AB Testing

Flash Cards

- https://apps.ankiweb.net/

Hypothesis Testing

Some Context

The Dawn of Scientific Reasoning

The story of how we came to understand the world around us is a long and fascinating one. It's a story of brilliant minds, bold experiments, and sometimes, surprising discoveries.

One of the earliest heroes of this story is **Galileo Galilei** (1564-1642). Often called the "father of experiments," Galileo insisted on testing his ideas through careful observation and measurement. Unlike many of his contemporaries, who relied on abstract theories and philosophical arguments, Galileo grounded his hypotheses in the real world. He used **deductive reasoning** to test his ideas, making predictions based on his theories and then seeing if those predictions held up in experiments. This approach revolutionized the study of physics and astronomy.

But deductive reasoning alone wasn't enough. **Francis Bacon** (1561-1626), writing in 1610, argued that scientists also needed to conduct experiments. He emphasized the importance of **replication**, the idea that an experiment should be repeatable under similar conditions. This was a crucial step in ensuring the reliability of scientific findings.

Isaac Newton (1643-1727), building on the work of Galileo and Bacon, further refined the scientific method. In 1721, he proposed that scientific laws should be based on data and should be considered accurate until new evidence comes along to challenge them. This emphasis on empirical evidence and the possibility of revision is a cornerstone of modern science.

A Curious Case of Boys and Girls

In the early 18th century, **John Arbuthnot** (1667-1735), a physician to Queen Anne, stumbled upon a curious puzzle. He wanted to know if more boys were born than girls. Examining baptismal records from 1629 to 1710, he found that indeed, more boys had been baptized.

Arbuthnot, assuming an equal chance of having a boy or a girl, realized that the odds of getting an excess of boys every year for 81 years in a row were astronomically small. It was like flipping a coin 81 times and getting heads every time!

He concluded that something must be influencing the sex ratio at birth. He initially suspected that the higher mortality rate among males might be a factor, but the data didn't support this. Arbuthnot, a man of faith, attributed the phenomenon to a "wise Creator" who ensured the survival of the human race by "bringing forth more Males than Females."

The Legacy of Arbuthnot's Discovery

Arbuthnot's work, though driven by religious beliefs, is considered a pioneering example of **hypothesis testing** or **significance testing**. He was essentially trying to determine if the observed difference in the number of boys and girls was due to chance or some other factor.

Modern studies have confirmed Arbuthnot's observation. The sex ratio at birth is slightly skewed towards males, with approximately 21 boys born for every 20 girls. The exact reasons for this are still being investigated, but it's clear that Arbuthnot's simple question about boys and girls opened up a whole new world of scientific inquiry.

Gauss and Measurement Error

In the 1800s, scientists started to formally recognize and study **measurement error**. This led to some important developments:

- **Gauss and the Normal Distribution:** Carl Friedrich Gauss connected the idea of measurement error to the normal distribution (bell curve), which was initially called the "error distribution." This showed how errors are often distributed around a true value.
- **Pierce and Outliers:** Benjamin Pierce used the normal distribution to identify outliers—measurements that fall far outside the expected range and might be due to errors.
- **Gosset and Fisher: Experiments and Variability (Beer and Cigarettes):** William Gosset (known as "Student") and Ronald Fisher extended these ideas to analyze variability in experiments. This laid the foundation for modern statistical methods used to test hypotheses and draw conclusions from data.

Essentially, these scientists helped us understand that **variability and error are inherent in measurement and experimentation**. They developed tools and techniques to account for this, allowing us to make more accurate inferences and draw more reliable conclusions from data.

Fisher, Smoking, and Cancer

The 1950s and 60s saw a fierce debate around the link between smoking and cancer. While evidence mounted suggesting a strong association, the lack of randomized controlled trials (ethically impossible in this case) allowed skeptics, notably statistician Ronald Fisher, to question the causal relationship.

Fisher, potentially influenced by his ties to the tobacco industry, argued that the observed association was spurious. This stance, despite growing evidence to the contrary, provided ammunition for cigarette companies to deny the harmful effects of their products.

By the late 50s, expert consensus solidified around smoking as a cause of cancer. However, the tobacco industry continued to exploit the uncertainty, fueled by Fisher's arguments, to protect their interests.

The formation of an advisory committee in the 60s further solidified the link, with many members personally affected by the findings and quitting smoking themselves. This era highlights the complex interplay of science, industry influence, and public perception in shaping our understanding of health risks.

The Normal Curve

```
# # the normal curve
# import numpy as np
# import matplotlib.pyplot as plt
# import scipy.stats as stats

# X = stats.norm.rvs(size=100000)
# X.sort()

# parameters = stats.norm.fit(X)
# mu, sigma = parameters

# fig, ax = plt.subplots()
# ax.plot(X, stats.norm.pdf(X, loc=mu, scale=sigma))
# ax.hist(X, bins=25, density=True, color='lightgray', alpha=0.5)
# ax.set_xlabel('x')
# ax.set_ylabel('pdf(x)')
# ax.set_xlim(-4, 4)
# ax.grid(True)
```

```
# plt.vlines(x=-3, ymin=0, ymax=stats.norm.pdf(-3, loc=mu, scale=sigma), color=
# plt.vlines(x=-2, ymin=0, ymax=stats.norm.pdf(-2, loc=mu, scale=sigma), color=
# plt.vlines(x=-1, ymin=0, ymax=stats.norm.pdf(-1, loc=mu, scale=sigma), color=
# plt.vlines(x=0, ymin=0, ymax=stats.norm.pdf(0, loc=mu, scale=sigma), color='b
# plt.vlines(x=1, ymin=0, ymax=stats.norm.pdf(1, loc=mu, scale=sigma), color='g
# plt.vlines(x=2, ymin=0, ymax=stats.norm.pdf(2, loc=mu, scale=sigma), color='r
# plt.vlines(x=3, ymin=0, ymax=stats.norm.pdf(3, loc=mu, scale=sigma), color='b
# # plt.axvline(np.percentile(X, 2.5), color='black')
# # plt.axvline(np.percentile(X, 97.5), color='black')

# plt.title('Normal Curve with 68-95-99 rule')
# plt.show()
```

Experimental Design

Experimental design in the context of **hypothesis testing** is the **structured, systematic framework** or plan used to collect (sample) and analyze data. Its primary purpose is to discover relationships between variables under controlled conditions, thereby providing the evidence needed to statistically test a **null hypothesis (H_0)** against an **alternative hypothesis (H_a)**.

In essence, a sound experimental design ensures that when you run your experiment, any observed effect is genuinely due to the factor you are manipulating (the independent variable) and not to confounding or extraneous variables. This control is critical for the validity of the hypothesis test.

Key Components and Role in Hypothesis Testing

Experimental design is vital because it determines how the data will be gathered, which directly influences the statistical tests you can use and the confidence you can place in the conclusion of your hypothesis test.

Component	Definition
Independent Variable (IV)	The factor that the researcher manipulates or controls (the "cause").
Dependent Variable (DV)	The factor that is measured for change (the "effect" or outcome).
Control	Procedures used to minimize the influence of extraneous variables (confounds).
Randomization	Randomly assigning participants/subjects to different experimental groups/treatments.
Replication	The repetition of the experiment's procedures, either within the experiment (e.g., sample

Repeated Random Sampling

1. The Sample Data Distribution

When you collect a single sample of data (the heights of 50 people, for example), the distribution (histogram) of that data will increasingly resemble the **population distribution** as your sample size (n) gets larger. Random sampling helps ensure that the data you collect is **representative** of the population.

---

2. The Sampling Distribution of a Statistic

The most critical impact of a larger sample size is on the **sampling distribution** of a statistic (like the sample mean, $\bar{x}$). The sampling distribution is the theoretical distribution of all possible sample means you could get from the population.

Here's how random sampling and sample size affect the sampling distribution:

- **The Distribution Becomes Narrower (Less Spread):** The **Standard Error** (the standard deviation of the sampling distribution) is calculated as $\frac{\sigma}{\sqrt{n}}$. As the sample size (n) increases, the denominator $\sqrt{n}$ gets larger, making the standard error smaller. This means the distribution of sample means tightens around the true population mean, leading to **greater precision** in your estimate.
- **The Distribution Becomes More Normal (Bell-Shaped):** The **Central Limit Theorem (CLT)** states that for a large enough sample size (n), the sampling distribution of the mean will be **approximately normal**, regardless of the shape of the original population distribution. This allows us to use standard statistical tests.

In short, when a larger sample size is important, random sampling helps by making your estimates **more precise** (narrower distribution) and **more reliable** (more normal distribution).

Monte Carlo

Monte Carlo methods are a broad class of computational algorithms that rely on **repeated random sampling** to obtain numerical results. They are particularly useful for solving problems that are difficult to solve using deterministic (exact) analytical methods, such as integration, optimization, and generating samples from complex probability distributions.

The core idea is: **Use randomness to estimate a deterministic value.**

How Monte Carlo Works (The General Steps)

1. **Define a Domain:** Establish a range of possible inputs (the "space" of your problem).

2. **Generate Samples:** Randomly and uniformly sample points within this domain. The more samples you take, the more accurate the result will be.
3. **Perform Calculation/Test:** Perform a test or calculation on each sample.
4. **Aggregate Results:** Sum or average the results of the individual samples to get the final estimate.

Python Example: Estimating Pi (π)

A classic example of Monte Carlo simulation is estimating the value of π by using random points inside a square that contains a circle.

1. **Define the Domain:** Consider a square that spans from $(-1, -1)$ to $(1, 1)$, with an area of $2 \times 2 = 4$.
2. **Define the Target Area:** Inscribe a quarter-circle (or a full circle) inside this square, centered at the origin $(0, 0)$. A quarter-circle in the first quadrant has a radius of $r = 1$ and an area of $\frac{1}{4}\pi r^2 = \frac{\pi}{4}$.
3. **The Ratio:** The ratio of the area of the quarter-circle to the area of the square is:

$$\frac{\text{Area}_{\text{Circle}}}{\text{Area}_{\text{Square}}} = \frac{\pi/4}{1} = \frac{\pi}{4}$$

4. **The Estimation:** If you randomly throw N darts at the square, the proportion of darts that land inside the circle (Hits/N) should approximate this area ratio ($\pi/4$).

$$\frac{\text{Hits}}{N} \approx \frac{\pi}{4}$$

$$\text{Estimated } \pi \approx 4 \times \frac{\text{Hits}}{N}$$

Key Takeaways

- **Accuracy vs. Samples:** The accuracy of the Monte Carlo estimate generally improves with the square root of the number of samples ($\sqrt{N}$). To double the accuracy, you need to quadruple the number of samples.
- **Applications:** Monte Carlo methods are widely used in finance (option pricing), physics (simulating particle interactions), and machine learning (e.g., Monte Carlo Tree Search in AI).

```
# # monte carlo example
# import random

# def estimate_pi_monte_carlo(N):
#     """
```

```
#     Estimates Pi using the Monte Carlo method.
#     N: The number of random samples (darts) to throw.
#     """
#     points_in_circle = 0
#     total_points = N

#     # 1. Generate Samples (randomly throw 'darts')
#     for _ in range(N):
#         # Generate random x and y coordinates between 0 and 1
#         x = random.uniform(0, 1)
#         y = random.uniform(0, 1)

#         # 2. Perform Calculation/Test: Check if the point is in the quarter-c
#         # The equation for a circle centered at (0,0) is x^2 + y^2 = r^2.
#         # Here, r=1, so we check if x^2 + y^2 <= 1.
#         distance_squared = x**2 + y**2

#         if distance_squared <= 1:
#             points_in_circle += 1

#     # 3. Aggregate Results: Calculate the estimate
#     # Ratio (Hits/N) approx pi/4  =>  pi approx 4 * (Hits/N)
#     estimated_pi = 4 * (points_in_circle / total_points)

#     return estimated_pi

# # Run the simulation with a large number of points
# NUM_SAMPLES = 1_000_000
# pi_estimate = estimate_pi_monte_carlo(NUM_SAMPLES)

# print(f"Number of Samples: {NUM_SAMPLES}")
# print(f"Estimated value of Pi: {pi_estimate}")
# print(f"Difference from math.pi: {abs(pi_estimate - 3.1415926535)}")
```

Bootstrapping

Bootstrapping is a powerful, non-parametric statistical technique used to estimate the **sampling distribution** of a statistic (like the mean, median, or standard deviation) by **sampling with replacement** from the observed data set. It allows you to estimate the variability and shape of the statistic's distribution, often used to construct **confidence intervals**.

How Bootstrapping Works

The core idea is that your single observed sample is the best estimate of the entire population distribution. Therefore, we treat the sample as the "population" and draw

smaller "resamples" from it to understand the sampling variation.

Here are the steps:

1. **Original Sample:** Start with your original dataset of size N.
2. **Resampling:** Draw a new sample (a **bootstrap sample**) of size N from the original data, *with replacement*. Because of replacement, a single data point can appear multiple times, or not at all, in the resample.
3. **Calculate Statistic:** Calculate the statistic of interest (e.g., the mean) for this resample.
4. **Repeat:** Repeat steps 2 and 3 a large number of times (e.g., 1,000 to 10,000 times) to create a distribution of the calculated statistic.
5. **Inference:** Use this **bootstrap distribution** to estimate properties like the bias, standard error, and confidence intervals of the statistic.

Python Example: Estimating a 95% Confidence Interval for the Mean

In the following example, we'll use bootstrapping to estimate a 95% confidence interval for the mean of a small dataset.

The output will give you an interval, for example, **(82.60, 89.20)**. This means we are 95% confident that the true population mean lies somewhere between 82.60 and 89.20.

```
# import numpy as np

# def bootstrap_confidence_interval(data, num_resamples, confidence_level):
#     """
#     Estimates a confidence interval for the mean using the bootstrapping meth
#     """
#     sample_size = len(data)
#     bootstrap_means = []

#     # 1. & 2. Resampling and Calculation
#     for _ in range(num_resamples):
#         # Create a bootstrap sample by sampling with replacement
#         # np.random.choice returns a random sample from the array
#         bootstrap_sample = np.random.choice(data, size=sample_size, replace=T

#         # Calculate the statistic (mean) for the bootstrap sample
#         bootstrap_mean = np.mean(bootstrap_sample)
#         bootstrap_means.append(bootstrap_mean)

#     # 3. Inference: Calculate the confidence interval

#     # Sort the list of bootstrap means
#     sorted_means = np.sort(bootstrap_means)
```

```
#     # Determine the percentiles for the confidence interval
#     # For a 95% CI, the lower bound is the 2.5th percentile, and the upper is
#     alpha = 1 - confidence_level # e.g., 0.05 for 95%
#     lower_percentile = alpha / 2  # 0.025
#     upper_percentile = 1 - (alpha / 2) # 0.975

#     # Use the numpy percentile function
#     lower_bound = np.percentile(sorted_means, lower_percentile * 100)
#     upper_bound = np.percentile(sorted_means, upper_percentile * 100)

#     return lower_bound, upper_bound, np.mean(data)

# # --- Example Usage ---
# # Original sample data (e.g., test scores)
# data = np.array([78, 85, 92, 88, 95, 80, 75, 90, 83, 86])

# # Parameters for the simulation
# NUM_RESAMPLES = 5000
# CONFIDENCE_LEVEL = 0.95

# lower, upper, original_mean = bootstrap_confidence_interval(
#     data, NUM_RESAMPLES, CONFIDENCE_LEVEL
# )

# print(f"Original Sample Mean: {original_mean:.2f}")
# print(f"Number of Bootstrap Resamples: {NUM_RESAMPLES}")
# print(f"{int(CONFIDENCE_LEVEL*100)}% Confidence Interval for the Mean: ({lowe
```

Test of Means

The test of means is a specific type of hypothesis test. It's used when you want to compare the means of two or more groups to see if there's a statistically significant difference between them.

