# AB Testing

THE KNOWLEDGE HOUSE

# Agenda - Schedule

1. Pandas Warm-Up
2. Hypothesis Testing Review
3. T-Testing & AB-Testing
4. Break
5. TLAB #3

# Agenda - Goals

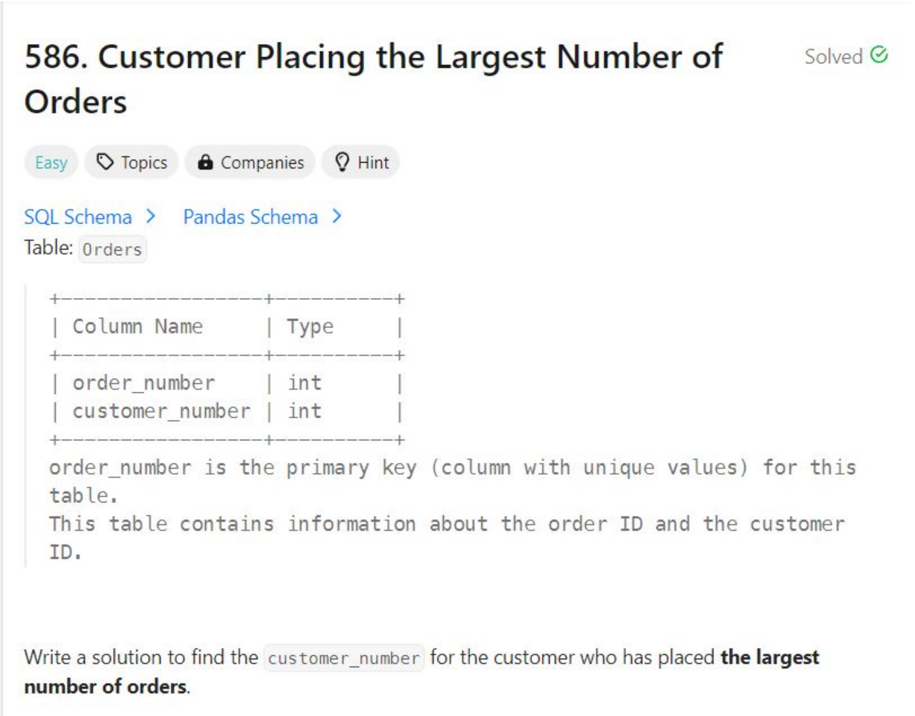
- **Explain the purpose of a two-sample t-test** and when it’s appropriate to use
- **Calculate a t-score** using sample means, variances, and sample sizes
- **Interpret a p-value** and explain what it tells us about statistical significance
- **Identify key real-world considerations when applying t-tests** in A/B testing, such as sample size

# Announcements

- Review Session on 5/1
- TLAB #3 due 5/14

*"be-leaf in yourself!"*

# Pandas Leetcode Warm-Up



## 586. Customer Placing the Largest Number of Orders

Table: `Orders`

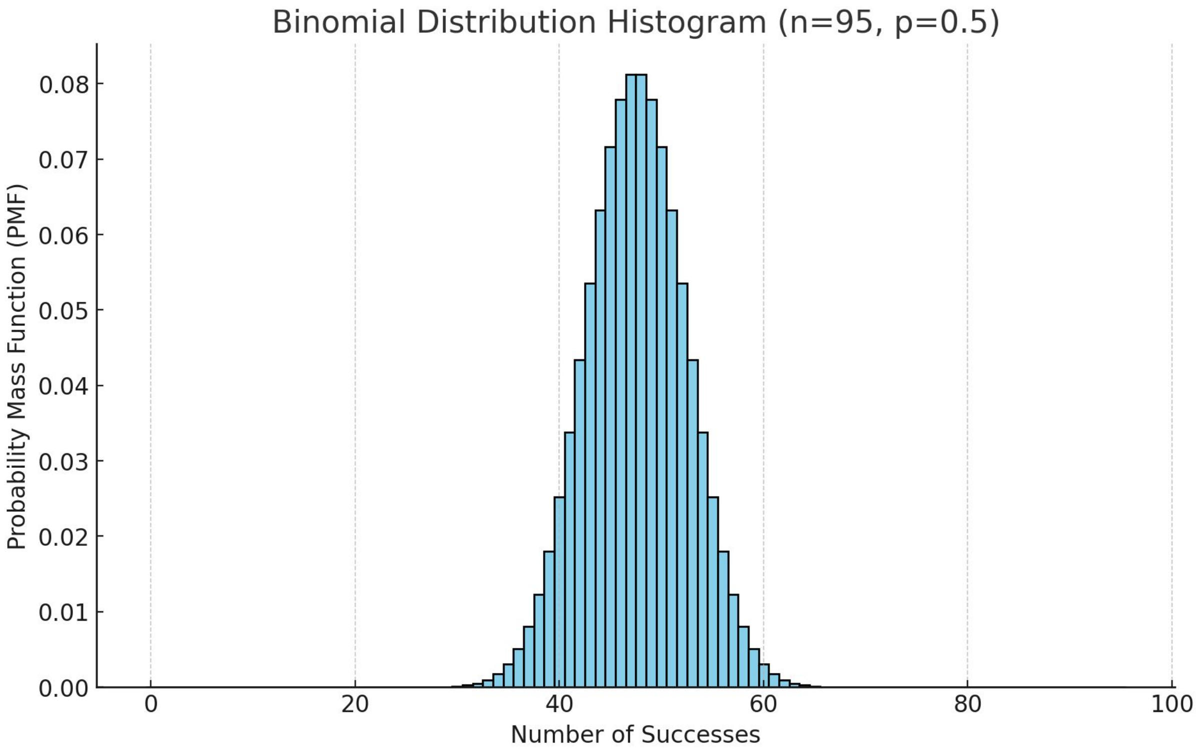
```
+-----------------+----------+
| Column Name     | Type     |
+-----------------+----------+
| order_number    | int      |
| customer_number | int      |
+-----------------+----------+
order_number is the primary key (column with unique values) for this
table.
This table contains information about the order ID and the customer
ID.
```

Write a solution to find the `customer_number` for the customer who has placed **the largest number of orders**.

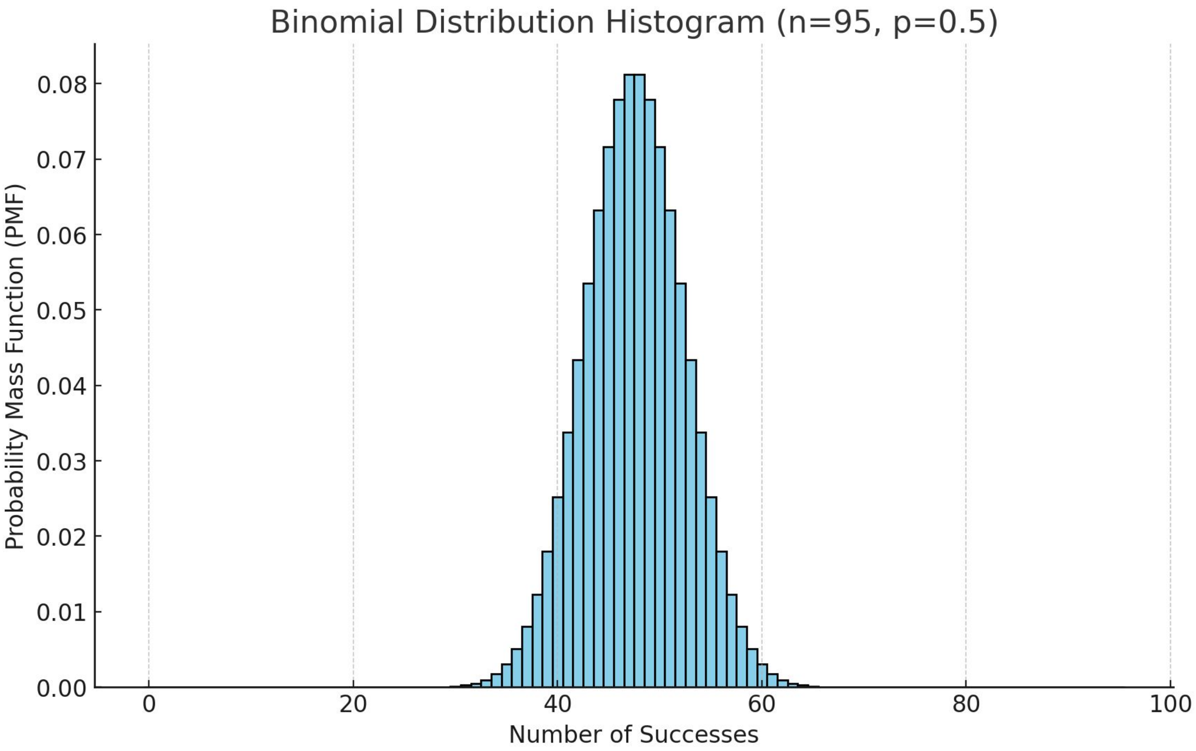
This will not be submitted, but don't forget... **the wheel** is watching.

