Introduction

Design of experiments is a **cornerstone** of **statistics**, with **applications** in almost all areas of **research**. The **goal** is to **design an experiment** to **confirm or reject a hypothesis**.

Data scientists often conduct **continual experiments**, particularly in **user interface** and **product marketing**.

This chapter **reviews traditional experimental design**, **common challenges** in **data science**, and some **oft-cited concepts** in **statistical inference**, explaining their **meaning** and **relevance** (or **lack of relevance**) to **data science**.

Whenever you see **references** to **statistical significance**, **t-tests**, or **p-values**, it is usually in the context of the **classical statistical inference pipeline** (see **Figure 3-1**).

The process starts with a **hypothesis**, such as:

- **"Drug A is better than the existing standard drug"**
- **"Price A is more profitable than the existing price B"**

An **experiment** (e.g., an **A/B test**) is designed to **test the hypothesis**, aiming to deliver **conclusive results**. The **data** is then **collected and analyzed**, and a **conclusion** is drawn.

The term **inference** reflects the intention to **apply the results** from a **limited dataset** to a **larger population** or **process**.

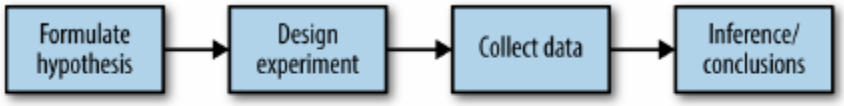


Figure 3-1. The classical statistical inference pipeline

I. A/B Testing

An **A/B test** is an **experiment** with **two groups** to determine which of **two treatments, products, or procedures** is **superior**.

Often, one group receives the **standard existing treatment** or **no treatment**, called the **control**.

A typical **hypothesis** is that a **new treatment** is **better than the control**.
Example: Testing a **new website design** against the **current design** to see which **increases clicks** more.

Key Terms for A/B Testing

Treatment
Something (drug, price, web headline) to which a subject is exposed.

Treatment group
A group of subjects exposed to a specific treatment.

Control group
A group of subjects exposed to no (or standard) treatment.

Randomization
The process of randomly assigning subjects to treatments.

Subjects
The items (web visitors, patients, etc.) that are exposed to treatments.

Test statistic
The metric used to measure the effect of the treatment.

A/B tests are common in web design and marketing, since results are so readily measured.

Ideally, subjects are randomized (assigned randomly) to treatments. In this way, you know that any difference between the treatment groups is due to one of two things:

- The effect of the different treatments
- Luck of the draw in which subjects are assigned to which treatments (i.e., the random assignment may have resulted in the naturally better-performing subjects being concentrated in A or B

You also need to pay attention to the test statistic or metric you use to compare group A to group B. Perhaps the most common metric in data science is a binary variable: click or no-click, buy or don't buy, fraud or no fraud, and so on. Those results would be summed up in a 2×2 table.

Table 3-1. 2×2 table for ecommerce experiment results

Outcome	Price A	Price B
Conversion	200	182
No conversion	23,539	22,406

If the metric is a continuous variable (purchase amount, profit, etc.) or a count (e.g., days in hospital, pages visited), the result might be displayed differently. If one were interested not in conversion but in revenue per page view, the results of the price test in Table 3-1 might look like this in typical default software output:
Revenue/page view with price A: mean = 3.87, SD = 51.10

Revenue/page view with price B: mean = 4.11, SD = 62.98
“SD” refers to the standard deviation of the values within each group.

Just because **statistical software** (like **R** or **Python**) generates **default output** does not mean all of it is **useful** or **relevant**.

For example, the **standard deviations** in the previous **price test** are not very **informative**: they suggest that some **values might be negative**, but **negative revenue** is not possible.

The **data** actually has a **small set of high values** (page views with **conversions**) and a **large number of 0-values** (page views with **no conversion**).

It is difficult to **summarize variability** of such data with a **single number**.
A better measure is the **mean absolute deviation** from the mean:

- **Price A:** 7.68
- **Price B:** 8.15

This is **more reasonable** than the **standard deviation** for **skewed data** with many **zeros**.

Why Have a Control Group?

Why not skip the control group and just run an experiment applying the treatment of interest to only one group, and compare the outcome to prior experience?

When you have a control group, it is subject to the same conditions (except for the treatment of interest) as the treatment group.

Blidning in studies

A **blind study** is one in which the **subjects** are **unaware** of whether they are receiving **treatment A** or **treatment B**. **Awareness** of the treatment can **affect the response**.

A **double-blind study** is one in which both the **investigators** and **facilitators** (e.g., **doctors** and **nurses**) are also **unaware** of which **subjects** are receiving which **treatment**.

Blinding is not possible when the **treatment is obvious**, for example: **cognitive therapy from a computer** versus therapy from a **psychologist**.

Important :

A/B testing in **data science** is often used in a **web context**.

- **Treatments** might include the **design of a web page**, the **price of a product**, the **wording of a headline**, or other elements.

- **Subjects** in the experiment are usually **web visitors**.
- **Outcomes** measured include **clicks**, **purchases**, **visit duration**, **number of pages visited**, or whether a **particular page** is visited.

In a **standard A/B experiment**, you must **decide on one metric ahead of time**. While multiple **behavior metrics** may be collected, if the goal is to **choose between treatment A and B**, a **single metric** or **test statistic** must be **predefined**.

Example: Choosing **web page design A or B** based on **click-through rate**.

Warning: Selecting a **test statistic after the experiment** introduces **researcher bias**.

Why Just A/B? Why Not C, D,...?

A/B tests are popular in **marketing** and **e-commerce**, but they are **not the only type of statistical experiment**.

- **Additional treatments** can be included.
- **Subjects** might have **repeated measurements**.
- **Pharmaceutical trials**—where **subjects are scarce, expensive, and acquired over time**—may be designed with **multiple opportunities** to **stop the experiment** and reach a **conclusion**.

Traditional experimental designs focus on answering a **static question** about the **efficacy of specified treatments**.

In **data science**, the focus is often **less on statistical significance** (e.g., **is price A better than price B?**) and more on **practical decision-making** and **continuous optimization**.

Example: A **website price test** might use **A/B/C variants** and **update prices dynamically** based on visitor behavior rather than waiting for a single final significance test.

Instead of asking the question:

"Is price A better than price B?", data scientists often ask: **"Which price, among multiple possible prices, is best?"**

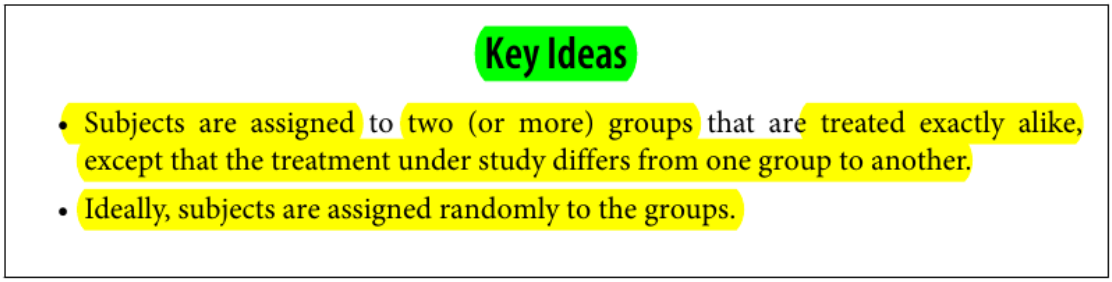
To answer this, a **relatively new type of experimental design** is used: the **multi-armed bandit**.

In **scientific and medical research** involving **human subjects**, researchers must obtain **participants' consent** and approval from an **institutional review board (IRB)**.

In **business experiments** conducted as part of **ongoing operations**, this is **rarely done**. For many cases such as **pricing experiments**, **headline testing**, or **offer selection**, this practice is **widely accepted**.

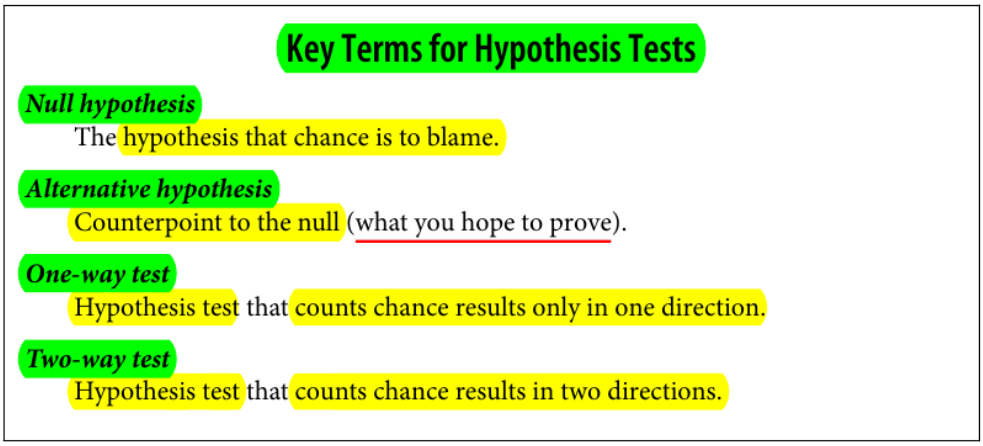
However, **Facebook** challenged this acceptance in **2014** when it conducted an experiment on the **emotional tone** of users' **newsfeeds**. Facebook used **sentiment analysis** to classify posts as **positive** or **negative**, then **manipulated** the balance shown to users. Some users saw **more positive posts**, while others saw **more negative posts**.

The experiment found that users exposed to **positive content** were more likely to **post positively**, and vice versa. Although the **effect size** was **small**, Facebook faced **strong criticism** for running the experiment **without users' knowledge**. Critics argued that exposing vulnerable users to **more negative content** could have caused **psychological harm**.



II. Hypothesis Tests

Hypothesis tests, also called significance tests help you learn whether random chance might be responsible for an observed effect.



An **A/B test** is usually designed with a **hypothesis** in mind, such as **"price B produces higher profit than price A."**

Why is a **hypothesis** needed? Why not simply choose the **treatment** that looks better after the experiment?