Here's a breakdown of how they connect:

Hypothesis Testing Framework:

- **Formulate Hypotheses:**
  - Null hypothesis (H0): States that there's no difference between the means of the groups.
  - Alternative hypothesis (Ha): States that there is a difference between the means.
- **Collect Data:** Gather data from samples representing each group.
- **Choose a Test:** Select an appropriate test of means based on the nature of your data and research question (e.g., t-test, ANOVA).

- **Calculate Test Statistic:** Compute the test statistic, which measures the difference between the sample means relative to the variability within the groups.
- **Determine P-value:** Calculate the probability of observing the obtained results (or more extreme results) if the null hypothesis were true.
- **Make a Decision:**
    - If the p-value is less than your significance level (alpha), reject the null hypothesis and conclude that there is a statistically significant difference between the means.
    - If the p-value is greater than or equal to alpha, fail to reject the null hypothesis, indicating that there's not enough evidence to support a difference.

Types of Tests of Means:

- **t-test:** Used to compare the means of two groups.
    - Independent samples t-test: For comparing means of two independent groups.
    - Paired samples t-test: For comparing means of two related groups (e.g., before-and-after measurements).
- **ANOVA (Analysis of Variance):** Used to compare the means of three or more groups.

Key Points:

- Tests of means are a subset of hypothesis tests, specifically designed for comparing means.
- The choice of which test of means to use depends on factors like the number of groups being compared, the nature of the data, and the assumptions of the test.
- Hypothesis testing provides the overall framework for conducting tests of means, guiding the decision-making process based on the p-value and significance level.

In essence, tests of means are tools within the broader hypothesis testing framework, providing specific methods for comparing means and drawing conclusions about differences between groups.

Many, if not most experiments are designed to compare means. The experiment may involve only one sample mean that is to be compared to a specific value. Or the experiment could be testing differences among many different experimental conditions, and the experimenter could be interested in comparing each mean with each of the other means.

https://onlinestatbook.com/2/tests_of_means/testing_means.html

Assumptions

- Groups are normally distributed (no significant outliers)
- Groups are independent
- Equal variance between groups
- Rule of thumb: equal variances if the ratio is less than 4 (larger / smaller)
- Scipy.ttest_ind equal_var: if True, perform a standard independent 2 sample t-test that assumes equal population variances. If False, perform Welch's t-test, which does not assume equal population variances. This is True by default.
- Scipy.ttest_ind alternative: two-sided, less, greater

The loc and scale parameters let you adjust the location and scale of a distribution. For example, to model IQ data, you'd build iq = scipy.stats.norm(loc=100, scale=15) because IQs are constructed so as to have a mean of 100 and a standard deviation of 15. Why don't we just call them mean and sd? It helps to have a more generalized concept because not every distribution has a mean.

https://stats.stackexchange.com/questions/560281/what-is-the-meaning-of-loc-and-scale-for-the-distributions-in-scipy-stats

Hypothesis Testing

Hypothesis testing is a statistical method used to determine if there is enough evidence in a sample of data to infer that a certain condition is true for the entire population.

Null and Alternative Hypotheses

- **Null Hypothesis (H_0):** A statement of no effect or no difference. It's what you're trying to disprove.
- **Alternative Hypothesis (H_a or H1):** A statement that contradicts the null hypothesis, proposing the effect or difference you're interested in.

Types of Tests

- **One-Tailed Test:** Used when you have a directional hypothesis (e.g., you expect one group to be greater than or less than the other).
    - Symbols: ≤ (less than or equal to) or ≥ (greater than or equal to) for the null hypothesis, and < (less than) or > (greater than) for the alternative hypothesis.
- **Two-Tailed Test:** Used when you're looking for a difference in either direction (greater than or less than).

- Symbols: = (equals) for the null hypothesis, and ≠ (not equals) for the alternative hypothesis.

Examples

- **One-Tailed:**
  - H_0: The average weight of apples in Group A is greater than or equal to the average weight of apples in Group B ($\bar{x}1 \geq \bar{x}2$).
  - H_a: The average weight of apples in Group A is less than the average weight of apples in Group B ($\bar{x}1 < \bar{x}2$).
- **Two-Tailed:**
  - H_0: The average height of men is equal to the average height of women ($\bar{x}1 = \bar{x}2$).
  - H_a: The average height of men is not equal to the average height of women ($\bar{x}1 \neq \bar{x}2$).

Important Notes

- **Decision:** In hypothesis testing, you either *reject* the null hypothesis (if there's enough evidence against it) or *fail to reject* it (if there's not enough evidence). You never "accept" the null hypothesis.
- **Business vs. Academia:** In business contexts, like A/B testing, the focus is often on finding practical, meaningful differences that drive success. Academic research may explore more subtle differences that might not have immediate business implications.

Key takeaway: Hypothesis testing helps you use data to make informed decisions about whether your observed results are likely due to chance or a real effect.

The Null Hypothesis

In scientific research, the null hypothesis (often denoted H_0) is the claim that the effect being studied does not exist. Note that the term "effect" here is not meant to imply a causative relationship.

The null hypothesis can also be described as the hypothesis in which no relationship exists between two sets of data or variables being analyzed. If the null hypothesis is true, any experimentally observed effect is due to chance alone, hence the term "null". In contrast with the null hypothesis, an alternative hypothesis is developed, which claims that a relationship does exist between two variables.

https://en.wikipedia.org/wiki/Null_hypothesis

https://medium.com/peter-flom-the-blog/should-you-even-do-a-hypothesis-test-c21607d31c4b

This passage describes the core concepts of **hypothesis testing**, a fundamental method in statistics developed in the 1920s. Here's a summary:

What is Hypothesis Testing?

Hypothesis testing is a way to make inferences about a population based on sample data. It involves formulating two competing hypotheses:

- **Null Hypothesis (H_0):** A statement of "no effect" or "no difference." It represents the status quo.
- **Alternative Hypothesis (H_a):** A statement that contradicts the null hypothesis, suggesting an effect or difference.

The Process:

1. **Formulate Hypotheses:** Clearly define the null and alternative hypotheses.
2. **Collect Data:** Gather data through experiments or observations.
3. **Analyze Data:** Calculate a test statistic that measures the difference between the sample data and what's expected under the null hypothesis.
4. **Calculate p-value:** The p-value is the probability of observing results as extreme as those obtained, assuming the null hypothesis is true.
5. **Make a Decision:**
    - If the p-value is below a predetermined significance level (often 0.05), we reject the null hypothesis in favor of the alternative hypothesis.
    - If the p-value is above the significance level, we fail to reject the null hypothesis.

Key Points from the Passage:

- **Hypotheses are Provisional:** Hypotheses are not absolute truths but working assumptions that can be tested.
- **The Null Hypothesis is Negative:** It denies change or effects. We aim to disprove it, not prove it.
- **P-value:** A measure of evidence against the null hypothesis. Smaller p-values indicate stronger evidence.
- **Statistical Significance:** When the p-value is below the significance level, we say the results are statistically significant, suggesting that the observed effect is unlikely due to chance alone.

Types of Tests:

- **One-tailed test:** Used when the alternative hypothesis specifies a direction of effect (e.g., greater than, less than).
- **Two-tailed test:** Used when the alternative hypothesis simply states a difference without specifying a direction.

In essence, hypothesis testing provides a framework for using data to make decisions about whether to reject or fail to reject a claim about a population.

The Alternative Hypothesis

The alternative hypothesis and null hypothesis are types of conjectures used in statistical tests, which are formal methods of reaching conclusions or making judgments on the basis of data. In statistical hypothesis testing, the null hypothesis and alternative hypothesis are two mutually exclusive statements.

"The statement being tested in a test of statistical significance is called the null hypothesis. The test of significance is designed to assess the strength of the evidence against the null hypothesis. Usually, the null hypothesis is a statement of 'no effect' or 'no difference'." Null hypothesis is often denoted as H0.

The statement that is being tested against the null hypothesis is the alternative hypothesis. Alternative hypothesis is often denoted as Ha or H1.

In statistical hypothesis testing, to prove the alternative hypothesis is true, it should be shown that the data is contradictory to the null hypothesis. Namely, there is sufficient evidence against null hypothesis to demonstrate that the alternative hypothesis is true.

https://en.wikipedia.org/wiki/Alternative_hypothesis

This passage delves into the **Neyman-Pearson approach to hypothesis testing**, contrasting it with Fisher's ideas and highlighting the ongoing debate between these two frameworks.

Neyman-Pearson vs. Fisher

While Fisher focused on the p-value as a measure of evidence against the null hypothesis, Neyman and Pearson introduced a decision-making framework with a focus on error rates and power.

Key Concepts in Neyman-Pearson:

- **Alternative Hypothesis:** Explicitly considered as a competing explanation for the data.
- **Decision Making:** The goal is to decide between rejecting the null hypothesis in favor of the alternative or failing to reject the null.

- **Type I Error (False Positive):** Rejecting the null hypothesis when it's actually true.
- **Type II Error (False Negative):** Failing to reject the null hypothesis when the alternative is true.
- **Size of the Test (Alpha):** The probability of making a Type I error, typically pre-specified (e.g., 0.05).
- **Power of the Test (Beta):** The probability of correctly rejecting the null hypothesis when the alternative is true.

The Trade-off:

Neyman and Pearson emphasized the importance of balancing Type I and Type II errors. Increasing power (reducing Type II errors) often comes at the cost of increasing the risk of Type I errors.

The Argument:

Fisher and Neyman-Pearson disagreed on the philosophical underpinnings and practical application of hypothesis testing. Fisher criticized the Neyman-Pearson approach for its focus on decision-making rather than evidence, while Neyman-Pearson argued that Fisher's p-value approach lacked a formal decision rule.

Resolution?

Despite the historical debate, modern practice often combines elements of both approaches. Confidence intervals (influenced by Neyman-Pearson) are used alongside p-values (Fisher's contribution) to provide a more complete picture of the data.

In essence, the Neyman-Pearson approach adds a layer of decision-making to hypothesis testing, considering error rates and power. This complements Fisher's p-value approach, and the combination of these perspectives provides a richer toolkit for statistical inference.

The Hypothesis Feud

The primary **feud with R. A. Fisher** regarding hypothesis tests was a bitter and long-standing disagreement with **Jerzy Neyman and Egon Pearson** over the fundamental **logic and purpose** of statistical testing.

Fisher developed the **test of significance** using the **p-value**, while Neyman and Pearson (N-P) developed the **hypothesis test** (or acceptance procedure), which introduced concepts like the alternative hypothesis, Type I (α) and Type II (β) errors, and statistical power.

Key Differences and Points of Conflict

The core of the dispute was a philosophical one about the goal of a statistical test:

- **Primary Goal:**
  - **Fisher:** The goal was **Scientific Inference**—to measure evidence *against* the null hypothesis (H_0).
  - **Neyman-Pearson:** The goal was **Inductive Behavior/Decision Making**—to provide a rule for making an 'accept' or 'reject' decision over the long run.
- **Hypotheses:**
  - **Fisher:** Only the **null hypothesis** (H_0) is required; the focus is on whether the data is sufficiently improbable under H_0.
  - **Neyman-Pearson:** Requires both a **null hypothesis** (H_0) and an explicit **alternative hypothesis** (H_1).
- **Outcome:**
  - **Fisher:** The **p-value**, which is a continuous measure of evidence.
  - **Neyman-Pearson:** A binary **decision** to "reject H_0" or "not reject H_0," based on a pre-selected significance level (α).
- **Error Rates:**
  - **Fisher:** Concerned with only the risk of rejecting a true H_0 (**Type I error**).
  - **Neyman-Pearson:** Explicitly addresses **two types of errors** (Type I and Type II), with the goal of maximizing statistical **power** ($1 - \beta$).

Fisher's Core Criticism

Fisher adamantly rejected the N-P framework, particularly its rigid decision-making structure, which he saw as inappropriate for scientific research.

He viewed the specification of a precise **alternative hypothesis** and the calculation of **statistical power** as unnecessary distractions from assessing the evidence against H_0. Most significantly, Fisher despised the rigid, automatic, and binary "accept/reject" **decisions** promoted by N-P, famously dismissing the N-P approach as only suitable for repetitive, industrial contexts like quality control, and not for the nuanced, inductive nature of scientific inquiry.

The **Null Hypothesis Significance Testing (NHST)** framework commonly used today is actually an **inconsistent hybrid** of Fisher's *p-value* idea and the N-P *fixed-alpha decision-making* framework, a blend that neither statistician fully endorsed.

Distribution of Sample Means

What are Sampling Distributions?

Imagine taking many random samples from a population and calculating a statistic (like the mean) for each sample. The distribution of all those sample statistics is the **sampling distribution**. It shows how much the statistic varies from sample to sample.

Why are they important for Hypothesis Testing?

- **Understanding Variability:** Sampling distributions help us understand how much our sample statistic might differ from the true population parameter due to random chance. This is crucial in hypothesis testing, where we're trying to infer something about the population based on a sample.
- **Calculating p-values:** The shape and spread of the sampling distribution allow us to calculate p-values. The p-value tells us how likely it is to observe our sample results (or more extreme) if the null hypothesis were true. This helps us decide whether to reject or fail to reject the null hypothesis.
- **Constructing Confidence Intervals:** Sampling distributions are also used to construct confidence intervals, which provide a range of plausible values for the population parameter.

Key Takeaways

- **Bootstrapping:** Modern computing allows us to create sampling distributions empirically using bootstrapping, even without knowing the exact population distribution.
- **Theoretical Distributions:** Probability theory provides formulas for common sampling distributions (like the t-distribution), simplifying calculations.
- **Central Limit Theorem:** As sample size increases, the sampling distribution of the mean tends towards a normal distribution, regardless of the original population distribution. This is the essence of the Central Limit Theorem, previously known as the "Law of Frequency of Error."

Connection to Hypothesis Testing:

In hypothesis testing, we use the sampling distribution to determine if our observed sample statistic is unusual enough to reject the null hypothesis. If our sample statistic falls in the extreme tails of the sampling distribution, it suggests that our results are unlikely to have occurred by chance alone, giving us evidence against the null hypothesis.

By understanding sampling distributions, we move from analyzing individual data points to understanding the behavior of statistics across different samples. This allows us to make more informed inferences about populations and test hypotheses effectively.

Standard Error

The standard error (SE) of a statistic (usually an estimate of a parameter) is the standard deviation of its sampling distribution or an estimate of that standard deviation. If the statistic is the sample mean, it is called the standard error of the mean (SEM). The sampling distribution of a mean is generated by repeated sampling from the same population and recording of the sample means obtained. This forms a distribution of different means, and this distribution has its own mean and variance. Mathematically, the variance of the sampling mean distribution obtained is equal to the variance of the population divided by the sample size. This is because as the sample size increases, sample means cluster more closely around the population mean.

https://en.wikipedia.org/wiki/Standard_error

The standard deviation reflects variability within a sample, while the standard error estimates the variability across samples of a population
https://www.scribbr.com/statistics/standard-error/

Derivation of the Standard Error of the Mean

The **Standard Error of the Mean ($\mathrm{SE}(\hat{\mu})$)** is the standard deviation of the sampling distribution of the sample mean ($\hat{\mu}$). It is derived by first finding the **Variance of the Sample Mean ($\mathrm{var}(\hat{\mu})$)**.