Take 10 minutes to work on the "Customer Placing the Largest Number of Orders" Leetcode problem: https://leetcode.com/problems/customer-placing-the-largest-number-of-orders/description/?envType=study-plan-v2&envId=30-days-of-pandas&lang=pythondata

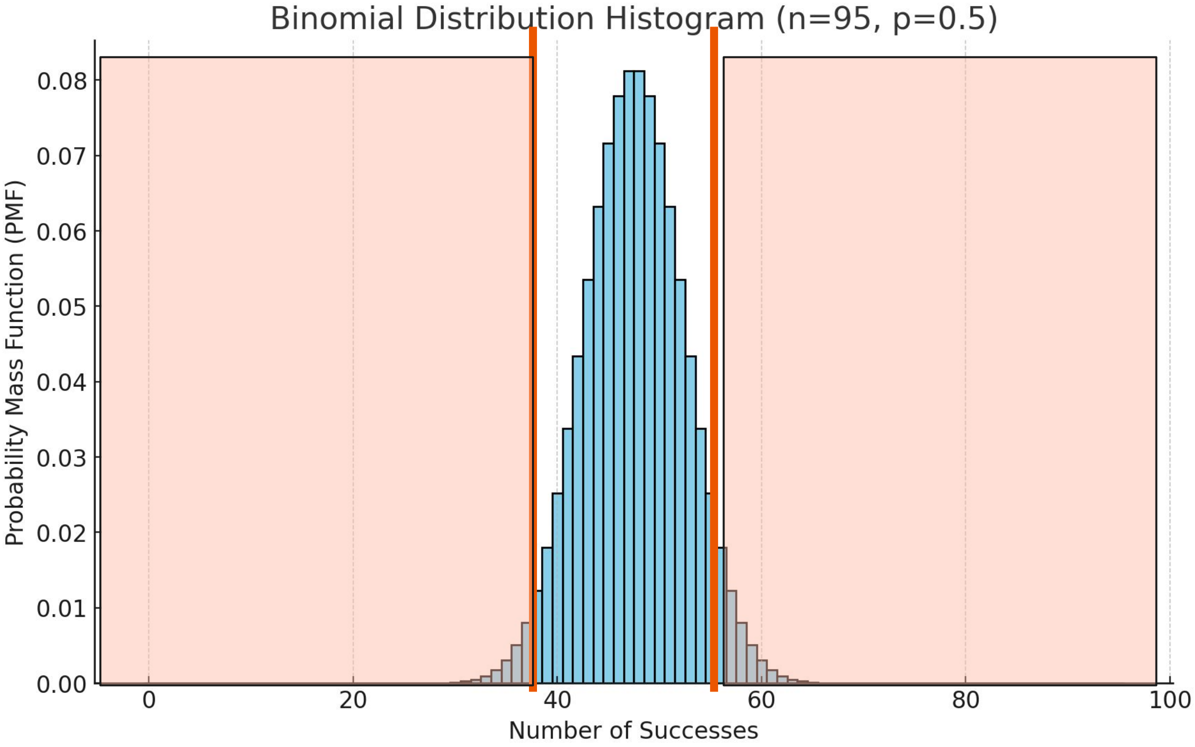
# Hypothesis Testing Review



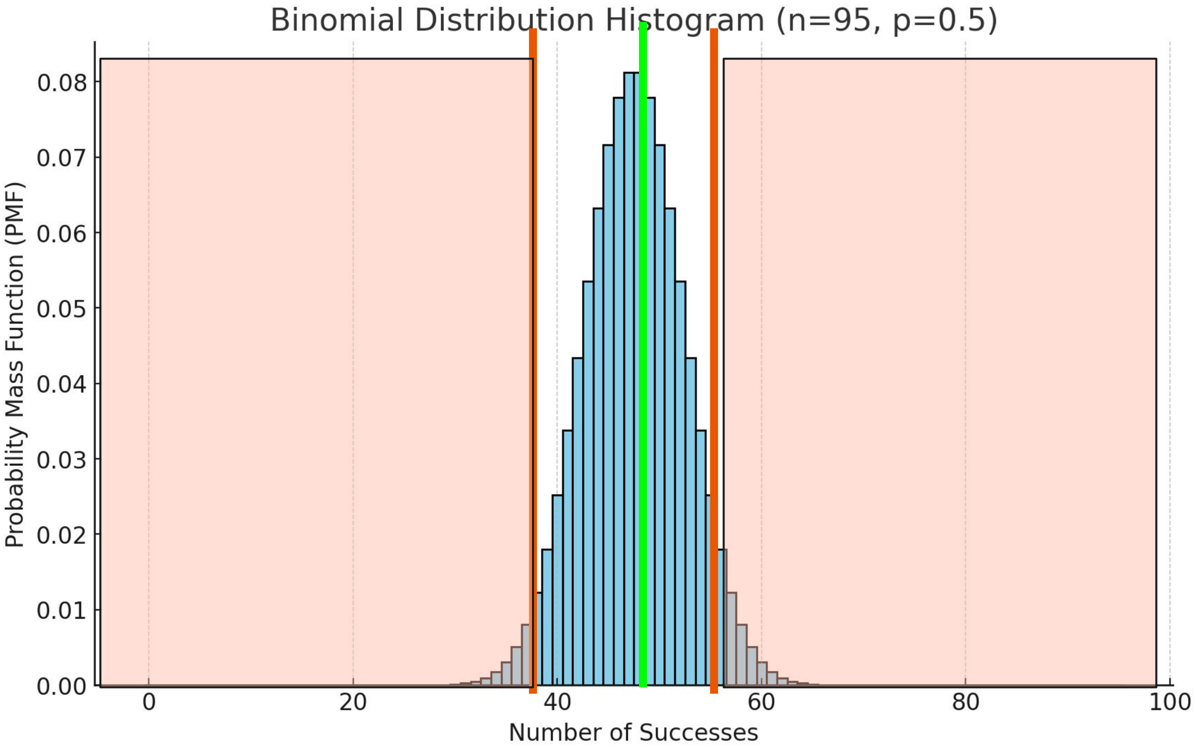
In week 5, we discussed the concept of hypothesis testing via our ESP experiment. Recall that we first assumed that ESP **did not** exist. This formed a binomial distribution. **Does anyone recall what we call this hypothesis?**



This was our distribution under the **null hypothesis**. This is the distribution that we expect our test-metric to follow assuming that there is no effect in our dataset.