The reason is that humans tend to **underestimate natural randomness**. We often fail to anticipate **extreme events** (sometimes called **"black swans"**) and tend to **see patterns** in **purely random outcomes**.

Statistical hypothesis testing was created to **protect researchers** from being **misled by random chance**, helping them distinguish **real effects** from **random variation**.

Example: A price may appear better in a short test due to **random fluctuations**, not because it truly performs better in the long run.

*A **black swan** is an **event that is rare, unexpected, and has a major impact**, and is often **explained only after it happens**.*

This experiment illustrates the **human tendency to underestimate randomness**.

Ask several **friends** to **invent** a sequence of **50 coin flips** by writing **Hs and Ts** at random. Then ask them to **actually flip a coin 50 times** and record the results. Put the **real flips** in one pile and the **made-up sequences** in another.

It is usually **easy to tell** which is which: the **real coin flips** contain **longer runs** of **Hs or Ts**. In **50 real flips**, seeing **five or six Hs (or Ts) in a row** is **not unusual**. However, when people **invent randomness**, they tend to **avoid long runs**, switching outcomes after **three or four** in a row to make it "look random."

The **key lesson** is that people expect randomness to look **more mixed** than it really is.

The **flip side** is that when we observe a **real-world equivalent**—for example, when **one headline outperforms another by 10%**—we often assume there is a **real underlying cause**, when the result may be due simply to **random chance**.

In a properly designed A/B test, you collect data on treatments A and B in such a way that any observed difference between A and B must be due to either:

- Random chance in assignment of subjects
- A true difference between A and B

A statistical hypothesis test is further analysis of an A/B test, or any randomized experiment, to assess whether random chance is a reasonable explanation for the observed difference between groups A and B.

Null Hypothesis : The test begins with a **baseline assumption** that the **treatments are equivalent**, and any observed difference is due to **chance**. This assumption is called the **null hypothesis**.

🥅 Our goal is to **reject** the null hypothesis by showing that the outcomes of **groups A and B** differ **more than chance alone would explain**.

Hypothesis tests by their nature involve not just a **null hypothesis** but also an **offsetting alternative hypothesis**. Here are some examples:

- **Null = "no difference between the means of group A and group B"**
  Alternative = "A is different from B" (could be bigger or smaller)
- **Null = "A ≤ B"**
  Alternative = "A > B"
- **Null = "B is not X% greater than A"**
  Alternative = "B is X% greater than A

Taken together, the **null and alternative hypotheses** must account for **all possibilities**.
The **nature of the null hypothesis** determines the **structure of the hypothesis test**.

II.1 One-Way Versus Two-Way Hypothesis Tests

In many **A/B tests**, option **A** is the current default, and option **B** is a new alternative. The assumption is that you will **keep A unless B clearly performs better**.

Your goal is to see if **B is truly better than A**.

- If **B looks better than A** just by **random chance**, you **might mistakenly switch to B** when it's not actually better. That's the mistake you want to **avoid**.
- If **B looks worse than A** by chance, you don't care, because your default action is to **stay with A anyway**. You're **not at risk of making a bad decision** here.

🙂 Therefore, you use a **directional (one-sided / one-tailed) hypothesis test**, where the **alternative hypothesis is "B > A."**

If you want a **hypothesis test** to protect against being **fooled by chance in either direction**, the **alternative hypothesis** is **bidirectional**: **"A is different from B"** (could be **larger or smaller**).

In this case, you use a **two-tailed (two-way) hypothesis test**, where **extreme results in either direction** contribute to the **p-value**.

A **one-tailed hypothesis test** often fits **A/B decision-making**, where a **default option** is kept unless the **new option proves better**.

However, **software** like **R** or **Python's scipy** usually reports **two-tailed tests** by default. Many **statisticians** prefer the **conservative two-tailed test** to avoid debate.

One-tailed vs two-tailed tests can be confusing, but in data science, the exact p-value precision is often less critical because the focus is on practical decisions.

Key Ideas

- A null hypothesis is a logical construct embodying the notion that nothing special has happened, and any effect you observe is due to random chance.
- The hypothesis test assumes that the null hypothesis is true, creates a "null model" (a probability model), and tests whether the effect you observe is a reasonable outcome of that model.

II.2.Resampling

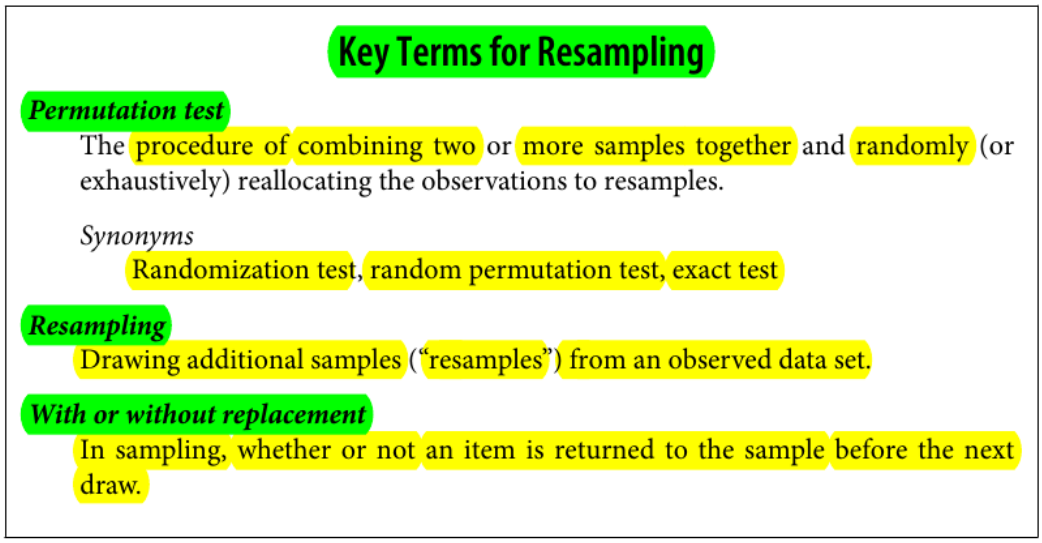
Resampling in **statistics** means repeatedly **sampling values from observed data** to assess **random variability** in a **statistic**.

It can also improve the **accuracy of machine-learning models**.
Example: In **decision trees**, predictions from multiple **bootstrapped datasets** can be **averaged** in a process called **bagging** (see **Random Forest**).

There are two main types of **resampling procedures**:

1. **Bootstrap** – used to **assess the reliability of an estimate**.
2. **Permutation tests** – used to **test hypotheses**, usually involving **two or more groups**.

Example: Using a **permutation test** to determine if **group A and B** differ significantly in an **A/B experiment**.



ii.3.Permutation Test

In a **permutation procedure**, two or more **groups** (e.g., A/B test groups) are involved. **Permute** means to **reorder a set of values**.
The first step is to **combine all groups' results** into a **single dataset**, representing the **null hypothesis** that the groups **do not differ**. Then, you **randomly draw new groups** from this combined set to see how much they differ by **chance**.

Permutation procedure steps:

1. **Combine** results from all groups into **one dataset**.
2. **Shuffle** the combined data and randomly draw a **resample** the same size as **group A**.
3. From the remaining data, draw a **resample** for **group B**.
4. Repeat for **groups C, D**, etc. Now you have **resamples matching original group sizes**.
5. Calculate the **statistic** (e.g., difference in group proportions) for the **resamples**; this is **one permutation iteration**.
6. **Repeat R times** to generate a **permutation distribution** of the test statistic.

After generating the **permutation distribution** of differences, compare the **observed difference** between groups to it:

- If the **observed difference** lies **within the range** of permuted differences, it could easily happen by **chance**, and we **cannot conclude a real effect**.

- If the **observed difference** lies **outside most of the permutation distribution**, it is **unlikely due to chance**, and we conclude the difference is **statistically significant**.

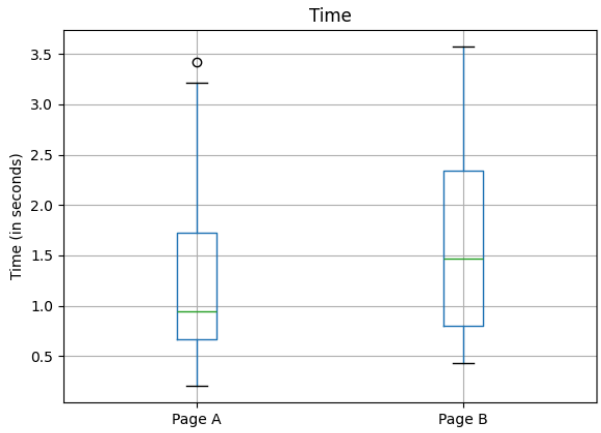
Example:
In an **A/B test** for **click-through rates**, if the **observed difference** between **Headline A and B** is larger than almost all differences from **permuted samples**, we infer that **Headline B truly performs better**, not just due to **random variation**.

Example: Web Stickiness

Scenario
- The company wants to know **which web page (A or B) is better** at leading to **sales**.
- **Actual sales** are too rare to measure quickly.
- They use a **proxy variable**: **average session time** (how long a visitor stays on the page).
- **Longer session time** = more interest = likely more sales.

Data
- **Page A:** 21 sessions
- **Page B:** 15 sessions
- Some sessions are recorded as **0** (because it was the last page visited) and are **removed**.

Comparing Pages
- **Boxplots** can be used to compare session times for **Page A** and **Page B**.
- Boxplots show the **distribution**, including **median, range, and variability**.

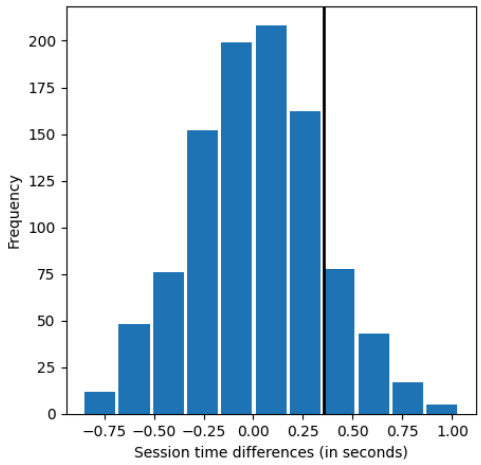


Observed difference:
- Page B sessions are on **average 35.67 seconds longer** than Page A.
- We want to know if this difference is **statistically significant** or could happen by **random chance**.