Step 1: Definition of the Sample Mean Variance

The sample mean is $\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} X_i$.

$$\mathrm{var}(\hat{\mu}) = \mathrm{var}\left(\frac{1}{N}\sum_{i=1}^{N} X_i\right)$$

Step 2: Applying the Expected Value Definition of Variance

The variance of any random variable Y is defined as the Expected Value of the squared difference between the variable and its mean: $\mathrm{var}(Y) = E[(Y - E[Y])^2]$. Since the sample mean $\hat{\mu}$ is an unbiased estimator, $E[\hat{\mu}] = \mu$.

$$= E\left[\left(\frac{1}{N}\sum_{i=1}^{N} X_i - \mu\right)^2\right]$$

Step 3: Factoring out $1/N$ from the Expected Value

We factor the constant $\frac{1}{N}$ out of the parentheses and move $\left(\frac{1}{N}\right)^2 = \frac{1}{N^2}$ outside the expectation E, using the property $E[cY] = cE[Y]$.

$$= E\left[\left(\frac{1}{N}\left(\sum_{i=1}^{N} X_i - N\mu\right)\right)^2\right]$$

$$= \frac{1}{N^2} E\left[\left(\sum_{i=1}^{N} X_i - N\mu\right)^2\right]$$

Step 4: Reverting to the Variance Notation

The term inside the expectation, $E\left[\left(\sum_{i=1}^{N} X_i - N\mu\right)^2\right]$, is the definition of the variance of the sum of the X_i's, because μ is the expected value of X_i, and $N\mu$ is the expected value of the sum $\sum X_i$: $E[\sum X_i] = \sum E[X_i] = N\mu$.

$$= \frac{1}{N^2}\operatorname{var}\left(\sum_{i=1}^{N} X_i\right)$$

Step 5: Variance of the Sum of Independent Variables

For independent and identically distributed (i.i.d.) variables, the variance of the sum is the sum of the variances, and $\operatorname{var}(X_i) = \sigma^2$:

$$\operatorname{var}\left(\sum_{i=1}^{N} X_i\right) = \sum_{i=1}^{N} \operatorname{var}(X_i) = N\sigma^2$$

Substituting this back into the equation:

$$= \frac{1}{N^2}(N\sigma^2)$$

Step 6: Final Variance of the Estimate

Simplify the term by canceling N:

$$\operatorname{var}(\hat{\mu}) = \frac{\sigma^2}{N}$$

Step 7: Calculating the Standard Error

The **Standard Error of the Mean ($\mathrm{SE}(\hat{\mu})$)** is the square root of the variance of the sample mean.

$$\mathrm{SE}(\hat{\mu}) = \sqrt{\mathrm{var}(\hat{\mu})} = \sqrt{\frac{\sigma^2}{N}} = \frac{\sigma}{\sqrt{N}}$$

This final formula explicitly shows that the variability of the sample mean decreases as the sample size N increases.

A Problem: Unknown Population Standard Deviation (σ)

The ideal formula for a 95% Confidence Interval for the population mean (μ) is:

$$95\% - CI = \left[\hat{\mu} - 1.96\frac{\sigma}{\sqrt{N}}, \quad \hat{\mu} + 1.96\frac{\sigma}{\sqrt{N}}\right]$$

This formula relies on the **population standard deviation (σ)**, which is almost always **unknown**. In a real experiment, you only have the **sample data** ($\{x_1, \ldots, x_N\}$) to work with.

The Solution: Replacing σ with the Sample Estimate ($\hat{\sigma}$ or s)

Since σ is unknown, the solution is to **replace it with an estimate** calculated from the sample data.

- **The Estimate:** The estimate for the population standard deviation (σ) is the **sample standard deviation**, denoted here as $\hat{\sigma}$ (or commonly as s).
- **The Formula:** The image provides the formula for this estimate:

$$\hat{\sigma} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

- **Key Detail ($N-1$):** Note the use of $N-1$ in the denominator (instead of N), which makes $\hat{\sigma}$ an **unbiased estimator** of σ (especially $\hat{\sigma}^2$ for σ^2).

Conclusion

When you replace σ with $\hat{\sigma}$ in the CI formula:

1. The standard error $\left(\frac{\sigma}{\sqrt{N}}\right)$ becomes the **estimated standard error** $\left(\frac{\hat{\sigma}}{\sqrt{N}}\right)$.
2. The critical value 1.96 **(from the Z-distribution)** is technically replaced by the critical value from the **t-distribution** (with $N-1$ degrees of freedom), though this critical value approaches 1.96 as the sample size (N) increases. This substitution accounts for the additional uncertainty introduced by estimating σ.

Margin of Error

In a normal distribution with a desired 95% confidence level, the margin of error is calculated as:

Margin of Error = Z * (σ / √n)

Where:

- **Z** is the critical value for a 95% confidence level. For a normal distribution, this is approximately **1.96**. (This value comes from the standard normal distribution table or can be calculated using statistical software).
- **σ** is the population standard deviation. If you don't know the population standard deviation, you can estimate it with the sample standard deviation (s).
- **n** is the sample size.

Explanation:

- The margin of error tells you how much your sample estimate (like the sample mean) is likely to differ from the true population parameter (the population mean).
- The Z value (1.96 for 95% confidence) reflects the number of standard deviations you need to go from the mean to capture the desired percentage of the distribution.
- The standard error (σ / √n) measures the variability of the sample mean.

Example:

Let's say you want to estimate the average height of adults in a city with 95% confidence. You take a sample of 100 adults and find the sample mean height is 5'8" with a standard deviation of 4 inches.

- Z = 1.96
- σ (estimated by s) = 4 inches
- n = 100

Margin of Error = 1.96 * (4 / √100) = 0.784 inches

This means you can be 95% confident that the true average height of adults in the city is within 0.784 inches of your sample mean (5'8").

Standard error: The standard error measures the variability or spread of the sample mean. It tells you how much you can expect the sample mean to vary from sample to sample.

Important Note: This formula assumes a normal distribution. If your data is not normally distributed, you may need to use a different critical value or a different method

altogether.

$\bar{x} \pm marginerror$

where the margin of error is a statistic expressing the amount of random sampling error in the results of a survey. The larger the margin of error, the less confidence one should have that a poll result would reflect the result of a census of the entire population.

https://en.wikipedia.org/wiki/Margin_of_error

Uncertainty

Check out the conclusion to this article: https://www.mathsisfun.com/data/confidence-interval.html

This passage explains the concept of **confidence intervals** and their role in statistical inference. Here's a summary:

What are Confidence Intervals?

A confidence interval is a range of plausible values for an unknown population parameter (like the population mean). It's calculated from sample data and expressed with a certain confidence level, usually 95%.

What does the confidence level mean?

The confidence level (e.g., 95%) indicates the long-run proportion of confidence intervals that would contain the true population parameter if we repeatedly sampled from the same population and calculated intervals in the same way.

How does this relate to uncertainty?

- **Aleatory uncertainty:** This is the inherent randomness before an event occurs (like flipping a coin). We can't eliminate this uncertainty.
- **Epistemic uncertainty:** This is uncertainty due to incomplete knowledge *after* an event. We use statistics to address epistemic uncertainty about populations.

How are confidence intervals constructed?

1. **Probability Theory:** We use probability theory to determine a range where we expect our sample statistic to fall with a certain probability (e.g., 95%).
2. **Inverting the Logic:** We then "invert" this logic to find the range of population parameters that would produce our observed statistic within that probability range. This gives us the confidence interval.

Key points from the passage:

- **Interpretation:** A 95% confidence interval means that if we repeatedly sampled and calculated intervals, 95% of those intervals would contain the true population parameter.
- **Assumptions:** Traditional methods based on probability theory often require assumptions about the population distribution (e.g., normality) or a large sample size.
- **Bootstrap:** The bootstrap approach offers a way to construct confidence intervals without these strict assumptions by resampling from the data itself.

In essence, confidence intervals provide a measure of uncertainty associated with our estimates. They help us express the range of plausible values for a population parameter based on our sample data.

Confidence Intervals

Confidence intervals are used to express how likely $\bar{x}$ falls within a range of values. If the hypothesized value falls in the tail outside of the one-directional area of interest, we reject the null hypothesis. If our hypothesized value falls outside of the two-tailed interval, we reject the null hypothesis.

In frequentist statistics, a confidence interval (CI) is a range of estimates for an unknown parameter. A confidence interval is computed at a designated confidence level. The 95% level is most common, but other levels (such as 90% or 99%) are sometimes used. The confidence level represents the long-run proportion of correspondingly computed intervals that end up containing the true value of the parameter. For example, out of all confidence intervals computed at the 95% level, 95% of them should contain the parameter's true value.

https://en.wikipedia.org/wiki/Confidence_interval

Calculating Confidence Intervals for Proportions

1. Estimate the Proportion (phat)

- `phat = x / n` where:
  - x is the number of successes (e.g., users who clicked)
  - n is the total number of trials (e.g., number of users)

2. Check for Normal Approximation

- If `N * phat > 5` and `N * (1 - phat) > 5`, you can use the normal distribution to approximate the binomial distribution.

3. Calculate the Margin of Error

- `Margin of Error = z * SE` where:
  - z is the critical value from the standard normal distribution (e.g., 1.96 for 95% confidence)
  - SE (Standard Error) = `sqrt(phat * (1 - phat) / n)`

4. Construct the Confidence Interval

- `Confidence Interval = phat ± Margin of Error`

Key Points

- **Factors Affecting Confidence Interval Width:**
  - **Distance of phat from 0.5:** The closer phat is to 0.5, the wider the interval (more uncertainty).
  - **Sample Size (n):** Larger sample sizes lead to narrower intervals (more precision).
- **Z-Score:** The z-score corresponds to the desired confidence level. For a 95% confidence interval, the z-score is approximately 1.96. This means that 95% of the time, the true population proportion will fall within 1.96 standard errors of the sample proportion.
- **Hypothesis Testing:** Confidence intervals are often used in hypothesis testing. If the confidence interval for the difference between two proportions (e.g., treatment and control groups) does not include 0, you can reject the null hypothesis that there is no difference.

This summary provides a clear and concise guide to calculating and interpreting confidence intervals for proportions, which is essential for analyzing A/B testing data and making informed decisions.

One Sided

```
# import numpy as np
# from scipy import stats
# import matplotlib.pyplot as plt

# x = np.arange(-3, 3, 0.001)

# pdf = stats.norm.pdf(x,loc=0,scale=1)
# plt.plot(x, pdf)

# z = 1.645 # 95% CI
```

```
# uci = z

# plt.fill_between(x, pdf, where=(x < uci), alpha=0.5)
# plt.text(uci, 0, uci, ha='center', fontsize=20)
# plt.text(-0.4, .2, f'95%', fontsize=20)
# plt.text(-1.1, .15, f'fail to reject', fontsize=20)
# plt.grid(True)
```

Note: A null hypothesis is either true or false. Unfortunately, we do not know which is the case, and we almost never will. It is important to realize that there is no probability that the null hypothesis is true or that it is false, because there is no element of chance.

http://strata.uga.edu/8370/lecturenotes/errors.html#:~:text=A%20null%20hypothesis%20is%20either,is%20no%20element%20of%20chance.

Two Sided

```
# import numpy as np
# from scipy import stats
# import matplotlib.pyplot as plt

# x = np.arange(-3, 3, 0.001)

# pdf = stats.norm.pdf(x,loc=0,scale=1)
# plt.plot(x, pdf)

# z = 1.96 # 95% CI
# lci = -z
# uci = z

# plt.fill_between(x, pdf, where=(lci < x) & (x < uci), alpha=0.5)
# plt.text(lci, 0, lci, ha='center', fontsize=20)
# plt.text(uci, 0, uci, ha='center', fontsize=20)
# plt.text(-0.4, .2, f'95%', fontsize=20)
# plt.grid(True)
```

Alpha (Significance Levels)

The significance level or alpha level is the probability of making the wrong decision when the null hypothesis is true. Alpha levels (sometimes just called “significance levels”) are used in hypothesis tests. Usually, these tests are run with an alpha level of .05 (5%), but other levels commonly used are .01 and .10.

https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/what-is-an-alpha-level/

Stephanie Glen. "Alpha Level (Significance Level): What is it?" From StatisticsHowTo.com: Elementary Statistics for the rest of us! https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/what-is-an-alpha-level/

One Sided

```
# import numpy as np
# from scipy import stats
# import matplotlib.pyplot as plt

# x = np.arange(-3, 3, 0.001)

# pdf = stats.norm.pdf(x,loc=0,scale=1)
# plt.plot(x, pdf)

# z = 1.645 # 95% CI
# uci = z

# plt.fill_between(x, pdf, where=(x > uci), alpha=0.5)
# plt.text(uci, 0, uci, ha='center', fontsize=20)
# plt.text(uci, .05, f'5%', fontsize=20)
# plt.grid(True)
```

Region of Rejection

The region of rejection, also known as the critical region, is a crucial concept in hypothesis testing. It's the range of values for the test statistic that would lead you to reject the null hypothesis.

Here's a breakdown:

1. **Hypothesis Testing:**
    - You start with a null hypothesis (H0), which is a statement of no effect or no difference.
    - You also have an alternative hypothesis (Ha), which contradicts the null hypothesis.
    - The goal of hypothesis testing is to determine if there's enough evidence to reject the null hypothesis in favor of the alternative hypothesis.
2. **Test Statistic:**
    - You calculate a test statistic from your sample data. This statistic measures how far your sample data deviates from what you'd expect if the null

hypothesis were true.

3. **Distribution of the Test Statistic:**
    - The test statistic follows a specific probability distribution (e.g., t-distribution, chi-squared distribution) under the assumption that the null hypothesis is true.
4. **Critical Value(s) and Significance Level:**
    - You choose a significance level (alpha), typically 0.05. This represents the probability of rejecting the null hypothesis when it's actually true (Type I error).
    - Based on the chosen alpha and the distribution of the test statistic, you determine the critical value(s). These values define the boundaries of the rejection region.
5. **Rejection Region:**
    - The rejection region is the area in the tails of the distribution beyond the critical value(s).
    - If your calculated test statistic falls within the rejection region, you reject the null hypothesis.
    - If the test statistic falls outside the rejection region, you fail to reject the null hypothesis.

Visual Example:

Imagine a normal distribution. For a two-tailed test with alpha = 0.05, the rejection region would be the 2.5% in each tail. The critical values would be the points on the x-axis that mark these boundaries (approximately ±1.96 standard deviations from the mean).

Key Points:

- The rejection region is determined by the chosen significance level (alpha) and the distribution of the test statistic.
- If the test statistic falls in the rejection region, it suggests that the observed data is unlikely to have occurred if the null hypothesis were true, leading you to reject the null hypothesis.
- The region of rejection helps make objective decisions in hypothesis testing by providing a clear criterion for rejecting or failing to reject the null hypothesis.

