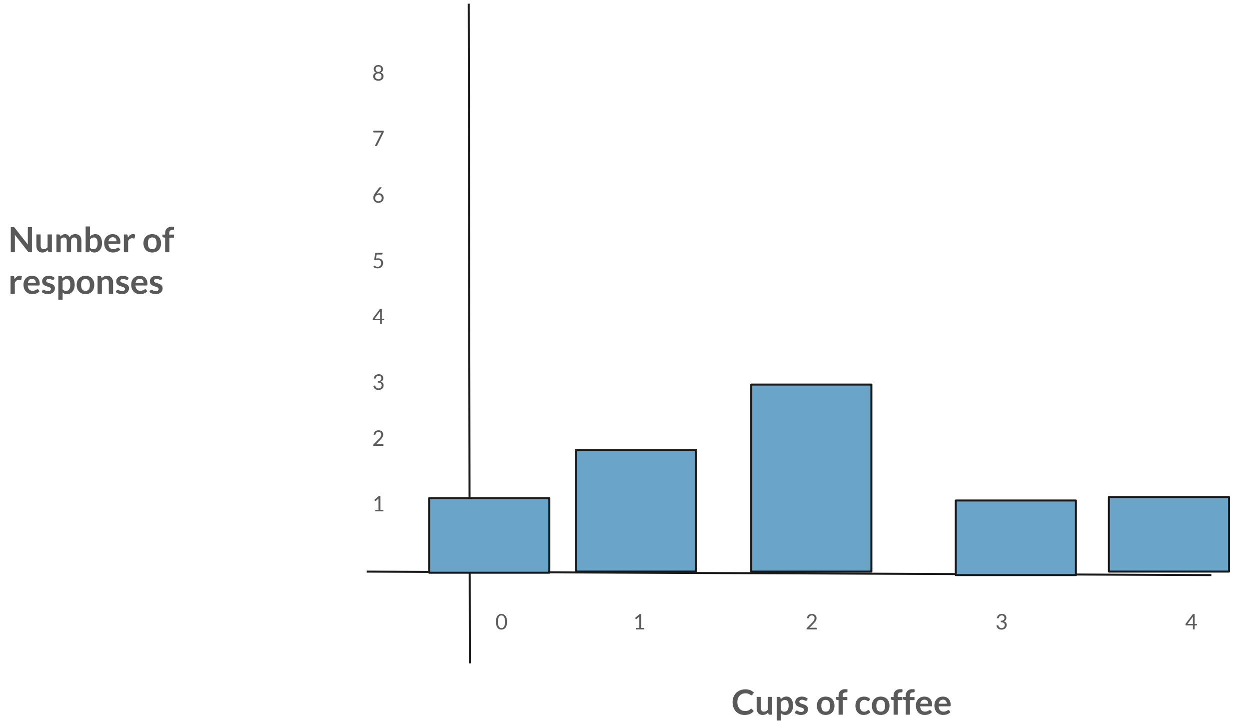
We then complete our data collection by surveying the rest of the 30 participants.



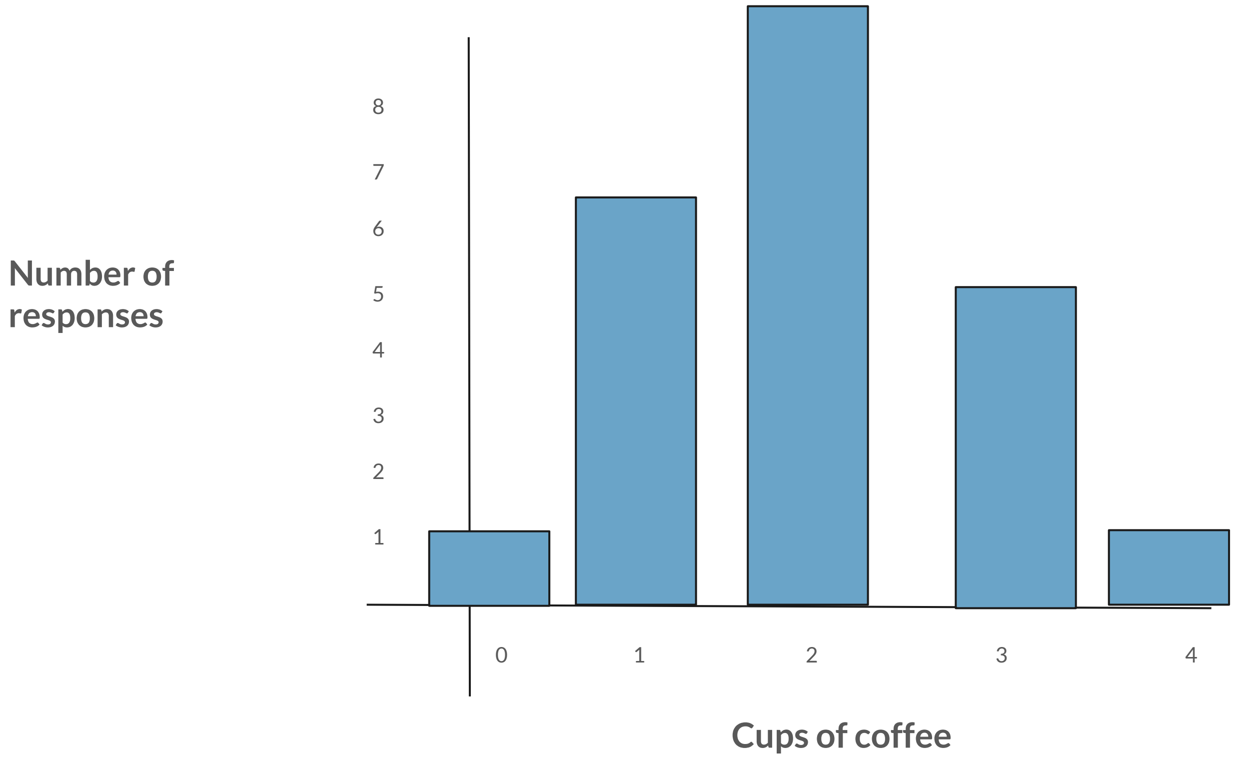
We then complete our data collection by surveying the rest of the 30 participants.



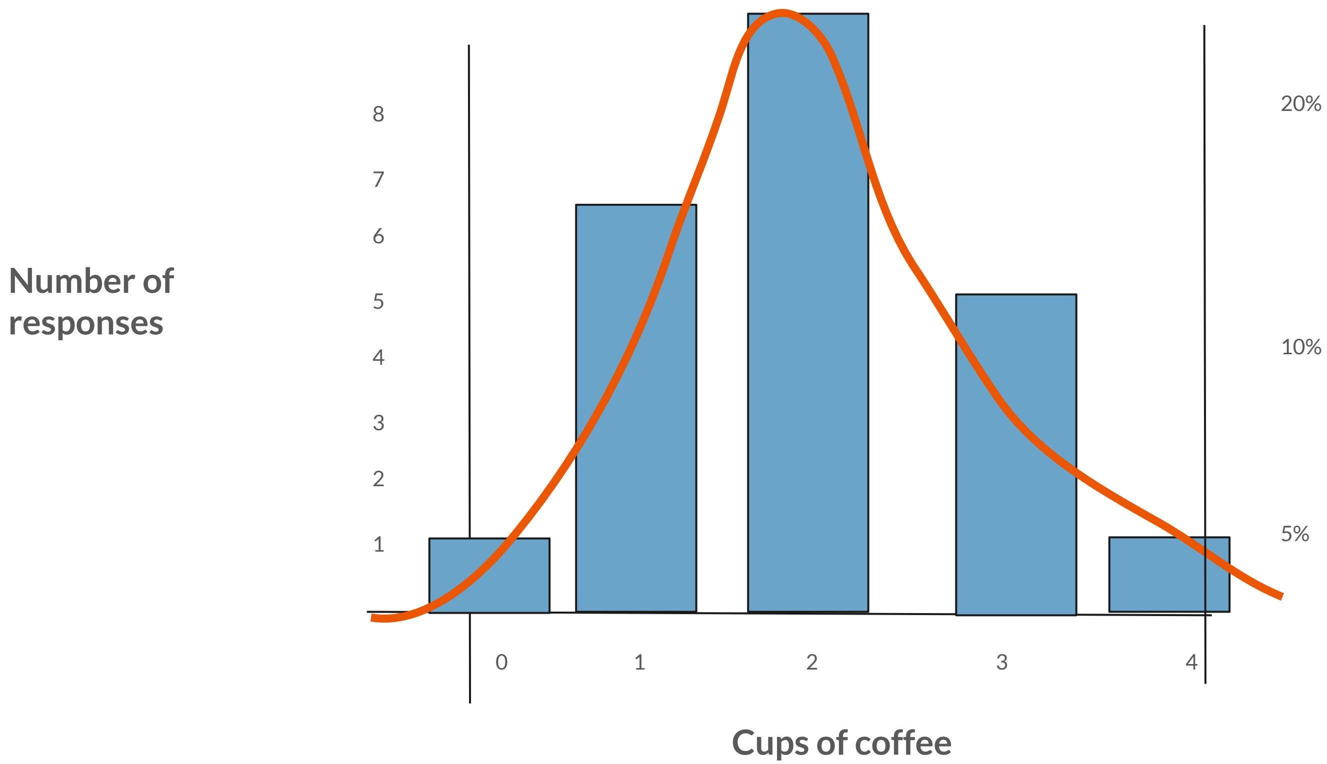
We then complete our data collection by surveying the rest of the 30 participants.



We then complete our data collection by surveying the rest of the 30 participants.



Skipping ahead, **what kind of distribution** do you notice we form after we are done collecting all of our data?



While this is technically a histogram, we could calculate the probability of pulling a specific "cups of coffee" from our dataset. This gives us a **probability distribution** which we see is the normal distribution.