Permutation test idea:
1. **Combine all 36 session times** into a single dataset.
2. **Randomly shuffle** the combined data.
3. **Divide** the shuffled data into:

    - **21 values** for Page A
    - **15 values** for Page B
4. **Calculate the difference in means** for this random assignment.
5. **Repeat many times** to create a **permutation distribution** of mean differences.
6. Compare the **observed difference (35.67)** to this distribution to see if it is **extreme**.

This function works by sampling (without replacement) nB indices and assigning them to the B group; the remaining nA indices are assigned to group A. The difference between the two means is returned. Calling this function R = 1,000 times and specifying nA = 21 and nB = 15 leads to a distribution of differences in the session times that can be plotted as a histogram.



The histogram, in Figure 3-4 shows that the mean difference of random permutations often exceeds the observed difference in session times (the vertical line = mean_b − mean_a 'observed difference'). ***This suggests that the observed difference in session time between page A and page B is well within the range of chance variation and thus is not statistically significant.***

ii.3.Exhaustive and Bootstrap Permutation Tests

In addition to the preceding random shuffling procedure, also called a random permutation test or a randomization test, there are two variants of the permutation test:

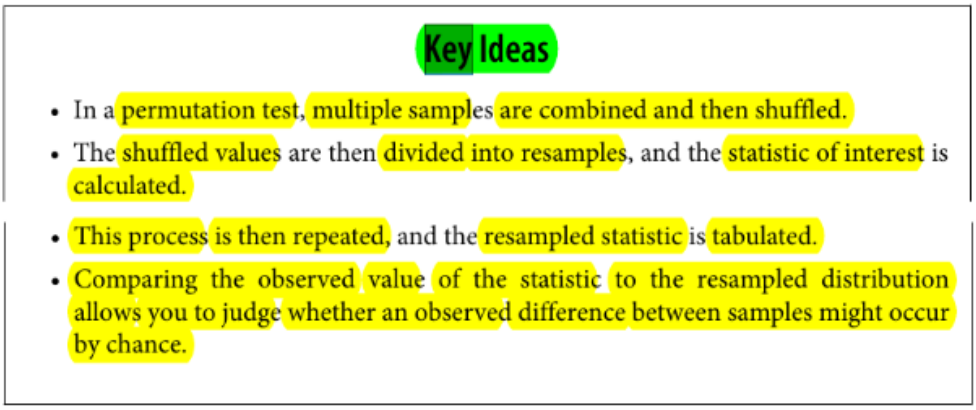
• An exhaustive permutation test (no replacement)
• A bootstrap permutation test (with replacement)

ii.4.Permutation Tests: The Bottom Line for Data Science

Permutation tests are **practical tools** to explore whether observed differences could happen by **random chance**.
Why they're useful in data science:

- **Easy to code, interpret, and explain**—no complicated formulas needed.
- Avoid the **false sense of certainty** that formula-based statistics sometimes give.
- **Flexible**:

- Works with **numeric or binary data**
- Works for **equal or unequal sample sizes**
- Does **not require normality assumptions**

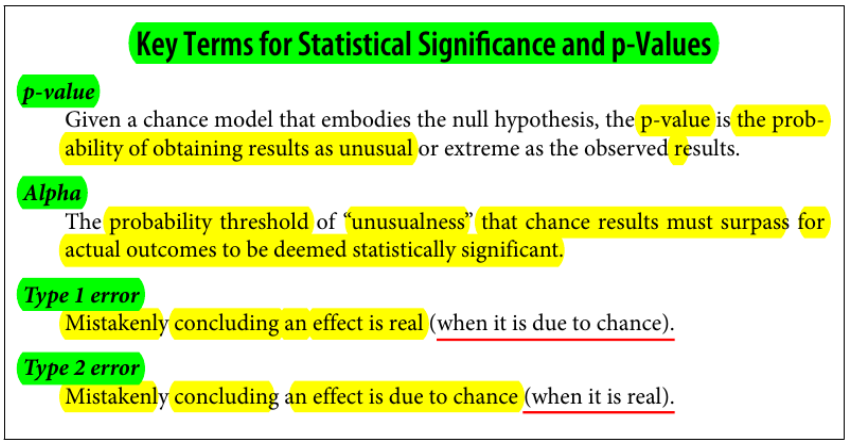


Key Ideas

- In a permutation test, multiple samples are combined and then shuffled.
- The shuffled values are then divided into resamples, and the statistic of interest is calculated.
- This process is then repeated, and the resampled statistic is tabulated.
- Comparing the observed value of the statistic to the resampled distribution allows you to judge whether an observed difference between samples might occur by chance.

III.Statistical Significance and p-Values

Statistical significance occurs when If the result is beyond the realm of chance variation,it is said to be statistically significant.



Key Terms for Statistical Significance and p-Values

p-value
Given a chance model that embodies the null hypothesis, the p-value is the probability of obtaining results as unusual or extreme as the observed results.

Alpha
The probability threshold of "unusualness" that chance results must surpass for actual outcomes to be deemed statistically significant.

Type 1 error
Mistakenly concluding an effect is real (when it is due to chance).

Type 2 error
Mistakenly concluding an effect is due to chance (when it is real).

Table 3-2. 2×2 table for ecommerce experiment results

Outcome	Price A	Price B
Conversion	200	182
No conversion	23,539	22,406

Scenario:

- Price A converts **0.8425%** of visitors (200 conversions out of 23,739)
- Price B converts **0.8057%** of visitors (182 conversions out of 22,588)
- Absolute difference: **0.0368 percentage points**
- Relative difference: **~4.6% better** for Price A

Key point:

Even with over **45,000 data points**, the actual **meaningful data—the conversions—are only in the hundreds**, so **statistical tests are still needed** to account for chance variation.

Permutation test idea:

- Null hypothesis: **both prices have the same conversion rate**

- Question: "If the true conversion rates were the same, could **random chance produce a difference as large as ~5%**?"
- Procedure: **resample conversions repeatedly** between the two prices and see how often a difference as big as the observed one occurs.

Takeaway for data science:

- High-volume data doesn't automatically remove the need for statistical tests.
- **Low-probability events** (like conversions <1%) mean that **effective sample size is determined by the number of actual events**, not total visits.

Procedure :

Permutation Test for Conversion Difference:

1. **Create the box of cards:**
    - 382 cards labeled **1** (conversion)
    - 45,945 cards labeled **0** (no conversion)
    - This represents the **combined conversion rate**: 382/46,327≈0.8246%
2. **Resample for Price A:**
    - Randomly draw **23,739 cards** (same size as Price A)
    - Count the number of 1s

3. **Resample for Price B:**
    - The remaining **22,588 cards** (same size as Price B)
    - Count the number of 1s

4. **Compute the difference:**
    - Difference in conversion rates between the two resamples

5. **Repeat steps 2–4** many times to build a distribution of differences
6. **Assess significance:**
    - Count how often the simulated difference is **≥ 0.0368%** (the observed difference)

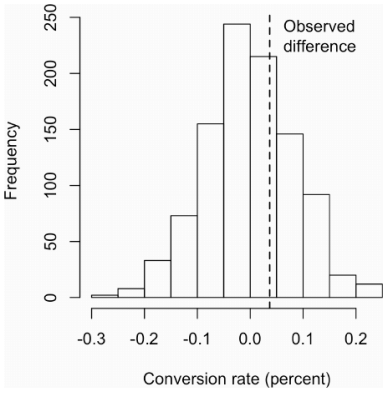


Figure 3-5. Frequency distribution for the difference in conversion rates between prices A and B

See the histogram of 1,000 resampled results in *Figure 3-5:* as it happens, in this case the observed difference of 0.0368% is well within the range of chance variation.

Explanation of ‘line within range of differences of permutations”:

Those ranges are obtained by chance by random, if the vertical line is within this range, it is mostly like this real difference is due to chance also, so we cant say that A is better than B, because difference is not statistically significant.

III.1.p-value

Simply looking at the graph is not a very precise way to measure statistical significance, so of more interest is the p-value.

In python, We can estimate a p-value from our permutation test by taking the proportion of times that the permutation test produces a difference equal to or greater than the observed difference.

Here p_value = 0.314. The **p-value of 0.308** means that if there were **no real difference** between prices A and B, a difference **as large or larger than 0.0368** would occur by **random chance about 31.4 % of the time**.

Because the data represent **counts of successes and failures** (conversions vs. no conversions), the situation naturally follows a **binomial distribution**. In such cases, we don’t need to rely on a **permutation test** (which reshuffles the data many times to estimate significance) to get a p-value.

Instead, we can use standard **statistical formulas or functions**—like:

```python
import numpy as np
from scipy import stats
# Data: [successes, failures] for each group
survivors = np.array([[200, 23739 - 200],
                      [182, 22588 - 182]])

# Perform chi-square test
chi2, p_value, df, _ = stats.chi2_contingency(survivors)

# Single-sided p-value
print(f'p-value for single sided test: {p_value / 2:.4f}')
```

Binomial/count data → chi-square test or normal approximation.

Non-binomial/continuous data → permutation test, bootstrap, or parametric tests depending on assumptions.

The normal approximation yields a p-value of 0.3498, which is close to the p-value obtained from the permutation test.

```
Observed difference: 0.0368%
p_value =  0.334
p-value for single sided test: 0.3498
```

III.2.Alpha

Statisticians **don’t let researchers decide after the fact** if a result is “too unusual.” Instead, a threshold, called **alpha (α)**, is set **before** the experiment. Common alpha levels are **5% or 1%**.

f a result would happen by chance **less than 5% of the time**, we call it statistically significant.

- If the **p-value < α (e.g., 0.05)** → we **reject H_0** → result is **statistically significant**.
- If the **p-value ≥ α** → we **fail to reject H_0** → result is **not statistically significant**.