Two Sided

```
# import numpy as np
# from scipy import stats
# import matplotlib.pyplot as plt

# x = np.arange(-3, 3, 0.001)

# pdf = stats.norm.pdf(x,loc=0,scale=1)
# plt.plot(x, pdf)

# z = 1.96 # 95% CI
# lci = -z
# uci = z

# plt.fill_between(x, pdf, where=(lci > x) | (x > uci), alpha=0.5)
# plt.text(lci, 0, lci, ha='center', fontsize=20)
# plt.text(uci, 0, uci, ha='center', fontsize=20)
# # plt.text(-0.4, .2, f'95%', fontsize=20)
# plt.text(lci-.5, 0.05, f'2.5%', ha='center', fontsize=20)
# plt.text(uci+.5, 0.05, f'2.5%', ha='center', fontsize=20)
# plt.grid(True)
```

Critical Z-Value for α=0.05 (Right-Side)

The **critical z-score** for the right tail of a standard normal distribution, specifically for a $\mathbf{95\%}$ confidence interval or a two-tailed hypothesis test with a $\mathbf{5\%}$ significance level ($\alpha = 0.05$).

1. Context and Symmetry

- **Goal:** Find the z-score that cuts off the top **2.5%** (0.025) of the distribution.
- **Two-Tailed Test/Confidence Interval:** In a standard two-tailed scenario, the total $\alpha = 5\%$ is split into two tails: 2.5% in the left tail and 2.5% in the right tail.
- **Symmetry:** Because the standard normal distribution is symmetric around $z = 0$, the z-score for the right tail (z_{right}) and the left tail (z_{left}) will have the same magnitude but opposite signs. Since the left z-score is -1.96, the right z-score is $+1.96$.

2. Area and the Cumulative Distribution Function (Φ)

The core mathematical step involves the cumulative distribution function (CDF), $\Phi(z)$, which gives the total area under the curve to the **left** of a given z-score.

- **Area Calculation:** If 2.5% is in the right tail, the area to the left of z_{right} must be:

$$\text{Area}_{\text{left}} = 1 - \text{Area}_{\text{right}} = 1 - 0.025 = \mathbf{0.975}$$

- **Mathematical Expression:** This relationship is expressed using the integral of the standard normal probability density function (PDF):

$$\mathbf{0.975} = \int_{-\infty}^{z_{\text{right}}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \, dz$$

  - The expression $\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$ is the **PDF** of the standard normal distribution.
  - or $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$
  - The integral from $-\infty$ up to z_{right} calculates the **cumulative area** up to that point.

- **CDF Notation:** This integral is mathematically defined as the **Cumulative Distribution Function (CDF)**, $\mathbf{\Phi(z_{right})}$.

$$\mathbf{0.975} = \mathbf{\Phi(z_{right})}$$

3. Finding the Z-Score (Inverse CDF)

To find the z-score itself, we must perform the **inverse** of the CDF, often called the **quantile function** or Φ^{-1}.

- **Inverse Function:** We are looking for the z-score corresponding to a cumulative area of 0.975: $$\mathbf{z_{\text{right}}} = \mathbf{\Phi^{-1}(0.975)}$$
- **Result:** Consulting a standard normal table or using statistical software (which performs the inverse calculation) yields the result:

$$\mathbf{z_{right}} = \mathbf{1.96}$$

This value, 1.96, is the critical z-score often used in statistical inference, establishing the boundary that separates the middle 95% of the distribution from the outer 5% of extreme values.

Test Statistic vs Critical Value

In order to make a decision whether to reject the null hypothesis a test statistic is calculated. The decision is made on the basis of the numerical value of the test statistic. There are two approaches how to derive at that decision: The critical value approach and the p-value approach.

https://www.geo.fu-berlin.de/en/v/soga/Basics-of-statistics/Hypothesis-Tests/Introduction-to-Hypothesis-Testing/Critical-Value-and-the-p-Value-Approach/index.html

Critical value example:

- 1.96 - if the test statistic is greater than the critical value we reject the null

P-Value

For a p value test:

- Get the test statistic
- Use it to determine the p-value
- Compare the p-value to the level of significance
- If the p-value is low the null must go! Reject H_0
- If the p-value is high the null must fly! Fail to reject H_0

"The p value is the evidence against a null hypothesis. The smaller the p-value, the stronger the evidence that you should reject the null hypothesis."

Stephanie Glen. "P-Value in Statistical Hypothesis Tests: What is it?" From StatisticsHowTo.com: Elementary Statistics for the rest of us! https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/p-value/

The lower the p value, the more predictive the feature is in principle. When we run tests, we are often concerned with the alpha level to help us reject or fail to reject the null hypothesis. The alpha level is (1 - our confidence interval), so if we wanted to have a confidence level of 95% we would use a alpha value of 5%. If our p value is less than the alpha value then the evidence points to rejecting the null hypothesis. If our p values is less than the alpha value then we can say our results are statistcally significant. We found something that is probably not the result of chance. But beware.

Types of Errors in Hypothesis Testing

When conducting hypothesis testing, there's always a risk of making an incorrect decision. These incorrect decisions are classified into two types of errors:

- **Type I Error (False Positive):** This occurs when you reject the null hypothesis when it is actually true. In other words, you conclude that there is a significant effect or difference when there really isn't.
- **Type II Error (False Negative):** This occurs when you fail to reject the null hypothesis when it is actually false. In other words, you miss a real effect or difference.

Alpha (α)

- Alpha is the probability of making a Type I error. It's the significance level you set for your hypothesis test, typically 0.05. This means you're willing to accept a 5% chance of rejecting the null hypothesis when it's true.

Beta (β)

- Beta is the probability of making a Type II error. It's influenced by factors like the sample size, the effect size, and the variability in the data.

Statistical Power (1 - β)

- Statistical power, also known as the power of a test, is the probability of correctly rejecting the null hypothesis when it is false. In other words, it's the probability of detecting a real effect.
- It's calculated as 1 - beta.
- Higher power means a lower chance of missing a real effect.

1 - Alpha

- 1 - alpha represents the confidence level. It's the probability of correctly failing to reject the null hypothesis when it is true.
- For example, if alpha is 0.05, then 1 - alpha is 0.95, or a 95% confidence level.

Relationship Between Concepts

These concepts are interconnected:

- **Alpha and Beta:** There's a trade-off between alpha and beta. Decreasing alpha (lowering the risk of a Type I error) generally increases beta (increasing the risk of a Type II error), and vice versa.
- **Power and Sample Size:** Increasing the sample size generally increases statistical power, as you have more data to detect a real effect.
- **Power and Effect Size:** Larger effect sizes are easier to detect, leading to higher power.

Practical Implications

Understanding these concepts is crucial for designing and interpreting hypothesis tests:

- **Choosing Alpha:** You need to balance the risks of Type I and Type II errors based on the consequences of each in your specific context.
- **Ensuring Adequate Power:** Aim for sufficient statistical power (often 80% or higher) to increase the chances of detecting a real effect if one exists.
- **Interpreting Results:** Consider the possibility of both types of errors when interpreting the results of a hypothesis test.

By carefully considering these factors, you can conduct more robust and reliable hypothesis tests.

```
# import numpy as np
# import matplotlib.pyplot as plt
# from scipy.stats import norm

# # --- 1. Define Distribution Parameters ---
# mu_0 = 0.0  # Mean for Null Hypothesis (H0)
# mu_A = 2.0  # Mean for Alternative Hypothesis (HA)
# sigma = 1.0 # Standard deviation for both distributions
# threshold = 1.2 # Critical value / Decision threshold

# # --- 2. Define X-axis Range and PDFs ---
# x = np.linspace(-4, 6, 500)
# pdf_H0 = norm.pdf(x, loc=mu_0, scale=sigma)
# pdf_HA = norm.pdf(x, loc=mu_A, scale=sigma)

# # --- 3. Plotting ---
# plt.figure(figsize=(10, 6))

# # Plot the curves
# plt.plot(x, pdf_H0, label='Null Hypothesis ($H_0$)', color='blue')
# plt.plot(x, pdf_HA, label='Alternative Hypothesis ($H_A$)', color='red')

# # Plot the threshold line
# plt.axvline(threshold, color='black', linestyle='--', linewidth=1.5, label='D

# # --- 4. Shading the Regions (TP, TN, FP, FN) ---

# # --- True Negatives (TN): H0, Left of threshold (Correct Negative) ---
# x_tn = x[x <= threshold]
# pdf_H0_tn = pdf_H0[x <= threshold]
# plt.fill_between(x_tn, pdf_H0_tn, color='blue', alpha=0.3, label='TN: Fail to

# # --- False Positives (FP): H0, Right of threshold (Type I Error) ---
# x_fp = x[x > threshold]
# pdf_H0_fp = pdf_H0[x > threshold]
# plt.fill_between(x_fp, pdf_H0_fp, color='red', alpha=1, label='FP: Type I Err

# # --- False Negatives (FN): HA, Left of threshold (Type II Error) ---
# x_fn = x[x <= threshold]
# pdf_HA_fn = pdf_HA[x <= threshold]
# plt.fill_between(x_fn, pdf_HA_fn, color='gray', alpha=1, label='FN: Type II E

# # --- True Positives (TP): HA, Right of threshold (Power) ---
# x_tp = x[x > threshold]
# pdf_HA_tp = pdf_HA[x > threshold]
# plt.fill_between(x_tp, pdf_HA_tp, color='green', alpha=0.3, label='TP: Reject



# # --- 5. Final Touches ---
# plt.title('Hypothesis Testing Errors and Power (Overlapping Distributions)',
# plt.xlabel('Test Statistic Value', fontsize=12)
```

```
    # plt.ylabel('Probability Density', fontsize=12)
    # plt.legend(loc='upper right', frameon=True, shadow=True, fontsize=9)
    # plt.grid(axis='y', alpha=0.5)
    # plt.tight_layout()
    # plt.show()
```

Comparison of TN, TP, FP, FN

Outcome	Decision	Reality	Label
TN	Fail to Reject H_0	H_0 is true	**TN (Correctly Fail to Reject H_0 when true)**
TP	Reject H_0	H_0 is false	**TP / Power (Correctly Reject H_0 when false)**
FP	Reject H_0	H_0 is true	**FP / Type I Error (α): Reject H_0 when true**
FN	Fail to Reject H_0	H_0 is false	**FN / Type II Error (β): Fail to Reject H_0 when false**

P Value vs Alpha

Alpha, the significance level, is the probability that you will make the mistake of rejecting the null hypothesis when in fact it is true. The p-value measures the probability of getting a more extreme value than the one you got from the experiment. If the p-value is greater than alpha, we fail to reject the null hypothesis.

https://www.spcforexcel.com/knowledge/basic-statistics/interpretation-alpha-and-p-value

Statistical Power (Sensitivity)

Power reduces false negatives, while a lower alpha reduces false positives.

1. **Power** $(1 - \beta)$ is your ability to detect a true effect, meaning it **reduces the probability of a False Negative (β)**.
2. A **lower alpha (α)** is a more conservative threshold for significance, meaning it **reduces the probability of a False Positive (α)**.

https://www.statisticshowto.com/wp-content/uploads/2015/04/statistical-power.png

Beta is directly related to the power of a test. Power relates to how likely a test is to distinguish an actual effect from one you could expect to happen by chance alone. Beta plus the power of a test is always equal to 1. Usually, researchers will refer to the power of a test (e.g. a power of .8), leaving the beta level (.2 in this case) as implied.

https://www.statisticshowto.com/beta-level/

The **statistical power of a binary hypothesis test** is the probability that the test correctly rejects the null hypothesis $H_{0}$$H_{0}$ when a specific alternative hypothesis $H_{a}$$H_{a}$ is true. It is commonly **denoted by $1-\beta$$1-\beta$** , and represents the chances of a "true positive" detection conditional on the actual existence of an effect to detect. Statistical power ranges from 0 to 1, and as the power of a test increases, the probability $\beta$$\beta$ of making a type II error by wrongly failing to reject the null hypothesis decreases.

https://en.wikipedia.org/wiki/Power_of_a_test

We say "**fail to reject the null hypothesis**" instead of "**accept the null hypothesis**" because in classical (frequentist) hypothesis testing, the methodology is designed *only* to assess evidence *against* the null hypothesis ($\mathbf{H_0}$$\mathbf{H_0}$).

This cautious language reflects the fundamental principles and limitations of the testing framework:

1. The Burden of Proof Principle (Falsification)

The core reason is that hypothesis testing operates under a philosophy of **falsification**, not proof.

- The $\mathbf{H_0}$$\mathbf{H_0}$ (the claim of no effect or no difference) is assumed to be **true by default**—like a defendant presumed **"innocent until proven guilty."**
- The statistical test measures how **unlikely** your observed data is, *assuming* $H_0$$H_0$ is true (this is the $p$$p$-value).
- If the data is **extremely unlikely** ($p < \alpha$$p < \alpha$), you have enough evidence to **reject $H_0$$H_0$** (finding the defendant "guilty").
- If the data is **not extremely unlikely** ($p > \alpha$$p > \alpha$), it simply means your data is *consistent* with $H_0$$H_0$. It does **not** prove that $H_0$$H_0$ is definitively true. It just means the evidence was insufficient to overturn the default assumption.

Therefore, **failing to reject** is the statistical equivalent of a **"not guilty"** verdict; it means there wasn't enough evidence to convict, not that innocence was proven.

2. The Issue of Statistical Power

You can fail to reject $H_0$$H_0$ for one of two reasons:

1. The null hypothesis ($H_0$$H_0$) is actually true (or close enough to it).
2. The **Alternative Hypothesis ($\mathbf{H_A}$$\mathbf{H_A}$) is true**, but your experiment had **insufficient statistical power** to detect the effect (this is a **Type**

II error, or a False Negative).

If you were to "**accept**" the null hypothesis when $p > \alpha$$p > \alpha$, you would incorrectly be claiming it is true, when in reality, your sample size might simply have been too small, or your measurements too noisy, to detect a real effect. Since the test cannot distinguish between the first reason and the second, the safest, most precise conclusion is simply that the data **failed to provide sufficient evidence for rejection.**

3. Testing a Point, Not an Area

The null hypothesis is usually a precise statement (e.g., the mean difference is exactly zero, $\mu = 0$$\mu = 0$).

- To "accept" $H_0$$H_0$ ($\mu=0$$\mu=0$) would require evidence that rules out *all other possibilities* (like $\mu=0.0001$$\mu=0.0001$ or $\mu=-0.00000001$$\mu=-0.00000001$).
- A single experiment, being based on a sample and having finite precision, can never **prove** that a value is *exactly* zero. It can only show that the value is close enough to zero that we cannot confidently rule out $H_0$$H_0$.

Because a single test can only **reject** (providing evidence of a difference) or **fail to reject** (indicating a lack of sufficient evidence), "fail to reject" is the only statement that accurately reflects the scope and limitations of the statistical procedure.

P-hacking

P-hacking refers to the practice of repeatedly analyzing data or manipulating test parameters until a statistically significant result (low p-value) is achieved. This can lead to false positives, where you mistakenly conclude that a change has an effect when it actually doesn't.

Why is this bad?

- **Misleading Results:** P-hacking can lead to implementing changes based on flawed data, wasting resources and potentially harming user experience.
- **Lack of Reproducibility:** P-hacked results are often not reproducible, as they are based on chance findings rather than true effects.

The Solution: Power Analysis

Instead of p-hacking, the recommended approach is to conduct a power analysis before running the A/B test. This involves determining the minimum sample size needed to detect a meaningful effect with a certain level of confidence.