Distributions Review

Let’s go over the names of distributions and what they might indicate about our real-world dataset before we move forward:

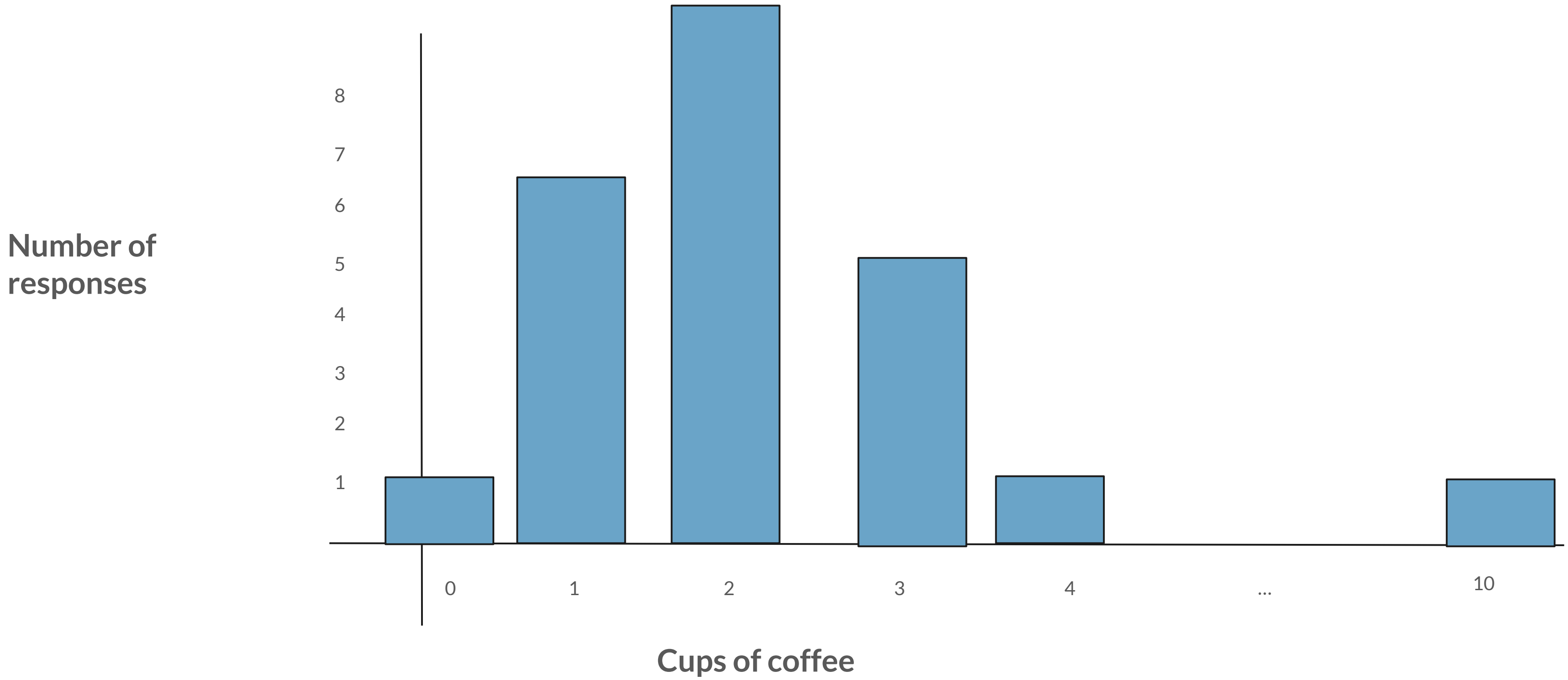
- Uniform distribution
- Binomial distribution
- Normal distribution

Can anyone provide context, **when** might we see these distributions?

Distributions Review

Remember, distributions don't happen by chance! Each distribution emerges from specific real-world phenomenon:

- Uniform distribution: Random games of chance/simulations
- Binomial distribution: Repeated games of chance
- Normal distribution: Real-life distributions of physical phenomenon (height of humans, weight of frogs, daily coffee consumption)



Thought experiment: What if you survey one more participant who indicates that they drink **10 cups of coffee a day**. Can we reasonably assume they are part of this distribution? **First, let's see how the z-score helps us determine outliers.**

Z-Score

Normal Distribution

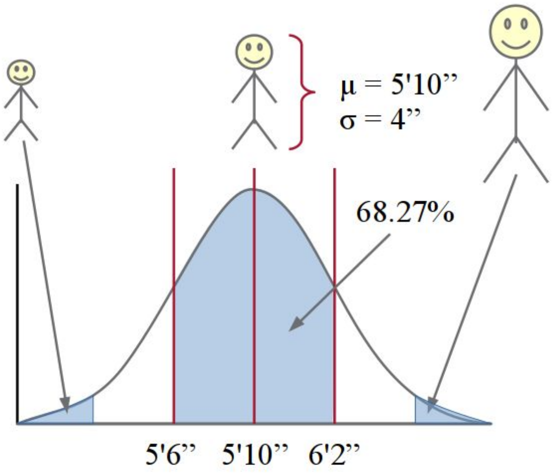
We see the **normal distribution** when observing measurements on real world quantities (*height, miles driven, mosquito wingspan, number of gun crimes in the US*).

So far, we've been discussing the perfect case of the **standard normal distribution**.

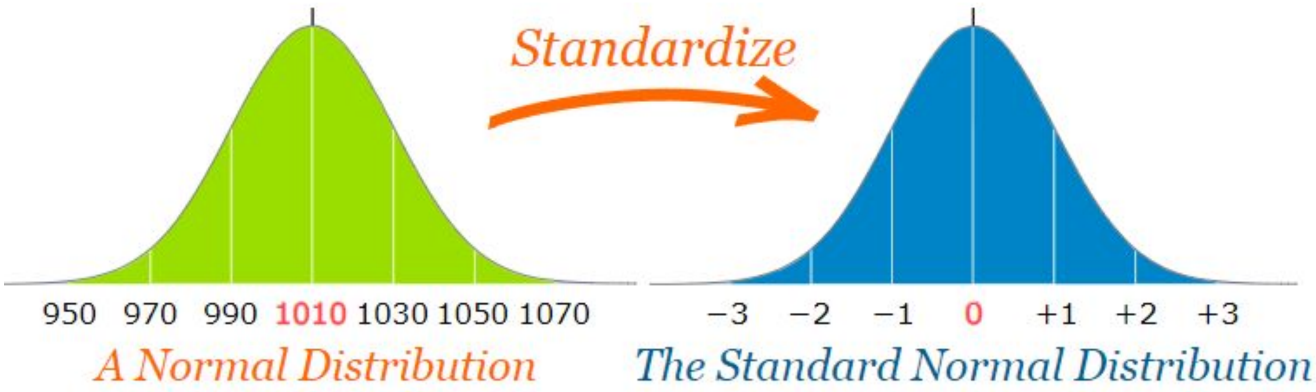
This is the case where the **mean** & **median** are 0, the **standard deviation** is 1.

However, we can turn our normal distributions into the **standard normal** by **standardizing our dataset.**

By converting our normal distribution to the standard normal, we are able to better determine when a sample is an outlier.



In essence, this tells us "**how many standard deviations is our sample from the mean.**"



$$z = \frac{x - \mu}{\sigma}$$

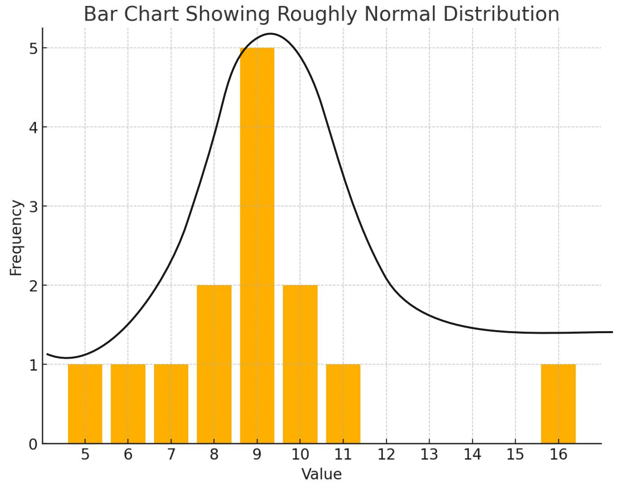
$\mu =$ Mean

$\sigma =$ Standard Deviation

We will revisit this process when we get introduced to **machine learning**.

But first, let's identify the calculation we use to standardize our data. We call this the **Z-score**. We calculate this "**score**" by subtracting each sample by the mean and **dividing** it by the standard deviation.

data = [5, 6, 7, 8, 8, 9, 9, 9, 9, 10, 10, 11, 16]



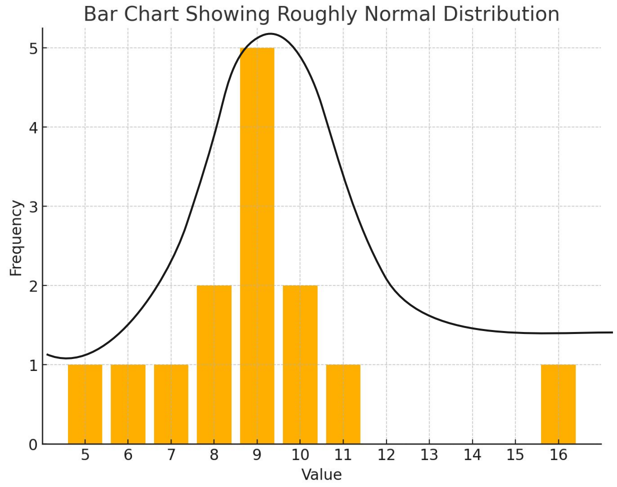
To see how the **Z-score** allows us to detect outliers, let's work with the following toy dataset of **customer wait-times at a coffee shop**.

Notice that we **form a normal distribution** (with a slight skew)...

data = [5, 6, 7, 8, 8, 9, 9, 9, 9, 10, 10, 11, 16]

Average = 9

Sample Standard Dev ~= 2.6



But this is still not a standard normal! I.e. our average is not 0 and our standard deviation is not 1.

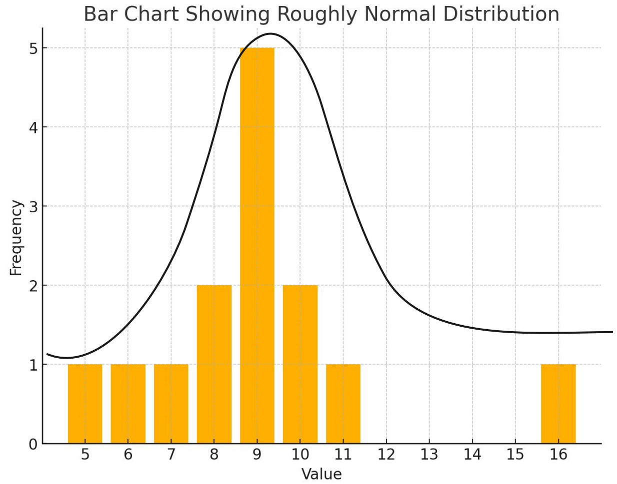
data = [5, 6, 7, 8, 8, 9, 9, 9, 9, 10, 10, 11, 16]

Average = 9

Sample Standard Dev ~= 2.6

$$z = \frac{x - \mu}{\sigma}$$

$\mu =$ Mean
$\sigma =$ Standard Deviation



Let's calculate the **z-score** of each sample, note that we apply this calculation to **each value** in our list to create a brand new list.