P-value controversy: Many researchers misuse p-values because they misunderstand what they actually mean. A p-value **does not measure the probability that a hypothesis is true**, nor does it measure the importance of a result.

Problem in practice: Some researchers "data mine" or try multiple hypotheses until they find one with a p-value below 0.05. This can make random chance look like a significant finding.

Example:

- A psychologist tests 20 different ways to correlate sleep with mood. Only one gives **p = 0.03**. They report it as "significant," ignoring that this could easily happen by chance.

Result: Journals started banning p-values because relying solely on them led to poor-quality research and false claims of discovery.

This shows why understanding **what a p-value really represents** is crucial.

The real problem is that people want more meaning from the p-value than it contains. Here's what we would like the p-value to convey: *The probability that the result is due to chance.*

We hope for a low value, so we can conclude that we've proved something. This is how many journal editors were interpreting the p-value. But here's what the p-value actually represents:

احتمالية حدوث نتائج متطرفة ,في ظل نموذج عشوائي

Subtle but important: A "statistically significant" p-value does **not** prove your hypothesis. It only tells you how compatible the data are with the null hypothesis.

ASA 2016 Principles:

1. P-values indicate **how incompatible the data are with a given model**.
2. P-values do **not** give the probability that the hypothesis is true or that the data are purely random.
3. Decisions should **not rely solely on a p-value threshold**.
4. Full reporting and transparency are essential for proper inference.
5. P-values **do not measure effect size or importance**.
6. Alone, a p-value is **not a reliable measure of evidence** for a model or hypothesis.

Practical significance vs. statistical significance:

- **Statistical significance**: A result is unlikely to occur by chance (e.g., $p < 0.05$).
- **Practical significance**: The result is large or meaningful enough to matter in real-world terms.

Key point: Large sample sizes can make **tiny, unimportant differences statistically significant**. Just because chance is ruled out doesn't mean the effect matters practically.

Example:

- A new website design increases conversion from 0.8057% → 0.8425% (difference = 0.0368%).
- With 45,000 users, this difference may be statistically significant, but in practical terms, it's **almost negligible**.

Takeaway: Always consider the **size and real-world impact** of the effect, not just the p-value.

Iii.3. Type 1 and Type 2 Errors

When testing a hypothesis, two types of errors can occur:

- **Type 1 error (false positive):** Concluding there is an effect when it's actually due to chance.
- **Type 2 error (false negative):** Concluding there is no effect when there actually is one. Often occurs because the **sample size is too small** to detect the effect.

Key point:

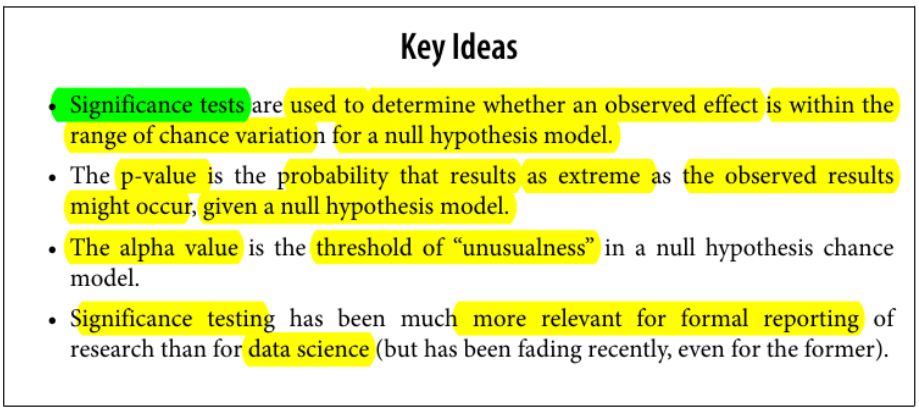
- A p-value above the significance threshold (e.g., >5%) does **not prove no effect**—it just means the effect is **not proven**.
- Significance tests are mainly designed to **minimize Type 1 errors**, protecting against being fooled by random chance.

Iii.4. Data Science and p-Values

For data scientists, p-values are **practical tools**, not absolute rules. They help assess whether an observed effect or model result is likely due to **normal chance variability**.

Key points:

- P-values are **informational**, not controlling—used as one factor in decision-making.
- They can guide choices in models, e.g., deciding whether to include or exclude a feature based on its p-value.
- Unlike scientific research, the goal isn't publication but **making data-driven decisions**.



Key Ideas

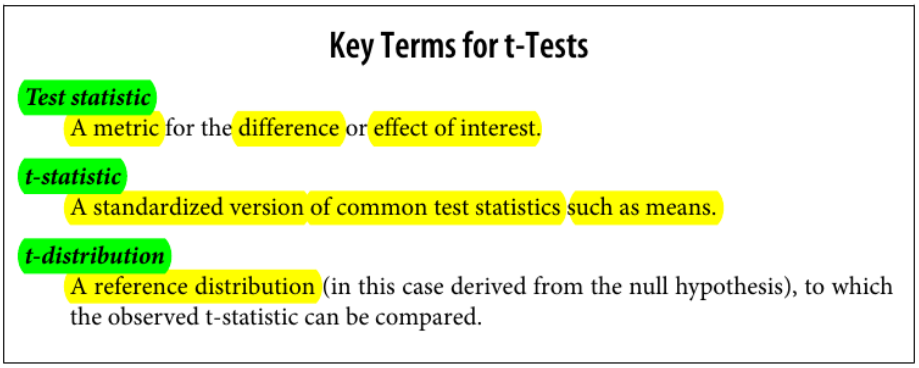
- Significance tests are used to determine whether an observed effect is within the range of chance variation for a null hypothesis model.
- The p-value is the probability that results as extreme as the observed results might occur, given a null hypothesis model.
- The alpha value is the threshold of "unusualness" in a null hypothesis chance model.
- Significance testing has been much more relevant for formal reporting of research than for data science (but has been fading recently, even for the former).

IV. t-test

There are numerous types of significance tests, depending on whether the data comprises count data or measured data.

There are many types, depending on the **data type** (counts vs. measurements), the **number of samples**, and what is being measured.

A common example is the **t-test**: Used to test whether a **sample mean** differs significantly from a reference value or another sample mean.



Key Terms for t-Tests

Test statistic
A metric for the difference or effect of interest.

t-statistic
A standardized version of common test statistics such as means.

t-distribution
A reference distribution (in this case derived from the null hypothesis), to which the observed t-statistic can be compared.

t-test require a **test statistic** that quantifies the effect of interest and helps decide if the observed effect is **beyond what random chance could produce**.
In **resampling tests** (e.g., permutation tests), the **data scale doesn't matter**—the null distribution is built from the data itself, and the observed test statistic is compared directly to it.

In the 1920s–1930s, resampling thousands of times wasn't practical. Statisticians realized that the **t-test** (based on Gosset's t-distribution) approximates the permutation distribution for numeric two-sample comparisons (gives t-statistic and p-value). To use it **independent of data scale**, the test statistic is standardized (observed difference means between A and B -> calculate one 't' -> conclusion).

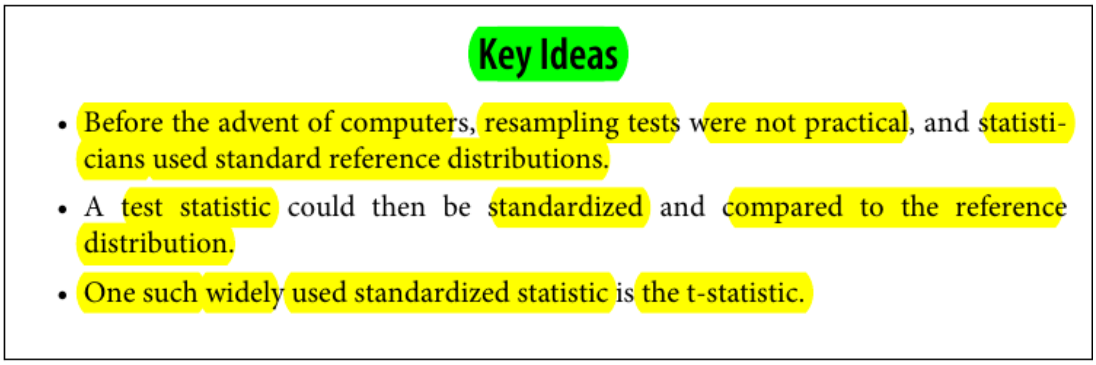
*In a two-sample t-test (like an A/B test), you usually get **one t-statistic** that summarizes the difference between the two groups. If you do **multiple comparisons** (e.g., testing several metrics or several pairs of groups), you'll get **multiple t-values**—one for each comparison. Each t-value is a **standardized measure of difference** for that specific pair of groups.*

The t-test doesn't replace a permutation test in principle, but in practice, for numeric two-sample comparisons.

Software like **R or Python** automatically standardizes your data and compares it to the t-distribution, giving you the **t-value** and **p-value** directly.

- The **alternative hypothesis** is that **Page B has a higher mean session time than Page A**.
- The **t-test p-value (0.1408)** is similar to the **permutation test p-values (0.121–0.126)**, showing both methods agree that the difference is **not statistically significant**.
- **Resampling methods** (like permutation tests) are flexible: they work regardless of whether data is numeric or binary, balanced or unbalanced, or has unequal variances.
- In contrast, **formula-based methods** (t-tests, confidence intervals) have many ***variations and rules***, which statisticians must navigate carefully.

- **For data scientists, the focus is practical: using these tests as tools to guide decisions, without needing to memorize all the formulaic nuances.**



V. multiple testing

"Torture the data long enough, and it will confess." This means that if you look at the data through enough different perspectives and ask enough questions, you almost invariably will find a statistically significant effect.

When you run **many significance tests**, even if *nothing real is happening*, you are likely to get **false positives** (Type 1 errors).

Example in the text:

- You have **20 predictors**, all random (no real effect).
- You test each at **α = 0.05**.
- Each test has a **5% chance** of being falsely "significant."

Key logic:

- Probability one test is **not** significant = 0.95
- Probability **all 20** are not significant = 0.9520≈0.360.95^{20} ≈ 0.360.9520≈0.36
- Probability **at least one false significant result** = 1−0.36=0.641 − 0.36 = 0.641−0.36=0.64 → **64% chance**

Meaning:

Even though each test uses α = 0.05, doing many tests **inflates the overall false-positive risk**.