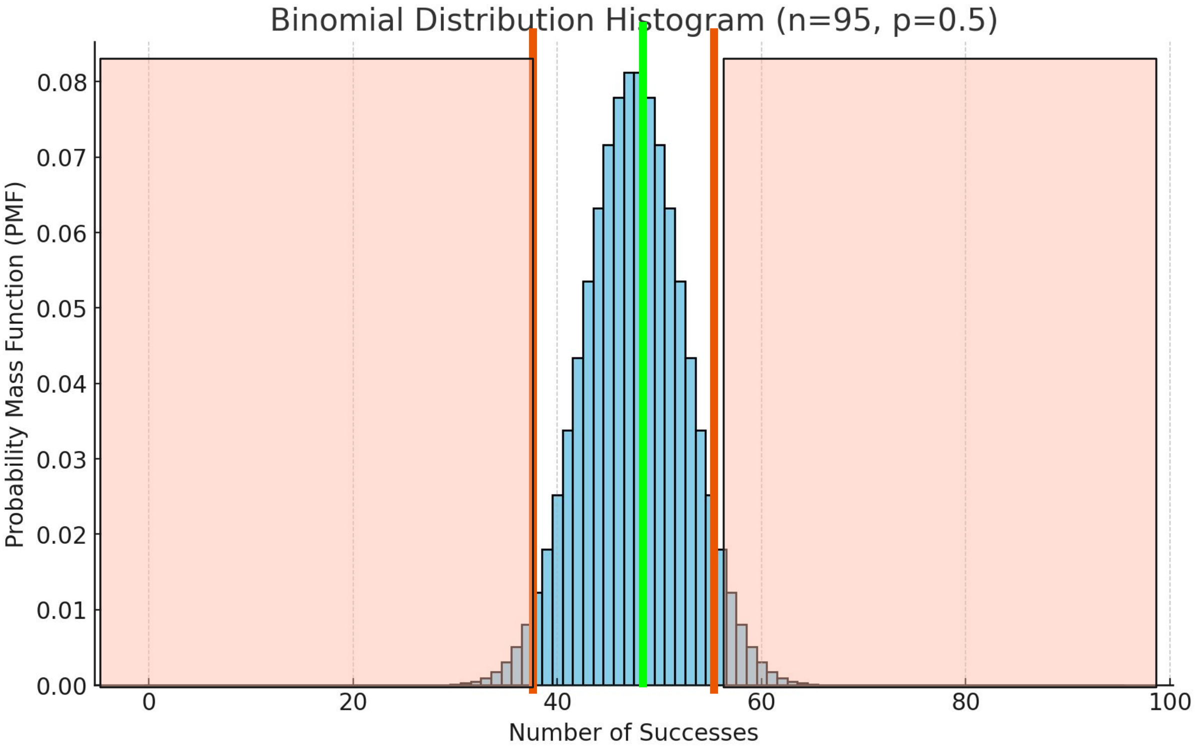


This distribution allowed us to identify the **critical regions of our distribution.** If we get a test-metric that falls into one of these areas, we can assuredly **reject** this distribution. The probability that this metric does not belong to this distribution is too high for us to tolerate!



Going back to our experiment, we got test-metric of 49 (*49 out of 95 of you correctly chose the cat*). **Does this fall within our critical region?**

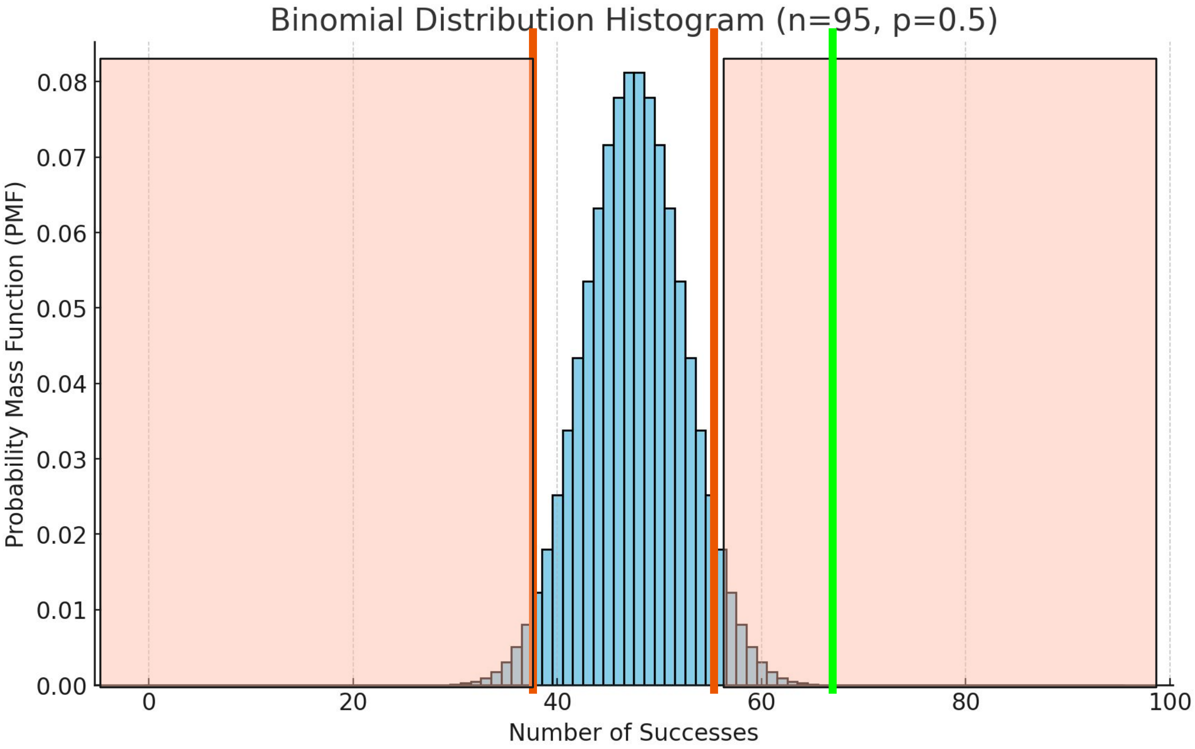
For this test-metric, we get a p-value of 0.83 (83% probability of us getting this test-metric given our null)



No! This means that we **cannot reject this null distribution.** We will see in later slides how this translates to a value we call the **p-value**. This tells us the probability that we will observe this value given the null hypothesis.

However, if we got X=67 we would get a p-value of 0.000039 (0.0039% of getting this metric given the null).

Would you be willing to put money down on something that has a 0.0039% chance of being true?



Notice that as we get a **smaller p-value**, we become **more confident that our null hypothesis is false.** This means that we most likely do have some effect in our dataset!

# We start with nothing (Null)

Let’s review our hypothesis testing terminology

**null hypothesis (H0)** - this is the base case, what our dataset would show if there is **no effect**

**alternative hypothesis (H1) -** this is what we think could be a possibility, this is what the dataset would look like **if there is an effect**

**probability (p) value -** the probability of obtaining the result which occurred assuming the null is true

## Null vs Alternative Hypothesis

Imagine our question about whether **caffeine affects our appetite.**

**H0 (null)**: caffeine **does *not*** have an effect on our appetite

**H1 (alternative)**: caffeine ***does*** have an effect on our appetite

Note, **H1** in this case does not say **decrease** or **increase**, just whether there is an effect period.

Let’s try to formulate our own null and alternative hypotheses.

## Null vs Alternative Hypothesis Examples

Does a new medication improve the recovery time of a patient?

**H0:**

**H1:**

Does the new banner ad we create improve the click-through rate?

**H0:**

**H1:**

# Null vs Alternative Hypothesis Examples

Does a new medication improve the recovery time of a patient?

**H0**: The new medication **does not** improve recovery time ($\mu0 = \mu1$)

**H1**: The new medication **does** improve the recovery time of a patient ($\mu1 < \mu0$)

Does the new banner ad we create improve the click-through rate?

**H0**: The new banner **does not** improve click-through rate ($\mu0 = \mu1$)

**H1**:The new banner **does** improve click-through rate ($\mu1 > \mu0$)

# p-values

A small p-value gives us a level of **confidence** that differences between **two groups** are due to **differences** between the groups instead of just **random chance**

*generally* we use a value of **$p < 0.05$** as a cut-off for ***rejecting the null hypothesis***

However, some researchers (especially in medicine) argue for even stricter p-values (0.005, 0.0005).

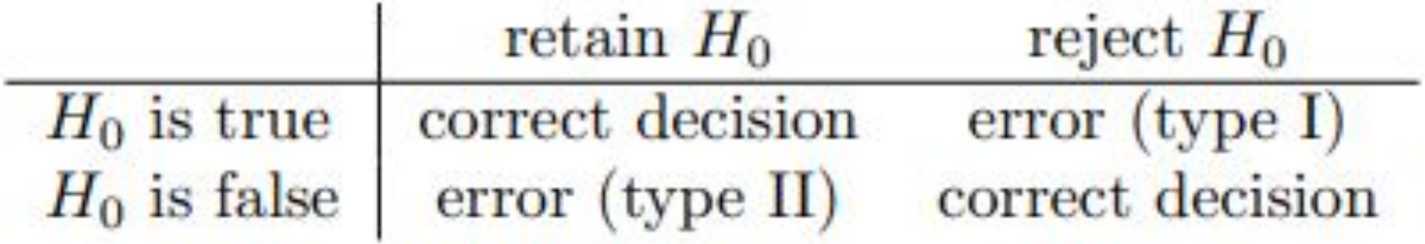
# p-values - Type of Error

If we **reject** the null hypothesis and are **wrong**, this is a **false positive (Type I error)**

If we **do not reject** the null hypothesis and are wrong, this is a **false negative (Type II error)**

We use a small p-value to reduce **false positives (Type I error)**

Reducing **Type II error** involves **increasing** the sample size (aka increasing the ***power*** of your dataset).



|  | 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**.