Does anyone recall the name of this pattern in Python?

data = [5, 6, 7, 8, 8, 9, 9, 9, 9, 10, 10, 11, 16]

Average = 9

Sample Standard Dev ~= 2.6

$$z = \frac{x - \mu}{\sigma}$$

$\mu =$ Mean
$\sigma =$ Standard Deviation

z_score = (5 - 9) / 2.6

z_score = ...

new_data = []

Let's start off with the first element (5)...

data = [5, 6, 7, 8, 8, 9, 9, 9, 9, 10, 10, 11, 16]

Average = 9

Sample Standard Dev ~= 2.6

$$z = \frac{x - \mu}{\sigma}$$

$\mu =$ Mean
$\sigma =$ Standard Deviation

z_score = (5 - 9) / 2.6

z_score = -1.53

new_data = []

Let's recall our interpretation of the z-score "**how many standard deviations is our sample from the mean.**" Using this definition, how many standard deviations is **5** from the mean???

data = [5, 6, 7, 8, 8, 9, 9, 9, 9, 10, 10, 11, 16]

Average = 9

Sample Standard Dev ~= 2.6

$$z = \frac{x - \mu}{\sigma}$$

$\mu =$ Mean
$\sigma =$ Standard Deviation

new_data = [-1.53, ...]

Using the **Z-score**, state that our first data point is -1.53 standard deviations from the mean. Because this is a negative value, we state that it is 1.53 standard deviations to the **left** of the mean.

data = [5, 6, 7, 8, 8, 9, 9, 9, 9, 10, 10, 11, 16]

Average = 9

Sample Standard Dev ~= 2.6

$$z = \frac{x - \mu}{\sigma}$$

$\mu =$ Mean
$\sigma =$ Standard Deviation

z_score = (6 - 9) / 2.6

z_score = -1.15

new_data = [-1.53, -1.15, ...]

We’re not done yet... let’s continue this calculation for the rest of the data points.

data = [5, 6, 7, 8, 8, 9, 9, 9, 9, 10, 10, 11, 16]

Average = 9

Sample Standard Dev ~= 2.6

$$z = \frac{x - \mu}{\sigma}$$

$\mu =$ Mean
$\sigma =$ Standard Deviation

z_score = (7 - 9) / 2.6

z_score = -0.76

new_data = [-1.53, -1.15, -0.76, ...]

We're not done yet... let's continue this calculation for the rest of the data points.

data = [5, 6, 7, 8, 8, 9, 9, 9, 9, 10, 10, 11, 16]

Average = 9

Sample Standard Dev ~= 2.6

$$z = \frac{x - \mu}{\sigma}$$

$\mu =$ Mean
$\sigma =$ Standard Deviation

z_score = (8 - 9) / 2.6

z_score = -0.38

new_data = [-1.53, -1.15, -0.76, -0.38, ...]

We're not done yet... let's continue this calculation for the rest of the data points.

data = [5, 6, 7, 8, 8, 9, 9, 9, 9, 10, 10, 11, 16]

Average = 9

Sample Standard Dev ~= 2.6

$$z = \frac{x - \mu}{\sigma}$$

$\mu =$ Mean
$\sigma =$ Standard Deviation

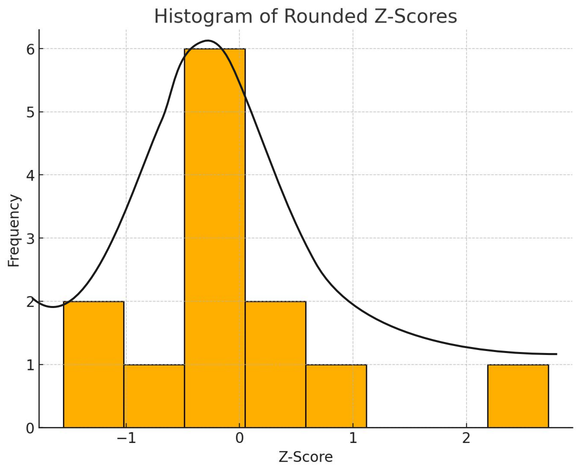
new_data = [-1.555, -1.166, -0.778, -0.389, -0.389, 0.0, 0.0, 0.0, 0.0, 0.389, 0.389, 0.778, 2.722]

Eventually we will complete the loop and calculate the z-score for each sample. This gives us our **standard normal distribution**.

new_data = [-1.555, -1.166, -0.778, -0.389, -0.389, 0.0, 0.0, 0.0, 0.0, 0.389, 0.389, 0.778, 2.722]

Average = 0.000077

Sample Standard Dev ~= 1.041

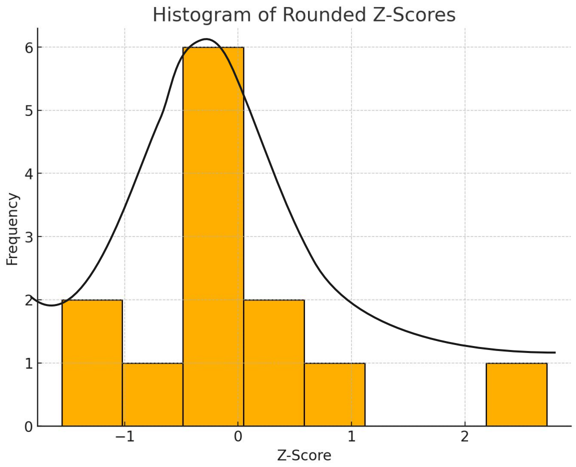


Eventually we will complete the loop and calculate the z-score for each sample. This gives us our **standard normal distribution**. Notice that the **mean and standard deviation** are **approaching 0 and 1 respectively** (the more samples we get, the closer we get to this value).

new_data = [-1.555, -1.166, -0.778, -0.389, -0.389, 0.0, 0.0, 0.0, 0.0, 0.389, 0.389, 0.778, 2.722]

Average = 0.000077

Sample Standard Dev ~= 1.041



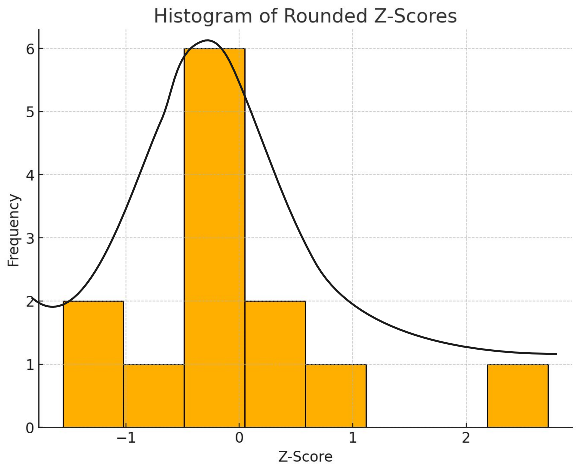
Most importantly, note that we can now directly observe when a sample is **truly an outlier**.

Thinking back to the **normal distribution**, within how many standard deviations do we **expect ~99% of our data to be**?

new_data = [-1.555, -1.166, -0.778, -0.389, -0.389, 0.0, 0.0, 0.0, 0.0, 0.389, 0.389, 0.778, 2.722]

Average = 0.000077

Sample Standard Dev ~= 1.041

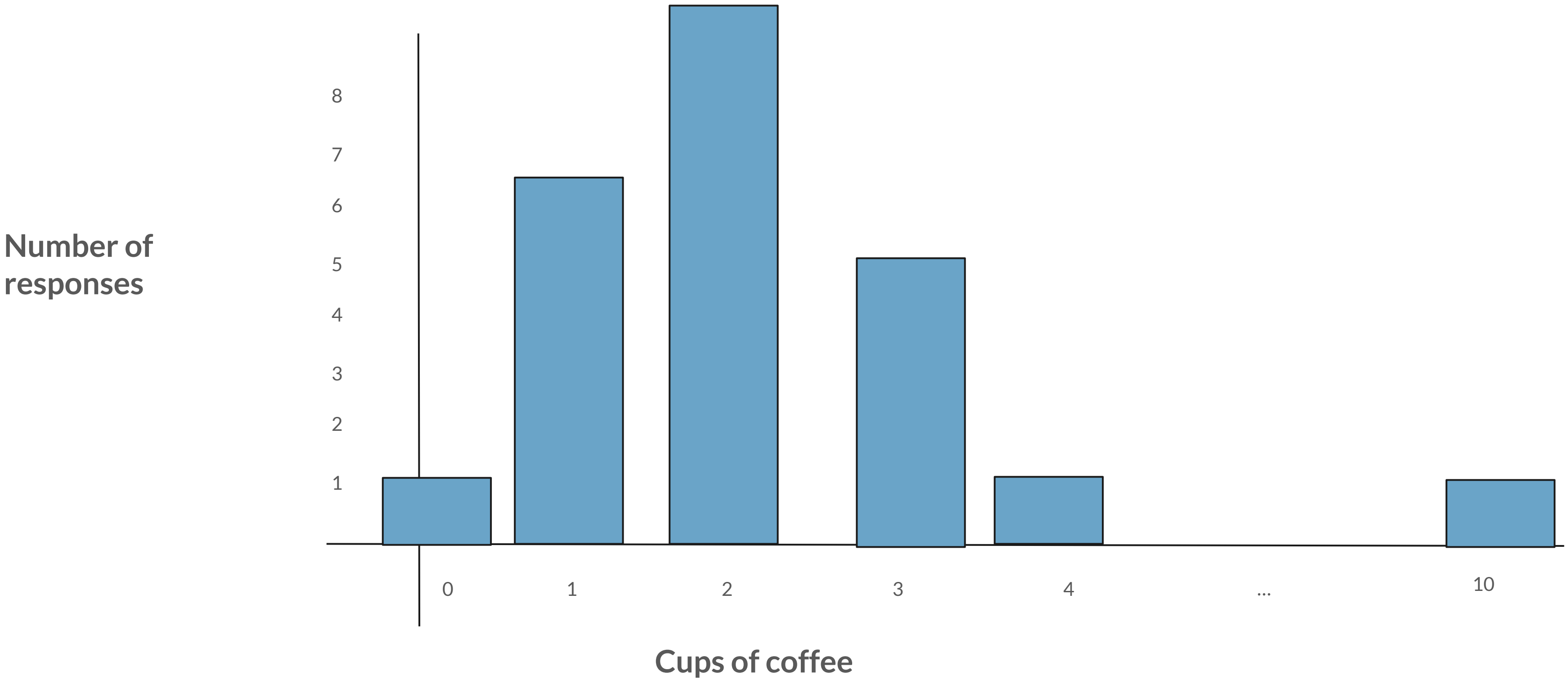


Keep in mind that we only have 13 samples. This is a tiny dataset, and as we collect more data, **we will need to update our z-scores.**