Key Components of Power Analysis

- **Significance Level (Alpha):** The probability of rejecting a true null hypothesis (false positive). It's the risk you're willing to take of concluding there's an effect when there isn't.
- **Statistical Power (1 - Beta):** The probability of correctly rejecting a false null hypothesis (true positive). It's the likelihood of detecting a real effect if one exists.
- **Minimum Detectable Effect (MDE):** The smallest effect size that is practically meaningful for your business.

How Power Analysis Helps

- **Determines Sample Size:** By specifying the desired significance level, power, and MDE, you can calculate the minimum sample size needed for a reliable A/B test.
- **Reduces Bias:** This approach reduces the risk of p-hacking by pre-determining the sample size and avoiding the temptation to stop the test early based on seemingly significant results.
- **Increases Confidence:** A well-powered A/B test gives you greater confidence that your results are valid and reproducible.

In Summary

The passage emphasizes the importance of conducting A/B tests with a robust methodology. By avoiding p-hacking and utilizing power analysis, you can increase the reliability and trustworthiness of your results, leading to better data-driven decisions.

A/B Test Statistics

- **Confidence Level:**
    - This represents the probability that the confidence interval contains the true population parameter. Commonly expressed as a percentage (e.g., 95%). A 95% confidence level means that if you repeated the experiment many times, 95% of the confidence intervals would contain the true value.
- **Margin of Error:**
    - This is the range of values above and below the sample statistic within which the true population parameter is likely to fall. It quantifies the uncertainty in your estimate.
- **Confidence Interval:**
    - This is the range of values, calculated from sample data, that is likely to contain the true population parameter with a certain level of confidence.
- **Type I Error (False Positive):**

    - This occurs when you reject the null hypothesis when it is actually true. In A/B testing, it means concluding that there is a significant difference between variations when there isn't.
- **Type II Error (False Negative):**
    - This occurs when you fail to reject the null hypothesis when it is actually false. In A/B testing, it means missing a real difference between variations.
- **p-Value:**
    - This is the probability of obtaining the observed results (or more extreme) if the null hypothesis were true. A small p-value indicates strong evidence against the null hypothesis.
- **Statistical Significance:**
    - This refers to the likelihood that an observed effect is not due to chance. It is typically determined by comparing the p-value to a predetermined significance level (alpha).
- **Statistical Power:**
    - This is the probability of correctly rejecting the null hypothesis when it is false. It represents the test's ability to detect a real effect.
- **Minimum Detectable Effect (MDE):**
    - This is the smallest effect size that you want to be able to detect with your A/B test. It helps determine the required sample size.
- **Practical Significance:**
    - While statistical significance indicates whether an effect is likely real, practical significance addresses whether the effect is meaningful in a real-world context. A statistically significant result may not be practically significant if the effect size is too small.
- **Sample Size and Duration:**
    - Sample size is the number of participants or observations in a study. Duration refers to the length of time the study is conducted. Both are critical for achieving statistically reliable results. Larger sample sizes and longer durations generally increase statistical power.

Threats to Validity

When conducting experiments, it's crucial to be aware of potential threats to their validity. These factors can introduce bias and distort the results, making it difficult to draw accurate conclusions. Here's a breakdown of the terms you provided:

- **Novelty Effect:**
  - This occurs when participants in an experiment react differently simply because something is new or unfamiliar. The introduction of a new feature, interface, or product can lead to a temporary surge in engagement or positive behavior, which may not be sustained over time.
  - It's important to allow sufficient time for the novelty effect to wear off before drawing conclusions about the long-term impact of a change.
- **Primary Effect:**
  - This is a cognitive bias where people tend to remember the first items in a series better than subsequent items. In an experiment, if participants are exposed to multiple treatments or variations, their responses to the initial ones may be disproportionately influential.
  - This is a form of an order effect.
- **Seasonality:**
  - This refers to predictable patterns in data that occur at specific times of the year. For example, retail sales tend to increase during the holiday season, and website traffic may fluctuate depending on school schedules.
  - Seasonality can confound experimental results if the study is conducted during a period of unusually high or low activity. It's important to account for seasonality when analyzing data and drawing conclusions.
- **Day of the Week:**
  - This is a specific form of seasonality, but focused on the weekly cycle. User behavior and activity levels can vary significantly depending on the day of the week. For example, website traffic may be higher on weekdays than on weekends.
  - Similar to seasonality, the day of the week can introduce bias into experimental results if the study is not designed to account for these variations.

Example

Scenario:

Imagine you have a website with a "Sign Up" button. You want to test whether changing the button's color from blue (version A) to green (version B) will increase the sign-up rate.

```
# import numpy as np
# import scipy.stats as stats
# from statsmodels.stats.proportion import proportions_ztest # Import the corre

# # Simulated data (replace with actual data)
# np.random.seed(42) # For reproducibility

# # Number of users (trials) in each group
# n_A = 1000
# n_B = 1000

# # Simulated sign-up rates (replace with actual data)
# conversion_rate_A = 0.10 # 10% for blue button
# conversion_rate_B = 0.12 # 12% for green button

# # Generate simulated sign-up data (1 = sign-up, 0 = no sign-up)
# signups_A = np.random.choice([1, 0], size=n_A, p=[conversion_rate_A, 1 - conv
# signups_B = np.random.choice([1, 0], size=n_B, p=[conversion_rate_B, 1 - conv

# # Calculate number of successes (sign-ups) and trials (users)
# count_A = np.sum(signups_A)
# count_B = np.sum(signups_B)

# # Calculate conversion rates from the simulated data.
# actual_conversion_rate_A = count_A / n_A
# actual_conversion_rate_B = count_B / n_B

# print(f"Version A (Blue): Successes = {count_A}, Trials = {n_A}, Rate = {actu
# print(f"Version B (Green): Successes = {count_B}, Trials = {n_B}, Rate = {act

# count = np.array([count_A, count_B])
# nobs = np.array([n_A, n_B])

# # The proportions_ztest function returns (z_statistic, p_value)
# z_stat, p_value = proportions_ztest(count, nobs, alternative='two-sided')

# print("\n--- A/B Test Results (Z-Test for Proportions) ---")
# print(f"Z-statistic: {z_stat:.4f}")
# print(f"P-value: {p_value:.4f}")

# # Determine statistical significance
# alpha = 0.05 # Significance level

# if p_value < alpha:
#     print("Result: Statistically significant. Reject the null hypothesis.")
#     if actual_conversion_rate_B > actual_conversion_rate_A:
```

```
#         print("Conclusion: Version B (Green) performed better.")
#     else:
#         print("Conclusion: Version A (Blue) performed better.")
# else:
#     print("Result: Not statistically significant. Fail to reject the null hyp
#     print("Conclusion: There is no strong evidence to suggest a difference be
```

Explanation

This Python code simulates and analyzes the results of an **A/B test** to determine if there is a statistically significant difference between the conversion rates of two versions, Version A and Version B. It uses statistical methods appropriate for comparing two binary proportions.

1. Setup and Simulation

- **Imports:** The code imports `numpy` for numerical operations, `scipy.stats` for general statistics, and `proportions_ztest` from `statsmodels.stats.proportion` for the core statistical test.
- **Seed:** `np.random.seed(42)` ensures that the simulated random data is the same every time the code runs, making the results **reproducible**.
- **Parameters:**
  - `n_A` and `n_B` define the sample size (number of users) in each group (1,000 each).
  - `conversion_rate_A` (10%) and `conversion_rate_B` (12%) are the true underlying rates used to generate the data.
- **Data Generation:** `np.random.choice` simulates the outcome for each user (1 for a sign-up, 0 for no sign-up) based on the defined conversion rates and sample sizes. This simulates the observed data from an actual experiment.
- **Success Counts:** `count_A = np.sum(signups_A)` and `count_B = np.sum(signups_B)` calculate the total number of successes (sign-ups) in each group.

---

2. Rate Calculation

- The **actual conversion rates** are calculated by dividing the number of successes by the number of trials (`count / n`).
- The code prints these observed rates to provide context for the statistical test.

---

3. Statistical Test (Two-Sample Z-Test for Proportions)

The central part of the analysis uses the **Two-Sample Z-Test for Proportions**, which is the standard statistical test for A/B testing with conversion data.

- **Test Function:** `proportions_ztest(count, nobs, alternative='two-sided')` is called, where:
  - `count` is an array containing `[count_A, count_B]`.
  - `nobs` is an array containing the total observations `[n_A, n_B]`.
  - `alternative='two-sided'` means the test is checking if the rates are different (either A > B or B > A).
- **Outputs:** The function returns the **Z-statistic** and the **P-value**.
  - The **Z-statistic** measures how many standard deviations the difference between the two conversion rates is from zero (the null hypothesis).
  - The **P-value** is the probability of observing a difference as extreme as the one calculated, assuming the null hypothesis (that there is no real difference between the versions) is true.

---

4. Conclusion and Interpretation

- **Significance Level ($\alpha$$\alpha$):** A standard alpha value of **0.05** is set.
- **Decision Rule:**
  - If the **P-value is less than 0.05**, the result is declared **statistically significant**. This means we have enough evidence to **reject the null hypothesis** and conclude that there is a real difference between the two versions. The code then checks which version had the higher observed rate to state the final conclusion.
  - If the **P-value is greater than or equal to 0.05**, the result is **not statistically significant**. This means we **fail to reject the null hypothesis**, concluding that the observed difference could be due to random chance.

RCTs and AB Testing

Free AB Testing Course: https://www.udacity.com/course/ab-testing--ud257

Randomized Controlled Trials (RCTs)

- **Core Concept:**
  - RCTs are considered the "gold standard" for determining the effectiveness of an intervention.

    - Participants are randomly assigned to either a "control" group (which receives a standard treatment or a placebo) or an "experimental" group (which receives the new intervention).
    - Randomization minimizes bias by ensuring that both groups are as similar as possible at the start of the experiment.
    - Outcomes are then compared between the groups to determine if the intervention had a significant effect.
- **Common Applications:**
    - Medical research (testing new drugs or treatments)
    - Social science research (evaluating the impact of social programs)
    - Any field where it is important to establish a causal relationship.

A/B Testing

- **Core Concept:**
    - A/B testing is a specific type of RCT commonly used in online environments.
    - It involves comparing two versions of a webpage, app, or other digital product (version A and version B) to see which performs better.
    - Users are randomly assigned to see either version A or version B.
    - Metrics such as click-through rates, conversion rates, or user engagement are then compared to determine which version is more effective.
- **Common Applications:**
    - Website optimization
    - Marketing campaigns
    - App development
    - User experience (UX) design

Similarities

- **Randomization:** Both RCTs and A/B testing rely on random assignment to minimize bias and ensure that groups are comparable.
- **Control Groups:** Both involve a control group (or version A) that serves as a baseline for comparison.
- **Causal Inference:** Both aim to establish a causal relationship between the intervention (or version B) and the observed outcomes.
- **Statistical Analysis:** Both use statistical methods to analyze the data and determine if the observed differences are statistically significant.

Key Takeaways

- A/B testing is essentially a specific, practical application of the broader concept of randomized controlled trials, particularly in the digital realm.
- They both provide a reliable way to get evidence based results.
- The goal of both is to remove bias from testing, and provide accurate data to make informed decisions.

https://vwo.com/ab-testing/

Common Tests Used for AB Testing

1. Tests for Categorical/Binary Data (Counts and Rates)

These tests are used for metrics that are binary (yes/no) and measured as a **proportion** or **rate** (e.g., conversion rate).

- **Z-Test for Proportions**
  - **Purpose:** Compares the proportions (rates) of two large groups (A and B) to see if the difference is statistically significant.
  - **Common Use Case:** The most frequent test for A/B testing conversion rates (Did a user convert: Yes/No?).
  - **Assumption:** Requires a large sample size.
- **Chi-Square ($\chi^2$$\chi^2$) Test**
  - **Purpose:** Checks if the distribution of counts across categories is independent of the group. It is mathematically very similar to the Z-Test for proportions when only two groups are compared.
  - **Common Use Case:** Used for Click-Through Rate (CTR) and other binary success/failure counts.
- **Fisher's Exact Test**
  - **Purpose:** A non-approximate alternative to the Chi-Square test.
  - **Common Use Case:** Used for categorical data when the overall **sample size is small** or when expected counts are very low, where the Chi-Square's approximation might be unreliable.

2. Tests for Continuous Data (Means and Averages)

These tests are used for numerical metrics that have a mean and standard deviation.

- **Welch's T-Test**

- **Purpose:** Compares the **means** (averages) of a numerical metric between two groups (A and B).
- **Common Use Case:** The standard test for continuous metrics because it does not require the two groups to have equal variances, which is often a safer assumption in real-world A/B experiments.
- **A/B Metric Examples:** Average Order Value (AOV), Revenue Per User (RPU), or Time Spent on Site.

- **Student's T-Test (Independent Samples)**
  - **Purpose:** Also compares the means of two groups, but specifically assumes the variances of the two groups are equal.
  - **Common Use Case:** Used less frequently than Welch's T-Test in industry unless there is a strong reason to assume equal variance.

3. Test for Multiple Variants ($\text{A/B/n}$$\text{A/B/n}$ Testing)

- **ANOVA (Analysis of Variance)**
  - **Purpose:** Compares the **means** of three or more independent groups (e.g., Variant A, Variant B, and Variant C) to determine if at least one group mean is different from the others.
  - **Common Use Case:** When running an A/B/C/D test on a new feature or design.
  - **Note:** If ANOVA shows a difference, you must use a **post-hoc test** (like Tukey's HSD) to find out *exactly which pairs* are different.

Chi-Square Tests

The **Chi-Square ($\chi^2$$\chi^2$) test** is a fundamental statistical tool in A/B testing used to determine if the **observed difference** in conversion rates between your variants (A and B) is likely due to a real effect or simply **random chance**.

The Chi-Square test is specifically designed for **categorical data**, which perfectly describes the outcomes of most A/B tests:

1. **Group/Variable 1:** The version shown (Category 1: Version A, Category 2: Version B).
2. **Outcome/Variable 2:** The result (Category 1: Converted/Success, Category 2: Did Not Convert/Failure).

The test compares the **Observed Frequencies** (your actual A/B test results) to the **Expected Frequencies** (what you would expect if there were **NO difference** between

the two versions, which is the **null hypothesis**).

- **Null Hypothesis ($H_0$$H_0$):** The conversion rate for Version A is equal to the conversion rate for Version B (the two variables—version and outcome—are independent).
- **Alternative Hypothesis ($H_a$$H_a$):** The conversion rates are different (the two variables are dependent).

The $\chi^2$$\chi^2$ statistic is calculated using the following formula, summed over all cells in your contingency table: $$\chi^2 = \sum \frac{(O - E)^2}{E}$$$$\chi^2 = \sum \frac{(O - E)^2}{E}$$ Where:

- $O$$O$ = **Observed** count in a cell.
- $E$$E$ = **Expected** count in a cell (calculated under the null hypothesis).

A **larger $\chi^2$$\chi^2$ value** indicates a greater deviation between what you observed and what you expected, making it more likely that the difference is statistically significant.

The Chi-Square test is most appropriate and commonly used in A/B testing for **conversion metrics** when the data is structured as a **contingency table**.

1. Simple A/B Tests (2x2 Tables)

This is the most common scenario: comparing two groups (A and B) on a binary outcome (e.g., converted or not converted).