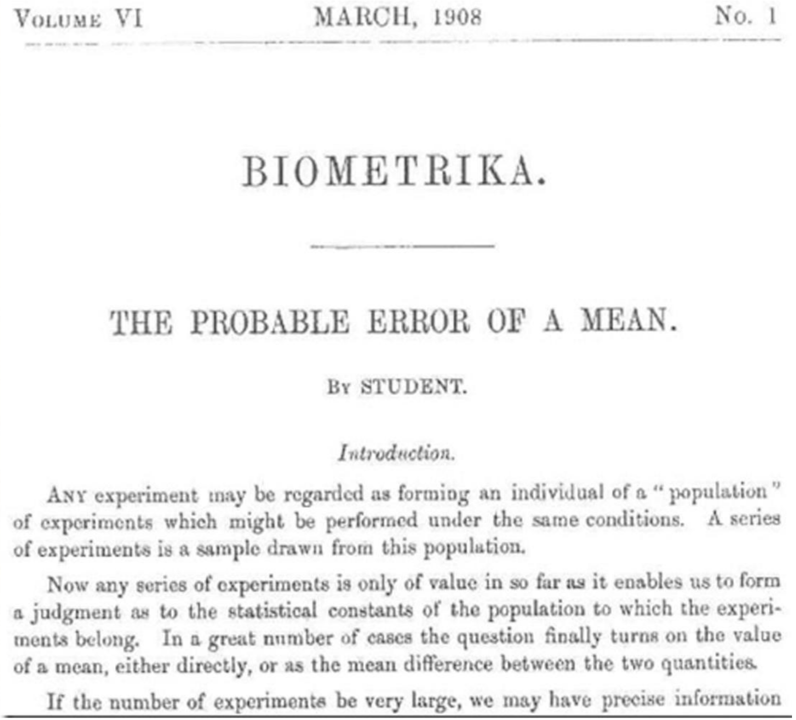
# Two Sample T-Test

## calculating p-values

To calculate p-values, we will choose a few different methods depending on what we are researching:

- **T-test**: used to compare the averages between 2 groups
- **ANOVA**: used to compare the averages between more than 2 groups
- **Chi-Square Test**: used to check for dependence between **categorical** variables

We will discuss the **T-test** today and how we apply this concept to performing **AB tests**.


Volume VI MARCH, 1908 No. 1

BIOMETRIKA.

THE PROBABLE ERROR OF A MEAN.

By STUDENT.

*Introduction.*

Any experiment may be regarded as forming an individual of a "population" of experiments which might be performed under the same conditions. A series of experiments is a sample drawn from this population.

Now any series of experiments is only of value in so far as it enables us to form a judgment as to the statistical constants of the population to which the experiments belong. In a great number of cases the question finally turns on the value of a mean, either directly, or as the mean difference between the two quantities.

If the number of experiments be very large, we may have precise information

Fun fact, the t-test was invented by **William Sealy Gosset** to compare yeast cells between batches of Guinness beer. Never underestimate human capacity for inventing mathematical formulas for the purpose of drinking, gambling, video games, and other vices.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$s^2 = \frac{\sum_{i=1}^{n_1}(x_i - \bar{x}_1)^2 + \sum_{j=1}^{n_2}(x_j - \bar{x}_2)^2}{n_1 + n_2 - 2}$$

This is the formula which **William Sealy Gosset** developed to calculate if there is a **significant difference** between groups. There’s a lot going on here at once, so let’s break down  these steps.

# Two Sample T-Test

The formula describes the steps below:

1. Calculate the **mean** for both groups and take the **difference** between the means for each group
2. Calculate the **pooled variance**
3. Multiple the **pooled variance** with **harmonic mean of group sizes,** take the **square root**
4. **Divide** the difference between means, with the result of the previous step
5. Calculate the degrees of freedom and define the critical value **alpha**
6. Find the p-value based on whether you care about **one tail** (one direction) or **two tail** (either direction)

It might seem overwhelming, but let’s walk through each step!

| Group A | Group B |
|---|---|
| 5 | 1 |
| 4 | 3 |
| 5 | 2 |
| 6 | 3 |
| 3 | 2 |

Let’s say we made a change to the TKH website and we want to see if our users spend **more or less** time on our website due to this change. We will give **group A the original TKH website**, and **group B the new version**. We record the number of minutes they spend on our site.

| Group A | Group B |
|---|---|
| 5 | 1 |
| 4 | 3 |
| 5 | 2 |
| 6 | 3 |
| 3 | 2 |

First we iron out our **null and alternative hypotheses**. Our **null** will state: both groups will spend an **equal amount** of time on the site. Our **alternative** will state: group A and group B will spend an **unequal amount of time** on the site (more or less). **What will be our first step?**

| Group A | Group B |
|---|---|
| 5 | 1 |
| 4 | 3 |
| 5 | 2 |
| 6 | 3 |
| 3 | 2 |

Let’s calculate the mean between groups: Group A spends an average of **4.6 minutes on our website**. Group B spends an average of **2.2 minutes on the site.** **Obviously there appears to be a difference, but is this difference actually due to this change or are we observing random variance in our data?**

$$s^2 = \frac{\sum_{i=1}^{n_1}(x_i - \bar{x}_1)^2 + \sum_{j=1}^{n_2}(x_j - \bar{x}_2)^2}{n_1 + n_2 - 2}$$

Numerator = (5-4.6)^2 + (4-4.6)^2 + (5-4.6)^2 + (6-4.6)^2+ (3-4.6)^2 + (1-2.2)^2 + (3-2.2)^2 + (2-2.2)^2 + (3-2.2)^2 + (2-2.2)^2
= 5.2 + 2.8
= 8

Denominator =5 + 5 - 2
= 10 - 2
= 8

s^2 = 8/8
= 1

Next, we calculate **pooled variance.** Notice that is simply a **modified variance formula** that takes into account the variance of both groups (hence the name pooled).

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

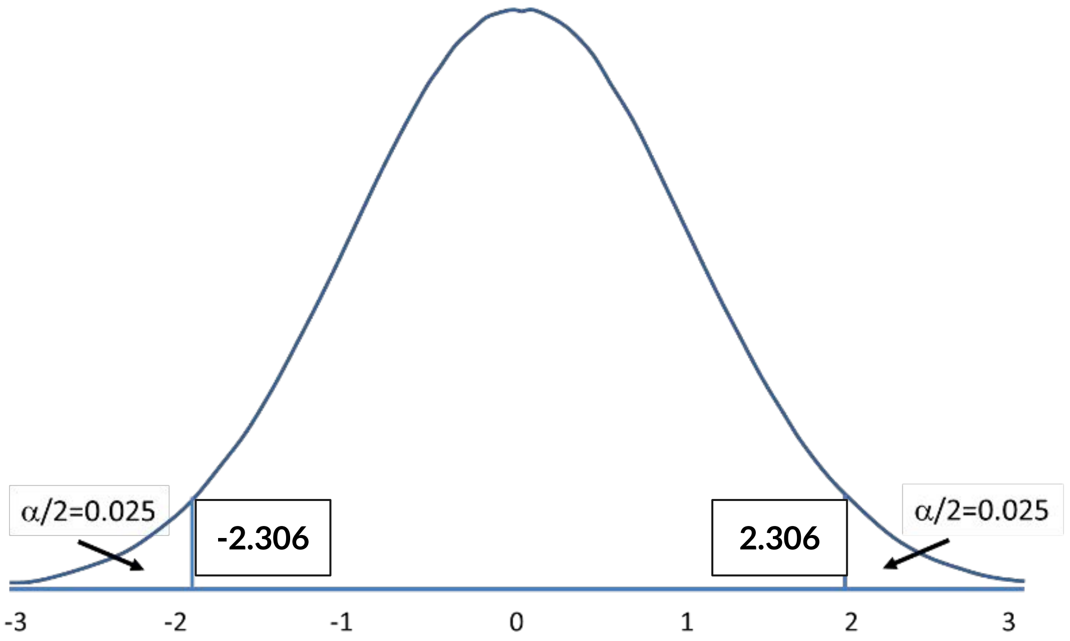
Denominator = 1 ( ⅕ + ⅕)
= 1 * 0.4
= 0.4