This is called **alpha inflation** (or multiple testing problem).

This means that when you test **many variables or many models**, you increase the chance of finding a **"significant" result that is actually just random noise**, not a real effect.

Why this happens:

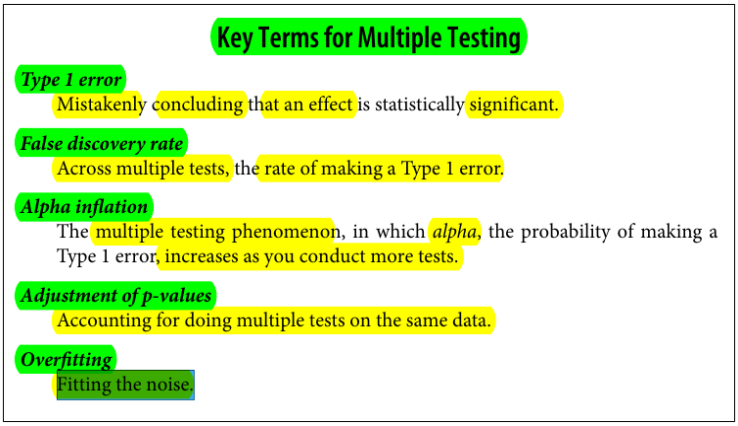
- Random data always contains patterns by chance.
- The more you search (more variables, more models), the more likely you are to **find a fake pattern**.
- This is called **overfitting**: the model fits the noise instead of the true signal.

Simple example:

- You try 100 random predictors on an outcome.
- Even if none truly matter, about **5 will look significant** at α = 0.05 just by luck.

Key idea for data science:

- "Significant" does **not** always mean "real."
- Use validation data, regularization, and multiple-testing corrections to avoid fitting noise.



In supervised learning tasks, a holdout set where models are assessed on data that the model has not seen before mitigates this risk. In statistical and machine learning tasks not involving a labeled holdout set, the risk of reaching conclusions based on statistical noise persist

When you run **multiple hypothesis tests**, the chance of getting a **false positive (Type 1 error)** increases. To control this, statistics uses **multiple-testing adjustments** that make it **harder** for any single test to be called "significant."

Why this is needed

- Each test has its own chance of being wrong.
- More tests ⇒ higher chance that **at least one** looks significant **just by chance**.

How adjustments work : They **lower the alpha level** for each test (i.e., raise the bar for significance).

Common methods

- **Bonferroni correction**
  - Simple and conservative
  - New alpha = α / number of tests
  - Example: α = 0.05, 3 tests → each test uses 0.0167

- **Tukey's HSD (Honest Significant Difference)**
  - Used when comparing **multiple group means**
  - Controls false positives across **all pairwise comparisons**
  - Conceptually similar to a **permutation approach** that looks at the **largest mean difference** expected by chance

There is what is called **data dredging (a.k.a. p-hacking)**—repeatedly searching the same data in many ways until something looks "significant."

Concise explanation (key ideas for data science):

- **More research ≠ better research.**
  Because of *multiplicity* (many tests, models, and questions), false findings are common. The Bayer example shows this clearly: most published results **failed to replicate**, meaning they were likely driven by chance, not real effects.

- **Why classic statistical corrections aren't enough for data science:**
  Methods like Bonferroni are designed for **narrow, predefined tests**. Data science work is usually exploratory, iterative, and high-dimensional, so those corrections are often **too rigid and impractical**.

What data scientists should do instead

1. Predictive modeling

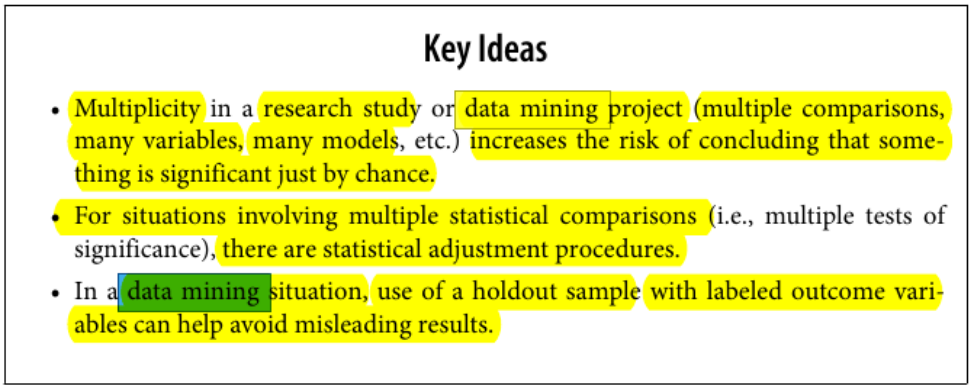
- Use **cross-validation** and **holdout/test sets**
- This checks performance on unseen data and reduces the risk of chance-driven models

2. Exploratory or non-predictive analysis (no holdout set)

- Stay aware: **the more you explore, the more chance can fool you**
- Use **resampling and simulation** (permutation tests, bootstrapping) to:
  - Estimate what "chance alone" would produce
  - Compare observed results against that benchmark

Bottom line

Data scientists don't eliminate multiplicity—they **manage it** using validation, resampling, and skepticism about "interesting" results found after heavy data exploration.



Key Ideas

- Multiplicity in a research study or data mining project (multiple comparisons, many variables, many models, etc.) increases the risk of concluding that something is significant just by chance.
- For situations involving multiple statistical comparisons (i.e., multiple tests of significance), there are statistical adjustment procedures.
- In a data mining situation, use of a holdout sample with labeled outcome variables can help avoid misleading results.

VI. Degrees of Freedom

What are *degrees of freedom* (DoF)?

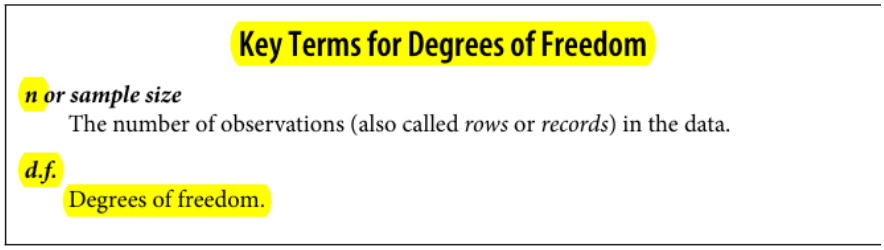
Degrees of freedom = the number of values that are free to vary after constraints are applied.

Simple example

- You have **10 numbers**.
- If you already know their **mean**, then:
  - Only **9 numbers can vary freely**

- The **10th is fixed** to keep the mean correct
  👉 So, **degrees of freedom = 10 − 1 = 9**

Why does this matter?
When we estimate statistics (like **variance**) from a **sample**, we lose one degree of freedom because the sample mean is estimated from the data.
If we divide by **n**, the variance estimate is **too small (biased downward)**.
Dividing by **n − 1** corrects this bias.



In traditional statistics, tests like the t-test or F-test use **standardized statistics** (e.g., t-values or F-values) to compare observed data to a reference distribution.

Degrees of freedom (df) is used in this standardization to account for how many values are free to vary, ensuring the standardized statistic correctly matches the shape of the reference distribution (t-distribution, F-distribution, etc.).

Scenario: You test whether a new webpage B keeps users longer than webpage A.

- Page A session times: `[50, 60, 55, 52, 48]`
- Page B session times: `[60, 65, 58, 62, 61]`

1. **Compute t-statistic**: Measures the difference between group means relative to the variability within groups. Suppose we get `t = 3.0`.
2. **Reference distribution**: Student's t-distribution with `df = nA + nB - 2 = 8` degrees of freedom.
   DoF **affects the shape** of distributions:
   a. Fewer DoF → wider, heavier tails
   b. More DoF → closer to normal distribution
3. **Compare t-value**: Check how extreme `t = 3.0` is in the t-distribution with 8 df.
4. **Result**: If only 1% of the t-values in this distribution are ≥ 3.0, p-value = 0.01 → difference is statistically significant.

So, the t-distribution tells us **how unusual our observed t-value is under the null hypothesis**.

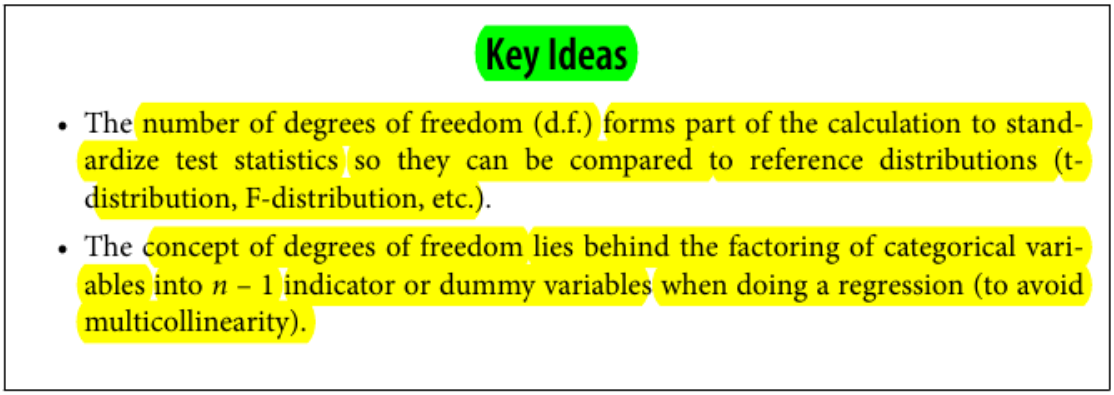
In data science, degrees of freedom and small-sample corrections matter less because:

1. **Data sets are usually large, so the difference between dividing by nnn or n−1n-1n−1 is negligible.**
2. **Formal statistical tests are used sparingly—resampling, simulations, or machine learning models are often more relevant.**

➡ In short: for most practical data science work, the small technical details of df or sample-size corrections don't change the results much.