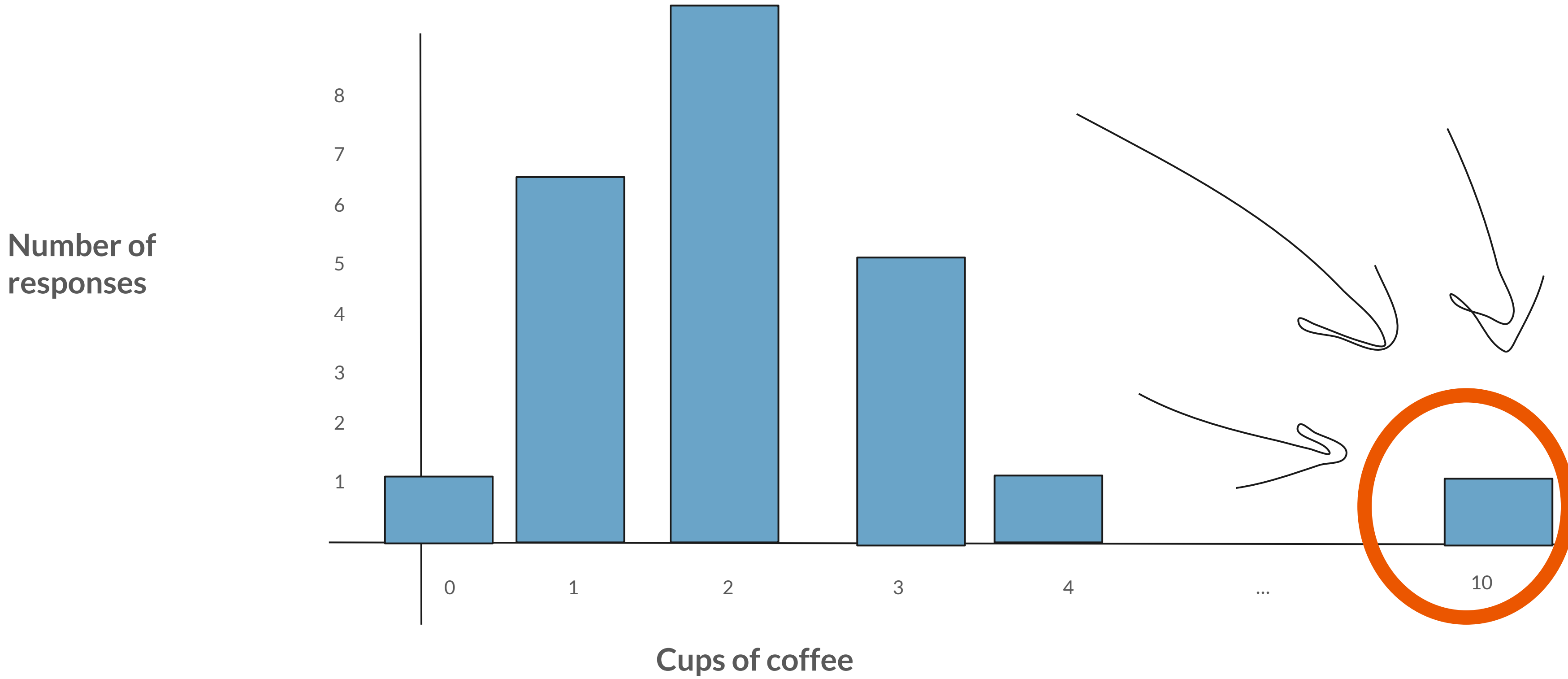
Remember z-scores change as you collect more data.

Within 3 standard deviations.

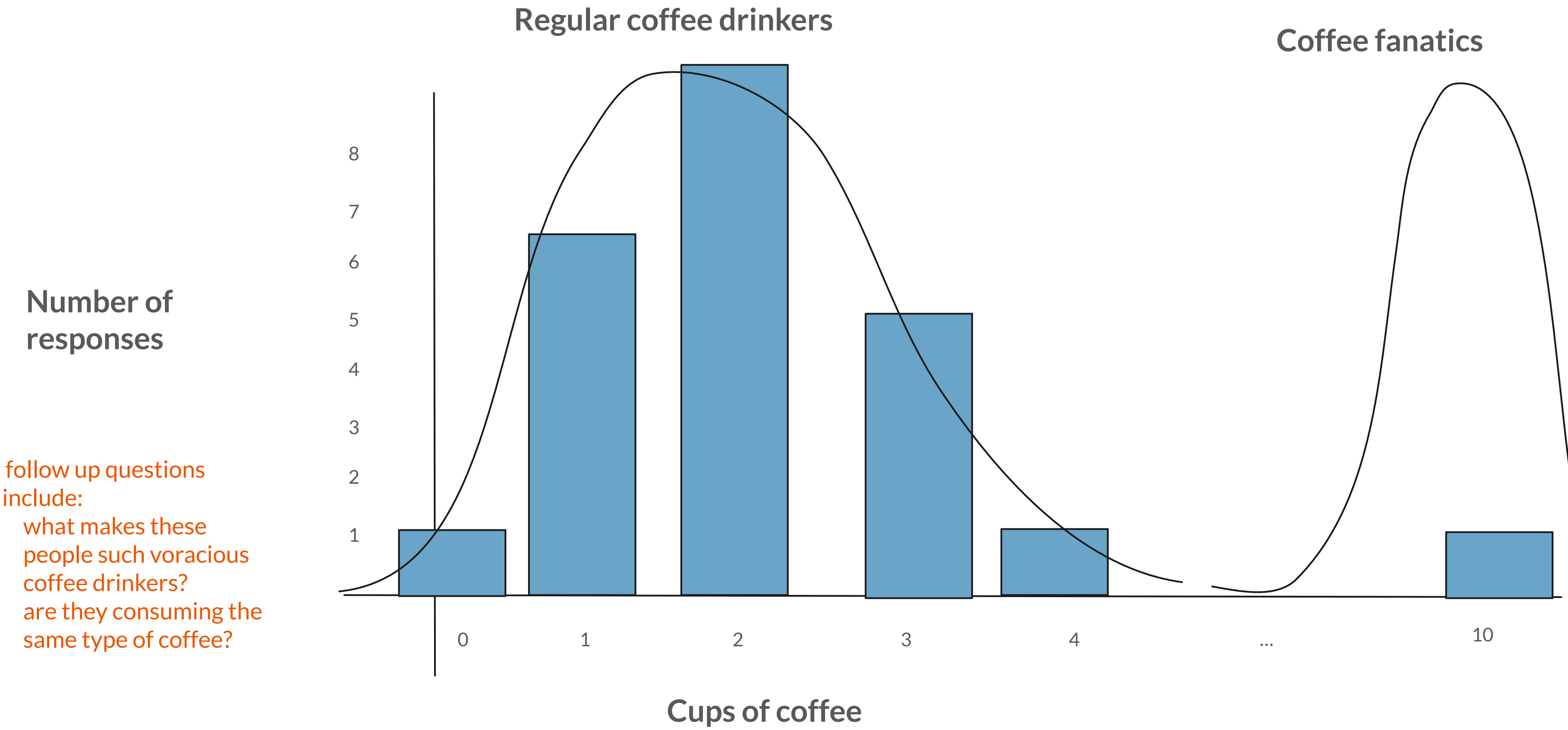
Do you see any z-score representing more than 3 standard deviations on **either the negative or positive side? No!** This means that, so far, none of this data represents an outlier.



But let's go back to our coffee study. Even without the proof of Z-scores, do you notice any values that are **obviously outliers**?



Clearly we have a 10-coffee a day outlier. If we were to calculate this z-score we **would get 8**. This means that this participant is 8 standard deviations away from the mean!



Other follow up questions might include:

- what makes these people such voracious coffee drinkers?
- are they consuming the same type of coffee?

Assuming this person is still alive, a question we should consider is: *is this person uniquely different from the average person we surveyed?* Or framed another way, **does this person belong to a different distribution of people? This is a question that companies try to answer every day.**

E-Commerce Specialist

Change Fashion Inc.
New York, NY 10018

From $3,500 a month | Full-time | Monday to Friday +1

Easily apply

ABOUT US: Change Fashion Inc., based in New York, is a dynamic and fast-growing womenswear brand. Backed by a proven supply chain and our own factory, we…

Insights Analyst, Global Programs and Incentives

Meta
New York, NY

$126,000 - $178,000 a year

The Global Programs and Incentives team is seeking an experienced professional to support the development of data insights that inform Meta's sales and…

SMB Monetization Strategy and Analytics Associate

TikTok
New York, NY

$123,000 - $210,000 a year

Responsibilities TikTok is the leading destination for short-form mobile video. At TikTok, our mission is to inspire creativity and bring joy. TikTok's global…

E-Commerce Specialist

Change Fashion Inc. | New York, NY 10018 | From $3,500 a month

Apply now

Marketing & Promotions:

- Collaborate with the marketing team to plan and execute online promotions, deals, and campaigns.
- Implement A/B testing to identify effective promotional strategies.

Customer Service:

- Respond promptly to customer inquiries and resolve any issues related to online orders.
- Maintain high customer satisfaction ratings by ensuring timely and professional service.