Outcome	Version A (Control)	Version B (Variant)	Total
Converted	Count $A_{success}$$A_{success}$	Count $B_{success}$$B_{success}$	Total Successes
Did Not Convert	Count $A_{failure}$$A_{failure}$	Count $B_{failure}$$B_{failure}$	Total Failures
Total Users	$N_A$$N_A$	$N_B$$N_B$	$N_{Total}$$N_{Total}

In this $2 \times 2$$2 \times 2$ case, the Chi-Square test is mathematically **equivalent** to the **two-sample Z-test for proportions**. Both tests will yield the same P-value, and the $\chi^2$$\chi^2$ statistic will be the square of the $Z$$Z$-statistic ($\chi^2 = Z^2$$\chi^2 = Z^2$) with one degree of freedom. Therefore, you can use either one.

2. A/B/C/D... Tests (Multiple Variants)

This is where the Chi-Square test shines and provides value beyond the Z-test. If you are testing **three or more versions (A, B, C, etc.) simultaneously**, the Chi-Square test is the correct choice to determine if **at least one** of the variants is performing significantly differently from the others.

3. Multiple Outcomes (Funnel Drop-off)

The test can also be extended to situations with multiple categorical outcomes, such as a multi-step funnel:

- Did not convert.
- Converted to Lead.
- Converted to Purchase.

Here, you would use a Chi-Square test (specifically, a **Test of Homogeneity**) to see if the overall distribution of users across these three outcome categories is the same for Version A and Version B.

```
# # remember the normal distribution?
# import numpy as np
# import matplotlib.pyplot as plt
# from scipy.stats import norm

# # 1. Define the parameters for the standard normal distribution
# mu = 0
# sigma = 1

# # 2. Create the x-axis range and the corresponding PDF curve (y-axis)
# x = np.linspace(mu - 4*sigma, mu + 4*sigma, 500)
# y = norm.pdf(x, mu, sigma)

# # 3. Create the plot
# plt.figure(figsize=(10, 6))
# plt.plot(x, y, color='black', linewidth=1)
# plt.title('The Empirical Rule (68-95-99.7) on the Normal Distribution')
# plt.xlabel('Z-score (Standard Deviations)')
# plt.ylabel('Probability Density Function (PDF)')
# plt.grid(True, linestyle='--', alpha=0.6)

# # 4. Define the boundaries for 1, 2, and 3 standard deviations
# # Shade the regions from the center outwards for visual clarity
# # 99.7% (3-sigma) region
# x_3sigma = x[(x >= mu - 3*sigma) & (x <= mu + 3*sigma)]
# y_3sigma = norm.pdf(x_3sigma, mu, sigma)
# plt.fill_between(x_3sigma, y_3sigma, color='skyblue', alpha=0.2, label='99.7%

# # 95% (2-sigma) region
# x_2sigma = x[(x >= mu - 2*sigma) & (x <= mu + 2*sigma)]
# y_2sigma = norm.pdf(x_2sigma, mu, sigma)
# plt.fill_between(x_2sigma, y_2sigma, color='skyblue', alpha=0.4, label='95% (

# # 68% (1-sigma) region
# x_1sigma = x[(x >= mu - 1*sigma) & (x <= mu + 1*sigma)]
# y_1sigma = norm.pdf(x_1sigma, mu, sigma)
# plt.fill_between(x_1sigma, y_1sigma, color='skyblue', alpha=0.7, label='68% (
```

```
# # Add lines for the boundaries
# boundaries = [-3, -2, -1, 1, 2, 3]
# for b in boundaries:
#     plt.axvline(b, color='gray', linestyle=':', linewidth=0.8)

# plt.legend(loc='upper right')
# plt.ylim(0, 0.45)
# plt.tight_layout()
```

Normal Distribution (Gaussian Distribution)

The normal distribution is defined by two parameters: the mean ($\mu$$\mu$) and the standard deviation ($\sigma$$\sigma$).

The Probability Density Function (PDF) is:

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

- **$x$$x$**: The variable value (your input).
- **$\mu$$\mu$ (Mean)**: The center of the distribution.
- **$\sigma$$\sigma$ (Standard Deviation)**: The spread of the distribution.
- **$\pi$$\pi$** ($\approx 3.14159...$$\approx 3.14159...$) and **$e$$e$** ($\approx 2.71828...$$\approx 2.71828...$) are mathematical constants.

The curve in the **Empirical Rule image** is a special case called the **Standard Normal Distribution**, where $\mu=0$$\mu=0$ and $\sigma=1$$\sigma=1$.

```
# import numpy as np
# import matplotlib.pyplot as plt
# from scipy.stats import chi2

# x = np.linspace(0, 9, 100)
# plt.title('The Chi Square Distribution')
# plt.plot(x, chi2(1).pdf(x), label=f'df = 1')
# plt.plot(x, chi2(2).pdf(x), label=f'df = 2')
# plt.plot(x, chi2(3).pdf(x), label=f'df = 3')
# plt.plot(x, chi2(4).pdf(x), label=f'df = 4')
# plt.plot(x, chi2(6).pdf(x), label=f'df = 6')
# plt.plot(x, chi2(9).pdf(x), label=f'df = 9')
# plt.ylim(0, 1)

# plt.legend();
```

Chi-Square ($\chi^2$$\chi^2$) Distribution

The Chi-Square distribution is defined by a single parameter: the **degrees of freedom ($k$$k$ or $df$$df$)**. It is always non-negative and is right-skewed.

The Probability Density Function (PDF) is:

$$f(x; k) = \frac{1}{2^{k/2} \Gamma(k/2)} x^{(k/2) - 1} e^{-x/2} \quad \text{for } x > 0$$$$f(x; k) = \frac{1}{2^{k/2} \Gamma(k/2)} x^{(k/2) - 1} e^{-x/2} \quad \text{for } x > 0$$

- **$x$$x$**: The $\chi^2$$\chi^2$ value (your input, which must be non-negative).
- **$k$$k$ (Degrees of Freedom, $df$$df$)**: The parameter that defines the shape of the curve.
- **$e$$e$** ($\approx 2.71828...$$\approx 2.71828...$): The mathematical constant.
- **$\Gamma$$\Gamma$ (Gamma Function)**: A generalization of the factorial function to real and complex numbers. For integer $n$$n$, $\Gamma(n) = (n-1)!$$\Gamma(n) = (n-1)!$.

The multiple curves in the **Chi-Square distribution image** show how the shape changes as the degrees of freedom ($df$$df$) increases. As $df$$df$ gets larger, the $\chi^2$$\chi^2$ distribution shifts right and becomes less skewed, eventually resembling a normal distribution.

The degrees of freedom (df) is the only parameter needed to define the shape of a chi-squared distribution because it encapsulates the essential information about the variability and constraints within the data being analyzed.

Here's a breakdown of why df is so crucial:

1. **Sum of Squared Variables:** The chi-squared distribution arises from the sum of squared standard normal random variables. Each squared standard normal variable contributes one degree of freedom.
2. **Constraints and Independence:** The degrees of freedom represent the number of independent pieces of information available in the data. When you have constraints or dependencies within the data, the degrees of freedom are reduced.
3. **Shape of the Distribution:** The degrees of freedom directly determine the shape of the chi-squared distribution. Here's how:
    - **Lower df:** The distribution is highly skewed to the right, with a long tail.
    - **Higher df:** The distribution becomes more symmetrical and approaches a normal distribution.
4. **Variability:** The degrees of freedom also reflect the variability of the distribution. Lower df means higher variability (and heavier tails).

In simpler terms:

Think of degrees of freedom as the amount of "wiggle room" or freedom you have in your data. If you have few degrees of freedom, your data is more constrained, and the distribution will be more skewed. As you gain more degrees of freedom, the constraints loosen, and the distribution becomes more like a normal curve.

Key Takeaway:

The degrees of freedom essentially "tell" the chi-squared distribution how much variability and constraint to exhibit. This single parameter is sufficient to define its shape and behavior.

```
# import numpy as np
# import matplotlib.pyplot as plt
# import scipy.stats as stats

# # Parameters
# df = 1  # Degrees of freedom
# alpha = 0.05  # Significance level
# chi2_stat = 0.779  # Example test statistic (from your previous output)

# # Generate x-values for the distribution
# x = np.linspace(0, 10, 500)

# # Calculate the chi-squared distribution PDF
# pdf = stats.chi2.pdf(x, df)

# # Calculate the critical value
# critical_value = stats.chi2.ppf(1 - alpha, df)

# # Plot the chi-squared distribution
# plt.plot(x, pdf, label=f"Chi-Square Distribution (df={df})")

# # Shade the rejection region
# plt.fill_between(x[x > critical_value], pdf[x > critical_value], color='red',

# # Mark the critical value
# plt.axvline(x=critical_value, color='black', linestyle='--', label=f"Critical

# # Mark the test statistic
# plt.axvline(x=chi2_stat, color='green', linestyle='--', label=f"Test Statisti

# # Add labels and title
# plt.xlabel("Chi-Squared Statistic")
# plt.ylabel("Probability Density")
# plt.title("Chi-Squared Distribution with Critical Value and Test Statistic")
# plt.legend()
# plt.show()
```

In the chi-squared test for A/B testing with conversions and non-conversions, the degrees of freedom (df) should be **(number of rows - 1) * (number of columns - 1)**.

Since we have:

- 2 rows (Group A and Group B)
- 2 columns (Conversions and Non-conversions)

The df should be (2-1) * (2-1) = 1.

Why 1 df in this case?

In this specific scenario, with a 2x2 contingency table, the constraint is that the total number of observations in each group is fixed. Once you know the number of conversions in one group, the number of non-conversions is automatically determined. This reduces the independent pieces of information by one, leading to 1 degree of freedom.

Key takeaway:

While the chi-squared distribution itself can have various degrees of freedom, in the context of a 2x2 A/B test with conversions and non-conversions, the df will always be 1 due to the inherent constraints in the data.

AB Terms

Overview

AB Cheat Sheet:

https://towardsdatascience.com/25-a-b-testing-concepts-interview-cheat-sheet-c998a501f911

Descriptive / Summary Statistics

- Population
- Sample
- Sample Mean
- Sample Variability

Experiment Design

- Null Hypothesis
- Key Metrics

- Lifetime Value (LTV)
- Objectives and Key Results (OKR)
- Overall Evaluation Criteria (OEC)
- Gaurdrail Metrics
- Randomization Unit
- Interference

A/B Test Statistics

- Confidence Level
- Margin of Error
- Confidence Interval
- Type I Error
- Type II Error
- p-Value
- Statistical Significance
- Statistical Power
- Minimum Detectable Effect
- Practical Significance
- Sample Size and Duration

Threats to Experiment Validity

- Novelty Effect
- Primary Effect
- Seasonality
- Day of the Week

Dictionary of Terms

https://marketingexperiments.com/a-b-testing/marketing-and-online-testing-dictionary

Descriptive / Summary Statistics

- **Descriptive/Summary Statistics:**
    - These are values that summarize and describe the main features of a dataset. They provide a concise overview of the data without making inferences about a larger population.
    - Descriptive statistics include:
        - Measures of central tendency (e.g., mean, median, mode).

        - Measures of variability or dispersion (e.g., range, variance, standard deviation).
        - Measures of frequency distribution (e.g., counts, percentages).
    - Essentially, they help you understand the "shape" of your data.

- **Population:**
    - In statistics, a population is the entire group of individuals, objects, or events that are of interest in a study.
    - It's the complete set of all possible observations.
    - For example, if you're studying the heights of all high school students in a country, then all high school students in that country constitute the population.

- **Sample:**
    - A sample is a subset of a population selected for study.
    - Because it's often impractical or impossible to study an entire population, researchers collect data from a sample and use it to make inferences about the population.
    - A good sample is representative of the population, meaning it reflects the characteristics of the population as a whole.

- **Sample Mean:**
    - The sample mean is the average of the values in a sample.
    - It's calculated by summing all the values in the sample and dividing by the number of values.
    - It's used as an estimate of the population mean.
    - It is often represented by the symbol $\bar{x}$.

- **Sample Variability:**
    - Sample variability refers to how spread out or dispersed the data points are in a sample.
    - It measures the extent to which the values in the sample differ from each other and from the sample mean.
    - Common measures of sample variability include:
        - **Variance:** The average of the squared differences from the mean.
        - **Standard deviation:** The square root of the variance, which provides a measure of variability in the same units as the data.
        - **Range:** The difference between the largest and smallest values.

- Sample variability is important because it indicates how representative the sample mean is of the data as a whole.

Experimental Design

- **Experiment Design:**
  - This is the process of planning a study to test a hypothesis. It involves determining how to manipulate variables, assign participants, and collect data to minimize bias and ensure valid results.
  - Key aspects include:
    - Defining the independent and dependent variables.
    - Establishing control groups.
    - Implementing randomization.
    - Choosing appropriate statistical methods.
- **Null Hypothesis:**
  - In hypothesis testing, the null hypothesis is a statement that there is no significant difference or effect. It's the default assumption that researchers aim to disprove.
- **Key Metrics:**
  - These are the specific, measurable values that are tracked to assess the performance or impact of an experiment or process. They are vital for determining whether objectives are being met.
- **Lifetime Value (LTV):**
  - LTV is a prediction of the total value a customer will bring to a business over the entire duration of their relationship. It's a crucial metric for understanding customer profitability.
- **Objectives and Key Results (OKR):**
  - OKR is a framework for setting and tracking goals.
  - Objectives are qualitative, aspirational goals.
  - Key Results are quantitative, measurable outcomes that indicate progress toward the objective.
- **Overall Evaluation Criteria (OEC):**
  - OEC refers to the primary metrics that will be used to judge the overall success of an experiment or project. It's the overarching measure of whether

the desired outcome was achieved.

- **Guardrail Metrics:**
    - These are metrics that are monitored to ensure that an experiment or change does not have unintended negative consequences. They act as "safety checks" to prevent harmful side effects.
- **Randomization Unit:**
    - This is the entity (e.g., individual, user, group) that is randomly assigned to different experimental conditions. It's the level at which randomization occurs.
- **Interference:**
    - In experimental contexts, interference occurs when the behavior of one experimental unit affects the behavior of another. This can compromise the validity of the results, as it violates the assumption of independent observations.

Overall Evaluation Criteria (OEC)

An Overall Evaluation Criterion (OEC) is a (usually composite) quantitative measure of the experiment's objective. Other names include Response or Dependent Variable, Outcome Variable, Evaluation metric, Performance metric.

https://www.analytics-toolkit.com/glossary/overall-evaluation-criterion/

Gaurdrail Metrics

Business metrics designed to indirectly measure business value and provide alerts about any potentially misleading or erroneous results and analysis.

https://www.split.io/glossary/guardrail-metrics/

Randomization Unit

A who or what randomly assigned to a group.

https://ianwhitestone.work/choosing-randomization-unit/

Data Leakage (Interference)

The behavior of the control group is influenced by the treatment given to the test group.

https://towardsdatascience.com/25-a-b-testing-concepts-interview-cheat-sheet-c998a501f911

SUTVA Assumptions

The Stable Unit Treatment Value Assumption (SUTVA) is a key assumption that is usually made in causal inference. Reference 1 gives a clear definition of SUTVA, which points out that SUTVA is really two assumptions rolled into one:

- The potential outcomes for any unit do not vary with the treatments assigned to other units.
- For each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes.

https://statisticaloddsandends.wordpress.com/2021/06/08/what-is-the-stable-unit-treatment-value-assumption-sutva/

The Stable Unit Treatment Value Assumption (SUTVA)

What is SUTVA?

