Module 3

Statistical Experiments and Significance Testing

Syllabus:

Statistical Experiments and Significance Testing: A/B testing, hypothesis testing, resampling, statistical significance & p-values, t-tests, multiple testing, degrees of freedom.

Textbook: **Chapter 3**

Introduction

- The ***design of experiments is a cornerstone*** of the practice of statistics, and it has applications in virtually all research areas.
- The goal is to ***design an experiment*** to **confirm** or **reject** a hypothesis.
- Data scientists often need to conduct continual experiments, particularly regarding user interface and product marketing.

Statistical significance, t-tests, and p-values are part of the classical inference process. It begins with a **hypothesis** (e.g., "Drug A is better than Drug B" or "Price A outperforms Price B"). An **experiment**—often an A/B test—is carefully designed to test this hypothesis. After **data collection and analysis**, researchers draw a **conclusion**. The idea of **inference** is to generalize results from the sample data to a broader population or process.

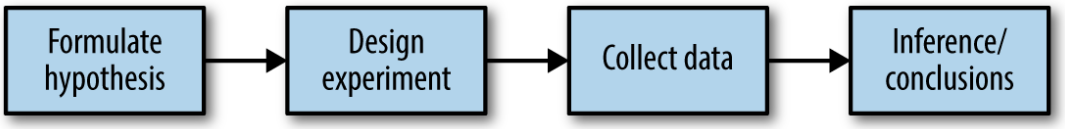


Figure 3-1. The classical statistical inference pipeline

3.1 A/B Testing

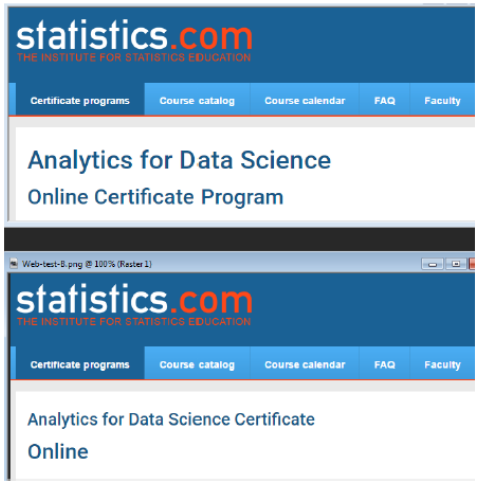
An **A/B test** is an experiment with two groups to determine which of two options performs better.

- One option is usually the **new treatment** (B).
- The other is the **control** (A) — often the standard practice or no treatment.
- The **hypothesis** typically states that the new treatment outperforms the control.

A/B tests are widely used in fields like web design and marketing because outcomes are easy to measure. Examples include:

- Comparing two **soil treatments** for seed germination
- Comparing two **therapies** for cancer suppression
- Comparing two **prices** for net profit

- Comparing two **web headlines** for clicks
- Comparing two **web ads** for conversions



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.

Figure 3-2. Marketers continually test one web presentation against another

A proper **A/B test** requires subjects (e.g., people, seeds, or web visitors) who can be assigned to one of two treatments. The critical step is **random assignment**, so that differences between groups come only from:

1. The **true effect** of the treatments, or
2. **Chance variation** from random assignment.

The choice of **test statistic or metric** is also essential. In data science, common metrics are binary outcomes (e.g., click/no-click, buy/don't buy, fraud/no fraud). These outcomes are often summarized in a **2×2 table**, which helps compare the performance of group A versus group B.

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

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

Why Have a Control Group?

Skipping a **control group** and comparing only to past experience is risky because other factors (seasonality, user behavior shifts, external events) may influence outcomes. A control group ensures that both groups face the same conditions, so any difference can be attributed to the treatment—or to chance.

In **data science**, A/B testing is often used in a **web context** (e.g., webpage design, product price, headline wording). The **subject** is usually the web visitor, and outcomes of interest include clicks, purchases, session duration, or page visits.

A **key principle** is to select **one main metric (test statistic) beforehand**—for example, conversion rate—since choosing metrics after running the experiment can introduce **researcher bias**.

Why Just A/B? Why Not C, D,⋯?

While **A/B tests** are very common in marketing and e-commerce, they are just one type of statistical experiment. Experiments can be more complex, with:

- **Additional treatments** (beyond just A and B),
- **Repeated measurements** on the same subjects, or
- **Early stopping rules** (common in clinical trials where subjects are costly or limited).

Traditional experiments focus on testing **whether one treatment is better than another** (e.g., "Is price A better than price B?").

By contrast, data science often asks: **"Which option, among many, performs best?"**

For this, a newer design is used: the **multi-arm bandit algorithm**, which balances experimentation (trying different options) with exploitation (favoring the best-performing option as data accumulates).

3.2 Hypothesis Tests

Hypothesis tests (or significance tests) are used to determine whether the results of an experiment could simply be due to **random chance**.

In an **A/B test**, we usually start with a hypothesis—for example, *"Price B yields higher profit than Price A."* Without a hypothesis, we risk being misled by random variation, since people tend to **see patterns where none exist** or underestimate the role of chance (e.g., "black swan" events).