Requirements

Experience:

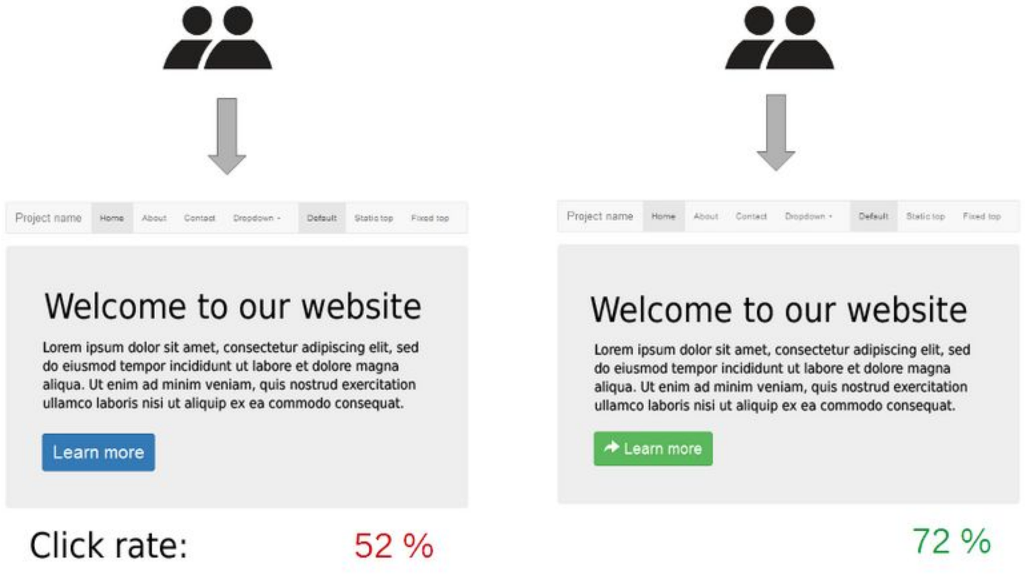
- Proven experience managing e-commerce operations on platforms such as Amazon and Walmart.
- Strong background in product listing optimization, inventory management, and platform-specific SEO.

Skills:

- Expertise in using tools like Amazon Seller Central, Walmart Marketplace, and relevant analytics software.
- Familiarity with e-commerce SEO and keyword optimization strategies.

Introduction to Hypothesis Testing

Did a green button make our customers engage with our website more?

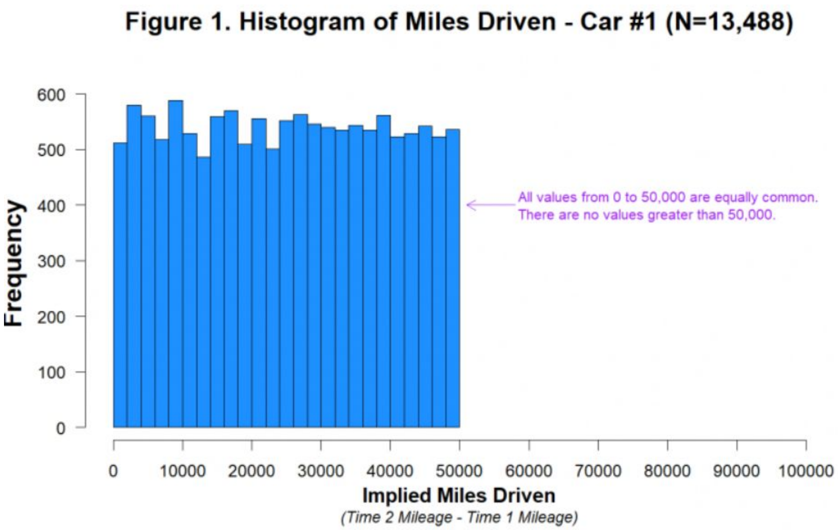


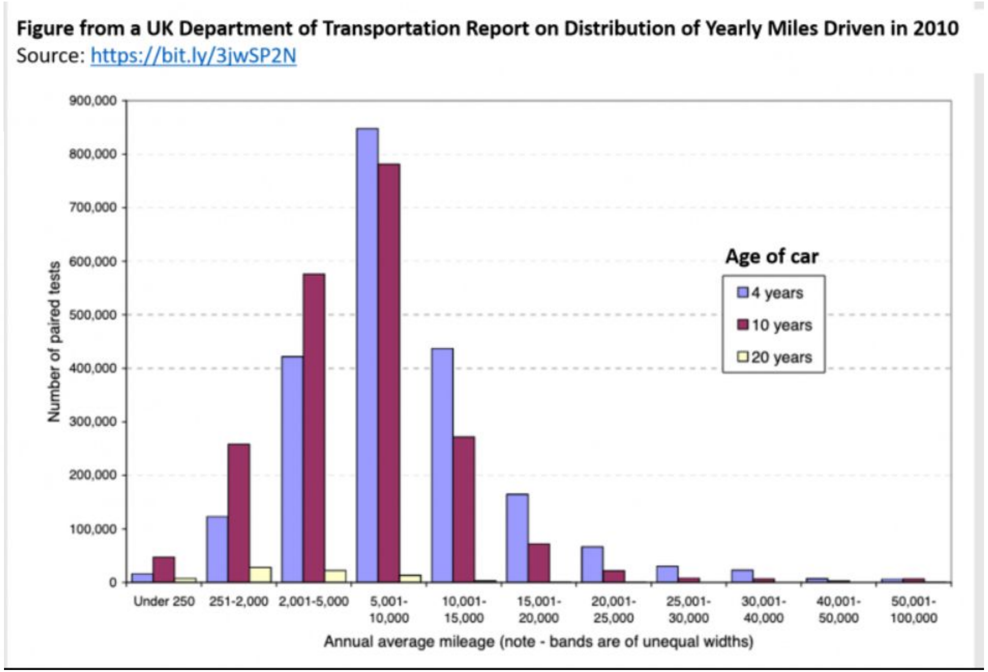
We do **AB testing** to check if users that used an alternative product responded/reacted differently.

Do changes in offer affect if a subscriber responds or not?

		Offer to Subscriber		
		20% Discount	Extended Subscription	Total
		Column %	Column %	Count
Responded to offer	Responded	3.2%	3.6%	1360
	Did not respond	96.8%	96.4%	38640
	Total	100.0%	100.0%	40000

We do use the **chi-squared test** to check if two categorical variables are dependent on one another.





Like we saw in foundations day 1, it's **important and profitable** to note the **difference between distributions**. One represents a fabricated dataset, while another represents a real-world dataset.

Introduction to Hypothesis Testing

Before we formally introduce **hypothesis testing**, let's identify some frameworks of **experiments** we could run:

- **One Sample Test:** Does one **group** contain **an average we expect**?
- **Two Sample Test:** Is one **group** different from **another group**?

Note that experiments like these must be **set up well before you perform data analysis** so that you can guarantee that the effects you observe are truly due to some **specific change**.

data = [6.1, 7.2, 6.4, 6.9, 5.8, 6.7, 7.0, 5.9, 6.3, 6.6]

Expected Avg = 8

Real Avg = 6.49

Does this value seem off from our expected average of 8?

How off is this calculated average from the expected average?

All of these questions can be answered using one-sample testing.

One sample testing example: Let’s say I want to check if data science students are **getting at least 8 hours of sleep per day**. I survey 10 students and get the following dataset and average.

We will not go over these calculations just yet. Instead we will get introduced to these formulas next Wednesday.

One-Sample Test

During **one sample testing**, we collect data and check if the sample set we collected has an **expected mean**.

This is not as simple as just calculating t**he mean of a sample** and comparing it with our **expected mean**. Keep in mind that since samples are **approximations** of the population, we expect a bit of variability in our calculations. **We need a special formula to answer this question.**

The **one-sample test** gives us the ability to determine if a calculated sample falls within some expected range of values.

original_site = [5.25, 4.93, 5.32, 5.76, 4.88, 4.88, 5.79, 5.38, 4.77, 5.27]

original_average = 5.223

What happened to the average amount of time on the site after we made this change?

Is this change **significant** or could it be due to random chance?

new_site = [5.27, 5.27, 5.62, 4.54, 4.64, 5.29, 5.09, 4.94, 5.24, 4.60]

new_average = 5.05

All of these questions can be answered using two-sample testing.

Two sample testing example: Let's say I made a **single** change to the TKH website (*I made the logo orange*) and I want to **measure how long users spend on the website (in minutes)**. I direct 10 users to the unmodified original website and 10 other users to the modified website.

Two-Sample Test

During **two sample testing**, we collect from two groups: ideally the **control** and the **experimental** group.

The control gets **no** “treatment”, while the experimental receives **some sort** of “treatment.”

The **two-sample test** gives us the ability to determine if some sort of change had a **real impact** on a population of users.

Introduction to Hypothesis Testing

These are just quick introductions to the testing methods we will get introduced to next week.

For now, let's understand how we form a hypothesis test using the domain of

EXTRA-SENSORY PERCEPTION

Example of Hypothesis Testing - ESP Study

Method of site acquisition:

Sealed envelope coupled with geographic coordinates.

The sealed envelope was given to the subject immediately prior to the interview. The envelope was not opened until after the interview. In the envelope was a 3 X 5 card with the following information:

The planet Mars.
Time of interest approximately
1 million years B.C.

Selected geographic coordinates, provided by the parties requesting the information, were verbally given to the subject during the interview.

If you think this is ridiculous, the CIA in 1984 didn't think so: https://www.cia.gov/readingroom/docs/cia-rdp96-00788r001900760001-9.pdf

Hypothesis Testing - ESP Testing

To introduce the concept of hypothesis testing, let's go through a live demonstration.

I am a researcher of **ghosts, clairvoyance**, and other **extra-natural phenomena**.

I want to prove that **extra sensory perception** exists. I will run the following experiment:

- I have two buttons.

- I randomly hide the words "Cat" and "Dog" behind these buttons
- I ask ~60 individuals to guess which button has the word "Cat" on it
- I calculate how often individuals are correct

Click here to run through this experiment: https://esp-experiment.streamlit.app/

Hypothesis Testing - ESP Testing

In order to perform this experiment, and then interpret the data I must perform the following steps:

1. Establish my **research hypothesis**, as well as **statistical hypothesis**.
2. Establish my **null and alternative hypothesis.**
3. Calculate my **test-statistic** and **sampling distribution**.
4. Calculate **p-value.**
5. Interpret findings (**as opposed to drawing conclusions**!)