SUTVA states that the outcome of a treatment on one individual should not be affected by the treatment assignment of other individuals. In simpler terms, there should be no interaction or interference between the treatment and control groups.

Why is SUTVA Important?

- **Valid Causal Inference:** SUTVA is essential for drawing valid conclusions about cause-and-effect relationships in A/B tests. If there's interference between groups, it becomes difficult to isolate the true effect of the treatment.
- **Unbiased Estimates:** Violations of SUTVA can lead to biased estimates of the treatment effect. This means your results may be inaccurate and misleading.

Example of SUTVA Violation

The passage provides a helpful example with Joe and Mary:

- Joe's blood pressure is the outcome of interest.
- The treatment is a drug that Mary might receive.
- If the drug causes Mary to cook with more salt, and Joe eats Mary's cooking, then Joe's blood pressure could be affected by Mary's treatment, even if Joe himself doesn't receive the drug. This is a violation of SUTVA.

How to Address SUTVA Violations

- **Identify Potential Interference:** Carefully consider the nature of your experiment and identify any potential sources of interference between groups.
- **Adjust Design:** If interference is likely, you may need to adjust your experimental design. This could involve:
    - Isolating groups to prevent interaction.
    - Accounting for the interference in your analysis (e.g., by including interaction terms in a statistical model).
    - Redesigning the experiment to measure the indirect effects as well.

In Summary

SUTVA is a critical assumption in A/B testing that ensures the treatment and control groups don't influence each other's outcomes. Violations of SUTVA can lead to biased and misleading results. By understanding and addressing potential SUTVA violations, you can improve the validity and reliability of your A/B testing conclusions.

Minimum Detectable Effect

Minimum Detectable Effect = Practical Significance Level

Minimum detectable effect (MDE) is a calculation that estimates the smallest improvement you are willing to be able to detect. It determines how "sensitive" an experiment is. Use MDE to estimate how long an experiment will take given the following:

- Baseline conversion rate
- Statistical significance
- Traffic allocation

https://support.optimizely.com/hc/en-us/articles/4410288881293-Use-minimum-detectable-effect-MDE-when-designing-an-experiment

Minimum effect is a business decision more than anything else, not really a data scientist decision. At work, it will typically be a product manager decision. After all, for that you need to take into account things like engineering costs, time, and opportunity-cost of not using those resources to run other tests. And that requires a comprehensive company vision which is typical of product managers, or VP/Director of product in smaller companies.

https://productds.com/wp-content/uploads/Sample_size.html

Practical, or Substantive, Significance

The fact that an estimated regression coefficient is “statistically significant” (i.e., you can reject the null hypothesis that the true β is 0 with a high level of confidence) does not mean that your independent variable is substantively important. Did it reach MDE?

A/B Testing and Statistical Significance

Power Analysis

- **Significance Level and Sample Size:** A lower significance level (alpha) means you require stronger evidence to reject the null hypothesis. This often necessitates a larger sample size to detect a true effect with higher confidence.
- **Statistical Power and Sample Size:** Higher statistical power (1 - beta) means a greater ability to detect a real effect. Increasing the sample size generally increases power.
- **MDE and Sample Size:** A smaller Minimum Detectable Effect (MDE) means you're looking for more subtle differences. Detecting smaller effects requires a larger sample size.

A/A Testing

- **Purpose:** A/A tests involve splitting users into two groups but giving them the *same* treatment. This helps assess random variability and understand how much difference can occur between groups simply due to chance.
- **Observations:**
    - Initially, you might see noticeable differences between the groups, even though they receive the same treatment.
    - Over time, these differences tend to decrease as the sample size increases and random variations average out.

Statistical Significance

- **Meaning:** Statistical significance indicates whether an observed effect is likely due to chance or a real difference. A low p-value (typically less than 0.05) suggests that the observed effect is unlikely to be due to random variation.
- **Misinterpretations:** The passage highlights common misunderstandings of statistical significance:
    - **Not Evidence of No Improvement:** A high p-value doesn't necessarily mean there's no improvement; it could just mean you don't have enough data to detect it.
    - **Confusing with Practical Significance:** Statistical significance doesn't guarantee practical significance. An effect might be statistically significant

but too small to be meaningful in a real-world context.
    - **Not the Likelihood of True Improvement:** Statistical significance doesn't tell you how likely the observed improvement is the true improvement. It only tells you how likely it is to observe the data if there were no improvement.
    - **Not the Likelihood of the Alternative Hypothesis:** Statistical significance doesn't directly tell you the probability of the alternative hypothesis being true or false.

Key Takeaways

- Power analysis helps determine the appropriate sample size for your A/B test.
- A/A testing helps assess random variability and understand the baseline differences between groups.
- Statistical significance should be interpreted carefully, avoiding common misinterpretations.
- Focus on both statistical and practical significance when making decisions based on A/B testing results.

Pooled Variance

Pooled variance and **pooled standard error** are used in statistics when you are comparing two or more samples and assume that the populations they came from have the **same variance**.

Pooled Variance

Pooled variance (also called combined, composite, or overall variance) is a single, weighted estimate of this common variance. It is calculated by "pooling" or combining the variance data from both samples to get a more robust estimate of the assumed shared population variance.

Pooled Standard Error

The **standard error of a sample** is simply the standard deviation of that sample, indicating how spread out the data is from the mean.

The **pooled standard error** is a measure that accounts for the variances of the two samples and assumes they are equal. It is called "pooled" because it uses the combined data from both samples to calculate a more precise estimate of the standard error for the difference between the two sample means.

The only conceptual difference between standard deviation and standard error is one of context:

- We refer to **standard deviations** when talking about a whole population.
- We refer to **standard errors** when talking about a sample.

Metrics

Discrete Metrics (Binomial Metrics)

Only two values are possible (0 or 1)

- Click-Through-Rate: Does a user click after seeing something
- Conversion Rate: Does a user become a customer after seeing something
- Click-Through-Probability: The probability a user clicks on the next step
- Bounce Rate: Percentage of people that land on your page and then leave

Continuous Metrics

- Average Revenue Per User
- Average Session Duration
- Average Order Value

Binomial Distribution

Successes and Fails

Formula:
$P(x: n,p) = \binom {n}{x} p^x (1 - p)^{(n-x)}$$P(x: n,p) = \binom {n}{x} p^x (1 - p)^{(n-x)}$

- n trials
- x successes

According to StatisticsHowTo (2022):

> A binomial distribution can be thought of as simply the probability of a SUCCESS or FAILURE outcome in an experiment or survey that is repeated multiple times. The binomial is a type of distribution that has two possible outcomes (the prefix “bi” means two, or twice). For example, a coin toss

> has only two possible outcomes: heads or tails and taking a test could have two possible outcomes: pass or fail.

https://www.statisticshowto.com/probability-and-statistics/binomial-theorem/binomial-distribution-formula/

Stephanie Glen. "Binomial Distribution: Formula, What it is, How to use it" From StatisticsHowTo.com: Elementary Statistics for the rest of us! https://www.statisticshowto.com/probability-and-statistics/binomial-theorem/binomial-distribution-formula/

```
# # define success - getting heads; A fair coins is flipped 10 times. What is t
# # vs getting 5 heads in a row 1/2^5
# from scipy import stats

# stats.binom.pmf(5, 10, .5)
```

Normal Distribution

Compare the formula for the normal distribution as shown below

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

with the python code below

1/(np.sqrt(2 * np.pi * sigma**2)) * np.exp(- (x - mu)**2 / (2 * sigma**2))

Let's break the code down:

- x = our set of numbers
- mu = mean
- sigma**2 (sigma squared) = variance of x (sigma = std)
- exp = exponential
- 1 is our numerator
- np.sqrt(2 * np.pi * sigma**2) = $\sqrt{2\pi\sigma^2}$$\sqrt{2\pi\sigma^2}$ NOTE: the two asterisk designate a power such as squared
- np.exp(- (x - mu)**2 / (2 * sigma**2)) = $\large{e^\frac{-(x-\mu)^2}{2\sigma^2}}$$\large{e^\frac{-(x-\mu)^2}{2\sigma^2}}$

Question:

"In an A/B test comparing two versions of a website landing page, we observed that the average time spent on the new version (treatment) was 2 minutes with a standard

deviation of 30 seconds. The control version had an average time spent of 1.8 minutes with a standard deviation of 45 seconds. Assuming the time spent on each page is normally distributed, is the difference in average time spent between the two versions statistically significant at a 95% confidence level?"

Why this can be answered with a normal distribution:

- **Assumption of Normality:** The question explicitly states that the time spent on each page is normally distributed. This is a key assumption for using a z-test or t-test, which are based on the normal distribution.
- **Comparing Means:** The question focuses on comparing the average time spent, which is a continuous variable. Tests of means (like the z-test or t-test) are appropriate for comparing continuous data that is normally distributed.
- **Confidence Level:** The question specifies a 95% confidence level, which directly relates to the critical value used in the normal distribution to determine statistical significance.

How to Answer the Question:

1. **Calculate the difference in means:** 2 minutes (treatment) - 1.8 minutes (control) = 0.2 minutes.
2. **Calculate the standard error of the difference:** This will depend on whether you're assuming equal variances or unequal variances between the groups. The formula involves the standard deviations and sample sizes of both groups.
3. **Calculate the test statistic (z or t):** Divide the difference in means by the standard error.
4. **Determine the p-value:** Use the test statistic and the normal distribution (or t-distribution if sample sizes are small) to find the p-value.
5. **Compare the p-value to alpha (0.05):** If the p-value is less than 0.05, the difference is statistically significant at the 95% confidence level.

Here's how to use the cumulative distribution function (CDF) and percent point function (PPF).

Scenario:

Recall that we're comparing the average time spent on two versions of a website landing page:

- Treatment: Average time = 2 minutes, Standard Deviation = 30 seconds
- Control: Average time = 1.8 minutes, Standard Deviation = 45 seconds

We want to determine if the difference in average time spent is statistically significant at a 95% confidence level, assuming the time spent on each page is normally distributed.

Using CDF and PPF

Here's how you can use the CDF and PPF in this scenario:

1. **Calculate the Test Statistic**
    - First, you'd calculate the test statistic (z or t) as mentioned. This involves calculating the difference in means and the standard error of the difference. Let's assume you've done that and obtained a z-score of `z_stat`.
2. **P-value using CDF**
    - To find the p-value, you can use the CDF of the standard normal distribution (`scipy.stats.norm.cdf`).
    - For a two-tailed test:
        - `p_value = 2 * (1 - stats.norm.cdf(abs(z_stat)))`
        - This calculates the area in both tails of the distribution beyond the absolute value of your test statistic.
3. **Critical Value using PPF**
    - Alternatively, you can find the critical value (z_critical) for a 95% confidence level using the PPF (`scipy.stats.norm.ppf`).
    - For a two-tailed test:
        - `alpha = 0.05`
        - `z_critical = stats.norm.ppf(1 - alpha/2)`
4. **Decision**
    - Compare the p-value to alpha (0.05) or compare the absolute value of the test statistic (`abs(z_stat)`) to the critical value (`z_critical`).
    - If `p_value < alpha` or `abs(z_stat) > z_critical`, you reject the null hypothesis and conclude that the difference in average time spent is statistically significant.

In Summary

The CDF and PPF are powerful tools in hypothesis testing. The CDF helps calculate the p-value, while the PPF helps determine the critical value. Both approaches can lead you to the same conclusion about whether to reject or fail to reject the null hypothesis.

```
# import numpy as np
# from scipy.stats import norm
# import matplotlib.pyplot as plt

# # --- Word Problem Scenario ---
```

```
# # A certain type of electronic component has a lifespan (in hours) that is
# # normally distributed with:
# MEAN = 5000  # hours (μ)
# STD_DEV = 500  # hours (σ)

# # --- PART 1: Find the Percentage (CDF) from a Standard Deviation (Z-score) -

# # Question: What percentage of components are expected to fail AFTER 6000 hou


# # 1. Convert the raw value (X) to a Z-score (standard deviations from the mea

# # print(f"Z-score for X = {X1} hours: Z = {Z1:.2f}")
```

```
# 2. Use the CDF to find the probability of being LESS than X1
# CDF(Z) gives P(X < x)


# print(f"P(X < {X1} hours) = {P_less_than_X1:.4f}")
```

```
# 3. Find the probability of being GREATER than X1
# P(X > x) = 1 - P(X < x)


# print(f"Probability (P > {X1} hours): {P_greater_than_X1:.4f}")
# print(f"Answer: {Percentage_X1:.2f}% of components will fail after 6000 hours
```

```
# # --- PART 2: Find the Standard Deviation (PPF) from a Percentage (Percentile

# # Question: The manufacturer wants to offer a warranty covering the shortest
# # 10% of component lifespans. What lifespan (X) marks this threshold?
# PERCENTILE = 0.10  # This is the 10th percentile

# # 1. Use the PPF (Inverse CDF) to find the Z-score corresponding to the 10th
# # PPF(P) gives the Z-score where P(X < x) = P

# # print(f"Z-score for the {PERCENTILE*100:.0f}th percentile: Z = {Z2:.2f}")
```

```
# 2. Convert the Z-score back to the raw value (X)
# Formula: X = MEAN + (Z * STD_DEV)


# print(f"Lifespan (X) for the {PERCENTILE*100:.0f}th percentile: {X2:.2f} hour
# print(f"Answer: The warranty threshold should be set at {X2:.0f} hours.\n")
```

```
# # --- Visualization (Optional) ---
# x = np.linspace(MEAN - 3 * STD_DEV, MEAN + 3 * STD_DEV, 100)
# pdf = norm.pdf(x, MEAN, STD_DEV)

# plt.figure(figsize=(10, 5))
# plt.plot(x, pdf, color='black')

# # Highlight P > 6000 (Z > +2.0)
# x_fill_cdf = np.linspace(X1, MEAN + 3 * STD_DEV)
# plt.fill_between(x_fill_cdf, norm.pdf(x_fill_cdf, MEAN, STD_DEV), color='red'
# plt.axvline(X1, color='red', linestyle='--', linewidth=1)

# # Highlight P < X2 (P < 10%)
# x_fill_ppf = np.linspace(MEAN - 3 * STD_DEV, X2)
# plt.fill_between(x_fill_ppf, norm.pdf(x_fill_ppf, MEAN, STD_DEV), color='blue
# plt.axvline(X2, color='blue', linestyle='--', linewidth=1)


# plt.title('Normal Distribution: CDF (Red) and PPF (Blue) Examples')
# plt.xlabel(r'Lifespan (Hours), $\mu={MEAN}, \sigma={STD_DEV}$')
# plt.ylabel('Probability Density')
# plt.legend()
# plt.grid(True)
# plt.show()
```

Conversion Rates

- Conversion rate: Conversion rates are calculated by simply taking the number of conversions and dividing that by the number of total ad interactions that can be tracked to a conversion during the same time period. For example, if you had 50 conversions from 1,000 interactions, your conversion rate would be 5%, since 50 ÷ 1,000 = 5%. https://support.google.com/google-ads/answer/2684489?hl=en
- Baseline conversion rate: Current conversion rate represented as a percentage
- A conversion can refer to any desired action that you want the user to take. This can include anything from a click on a button to making a purchase and becoming a customer. Websites and apps often have multiple conversion goals, and each will have its own conversion rate.

https://www.optimizely.com/optimization-glossary/conversion-rate/

One Tail vs Two Tail AB Tests in Terms of CVR

- Both Control and Treatment have equal conversion rates:
    - null $Control (A) = Treatment (B)$

    - alt $Control (A) \ne Treatment (B)$
- Treatment group's conversion rate is no better than the Control group's and could be worse:
    - null $Control (A) \geq Treatment (B)$
    - alt $Control (A) \lt Treatment (B)$
- One tail tests look for an improvement in customer experience

https://dominicsando.medium.com/why-two-sided-testing-is-reducing-your-a-b-testing-programs-impact-by-25-11d72276446a

Lift

Lift indicates if the treatment is better than the control. Lift is a percentage of how the treatment compares to the null hypothesis (control group).