In a well-designed A/B test, any observed difference between groups A and B must come from:

1. **Random chance** in subject assignment, or
2. A **true difference** in treatments.

A **statistical hypothesis test** then helps assess whether the observed difference is likely *due to chance*—or *whether it's strong evidence that one treatment really is better*.

The Null Hypothesis

Hypothesis tests guard against our natural tendency to mistake random fluctuations for meaningful patterns.

- We begin with a **baseline assumption**: the treatments are equivalent, and any observed difference is due to chance. This is the **null hypothesis (H_0)**.
- The goal is to see if the evidence is strong enough to **reject H_0**, showing that groups A and B differ more than chance alone could explain.

One method is a **resampling permutation test**:

1. Shuffle together the outcomes from both groups.
2. Randomly split them into new groups of the same sizes as A and B.
3. Measure the difference in outcomes.
4. Repeat many times to build a distribution of differences under the null model.
5. Compare the observed difference to this distribution.

If the observed difference is more extreme than what chance produces most of the time, we reject the null hypothesis.

Alternative Hypothesis

Every **hypothesis test** involves two parts:

- **Null hypothesis (H_0):** Assumes no effect, no difference, or a boundary condition (e.g., $A = B$, $A \leq B$).
- **Alternative hypothesis (H_1):** Represents what we hope to show (e.g., $A \neq B$, $A > B$, B is X% greater than A).

Together, **H_0 and H_1 cover all possibilities**. The exact form of the null hypothesis determines the type and structure of the test (e.g., one-tailed vs. two-tailed).

One-Way Versus Two-Way Hypothesis Tests

In A/B testing, we often compare a **new option (B)** against a **default option (A)**.

- If we only care whether **B is better than A**, we use a **one-tailed (directional) test**. Only extreme outcomes favoring B count toward the p-value.
- If we care about differences in **either direction**, we use a **two-tailed (bidirectional) test**, where extreme outcomes in both directions contribute to the p-value.

One-tailed tests align well with A/B decision-making, since the default is retained unless the new option proves superior. However, software often defaults to **two-tailed tests**, which are more conservative. For practical data science, the exact tail choice is often **less critical**, as p-value precision is not usually a major concern.

(A **p-value** is a probability that measures how likely it is to observe the data (or something more extreme) **if the null hypothesis (H_0) is true**.)

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.

3.3 Resampling

Resampling in statistics involves repeatedly drawing samples from observed data to **assess random variability** in a statistic. It's also used to **improve machine learning models**, for example, averaging predictions from multiple bootstrapped decision trees in **bagging**.

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

1. **Bootstrap:** Evaluates the reliability or variability of an estimate.
2. **Permutation tests:** Used for hypothesis testing, usually when comparing **two or more groups**.

Note: Resampling lets us understand how much results might **fluctuate by chance**, without relying strictly on theoretical distributions.

Permutation Test

Key Terms for Resampling

Permutation test
The procedure of combining two or more samples together and randomly (or exhaustively) reallocating the observations to resamples.

Synonyms
Randomization test, random permutation test, exact test

Resampling
Drawing additional samples ("resamples") from an observed data set.

With or without replacement
In sampling, whether or not an item is returned to the sample before the next draw.

Permutation Test

A **permutation test** evaluates whether differences between groups (e.g., in an A/B test) could be due to **chance**, under the null hypothesis that all groups are equivalent.

Steps:

1. **Combine groups:** Merge all observations from groups A, B (and C, D, … if any) into a single dataset.
2. **Shuffle & resample:** Randomly draw, without replacement, a resample of the same size as group A.

3. **Draw remaining groups:** From the leftover data, randomly draw resamples for groups B, C, D, etc., matching the original sizes.
4. **Compute statistic:** Calculate the test statistic (e.g., difference in means or proportions) for the resampled groups—this is one permutation iteration.
5. **Repeat:** Perform steps 2–4 **R times** to build a **permutation distribution** of the test statistic under the null hypothesis.
6. **Compare observed difference:**
   - If the observed difference falls **within** the permutation distribution → likely due to chance → **not significant**.
   - If the observed difference falls **outside** most of the permutation distribution → unlikely due to chance → **statistically significant**.

Note: The permutation test simulates what random chance could produce by repeatedly reshuffling the data, letting you see whether the actual observed difference is unusual.

Example: Suppose you run an **A/B test** to see if a new webpage design (**Group B**) increases click-through rate compared to the old design (**Group A**).

- **Group A (old design):** [3, 4, 5, 2, 6] (clicks per 100 users)
- **Group B (new design):** [7, 8, 6, 9, 10]

Observed difference in means:

- Mean(A) = (3+4+5+2+6)/5 = **4.0**
- Mean(B) = (7+8+6+9+10)/5 = **8.0**
- Difference = 8.0 – 4.0 = **4.0**

We want to test:

- **Null hypothesis (H_0):** The two groups are equivalent (differences are just due to chance).
- **Alternative (H_1):** Group B's mean is higher.