While this experiment is purely introductory, the same pattern follows for all other hypothesis testing methods.

What will be my research hypothesis here, and what is my statistical hypothesis?

Hypothesis Testing - 1) Establish Hypothesis

We have some idea about the world. We want to see if our data **supports or refutes this idea**.

To do this, we must first identify our **research** and **statistical hypothesis**:

Research hypothesis: testable scientific claim

*ex: drinking magnesium supplements **lowers stress***

Statistical hypothesis: correspond to specific claims about the characteristics of the data

*ex: experimental group mean stress levels **will be less than control group***

Hypothesis Testing - 1) Research vs Stats

In this ESP example, I have the following hypothesis.

Research hyp: ESP exists

Statistical hyp: The ratio of correct answers will **NOT be 0.5** (chance). This means it could be either greater than or less than 0.5

ESP either works in the positive direction (ESP tells where the correct button is),

ESP works in the negative (ESP gives a signal where the correct button is, but humans interpret that as the wrong answer).

Hypothesis Testing - 2) Null vs Alt

In order to prevent ourselves from "cheating" and drawing conclusions where they do not exist, we must next establish a **null and alternative hypothesis**.

Null Hypothesis: the **exact opposite** of what I am testing,

Alternative Hypothesis: the thing that I am trying to **prove**,

Because what I'm about to do is invent a new statistical hypothesis (the "null" hypothesis, ***H0****) that corresponds to t****and then focus exclusively on that,*** *almost to the neglect of the thing I'm actually interested in (which is now called the "alternative" hypothesis,* ***H1****).*

Hypothesis Testing - 2) Null vs Alt

The important thing to recognise is that the goal of a hypothesis test **is not to show** that the **alternative hypothesis is (probably) true**;

the goal is to show that **the null hypothesis is (probably) false**.

What is the null and alt of my ESP statistical hypothesis (in the context of correct vs wrong guesses).

Hypothesis Testing - 2) Null vs Alt

In this case, if ESP did not exist, then our supposed ratio of correct guesses should be 0.5 (chance). This is our null hypothesis (data shows no effect).

$$H0: theta = 0.5$$

However if ESP exists (in either direction), then we don't necessarily want to name a specific value that our correct ratio should go, but rather state that the ratio of correct guesses is NOT 0.5 (not chance). This is our alternative hypothesis (data might show effect).

$$H1: theta != 0.5$$

Hypothesis Testing - 2) Null vs Alt

The **null hypothesis is the defendant**, the **researcher is the prosecutor**, and the **statistical test itself is the judge**.

Just like a trial, there is a presumption of innocence:

the null hypothesis is deemed **to be true** unless you, the researcher, can **prove beyond a reasonable doubt that it is false**.

Basically we are stating: *the effect we are looking for does not exist*, and then we check if our data agrees with this hypothesis.

Hypothesis Testing - 2) Null vs Alt

Your goal when doing so is to **maximise the chance that the data actually shows an effect.**

The catch is that the statistical test **sets the rules of the trial**,'

and those rules are designed to **protect the null hypothesis** – specifically to ensure that if the **null hypothesis is actually true**, the chances of a false **conviction are guaranteed to be low**.

	retain H_0	reject H_0
H_0 is true	correct decision	error (type I)
H_0 is false	error (type II)	correct decision

By performing hypothesis testing in this fashion we are attempting to minimize our chances of making a **type I error**, where the **Null is true** (there is **no effect in your dataset**), but you erroneously assume there is an effect.

We also control for this error using something called **significance level**.

Hypothesis Testing - 3) Test-Statistic & Distribution

Which ratio of correct guesses would lead us to concede that the null is incorrect?

For example we tested **N = 100 people, and X = 53** of them got the question right,...

we'd probably be forced to concede that the data are **quite consistent with the null hypothesis**.

Hypothesis Testing - 3) Test-Statistic & Distribution

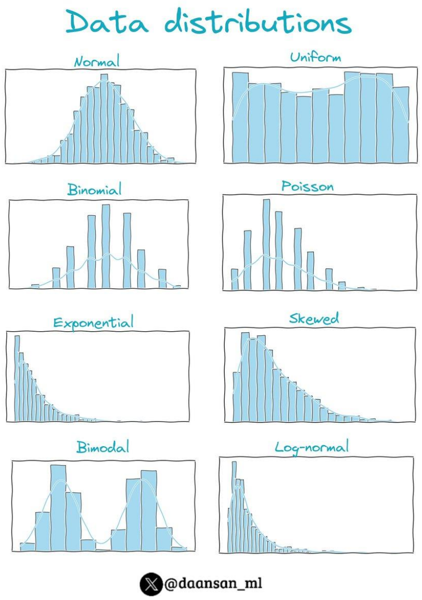
What about if **X = 99** of our participants got the question right or only **X = 3** people got the answer right,...

Hypothesis Testing - 3) Test-Statistic & Distribution

What about if **X = 99** of our participants got the question right or only **X = 3** people got the answer right,...

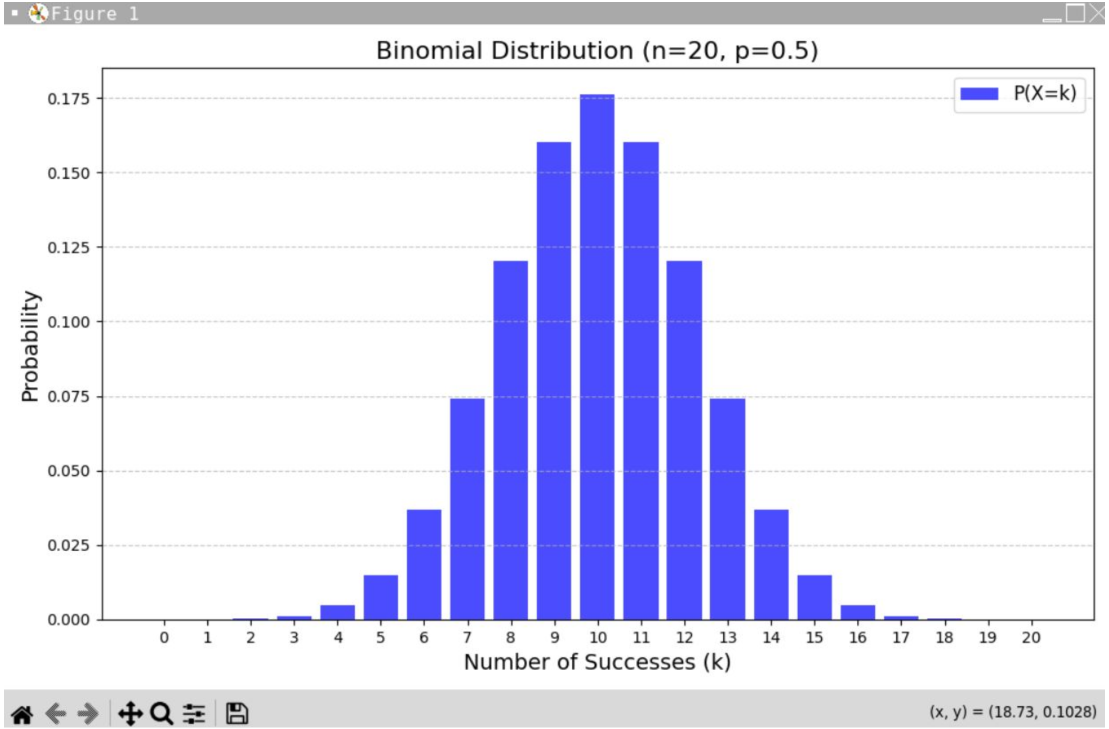
we'd be similarly confident that **the null was wrong.** This metric (**X**) is the **test-statistic** which we will use to draw conclusions on our null hypothesis.

If X is super large (or super small), we consider that as evidence that the null is false. However if **X** is around 50, then we state that the null is most likely not false.

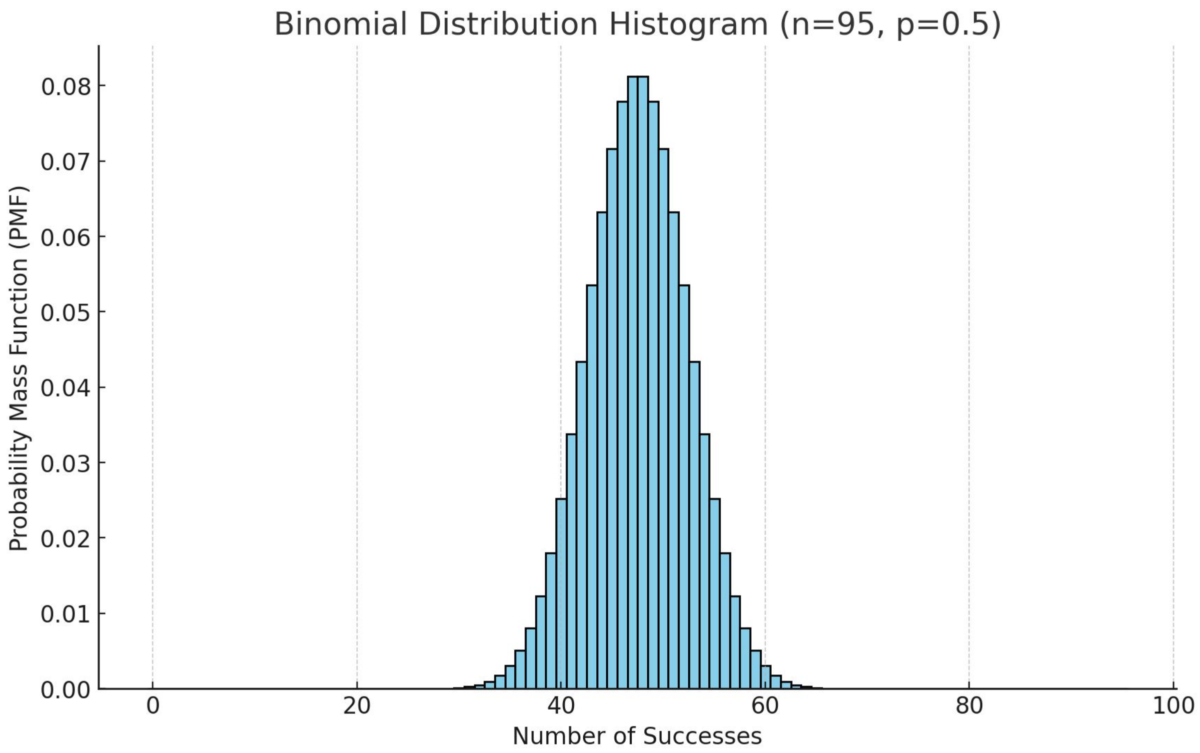


*Let's suppose for the moment that the **null hypothesis really is true**: ESP doesn't exist, and the true probability that anyone picks the correct colour is exactly **θ = 0.5**.*