**t** = 4.6 - 2.2/ (sqrt(0.4))
= 2.4 / 0.632 = 3.79

So we are given a final t-score of **3.79**

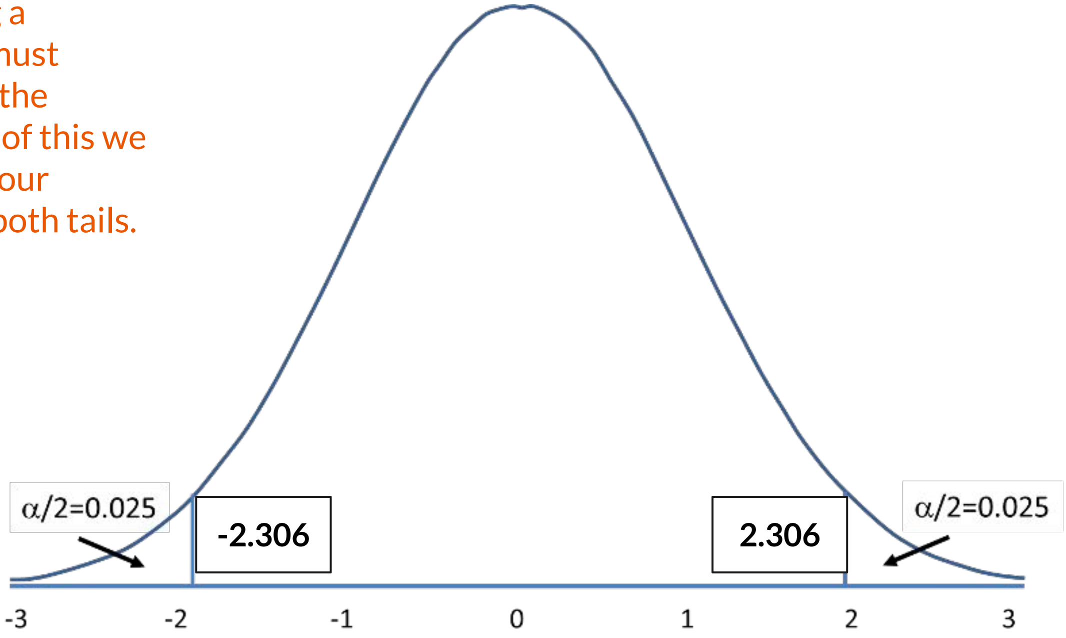
**What does that mean?**

Next we multiply the pooled variance with the **harmonic mean of sample sizes**. While this sounds like a high-minded formula, just keep in mind that this is the reciprocal of each groups sample size (1/n) added together.



The t-score by itself is meaningless, we have to first decide at what value are we willing to reject the null hypothesis, this is known as the “**alpha value**”

Because we are doing a two-sided t-test, we must consider both tails of the distribution. Because of this we use 0.05/2 (0.025) as our significance level for both tails.

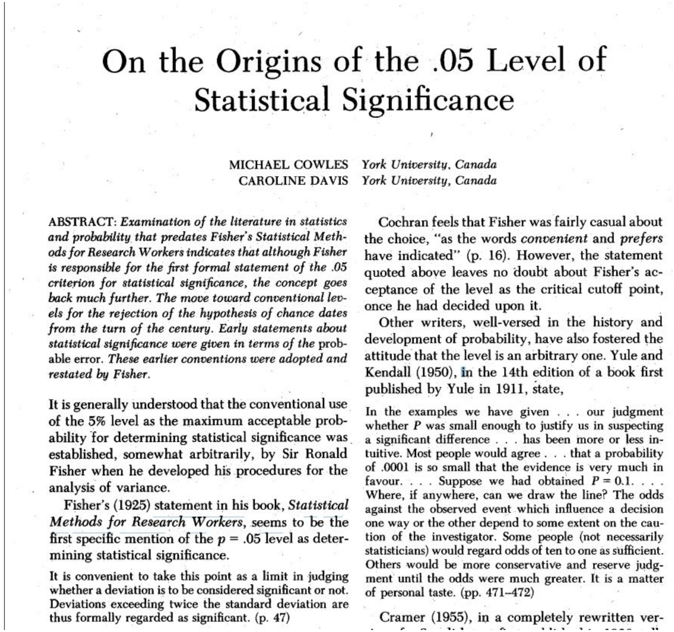


The exact t-score value needed to “beat” **the alpha** changes depending on the size of your dataset and the alpha score.  For this **two-sided t-test**, and an **alpha of 0.05 the t-score should be at least 2.306**

# An Aside - P-value

We usually state that our alpha (or significance level) is **0.05**. That is, if we calculate a p-value less than this alpha, we state that our p-value is significant and **we must reject the null hypothesis**.

The smaller the alpha, the more **confident that you are not making a Type I error**.



## On the Origins of the .05 Level of Statistical Significance

MICHAEL COWLES *York University, Canada*
CAROLINE DAVIS *York University, Canada*

ABSTRACT: *Examination of the literature in statistics and probability that predates Fisher's Statistical Methods for Research Workers indicates that although Fisher is responsible for the first formal statement of the .05 criterion for statistical significance, the concept goes back much further. The move toward conventional levels for the rejection of the hypothesis of chance dates from the turn of the century. Early statements about statistical significance were given in terms of the* probable error. *These earlier conventions were adopted and restated by Fisher.*

It is generally understood that the conventional use of the 5% level as the maximum acceptable probability for determining statistical significance was established, somewhat arbitrarily, by Sir Ronald Fisher when he developed his procedures for the analysis of variance.

Fisher's (1925) statement in his book, *Statistical Methods for Research Workers*, seems to be the first specific mention of the $p = .05$ level as determining statistical significance.

> It is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not. Deviations exceeding twice the standard deviation are thus formally regarded as significant. (p. 47)

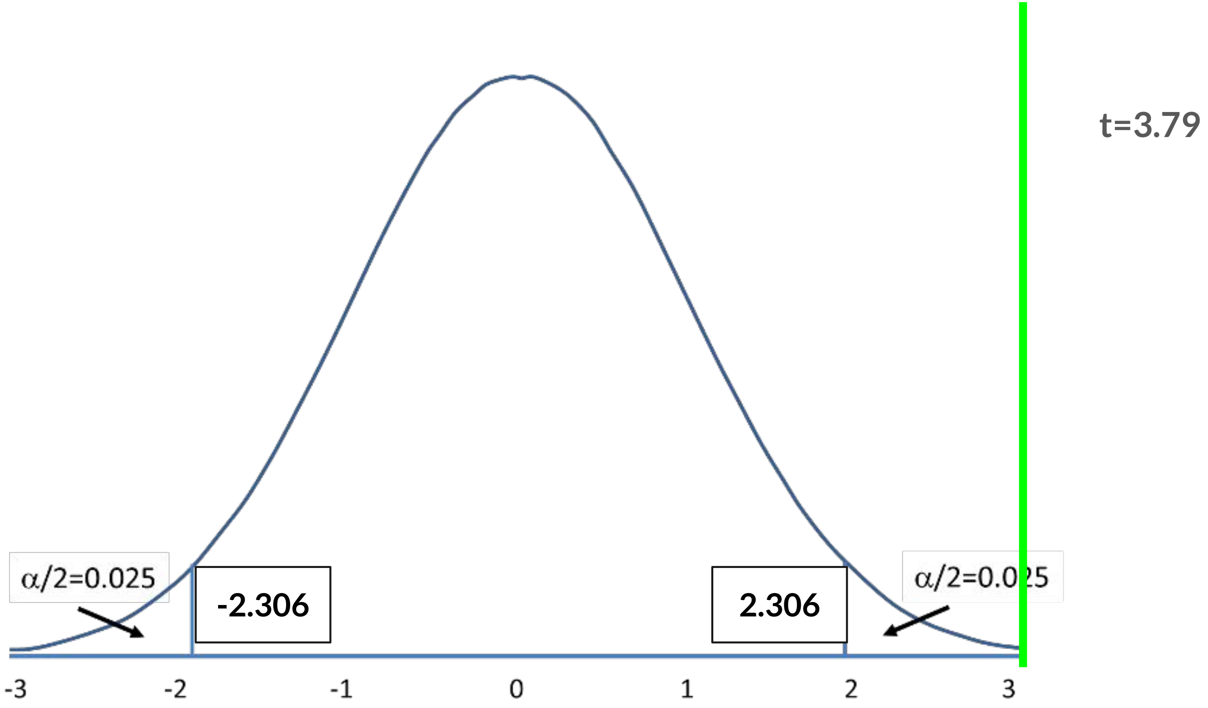
Cochran feels that Fisher was fairly casual about the choice, "as the words *convenient* and *prefers* have indicated" (p. 16). However, the statement quoted above leaves no doubt about Fisher's acceptance of the level as the critical cutoff point, once he had decided upon it.