Instead, data scientists often rely on:

- **Resampling methods** (permutation tests, bootstrap)
- **Simulations** to estimate variability
- **Cross-validation** to check models
- **Visualizations** to explore differences or trends

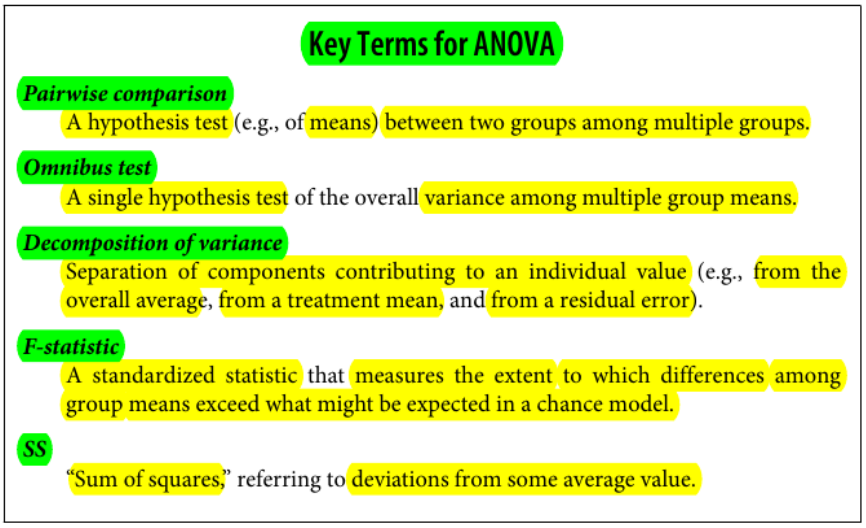


Key Ideas

- The number of degrees of freedom (d.f.) forms part of the calculation to standardize test statistics so they can be compared to reference distributions (t-distribution, F-distribution, etc.).
- The concept of degrees of freedom lies behind the factoring of categorical variables into $n - 1$ indicator or dummy variables when doing a regression (to avoid multicollinearity).

VII. ANOVA

Suppose that, instead of an A/B test, we had a comparison of multiple groups, say A/B/C/D, each with numeric data. The statistical procedure that tests for a statistically significant difference among the groups is called analysis of variance, or ANOVA.



Key Terms for ANOVA

Pairwise comparison
A hypothesis test (e.g., of means) between two groups among multiple groups.

Omnibus test
A single hypothesis test of the overall variance among multiple group means.

Decomposition of variance
Separation of components contributing to an individual value (e.g., from the overall average, from a treatment mean, and from a residual error).

F-statistic
A standardized statistic that measures the extent to which differences among group means exceed what might be expected in a chance model.

SS
"Sum of squares," referring to deviations from some average value.

This example describes a small web experiment with **four pages** and **five visitors per page**, measuring **"stickiness"** (time spent on the page). Key points:

Table 3-3. Stickiness (in seconds) of four web pages

	Page 1	Page 2	Page 3	Page 4
	164	178	175	155
	172	191	193	166
	177	182	171	164
	156	185	163	170
	195	177	176	168
Average	172	185	176	162
Grand average				173.75

Each column in the table is an **independent set of data** (different visitors).

Visitors are **not randomly sampled** from the entire population—they come as they arrive. This means there could be **bias**: time of day, device, season, etc., might affect results. When reviewing results, these potential biases should be considered, because they can influence which page appears "stickier."

The more such pairwise comparisons we make, the greater the potential for being fooled by random chance (type1 error). Instead of testing all pairs separately, we can do **one overall test** (ANOVA) to ask:

"Could all pages have the same average stickiness, with the observed differences due to random chance?"

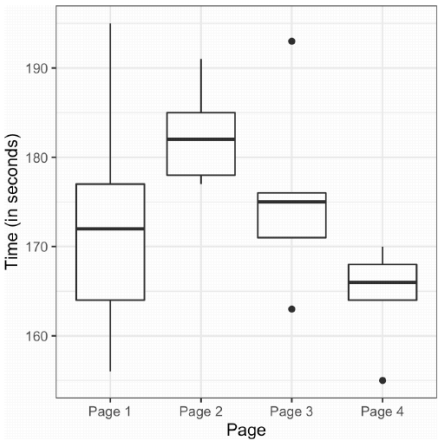


Figure 3-6. Boxplots of the four groups show considerable differences among them

This explains **ANOVA** using a **resampling (permutation) approach**:

1. Combine all data from the four pages.
2. Shuffle and randomly split into four groups of five (same as original).
3. Calculate the mean of each group.
4. Calculate the variance among these four means.
5. Repeat many times (e.g., 1,000) to build a **null distribution** of variances.
6. Compare the **observed variance** to this null distribution:
    - The **p-value** = proportion of resampled variances ≥ observed variance.

Since 9.3% > 5% (the usual significance threshold), the observed differences are **not statistically significant**.
Conclusion: the variation among the four pages **could easily happen by chance**, so we don't have strong evidence that any page is truly stickier than the others.

VII.1 F statistics

Just like a t-test replaces a permutation test for comparing two group means, ANOVA uses the F-statistic to test multiple group means.
F-statistic = variance between group means ÷ variance within groups (residual error).
A higher F means group differences are large relative to random variation → more likely significant.
If data is normally distributed, the F-statistic follows an F-distribution, allowing us to compute a p-value to assess significance.
In short: ANOVA formalizes the comparison of several means, just as t-tests do for two, using variance ratios instead of resampling.

Degree of freedom (the example)

Simple example (one mean)

You have **5 numbers** with a known mean.

- You can choose **4 numbers freely**

👉 **Degrees of freedom = 5 − 1 = 4**

Apply this to ANOVA

1️⃣ **Total degrees of freedom**

You have **20 observations** total.

- Estimating the **grand mean** uses **1 df**

👉 **Total df = 20 − 1 = 19**

Between-groups (Page)

You have **4 group means**.

- Once 3 group means are known **and** the grand mean is fixed,
- the 4th group mean is forced

👉 **df_between = 4 − 1 = 3**

Within-groups (Residual)

Each group has **5 observations**.

- Estimating **1 mean per group** uses 1 df
- Remaining df per group = 5 − 1 = 4
- For 4 groups: 4 × 4 = 16

👉 **df_residual = 16**

VIII. Two-way ANOVA

One-way ANOVA

- **One factor only** (one reason for differences)
- Example: **web page (A/B/C/D)**
- Question: *Do average session times differ between pages?*

Two-way ANOVA

- **Two factors** vary at the same time

- Example:
    - Factor 1: **Page (A/B/C/D)**
    - Factor 2: **Day type (Weekend vs Weekday)**

Now we ask **three questions**:

1. Do pages differ on average? (**page effect**)
2. Do weekends differ from weekdays overall? (**day effect**)
3. Does the weekend effect depend on the page? (**interaction effect**)

Interaction effect (key idea)

An **interaction** means:

The effect of one factor depends on the level of the other factor.

Example:

- Page A performs much better on weekends
- Page B shows no weekend change

➡ This difference is **not explained by page alone or day alone**, but by their **interaction**.

Why this matters

Two-way ANOVA:

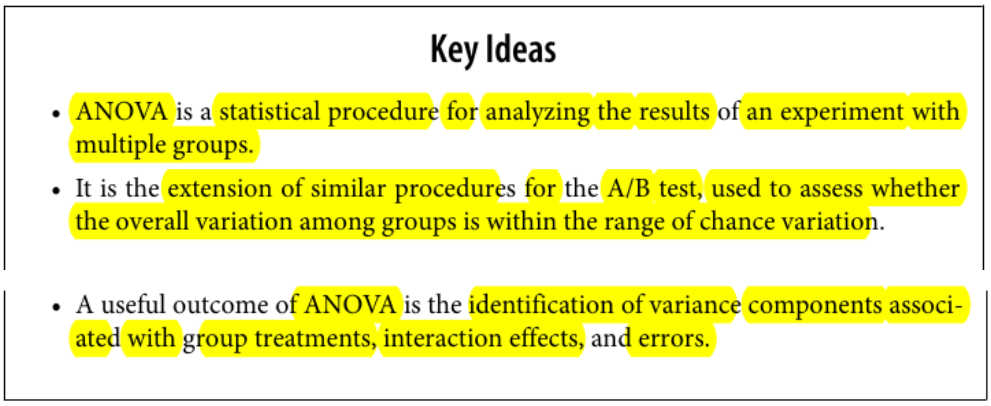
- Still decomposes variance (like one-way ANOVA)
- But splits it into:
    - Page effect
    - Day effect
    - Page × Day interaction
    - Residual error

Big picture

ANOVA → Two-way ANOVA → Regression

All are part of the **same modeling idea**:

Explain variation in data using multiple factors and their combined effects.



Key Ideas

- ANOVA is a statistical procedure for analyzing the results of an experiment with multiple groups.
- It is the extension of similar procedures for the A/B test, used to assess whether the overall variation among groups is within the range of chance variation.
- A useful outcome of ANOVA is the identification of variance components associated with group treatments, interaction effects, and errors.

VIII. Chi-Square Test

When web experiments test **more than two versions at once** (A/B/C/D, etc.), and the data are **counts** (e.g., clicks vs. no clicks), we often use the **chi-square test**.

The **chi-square test** checks whether the observed counts differ from what we would expect **by chance** under a **null hypothesis**.

✅ Chi-square test (categorical / count data)

📌 **Question** **Does page version affect conversion?**

📊 **Data (counts)**

Page	Converted	Not converted
A	200	23,539
B	182	22,588

- Outcome: **Converted or not** (yes / no)
- Data type: **Counts**
- Goal: Check **association** between *Page* and *Conversion*

Interpretation

- Small p-value → page version affects conversion
- Large p-value → difference could be due to chance

✅ ANOVA (numeric data)

📌 **Question : Do visitors spend different amounts of time on different pages?**

📊 **Data (numeric)**

Page A	Page B	Page C
164	178	175
172	191	193
177	182	171
156	185	163
195	177	176

- Data type: **Continuous**
- Goal: Compare **means across groups**

1. Actual data

	Headline A	Headline B	Headline C
Click	11.33	11.33	11.33
No-click	988.67	988.67	988.67

2. expected data(there is no difference on click number)

	Headline A	Headline B	Headline C
Click	0.792	−0.990	0.198
No-click	−0.085	0.106	−0.021

3. Pearson residuals: R measures the extent to which the actual counts differ from these expected counts.

	Headline A	Headline B	Headline C
Click	0.792	−0.990	0.198
No-click	−0.085	0.106	−0.021

4.The chi-square statistic $X = \sum_{i}^{r}\sum_{j}^{c} R^2$ where r and c are the number of rows and columns, respectively.