*What would we expect the data to look like? (i.e. which **probability distribution** do we of games of random chance with multiple trials?)*

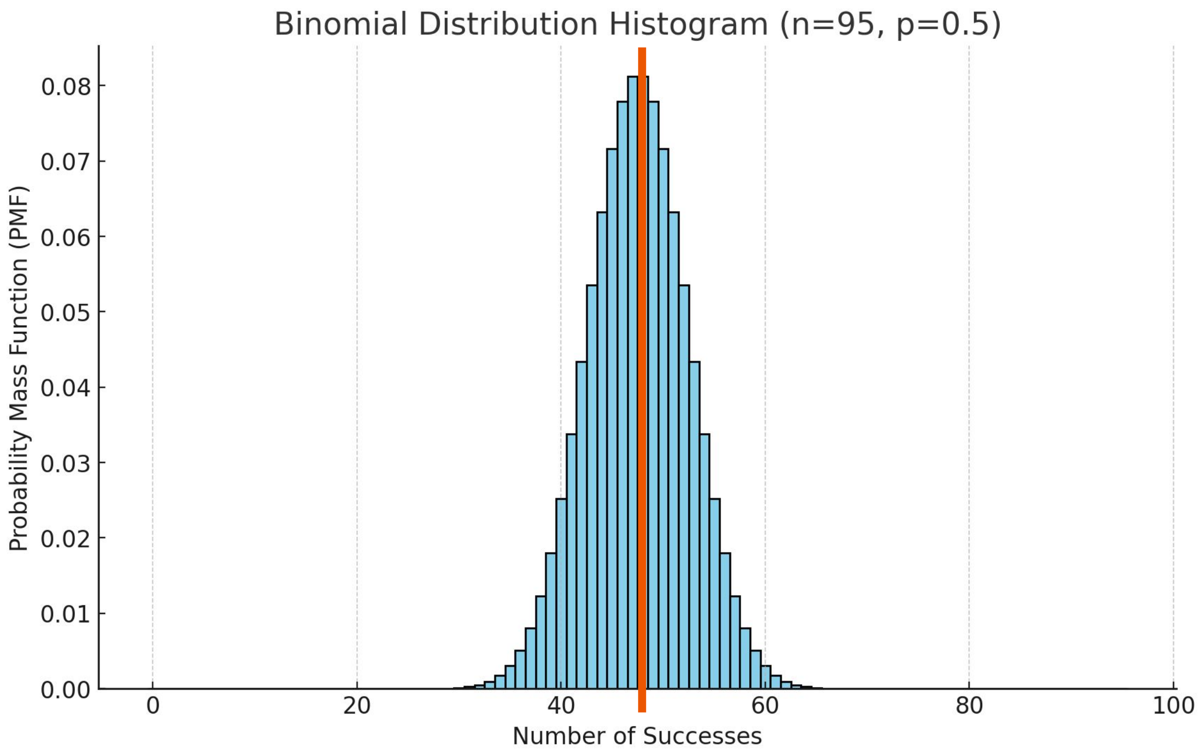


We will use the binomial distribution as we are discussing **repeated trials of games of random chance**. This is our sampling distribution of the **test statistic**.



As of data collection, we have N=95 (*95 responses*). This gives us **the following simulated binomial distribution**.

We state: if ESP did not exist, this is what our **distribution of X** would look like.



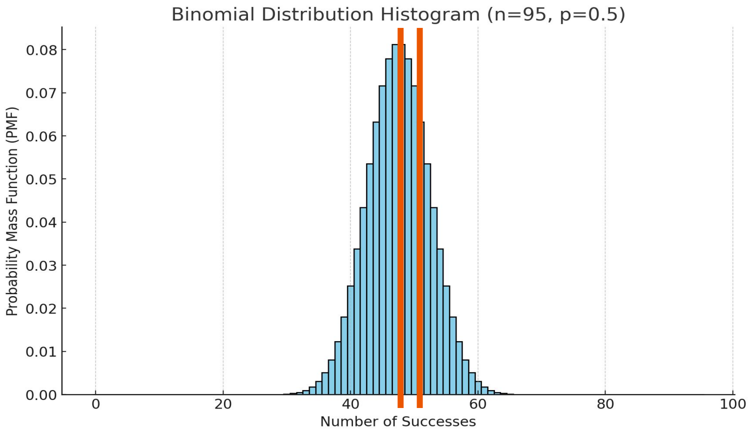
Notice that the average of this dataset is around ~48. **Therefore our Null Hypothesis will state that X = 48.**

Hypothesis Testing - 4) Critical Decisions

Using this distribution and the metric we've calculated from our survey, we can now figure **out if ESP exists in our dataset**.

Our collected **X is 49** (*we correctly figured out where the cat is 49 times*)

Is this pretty darn close to X=48???



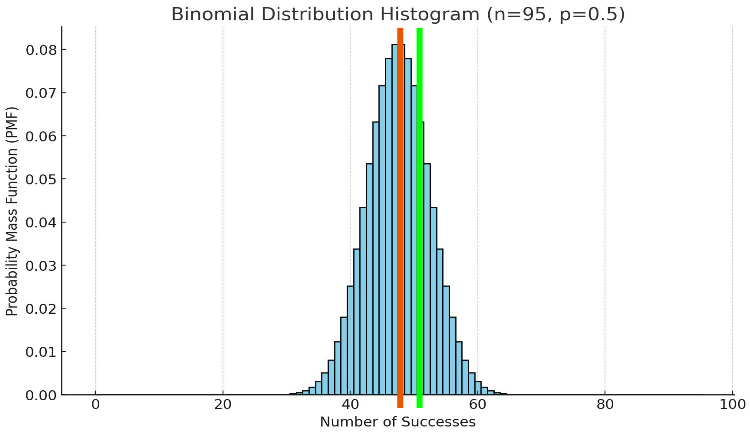
Hypothesis Testing - 4) Critical Decisions

Yes!

In fact if we were to calculate the z-score of **X=49** we would get **0.308**

This means that our collected value is only 0.308 standard deviations away from the mean.

This is a completely reasonable amount that we **should expect** if ESP does not exist and we randomly sampled our population.



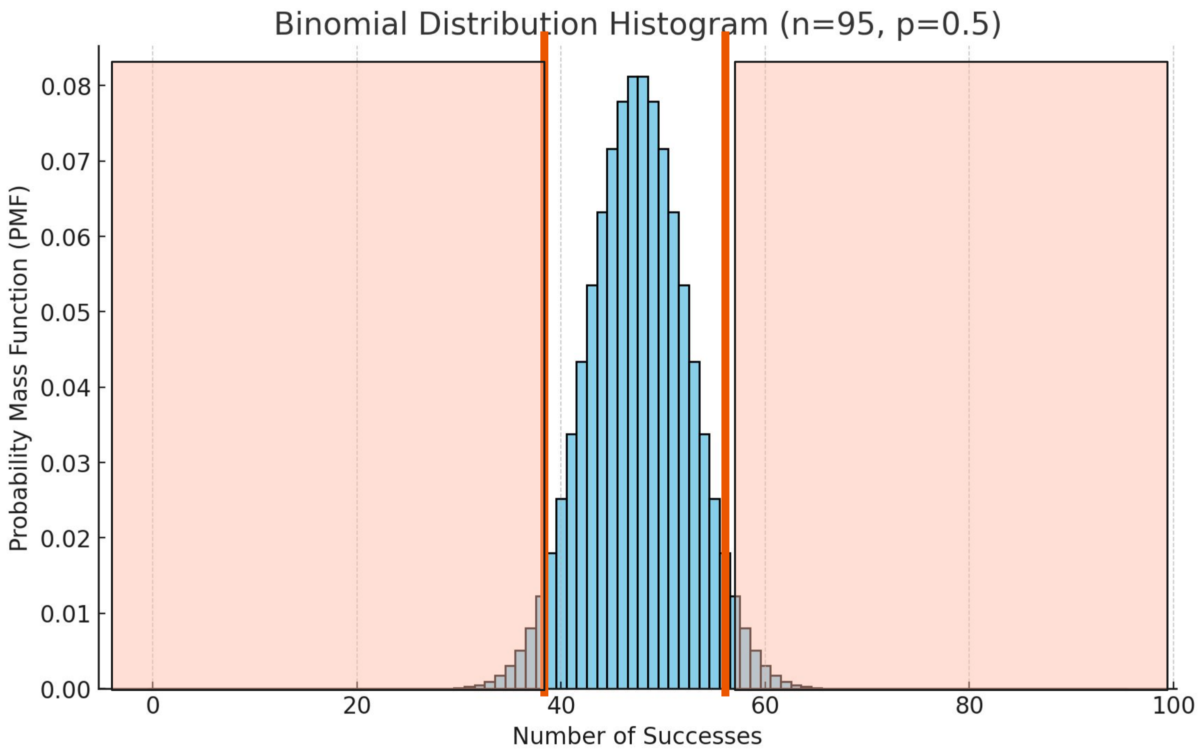
Hypothesis Testing - 4) Critical Decisions

The **critical region** corresponds to **those values of X** for which we would **reject the null hypothesis**.

As we can observe, getting values **near the mean** are expected and we should not consider those values to be part of the critical region.

Next week, we will formally identify what this critical region is.