Other writers, well-versed in the history and development of probability, have also fostered the attitude that the level is an arbitrary one. Yule and Kendall (1950), in the 14th edition of a book first published by Yule in 1911, state,

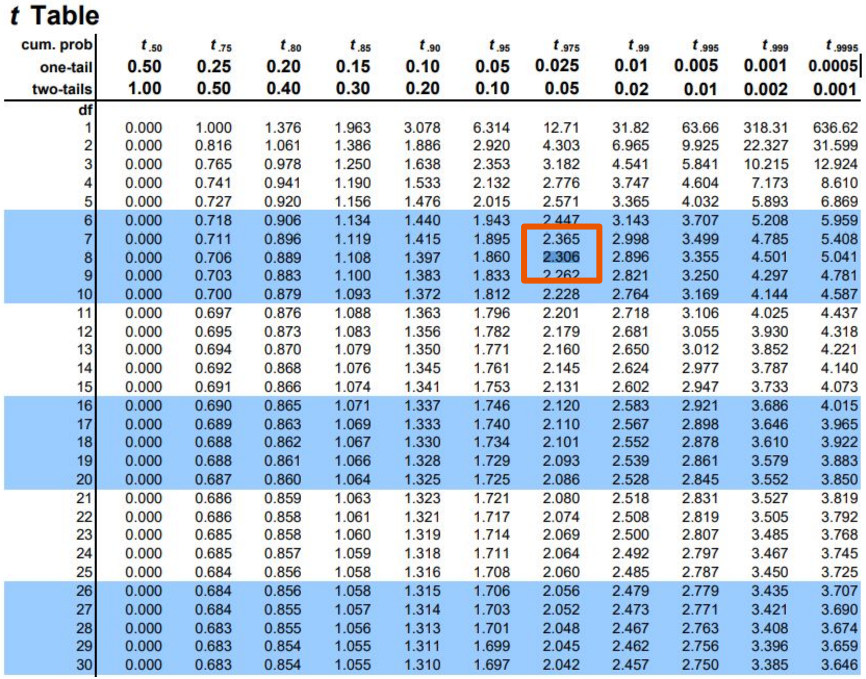
> In the examples we have given . . . our judgment whether $P$ was small enough to justify us in suspecting a significant difference . . . has been more or less intuitive. Most people would agree . . . that a probability of .0001 is so small that the evidence is very much in favour. . . . Suppose we had obtained $P = 0.1$. . . . Where, if anywhere, can we draw the line? The odds against the observed event which influence a decision one way or the other depend to some extent on the caution of the investigator. Some people (not necessarily statisticians) would regard odds of ten to one as sufficient. Others would be more conservative and reserve judgment until the odds were much greater. It is a matter of personal taste. (pp. 471–472)

Cramer (1955), in a completely rewritten ver-

https://www2.psych.ubc.ca/~schaller/528Readings/CowlesDavis1982.pdf



In this case, since our t-score > 2.306 we can immediately state that our p-value is significant (we reject our null hypothesis that these two groups are the same).
**We can state that our p-value is less than 0.5.**



## *t* Table

| cum. prob | $t_{.50}$ | $t_{.75}$ | $t_{.80}$ | $t_{.85}$ | $t_{.90}$ | $t_{.95}$ | $t_{.975}$ | $t_{.99}$ | $t_{.995}$ | $t_{.999}$ | $t_{.9995}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| one-tail | 0.50 | 0.25 | 0.20 | 0.15 | 0.10 | 0.05 | 0.025 | 0.01 | 0.005 | 0.001 | 0.0005 |
| two-tails | 1.00 | 0.50 | 0.40 | 0.30 | 0.20 | 0.10 | 0.05 | 0.02 | 0.01 | 0.002 | 0.001 |
| df | | | | | | | | | | | |
| 1 | 0.000 | 1.000 | 1.376 | 1.963 | 3.078 | 6.314 | 12.71 | 31.82 | 63.66 | 318.31 | 636.62 |
| 2 | 0.000 | 0.816 | 1.061 | 1.386 | 1.886 | 2.920 | 4.303 | 6.965 | 9.925 | 22.327 | 31.599 |
| 3 | 0.000 | 0.765 | 0.978 | 1.250 | 1.638 | 2.353 | 3.182 | 4.541 | 5.841 | 10.215 | 12.924 |
| 4 | 0.000 | 0.741 | 0.941 | 1.190 | 1.533 | 2.132 | 2.776 | 3.747 | 4.604 | 7.173 | 8.610 |
| 5 | 0.000 | 0.727 | 0.920 | 1.156 | 1.476 | 2.015 | 2.571 | 3.365 | 4.032 | 5.893 | 6.869 |
| 6 | 0.000 | 0.718 | 0.906 | 1.134 | 1.440 | 1.943 | 2.447 | 3.143 | 3.707 | 5.208 | 5.959 |
| 7 | 0.000 | 0.711 | 0.896 | 1.119 | 1.415 | 1.895 | 2.365 | 2.998 | 3.499 | 4.785 | 5.408 |
| 8 | 0.000 | 0.706 | 0.889 | 1.108 | 1.397 | 1.860 | 2.306 | 2.896 | 3.355 | 4.501 | 5.041 |
| 9 | 0.000 | 0.703 | 0.883 | 1.100 | 1.383 | 1.833 | 2.262 | 2.821 | 3.250 | 4.297 | 4.781 |
| 10 | 0.000 | 0.700 | 0.879 | 1.093 | 1.372 | 1.812 | 2.228 | 2.764 | 3.169 | 4.144 | 4.587 |
| 11 | 0.000 | 0.697 | 0.876 | 1.088 | 1.363 | 1.796 | 2.201 | 2.718 | 3.106 | 4.025 | 4.437 |
| 12 | 0.000 | 0.695 | 0.873 | 1.083 | 1.356 | 1.782 | 2.179 | 2.681 | 3.055 | 3.930 | 4.318 |
| 13 | 0.000 | 0.694 | 0.870 | 1.079 | 1.350 | 1.771 | 2.160 | 2.650 | 3.012 | 3.852 | 4.221 |
| 14 | 0.000 | 0.692 | 0.868 | 1.076 | 1.345 | 1.761 | 2.145 | 2.624 | 2.977 | 3.787 | 4.140 |
| 15 | 0.000 | 0.691 | 0.866 | 1.074 | 1.341 | 1.753 | 2.131 | 2.602 | 2.947 | 3.733 | 4.073 |
| 16 | 0.000 | 0.690 | 0.865 | 1.071 | 1.337 | 1.746 | 2.120 | 2.583 | 2.921 | 3.686 | 4.015 |
| 17 | 0.000 | 0.689 | 0.863 | 1.069 | 1.333 | 1.740 | 2.110 | 2.567 | 2.898 | 3.646 | 3.965 |
| 18 | 0.000 | 0.688 | 0.862 | 1.067 | 1.330 | 1.734 | 2.101 | 2.552 | 2.878 | 3.610 | 3.922 |
| 19 | 0.000 | 0.688 | 0.861 | 1.066 | 1.328 | 1.729 | 2.093 | 2.539 | 2.861 | 3.579 | 3.883 |
| 20 | 0.000 | 0.687 | 0.860 | 1.064 | 1.325 | 1.725 | 2.086 | 2.528 | 2.845 | 3.552 | 3.850 |
| 21 | 0.000 | 0.686 | 0.859 | 1.063 | 1.323 | 1.721 | 2.080 | 2.518 | 2.831 | 3.527 | 3.819 |
| 22 | 0.000 | 0.686 | 0.858 | 1.061 | 1.321 | 1.717 | 2.074 | 2.508 | 2.819 | 3.505 | 3.792 |
| 23 | 0.000 | 0.685 | 0.858 | 1.060 | 1.319 | 1.714 | 2.069 | 2.500 | 2.807 | 3.485 | 3.768 |
| 24 | 0.000 | 0.685 | 0.857 | 1.059 | 1.318 | 1.711 | 2.064 | 2.492 | 2.797 | 3.467 | 3.745 |
| 25 | 0.000 | 0.684 | 0.856 | 1.058 | 1.316 | 1.708 | 2.060 | 2.485 | 2.787 | 3.450 | 3.725 |
| 26 | 0.000 | 0.684 | 0.856 | 1.058 | 1.315 | 1.706 | 2.056 | 2.479 | 2.779 | 3.435 | 3.707 |
| 27 | 0.000 | 0.684 | 0.855 | 1.057 | 1.314 | 1.703 | 2.052 | 2.473 | 2.771 | 3.421 | 3.690 |
| 28 | 0.000 | 0.683 | 0.855 | 1.056 | 1.313 | 1.701 | 2.048 | 2.467 | 2.763 | 3.408 | 3.674 |
| 29 | 0.000 | 0.683 | 0.854 | 1.055 | 1.311 | 1.699 | 2.045 | 2.462 | 2.756 | 3.396 | 3.659 |
| 30 | 0.000 | 0.683 | 0.854 | 1.055 | 1.310 | 1.697 | 2.042 | 2.457 | 2.750 | 3.385 | 3.646 |