Formula:

- (test - control) / control (* 100)

https://henrykpano.medium.com/a-b-testing-calculating-lift-rate-of-a-test-3d071514deb4

Designing an A/B Test for an E-commerce Product Page

Scenario:

Imagine you're on the product team of an online e-commerce store. The UX designer has created a new version of the product page, aiming to increase the conversion rate (the percentage of users who purchase a product). The current conversion rate is 13%, and the team wants to see if the new design can increase it to 15%.

1. Formulate Hypotheses

- We'll use a two-tailed test since the new design could perform better or worse than the current one.
- **Null Hypothesis (H0):** The conversion rate of the new design is the same as the old design (p = p0).
- **Alternative Hypothesis (Ha):** The conversion rate of the new design is different from the old design (p ≠ p0).

2. Set Significance Level and Confidence Level

- **Alpha (α):** 0.05. This means there's a 5% chance of incorrectly rejecting the null hypothesis (false positive).
- **Confidence Level:** 95%. This means we want to be 95% confident that our results accurately reflect the true difference in conversion rates.

3. Define Groups and Variables

- **Control Group:** Users who see the old product page design.
- **Experimental Group:** Users who see the new product page design.
- **Independent Variable:** The design of the product page (old vs. new).
- **Dependent Variable:** The conversion rate (whether the user buys the product).

Why Two Groups?

Having both a control and experimental group allows us to control for external factors (like seasonality) that could influence the results. By comparing the two groups, we can isolate the effect of the design change on the conversion rate.

4. Determine Sample Size with Power Analysis

- To ensure reliable results, we'll use a power analysis to determine the necessary sample size for each group.
- **Power (1 - β):** 0.8 (80%). This is the probability of detecting a real difference in conversion rates if one exists.
- **Alpha (α):** 0.05. Alpha (α) determines the significance level of your statistical test and represents the maximum risk you are willing to take of making a Type I error (a false positive).
- **Effect Size:** The difference between the current conversion rate (13%) and the desired conversion rate (15%).
- **Sample Size:** The minimum number of units (e.g., users, visitors, clicks, or transactions) required for each group (Control Group A and Variant Group B) to ensure that the experiment's results are reliable and statistically valid.

Test of Proportions

$z = \frac{\hat{p} - p}{\sqrt{\frac{p (1 - p)}{n}}}$$z = \frac{\hat{p} - p}{\sqrt{\frac{p (1 - p)}{n}}}$

```
# # https://www.kaggle.com/datasets/zhangluyuan/ab-testing?select=ab_data.csv
# import pandas as pd
# import numpy as np
# import scipy.stats as stats
# from scipy import stats
# from statsmodels.stats.proportion import proportions_ztest, proportion_confir
```

```
# import statsmodels.stats.api as sms
# from math import ceil

# df = pd.read_csv('https://raw.githubusercontent.com/gitmystuff/Datasets/main/
# print(df.shape)
# print(df.head())
# print()

# effect_size = sms.proportion_effectsize(0.13, 0.15)
# # effect_size = sms.proportion_effectsize(0.116, 0.136)
# sample_size = sms.NormalIndPower().solve_power(
#     effect_size, # minimum detectable effect
#     power=0.8,
#     alpha=0.05
#     )

# sample_size = int(sample_size)
# print('Effect Size: ', effect_size)
# print('Sample Size: ', sample_size)
# print()

# # delete repeat users
# session_counts = df['user_id'].value_counts(ascending=False)
# repeat_users = session_counts[session_counts > 1].count()
# print(f'Repeat users: {repeat_users}')

# users_to_drop = session_counts[session_counts > 1].index
# df = df[~df['user_id'].isin(users_to_drop)]
# print(df.shape)
# print()

# # create comparison groups
# control_sample = df[df['group'] == 'control'].sample(n=round(sample_size), ra
# treatment_sample = df[df['group'] == 'treatment'].sample(n=round(sample_size)

# ab_test = pd.concat([control_sample, treatment_sample], axis=0)
# ab_test.reset_index(drop=True, inplace=True)

# # print(ab_test.head())
# # print()
# print(ab_test['group'].value_counts())
# print()

# # conversion rates
# conversion_rates = ab_test.groupby('group')['converted']
# std_p = lambda x: np.std(x, ddof=0) # standard deviation
# se_p = lambda x: stats.sem(x, ddof=0) # standard error

# # conversion_rates = conversion_rates.agg([np.mean, std_p, se_p])
# # conversion_rates.columns = ['conversion_rate', 'std_deviation', 'std_error'
# # conversion_rates.style.format('{:.3f}')
```

```
# # test of proportions; never use t
# control_group = ab_test[ab_test['group'] == 'control']['converted']
# treatment_group = ab_test[ab_test['group'] == 'treatment']['converted']
# n_con = control_group.count()
# n_treat = treatment_group.count()
# successes = [control_group.sum(), treatment_group.sum()]
# n_obs = [n_con, n_treat]

# _, pval = proportions_ztest(successes, nobs=n_obs)
# (lower_con, lower_treat), (upper_con, upper_treat) = proportion_confint(succe

# print(f'p-value: {pval:.3f}')
# print(f'control group ci (95%): [{lower_con:.3f} - {upper_con:.3f}]')
# print(f'treatment group ci (95%): [{lower_treat:.3f} - {upper_treat:.3f}]')
```

```
# effect_size = sms.proportion_effectsize(0.13, 0.15)
# effect_size = sms.proportion_effectsize(0.116, 0.136)
```

Conclusions

- compare the p-value to $\alpha$$\alpha$ (0.05)
- check treatment ci upper bound - is it more than the desired rate?
- significant, statistically significant, practically significant

Complete the Following with New Numbers

Interpret the outcome regarding the treatment's effect and the new design's viability.

- Is it smaller than $\alpha$$\alpha$ (0.05)?
- What is the treatment upper CI compared to the desired output?
- Is it significant? Statitistically significant? Practically significant?
- Does the new design work?

Core A/B Testing Concepts

- **A/B Test** - A randomized experiment comparing two versions (A = control, B = variant) to measure the effect on a key metric.
- **Control Group** - The group that experiences the original version or baseline condition.

- **Treatment (Variant) Group** - The group that receives the changed version to test its effect.
- **Randomization** - Assignment of users to control or treatment groups by chance, ensuring comparability.
- **Randomization Unit** - The level at which random assignment is applied.
- **Interference** - When treatment effects spill over across units, violating independence.
- **Experiment Design** - The structured plan for assigning treatments, collecting data, and analyzing results.
- **Experimentation Platform** - Software system managing randomization, tracking, metrics, and analysis.

Descriptive Statistics and Population

- **Population** - The entire set of individuals or items of interest in a study.
- **Sample** - A subset of the population used to draw conclusions about the whole.
- **Sample Mean** - The arithmetic average of values in a sample, used to estimate the population mean.
- **Sample Variability** - The degree to which values in a sample differ from each other and from the mean.

Hypothesis, Metrics, and Evaluation

- **Null Hypothesis** - The default assumption that there is **no effect or difference** between groups.
- **Key Metrics** - The primary performance measures used to evaluate the success of an experiment.
- **Overall Evaluation Criteria (OEC)** - The agreed-upon primary metric that determines success.
- **Guardrail Metrics** - Secondary metrics monitored to ensure no harm to critical aspects of the business.
- **Lifetime Value (LTV)** - The predicted total revenue or profit a customer will generate over their relationship with a business.
- **Objectives and Key Results (OKR)** - A goal-setting framework linking broad objectives to measurable outcomes.

Statistical Concepts and Errors

- **Sample Size / Power Analysis** - Calculation of how many users are needed to detect a given effect with high probability (power).
- **Minimum Detectable Effect (MDE)** - The smallest effect size a test is designed to detect with adequate power.
- **Effect Size** - The magnitude of the difference in outcomes between groups.
- **Confidence Interval (CI)** - A range of plausible values for the effect size, often at 95% confidence.
- **Confidence Level** - The probability that a confidence interval captures the true population parameter.
- **Margin of Error** - The maximum expected difference between the sample estimate and the true population value.
- **p-Value** - The probability of observing the data (or more extreme) if the null hypothesis of "no difference" is true.
- **Statistical Significance** - A determination that observed results are unlikely due to chance, given a threshold ($\alpha$$\alpha$).
- **Statistical Power** - Probability that the test will detect a true effect when it exists.
- **Type I Error (False Positive)** - Concluding there is an effect when none exists. Incorrectly rejecting a true null hypothesis.
- **Type II Error (False Negative)** - Failing to detect a true effect. Failing to reject a false null hypothesis.
- **Practical Significance** - Whether a statistically significant effect is large enough to matter in practice.

Experimental Design Variants

- **A/A Test** - Test where both groups see the same version, used to check randomization and instrumentation.
- **A/B/n Test** - Extension of A/B testing to multiple variants (n > 2).
- **Multivariate Test (MVT)** - Tests multiple factors and their interactions simultaneously.
- **Bandit Algorithm** - Adaptive alternative to A/B testing that dynamically allocates more traffic to better-performing variants.
- **Crossover Design** - Participants switch between conditions to control for individual differences.

Tests Commonly Used in A/B Analysis

- **Two-Proportion z-Test** - Compares two independent proportions, such as conversion rates.
- **Chi-Square Test of Independence** - Compares observed vs. expected frequencies in a contingency table.
- **Fisher's Exact Test** - Exact test for independence in small 2x2 tables.
- **t-Test (Two-Sample)** - Tests whether the means of two groups differ.
- **Welch's t-Test** - Adjusted t-test that does not assume equal variance.
- **ANOVA (Analysis of Variance)** - Compares means across 3+ groups (A/B/n testing).
- **Sequential Testing / SPRT** - Allows tests to be monitored continuously with stopping rules.

Advanced and Related Terms

- **Multiple Testing / Bonferroni Correction** - Adjustments to control false positives when running many simultaneous tests.
- **False Discovery Rate (FDR)** - Expected proportion of false positives among declared significant results.
- **Bayesian A/B Testing** - Uses Bayesian inference to estimate posterior probabilities of variant superiority.
- **Uplift Modeling** - Predicts differential treatment effects at the individual level.
- **Interim Analysis / Peeking Problem** - Looking at test results before the planned stopping time, inflating false positives.

Threats to Experiment Validity

- **Novelty Effect** - A temporary change in user behavior when first exposed to a new feature.
- **Primary Effect** - The stable, long-term impact of a treatment after novelty fades.
- **Seasonality** - Natural fluctuations in user behavior due to time of year or external cycles.
- **Day of the Week** - Systematic differences in behavior between weekdays and weekends.
- **Sample Size and Duration** - The number of participants and time needed for adequate power and stable results.

Comprehensive Guide to Statistical Tests

Tests of Normality

- **Shapiro–Wilk Test** Definition: Tests whether sample data are drawn from a normally distributed population. Scenario: Before applying a t-test on exam scores, you check if the distribution of scores approximates normality.
- **Kolmogorov–Smirnov Test (with Lilliefors correction)** Definition: Compares sample distribution with a reference normal distribution. Scenario: A quality-control engineer tests if machine-part diameters follow normal distribution assumptions required for tolerance models.
- **Anderson–Darling Test** Definition: Emphasizes tails of the distribution when testing for normality. Scenario: In finance, you want to check if daily returns are normally distributed, focusing especially on outliers in tails.

Tests of Equal Variance (Homogeneity of Variance)

- **Levene's Test** Definition: Tests if multiple groups have equal variances. Scenario: Comparing math test variability across three teaching methods.
- **Bartlett's Test** Definition: Sensitive test for homogeneity of variances, assumes normality. Scenario: Testing whether experimental groups in a chemistry trial have equal measurement precision.
- **Brown–Forsythe Test** Definition: Robust test for equality of variances using median instead of mean. Scenario: Applied in psychology to compare mood ratings variance across treatments with non-normal data.

Tests of Means

- **One-Sample t-Test** Definition: Compares sample mean against a known or hypothesized population mean. Scenario: Checking if the average daily commute time in your city differs from the national average of 30 minutes.
- **Independent Samples t-Test** Definition: Compares means between two independent groups. Scenario: Evaluating whether two different fertilizers produce different average crop yields.
- **Paired Samples t-Test** Definition: Compares means from the same group at two time points. Scenario: Testing weight of patients before and after a 6-week exercise program.

- **Welch's t-Test** Definition: Adjusted t-test when groups have unequal variances. Scenario: Comparing salaries of two professions with different sample sizes and variances.

Tests of Proportions

- **Chi-Square Test of Independence** Definition: Assesses whether two categorical variables are associated. Scenario: Testing if voting preference is related to gender in a survey.
- **Chi-Square Goodness-of-Fit Test** Definition: Compares observed frequency distribution with expected distribution. Scenario: Determining if die rolls follow uniform distribution.
- **Fisher's Exact Test** Definition: Exact test for small sample categorical data. Scenario: In a clinical trial with 20 patients, testing if a treatment group has different recovery rates than control.
- **One-Proportion z-Test** Definition: Tests if a sample proportion equals a hypothesized proportion. Scenario: Testing if 60% of voters favor a candidate vs. the claimed 50%.
- **Two-Proportion z-Test** Definition: Compares proportions between two groups. Scenario: Comparing click-through rates between two website designs.

ANOVA and Related

- **One-Way ANOVA** Definition: Compares means across 3+ groups using variance analysis. Scenario: Testing average exam scores across four different teaching methods.
- **Two-Way ANOVA** Definition: Examines effects of two independent factors on one dependent variable, with or without interaction. Scenario: Testing effects of diet type and exercise plan on weight loss.
- **Repeated Measures ANOVA** Definition: Compares means of the same subjects across multiple conditions or time points. Scenario: Measuring cognitive performance of students at three different times of day.
- **MANOVA (Multivariate ANOVA)** Definition: Tests differences in group means across multiple dependent variables simultaneously. Scenario: Studying impact of medication on both blood pressure and cholesterol.

Regression and Model Fit

- **F-Test for Overall Regression** Definition: Tests whether regression model explains significant variance in outcome. Scenario: Testing whether advertising spend, price, and seasonality together predict sales.
- **Durbin–Watson Test** Definition: Tests for autocorrelation in regression residuals. Scenario: Checking for serial correlation in stock return predictions.
- **Hosmer–Lemeshow Test** Definition: Goodness-of-fit test for logistic regression. Scenario: Evaluating whether predicted probabilities of disease match observed outcomes in patient groups.

Nonparametric Tests

- **Mann–Whitney U Test** Definition: Nonparametric alternative to independent t-test. Scenario: Comparing satisfaction scores between two restaurants when data