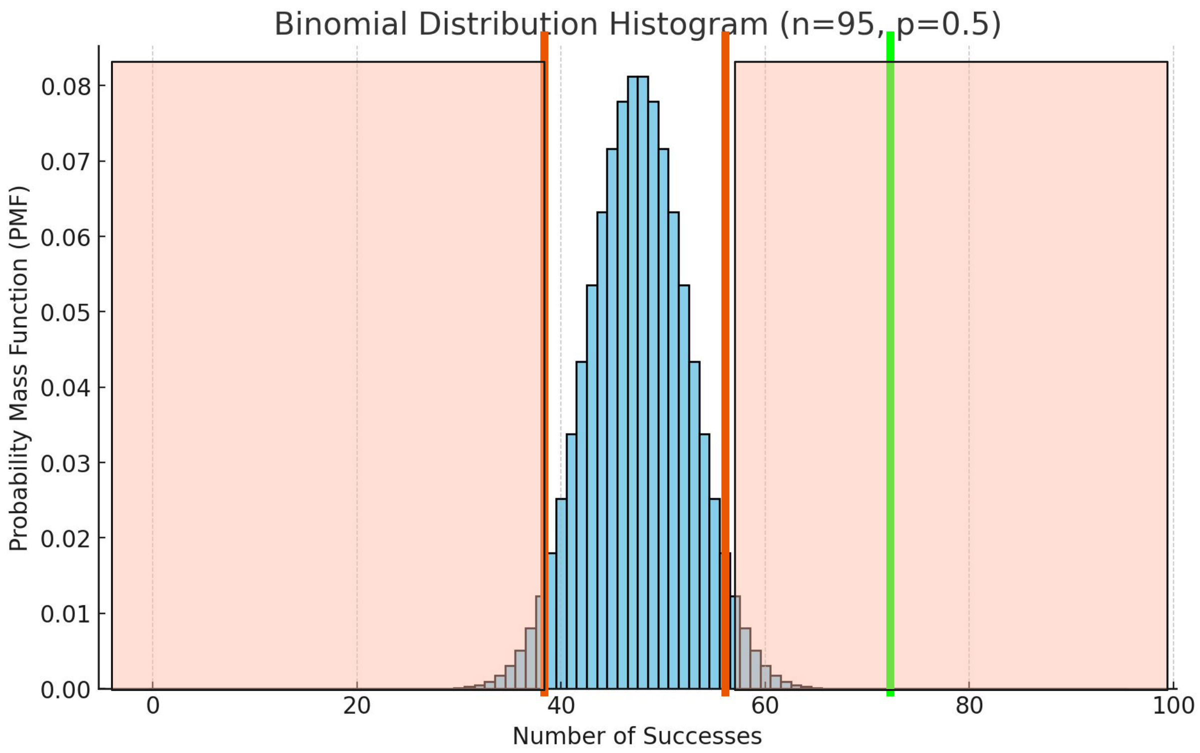
For now we will state that the **critical region** falls anywhere past we get a z-score of 2.



If we get an X value that falls in the red, most likely our null hypothesis is false!

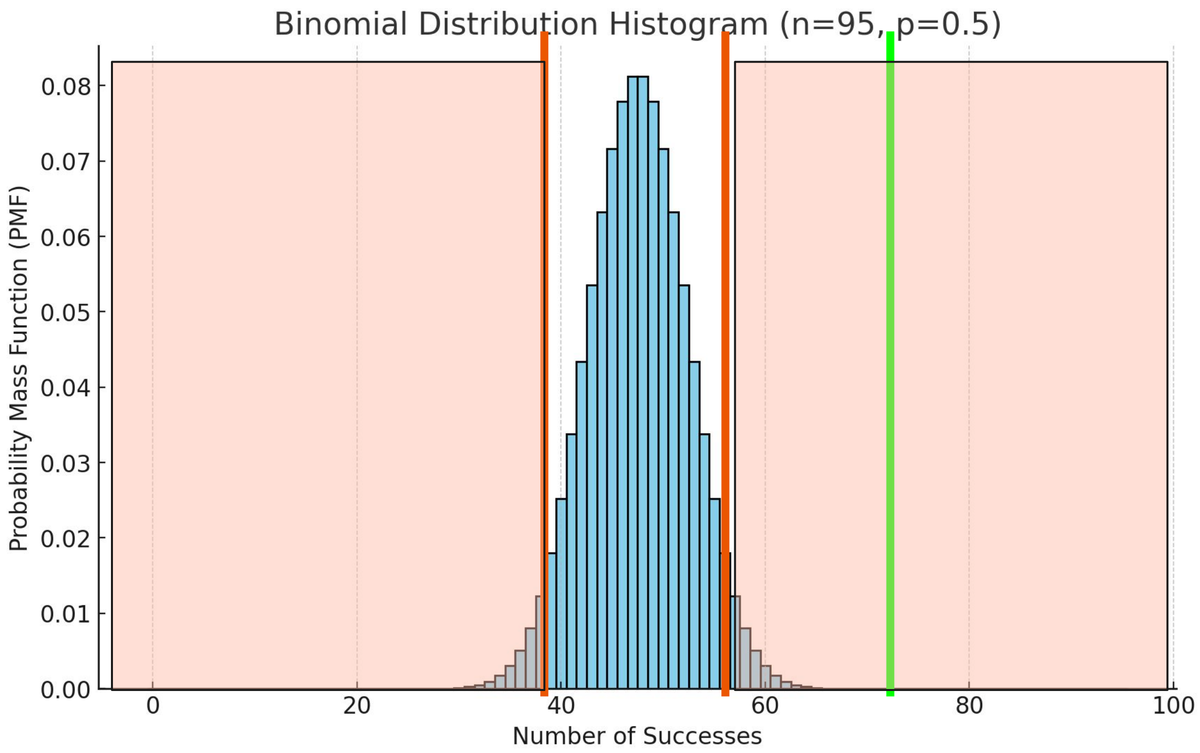
In this distribution, we see that **Z=-2** when **X=38**, and **Z=2** when **X=57**.

Notice that this demarcates **two areas under the probability distribution** (near the tails) where we **reject the null hypothesis.**

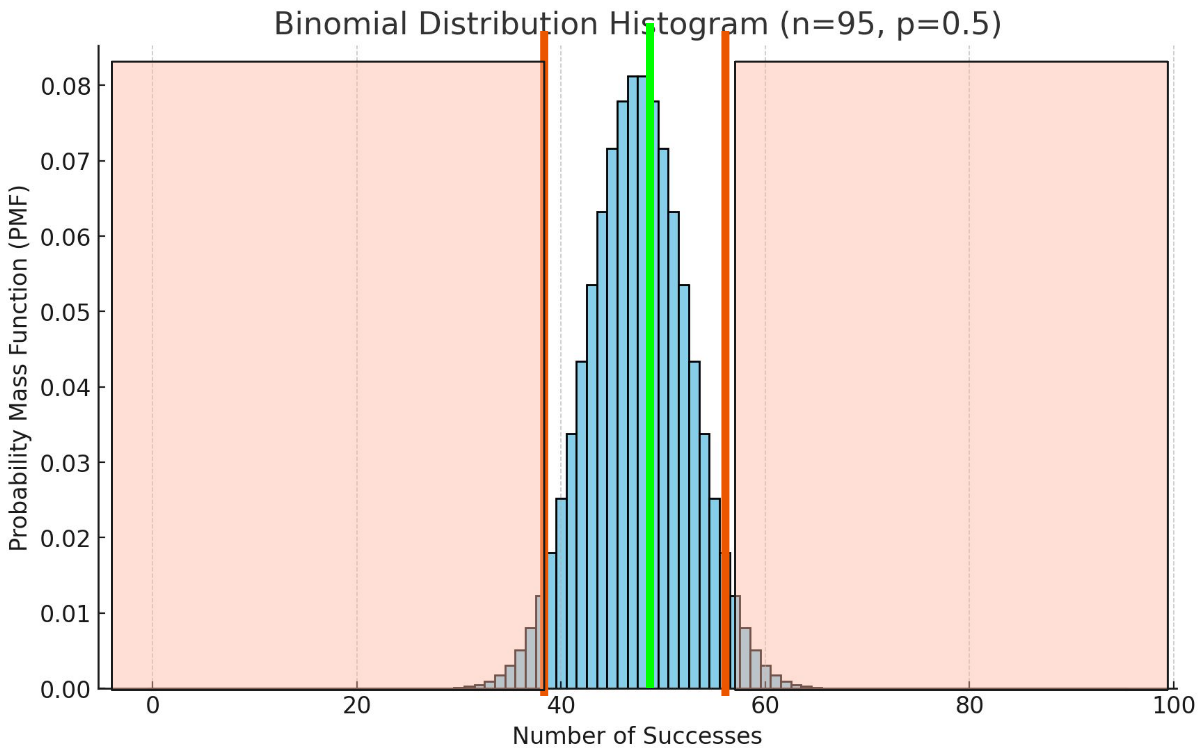


Let's say we get that **X=67**! Wow this falls past the critical region and we get a **Z-score of 4.**

For those of you familiar with calculus, keep in mind that we can calculate the area under the curve (*which is then expressed as a probability*)



The probability that **X=67** belongs to this distribution (where we assume the null is true) is **0.0039%.** This is extremely small, therefore we would state that **we reject the null hypothesis.**



However, keep in mind that we got **X=49**. Not only does this not fall in the critical region, but the likelihood of getting this value in our assumed distribution is **83.76%.** This shows that X=49 is a very typical amount and that we cannot reject the null hypothesis.

Hypothesis Testing - 5) Interpret Findings

Our **presumption of innocence held** (the X value we calculated fit well with our null distribution).

Therefore we **cannot** reject the null hypothesis that **X=48.**

This corresponds with the statement **ESP does not exist.**

Note that we are only interpreting what the data tells us, and this hinges on how well we set up the experiment.

Hypothesis Testing - ESP Testing

Again let's review the steps that we just took when running our hypothesis test:

1. Establish my **research hypothesis**, as well as **statistical hypothesis**.
2. Establish my **null and alternative hypothesis.**
3. Calculate my **test-statistic** and **sampling distribution**.
4. Calculate **p-value.**
5. Interpret findings

As we will see in next weeks class, we will use this same framework to perform data analysis in the workplace.

Wrap-Up

Lab (Due 04/21)

Vancouver, Canada

You are a growth analyst at a Vancouver-based consulting firm called Monica Group. Your manager is spearheading the completion of a a new analytical tool which will automatically label if a review is positive, neutral, negative, or irrelevant.

You will be kicking off completion of this milestone by independently implementing a minimal-viable-product. **This will be a Python pipeline that ingests a text-file of review data and interfaces with the Open AI API in order to automatically label each review.**

Thursday Review Session

See you all at the Thursday Review Session!

- Say hi to Cohort B
- Review GitHub
- Work on TLAB #2
- ...and more!!!

If you understand what you're doing, you're not learning anything. - Anonymous

jupyter

Jupyter: scratchpad of the data scientist