The chi-square statistic for this example is 1.666. Is that more than could reasonably occur in a chance model?

We can test with this resampling algorithm:

1. Constitute a box with 34 ones (clicks) and 2,966 zeros (no clicks).
2. Shuffle, take three separate samples of 1,000, and count the clicks in each.
3. Find the squared differences between the shuffled counts and the expected
4. counts and sum them.
5. Repeat steps 2 and 3, say, 1,000 times.
6. How often does the resampled sum of squared deviations exceed the observed?
7. That's the p-value.

VIII.1. Chi-Square Test: Statistical Theory

When we use a **chi-square test**, we calculate a **chi-square statistic** from the data.
Statistical theory shows that **when sample sizes are large**, this statistic follows (approximately) a **chi-square distribution**.
This allows us to:

- Compare our computed chi-square value
- To a **standard chi-square distribution**
- And get a **p-value** without resampling

The **shape** of the chi-square distribution depends on the **degrees of freedom**. df=(r−1)(c−1)
Shape of the chi-square distribution

- The **chi-square distribution is not symmetric**
- It is **right-skewed** (long tail to the right)
- Values are **always ≥ 0** (because it's based on squared differences)

Interpretation

- **Small chi-square** → differences easily explained by chance → large p-value
- **Large chi-square** → differences unlikely under the null → small p-value

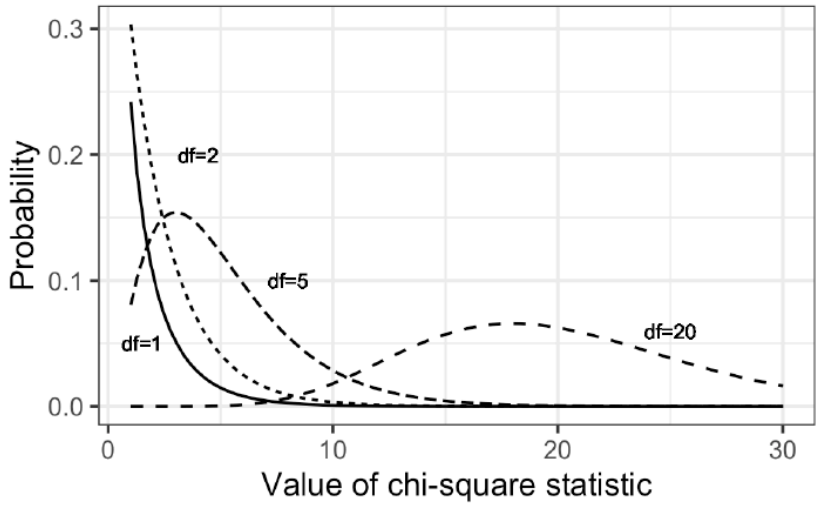


Figure 3-7. Chi-square distribution with various degrees of freedom

VIII.2. Fisher's Exact Test

	Success	Failure
Treatment A	2	3
Treatment B	1	4

Yes, exactly. **Fisher's exact test** is used when the counts in a contingency table are very small (typically ≤5). Unlike the chi-square test, it doesn't rely on approximations; it calculates the exact probability of observing the data (or more extreme) under the null hypothesis.

VIII.3. Relevance for Data Science

For **data scientists**, the goal is usually **finding the best option**, not just proving significance. For example, in an A/B/C test for webpage headlines, instead of just asking "Is one headline significantly better?", you want to **quickly identify the headline with the highest click rate**. Here, **multi-armed bandit algorithms** can adaptively allocate traffic to better-performing options, giving a more practical solution than relying solely on statistical significance.

One data science application of the chi-square test, especially Fisher's exact version, is in determining appropriate sample sizes for web experiments. These experiments often have very low click rates, and despite thousands of exposures, count rates might be too small to yield definitive conclusions in an experiment. In such cases, Fisher's exact test, the chi-square test, and other tests can be useful as a component of power and sample size calculations (see "Power and Sample Size" on page 135).

Chi-square tests in research are often used to **find statistically significant results** for publication.
In **data science**, they are more of a **filter** to decide if a feature or effect is worth further study.
Examples:

- **Spatial statistics:** Check if crimes are unusually concentrated in certain areas compared to random chance.

- **Machine learning:** Identify features where a class occurs much more or less than expected, helping select important features automatically.

Key Ideas

- A common procedure in statistics is to test whether observed data counts are consistent with an assumption of independence (e.g., propensity to buy a particular item is independent of gender).
- The chi-square distribution is the reference distribution (which embodies the assumption of independence) to which the observed calculated chi-square statistic must be compared.

X.Multi-Arm Bandit Algorithm

Multi-arm bandits offer an approach to testing, especially web testing, that allows explicit optimization and more rapid decision making than the traditional statistical approach to designing experiments.

Key Terms for Multi-Arm Bandits

Multi-arm bandit
An imaginary slot machine with multiple arms for the customer to choose from, each with different payoffs, here taken to be an analogy for a multitreatment experiment.

Arm
A treatment in an experiment (e.g., "headline A in a web test").

Win
The experimental analog of a win at the slot machine (e.g., "customer clicks on the link").

The traditional A/B testing approach has **limitations**:

1. Results may be **inconclusive** if the sample is too small ("effect not proven").
2. We cannot act on promising results **before the experiment ends**.
3. We cannot **adapt or change** the experiment based on new incoming data.

With modern computing, **flexible approaches** (like multi-arm bandits) allow ongoing learning and optimization.

- Example: Instead of waiting to finish testing web pages A and B, a system can **gradually show more users the better-performing page** while still testing alternatives.

Multi-armed bandits are algorithms for testing multiple options at once and quickly identifying the best one.

- Named after slot machines: a **single-armed bandit** is a classic slot machine.
- A **multi-armed bandit** has multiple "arms," each paying out at different rates. The goal is to **learn which arm gives the best reward** while still exploring the others.

Example: Testing four website headlines (A, B, C, D). Instead of showing each headline to the same number of visitors, the algorithm gradually shows **more users the better-performing headlines**, improving overall clicks while still learning.

In a **multi-armed bandit** scenario, the goal is to maximize your total reward by figuring out which arm is the best **as quickly as possible**.

- You **don't know the payout rates** of the arms in advance.
- You only see the result of each pull (win or loss).
- Each win has the **same value**, no matter the arm.

Example: You have 3 website headlines (A, B, C). Each visitor click is a "win." You don't know which headline gets the most clicks, so you try them all, but gradually **show more visitors the better-performing headlines** to maximize total clicks while still learning.

Imagine you're running a website test with three headlines: A, B, and C. After a short trial, it looks like **headline A is winning**.

One way is to **go all-in on A** immediately. If A really is the best, you win big—but if B or C is actually better, you'll never discover it.

The other extreme is to **treat them all equally**, giving each headline the same exposure. This explores all options fully, but you're also showing what seems to be inferior headlines to many users, wasting potential clicks.

Bandit algorithms take a smart middle path:

- Start favoring A slightly, but don't abandon B and C completely.
- If A keeps outperforming, you gradually show it more.
- If B or C starts catching up, you shift some traffic to them.

Example: Suppose:

- After 100 views: A gets 60 clicks, B 20, C 20.
- You show A to 60% of users, B and C to 20% each.
- If C suddenly gets more clicks in the next 100 views, you increase its share.

This way, you **exploit early winners** while still **exploring to find the true best option**, maximizing total clicks over time.

Imagine you're running a website with **four different banners**: Red, Blue, Green, and Yellow. Each banner is an "arm" of the slot machine, and each visitor either clicks (a win) or doesn't click.

At first, you **show all banners equally**—25% of visitors see each. After a few hundred visitors, you notice **Green is getting more clicks** than the others.

A **bandit algorithm** will now **adjust the display rates**: it might show Green to 40% of visitors, Red and Blue 20% each, and Yellow 20%. If Green keeps outperforming, it will gradually get more exposure. If Yellow suddenly starts getting more clicks, the algorithm can **shift traffic to Yellow** without abandoning the others entirely.

The **parameters of the algorithm** control:

- **How quickly you adjust** the display rates based on performance.
- **Minimum exposure** each variant gets, so you don't miss potential winners.

What epsilon-greedy algorithm does, in simpler terms with an example:

- **Goal:** Decide whether to show **offer A** or **offer B** to visitors in a way that balances **exploring** both options and **exploiting** the one that's performing best.

Step by step:

1. **Pick a random number** between 0 and 1.
2. **Exploration (probability = epsilon):**
    - If the random number is less than epsilon (say epsilon = 0.1 → 10% chance), **try both offers randomly**:
        - Flip a coin: heads → show A, tails → show B.
    - This ensures you **still occasionally test the weaker option** in case it improves.
3. **Exploitation (probability = 1 – epsilon):**
    - If the random number is greater than or equal to epsilon (90% of the time), **show the offer that currently has the best click rate**.
    - This maximizes the overall "wins" by favoring the better-performing offer.

Think of **epsilon** as the "curiosity dial" in the epsilon-greedy algorithm: it controls **how much you explore versus exploit**.