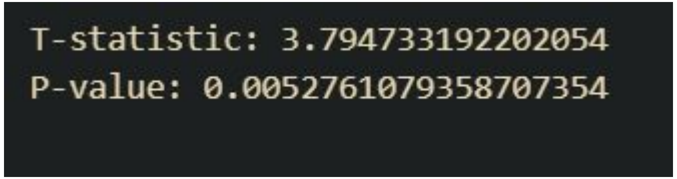
You might be wondering where we got this **critical value from**. Back in "the day" we would look up the critical score table and find what critical score we need to "beat" in order for our t-score to be significant (*and subsequently reject the null hypothesis*)

```
from scipy import stats

group1 = [5, 4, 5, 6, 3]
group2 = [1, 3, 2, 3, 2]

t, p = stats.ttest_ind(group1, group2, alternative='two-sided')


print("T-statistic:", t)
print("P-value:", p)
```



```
T-statistic: 3.794733192202054
P-value: 0.0052761079358707354
```

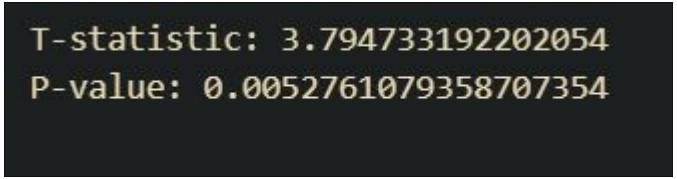
However you are most likely going to find yourself automatically calculating this value via some Python method or through a calculator. **Is our p-value less than our alpha of 0.05?**

```
from scipy import stats

group1 = [5, 4, 5, 6, 3]
group2 = [1, 3, 2, 3, 2]

t, p = stats.ttest_ind(group1, group2, alternative='two-sided')


print("T-statistic:", t)
print("P-value:", p)
```



```
T-statistic: 3.794733192202054
P-value: 0.0052761079358707354
```

However you are most likely going to find yourself automatically calculating this value via some Python method or through a calculator. **Is our p-value less than our alpha of 0.05? Undoubtedly yes!**

| Group A | Group B |
|---|---|
| 5 | 1 |
| 4 | 3 |
| 5 | 2 |
| 6 | 3 |
| 3 | 2 |

**Group A Avg = 4.6** **Group B Avg = 2.2**

So now that we've discovered that the difference between these two groups is significant, we must **reject the null hypothesis that these two groups behave similarly.**

The change in our website appears to have caused a drop in average usage of our website. **Can you think of any downsides to our collected data? Could we improve this experiment?**

# Applications to AB Testing

# AB Testing

What we just walked through is something called **AB Testing.** We do this when we want to **check if a change to our website/product** will lead to better "outcomes."

This is where we split a sample of our users into two groups: the **control** (Group A) and the **experimental** (Group B). We introduce Group B to a new iteration/version of our website or product, while we keep Group A on the same platform.

We then select one specific metric to measure such as **click-through rate, amount purchased, amount spent on website, or some other KPI** which our managers/stakeholders care about.

# AB Testing

Just like in this previous example, we perform a t-test on these two groups to check if the difference **between these groups are statistically significant**.

However, just like we observed, we don’t want to just state that an update to our product is preferable because the **change is significant**.

**What else must we be on the lookout for? Think back to our TKH Website example.**

# AB Testing - Additional Considerations

We must be aware of many things when performing **real-world confirmatory data analysis:**

- In which **direction** is our change (positive or negative)? If this is a beneficial change, is this effect large enough to spend $$$ on rolling out this change?
- Have we run this A/B Test **multiple times** to account for false positives?
- Are we just observing the **novelty effect**?

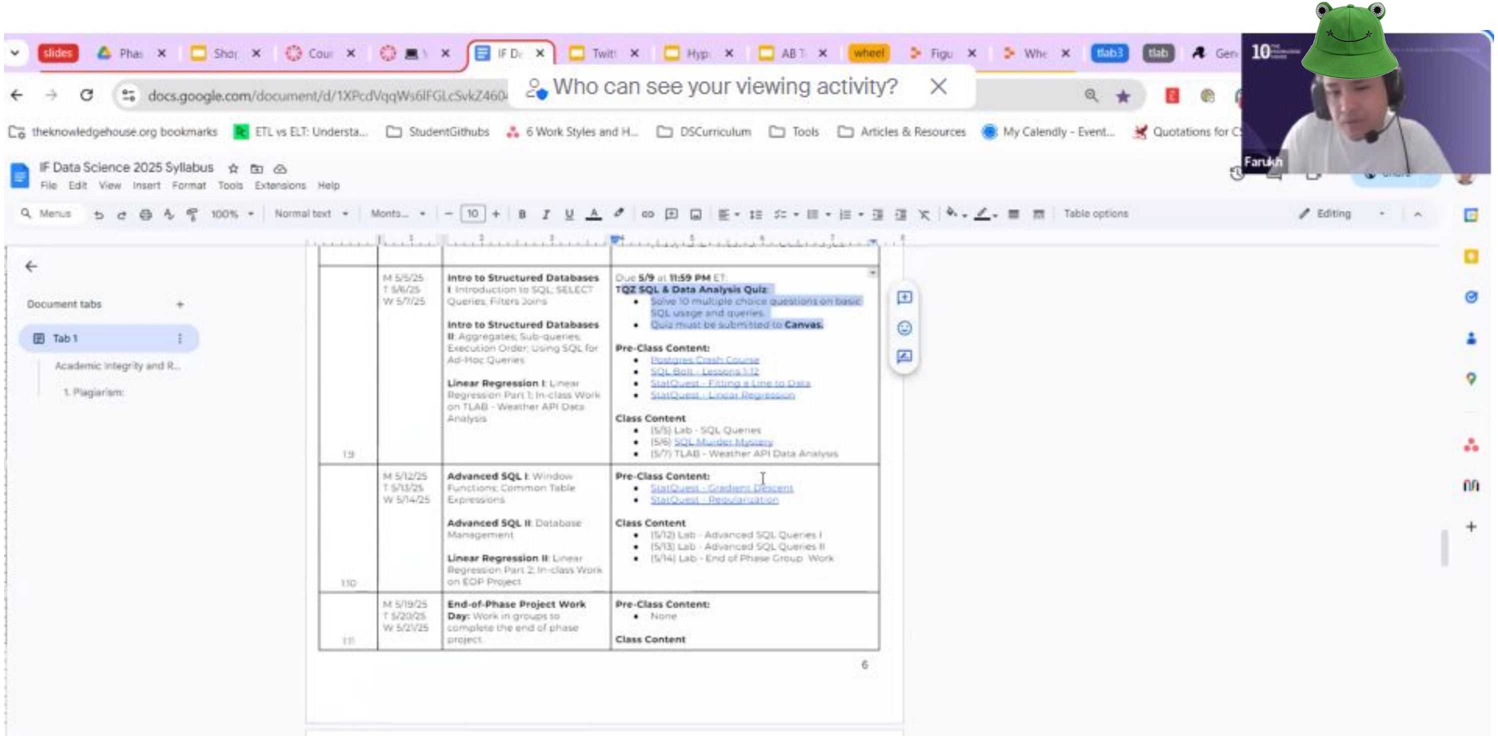
# Novelty Effect

The novelty effect describes a **temporary boost or drop in metrics** simply because something is *new*.

People could just be spending more/exploring our website more because:

- They are getting **used to the new interface** (*confused*)
- A small change introduces some **unpredictability** which they look at **longer** (excited).

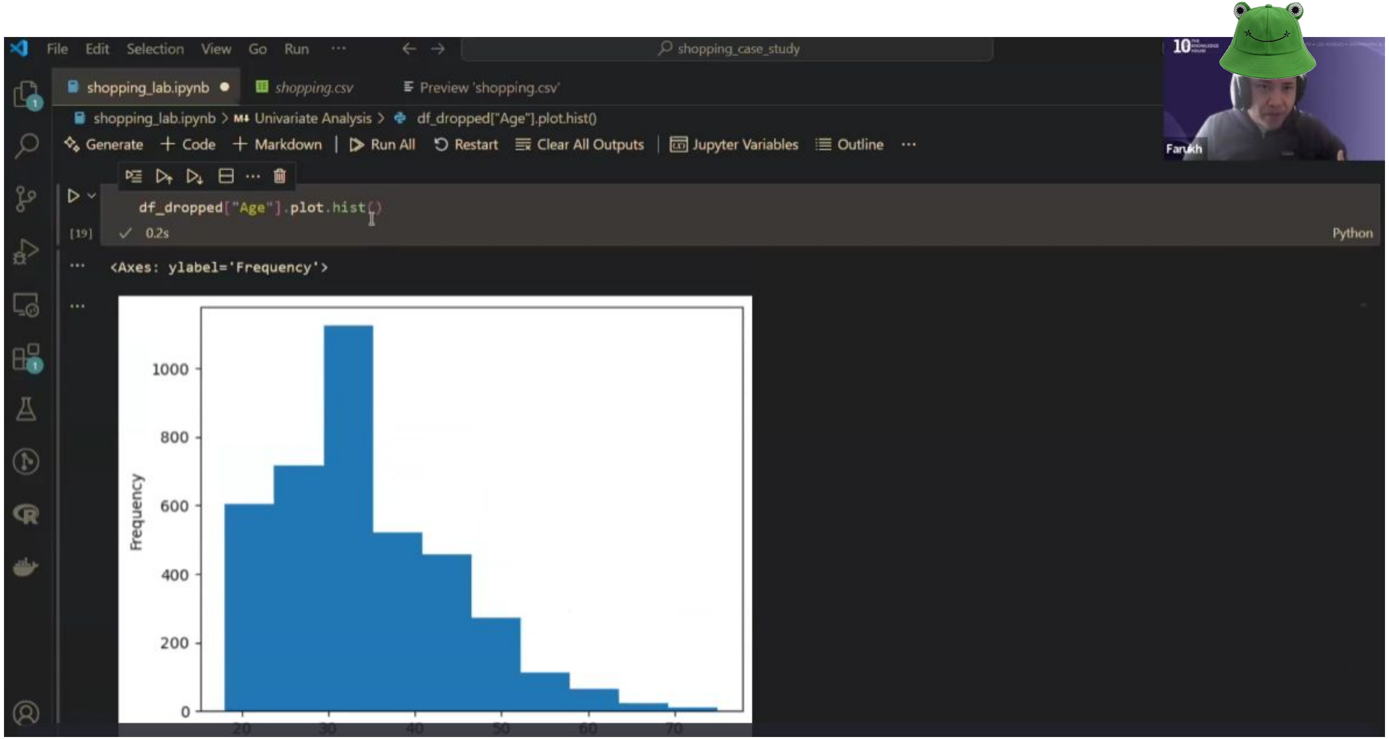
For example, let’s say in order to boost **Zoom engagement…**



| | | | |
|---|---|---|---|
| | | | |
| 1.9 | M 5/5/25<br>T 5/6/25<br>W 5/7/25 | **Intro to Structured Databases I**: Introduction to SQL; SELECT Queries; Filters Joins<br>**Intro to Structured Databases II**: Aggregates; Sub-queries; Execution Order; Using SQL for Ad-Hoc Queries<br>**Linear Regression I**: Linear Regression Part 1; In-class Work on TLAB - Weather API Data Analysis | Due **5/9** at **11:59 PM** ET:<br>**TQZ SQL & Data Analysis Quiz**:<br>• Solve 10 multiple choice questions on basic SQL usage and queries.<br>• Quiz must be submitted to **Canvas.**<br>**Pre-Class Content:**<br>• Postgres Crash Course<br>• SQL Bolt - Lessons 1-12<br>• StatQuest - Fitting a Line to Data<br>• StatQuest - Linear Regression<br>**Class Content**<br>• (5/5) Lab - SQL Queries<br>• (5/6) SQL Murder Mystery<br>• (5/7) TLAB - Weather API Data Analysis |
| 1.10 | M 5/12/25<br>T 5/13/25<br>W 5/14/25 | **Advanced SQL I**: Window Functions; Common Table Expressions<br>**Advanced SQL II**: Database Management<br>**Linear Regression II**: Linear Regression Part 2; In-class Work on EOP Project | **Pre-Class Content:**<br>• StatQuest - Gradient Descent<br>• StatQuest - Regularization<br>**Class Content**<br>• (5/12) Lab - Advanced SQL Queries I<br>• (5/13) Lab - Advanced SQL Queries II<br>• (5/14) Lab - End of Phase Group Work |
| 1.11 | M 5/19/25<br>T 5/20/25<br>W 5/21/25 | **End-of-Phase Project Work Day:** Work in groups to complete the end of phase project. | **Pre-Class Content:**<br>• None<br>**Class Content** |

6

… we make staff wear **fun hats** during the Zoom lecture. At first you guys might say to yourselves *“Woah whats this guys deal? I wonder what else he has up his sleeve.”* and we get a **massive uptick** in Zoom chat participation.



```
df_dropped["Age"].plot.hist()
```

```
<Axes: ylabel='Frequency'>
```

I might be tempted to bring this to my managers and state "Look, fun hats boost engagement!" But after the second week, everyone gets **bored** of the fun hats and **Zoom chats drops down to previous levels.** TKH's investment of **$45 (4.99 * 9 staff)** into the fun-hat fund has just been wasted.

# Novelty Effect - Patches

To offset the novelty effect, we might want to consider:

- Run the test for a **long-enough time** to offset novelty
- **Segment earlier results** to check which results reflect **novelty**

Talk is cheap though, **and experience is AB testing is only done through AB testing**.

We might try to run our own AB test in Phase 2…

# TLAB #3

# Lab (Due 5/14)

Please ensure that your EDA is completed **independently**.

*Minas Gerais, Brazil*

You are a data engineer at a Brazil-based weather prediction startup called Curu-Sight. The goal of this startup is to analyze weather trends in Brazil and predict the output of non-durable consumer goods at harvest time.

You will analyze a dataset that contains averages calculated based on rainfall, temperature, humidity, and wind metrics collected during the coffee growing season.

You will also analyze a dataset that contains Minas Gerais' crop output. **You will then combine these two datasets to explore how the weather influences coffee growth.**

# Thursday

On Thursday we will be meeting for our review session. We will:

- Review GitHub
- Review TLAB #3
- Work together on TLAB #3

Remember! Make this EDA your own. You can discuss **findings & documentation**, but you **cannot share code**.