- **Epsilon = 1 → Full exploration:**
    - Every visitor is shown **A or B randomly**, just like a traditional A/B test.
    - You gather data on both options equally, but you don't take advantage of early winners.
    - **Example:** If you have 1,000 visitors, about 500 see A and 500 see B, regardless of which performs better.
- **Epsilon = 0 → Full exploitation:**

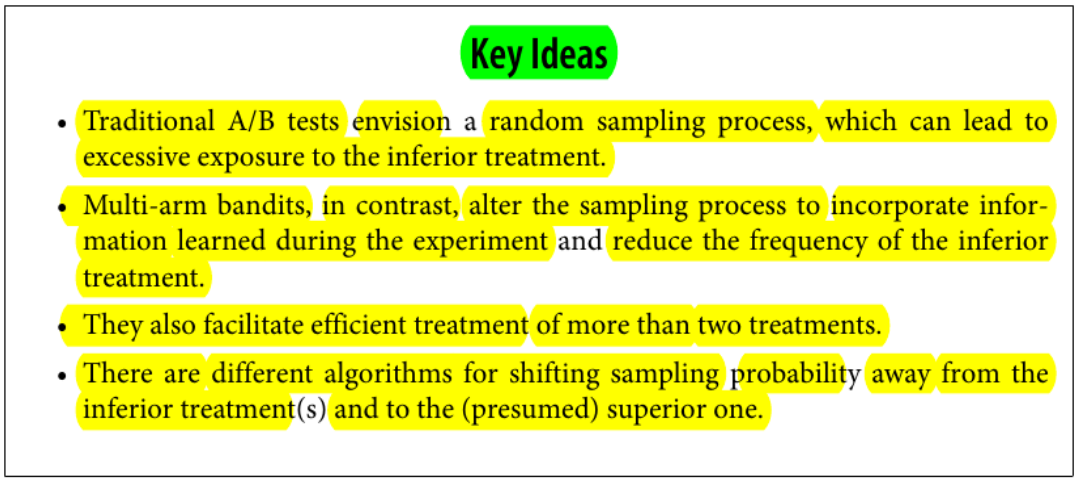
- The algorithm always picks the option that **currently looks best**.
- No more experimenting; you stick with the "leader" based on past results.
- **Example:** If A has a slightly higher click rate after the first 100 visitors, the next 900 will all see A.
- Risk: if B is actually better but got unlucky early, you **never discover it**.

A more sophisticated algorithm uses "Thompson's sampling."**Thompson's sampling** is like having a "guessing hat" for each headline:

1. **Start with a prior belief** about each headline's success (e.g., we think all have about a 50% chance to get clicks). This is modeled with a **beta distribution**.
2. **Show a headline to a visitor** and observe the result (click or no click).
3. **Update your beliefs** about each headline based on observed clicks. Headlines that perform better have their probability distributions "shifted" higher.
4. **Next visitor:** Randomly sample from each headline's updated distribution and pick the one with the **highest sampled value**.

Bandit algorithms shine when testing **three or more treatments** (e.g., headlines, offers, or colors) because they **adaptively shift traffic toward the best-performing option** as data comes in.

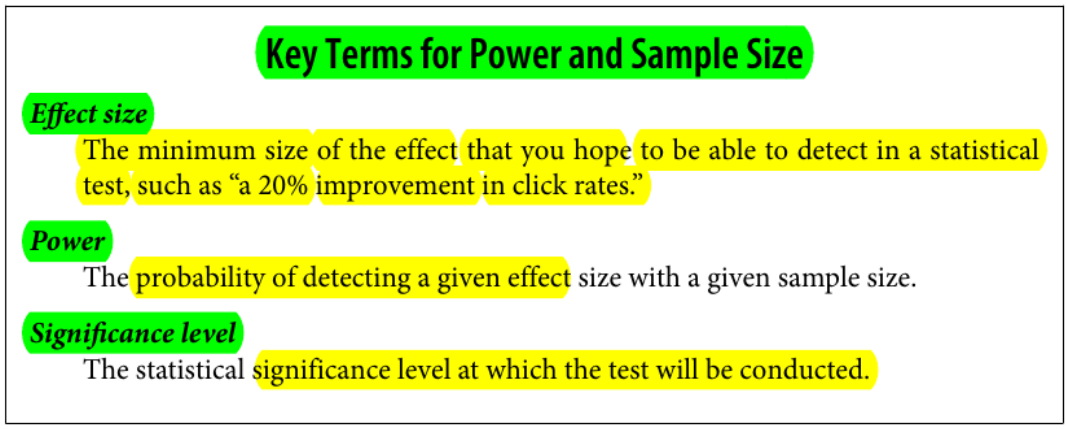
In contrast, **traditional A/B tests** get complicated with 3+ treatments: you must compare many pairs, adjust for multiple testing, and wait for a fixed sample size—making decision-making slower and less efficient.



XI.Power and Sample Size

Imagine you're running a web test to see which headline gets the most clicks. How long should you keep the test running? There's no fixed rule. The answer depends on **how often visitors actually click**.

- If clicks happen frequently, you'll gather enough data quickly to see which headline performs best.
- If clicks are rare, you'll need to show the headlines to many more visitors before differences become clear.



Think of it like comparing baseball hitters. If one batter hits .350 and another hits .200, the difference is obvious—you don't need many at-bats to see who's better. But if one hits .300 and another .280, the gap is smaller, so you need many more at-bats to confidently tell them apart.

In experiments, the same principle applies: the **larger the real difference between treatments**, the easier it is to detect. **Smaller differences require more data** to achieve a statistically reliable result. It's about making sure your sample size is enough to "see" the true effect, not just random variation.

Power is basically your experiment's "chance of success" at detecting a real effect.Imagine two baseball hitters: one hits .330, the other .200. If you give each 25 at-bats, there's a 75% chance (power = 0.75) that a statistical test will correctly show that the better hitter really is better. The **effect size** here is the difference in batting averages (0.130).

So, power tells you how likely your experiment is to detect a true difference, given your sample size and the size of the effect.

How to estimate power ?

Imagine you want to run an A/B test, but collecting data costs time or money. You don't want to waste effort only to get inconclusive results. One intuitive way to estimate **power** is to simulate the experiment before actually running it:

1. **Start with a hypothetical sample** reflecting what you expect. For example, say a baseline conversion rate is 20% → box with 20 ones (successes) and 80 zeros (failures). Or for time-on-site, a box with observed times.

2. **Add the effect you want to detect** to create a second sample. For instance, if you hope the new design increases conversion to 33%, make a box with 33 ones and 67 zeros, or add 25 seconds to each time value in the second sample.

3. **Draw a bootstrap sample** of size n from each box. This mimics randomly selecting users for your experiment.

4. **Run a hypothesis test** (permutation or formula-based) on these two samples and see if the difference is statistically significant.

5. **Repeat many times** (e.g., 1000 iterations). The fraction of times the test is significant is your **estimated power**.

Example: If out of 1000 simulated experiments, 750 show a significant difference, the power is 0.75, meaning a 75% chance your real experiment would detect the effect.

XII.Sample Size

Imagine you're testing a new online ad against the current one. You want to know **how many clicks you need** to decide if the new ad is better.

- If you only care about **huge differences**—say, the new ad gets 50% more clicks than the old one—you don't need a large sample; even a few dozen clicks might make the difference obvious.

- If you care about **small differences**—say, just a 10% improvement—you'll need a much **larger sample** to be confident that the difference is real and not just due to chance.

Here, the **effect size** (the minimum improvement you care about) drives the **sample size**. For example, if your policy says a new ad must outperform the old one by at least 10%, you simulate or calculate how many ad exposures are needed so that your experiment has a good chance (high power) of detecting that 10% difference.

In short: **bigger effect → smaller sample; smaller effect → bigger sample**.

For example, suppose current click-through rates are about 1.1%, and you are seeking a 10% boost to 1.21%. So we have two boxes: box A with 1.1% ones (say, 110 ones and 9,890 zeros), and box B with 1.21% ones (say, 121 ones and 9,879 zeros). For starters, let's try 300 draws from each box (this would be like 300 "impressions" for each ad). Suppose our first draw yields the following:

- Box A: 3 ones
- Box B: 5 ones

Right away we can see that any hypothesis test would reveal this difference (5 versus 3) to be well within the range of chance variation. This combination of sample size (n = 300 in each group) and effect size (10% difference) is too small for any hypothesis test to reliably show a difference.

So we can try increasing the sample size (let's try 2,000 impressions), and require a larger improvement (50% instead of 10%).

For example, suppose current click-through rates are still 1.1%, but we are now seeking a 50% boost to 1.65%. So we have two boxes: box A still with 1.1% ones (say, 110 ones and 9,890 zeros), and box B with 1.65% ones (say, 165 ones and 9,868 zeros). Now we'll try 2,000 draws from each box. Suppose our first draw yields the following:

- Box A: 19 ones
- Box B: 34 ones

A significance test on this difference (34–19) shows it still registers as "not significant" (though much closer to significance than the earlier difference of 5–3). To calculate power, we would need to repeat the

previous procedure many times, or use statistical software that can calculate power, but our initial draw suggests to us that even detecting a 50% improvement will require several thousand ad impressions.

In summary, for calculating power or required sample size, there are four moving parts:

- Sample size
- Effect size you want to detect
- Significance level (alpha) at which the test will be conducted
- Power

Specify any three of them, and the fourth can be calculated. Most commonly, you would want to calculate sample size, so you must specify the other three. With R and Python, you also have to specify the alternative hypothesis as "greater" or "larger" to get a one-sided test; see "One-Way Versus Two-Way Hypothesis Tests" on page 95 for more discussion of one-way versus two-way tests.

A **clean, complete Python example** using **statsmodels** that computes the required sample size to detect an increase from **1.1% to 1.21%** with **80% power** and **α = 0.05**. 9.sample_size.py

Key Ideas

- Finding out how big a sample size you need requires thinking ahead to the statistical test you plan to conduct.
- You must specify the minimum size of the effect that you want to detect.
- You must also specify the required probability of detecting that effect size (power).
- Finally, you must specify the significance level (alpha) at which the test will be conducted.

Summary

The principles of experimental design—randomization of subjects into two or more groups receiving different treatments—allow us to draw valid conclusions about how well the treatments work. It is best to include a control treatment of "making no change." The subject of formal statistical inference—hypothesis testing, p-values, t-tests, and much more along these lines—occupies much time and space in a traditional statistics course or text, and the formality is mostly unneeded from a data science perspective. However, it remains important to recognize the role that random variation can play in fooling the human brain. Intuitive resampling procedures (permutation and bootstrap) allow data scientists to gauge the extent to which chance variation can play a role in their data analysis.