Steps in the Permutation Test:

1. **Combine groups**

Merge all values:

3,4,5,2,6,7,8,6,9,10

2. **Shuffle & resample**

Randomly shuffle these 10 numbers. Suppose one shuffle gives: 7,4,6,10,3,9,2,6,8,5

3. Draw groups

- Take first 5 numbers → pretend this is Group A′: [7, 4, 6, 10, 3]
- Take remaining 5 → pretend this is Group B′: [9, 2, 6, 8, 5]

4. Compute statistic

- Mean(A′) = (7+4+6+10+3)/5 = 6.0
- Mean(B′) = (9+2+6+8+5)/5 = 6.0
- Difference = 0.0

That's **one permutation iteration**.

6. **Repeat many times (say 10,000)**

Each time, shuffle and split, then compute the mean difference. This gives us a **distribution of differences under H_0** (the "permutation distribution").

6. Compare observed difference (4.0)

- If many shuffled differences reach values close to 4.0, then our observed difference could easily occur by chance → **not significant**.
- If most shuffled differences cluster around 0, and only a very small proportion are ≥ 4.0, then the observed difference is unlikely under H_0 → **statistically significant**.

Example: Web Stickiness

The company wants to know which of two website presentations (Page A vs. Page B) leads to more sales. But since sales are rare and take a long time to track, they use a **proxy variable** instead of actual sales. A proxy variable is a stand-in measure that's easier or quicker to collect but still related to the real outcome.

- **Proxy chosen:** Average session time on the service's detailed landing page.
- **Reason:** If visitors spend more time on the page, it's likely they're more interested, and this could lead to more sales later.
- **Data challenge:** Google Analytics logs the final page visit of a session as zero seconds, so those need to be cleaned out.
- **Final dataset:** 36 usable sessions (21 for Page A, 15 for Page B).
- **Analysis method:** Compare the two groups visually using **side-by-side boxplots** of session times (with ggplot), to see if one page tends to keep users engaged longer.

The boxplot, shown in Figure 3-3, indicates that page B leads to longer sessions than page A.

Note: Instead of waiting months for sales data, the company uses session time as a **quick proxy** to compare which page design is more effective.

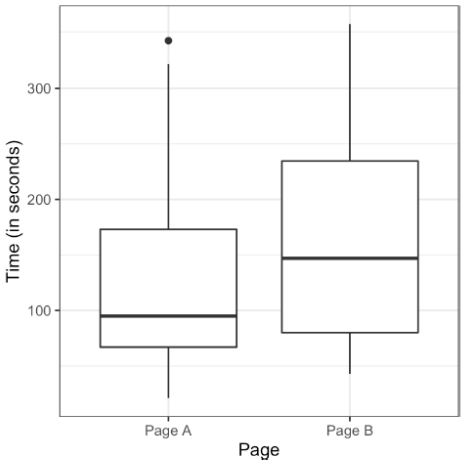


Figure 3-3. Session times for web pages A and B

The pandas boxplot command uses the keyword argument by to create the figure:

```
ax = session_times.boxplot(by='Page', column='Time')
ax.set_xlabel('')
ax.set_ylabel('Time (in seconds)')
plt.suptitle('')
```

In *Python*, we filter the pandas data frame first by page and then determine the mean of the Time column:

```
mean_a = session_times[session_times.Page == 'Page A'].Time.mean()
mean_b = session_times[session_times.Page == 'Page B'].Time.mean()
mean_b - mean_a
```

Page B's session times are, on average, **35.67 seconds longer than Page A's**. To check if this difference is **real** or could just be due to random variation, we use a **permutation test**.

- **Step 1:** Combine all 36 session times (from both pages).
- **Step 2:** Randomly shuffle them.
- **Step 3:** Split them into two groups — one of size 21 (Page A) and one of size 15 (Page B).
- **Step 4:** Compute the difference in means for these shuffled groups.
- **Step 5:** Repeat this many times to build a distribution of differences under the assumption that there's no real difference (null hypothesis).
- **Step 6:** See where the observed difference (35.67 sec) falls in this distribution to decide if it's statistically significant.

The *Python* version of this permutation test is the following:

```
def perm_fun(x, nA, nB):
        n = nA + nB
        idx_B = set(random.sample(range(n), nB))
        idx_A = set(range(n)) - idx_B
        return x.loc[idx_B].mean() - x.loc[idx_A].mean()
```

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

In *Python*, we can create a similar graph using matplotlib:

```
perm_diffs = [perm_fun(session_times.Time, nA, nB) for _ in range(1000)]
fig, ax = plt.subplots(figsize=(5, 5))
ax.hist(perm_diffs, bins=11, rwidth=0.9)
ax.axvline(x = mean_b - mean_a, color='black', lw=2)
ax.text(50, 190, 'Observed\ndifference', bbox={'facecolor':'white'})
ax.set_xlabel('Session time differences (in seconds)')
ax.set_ylabel('Frequency')
```

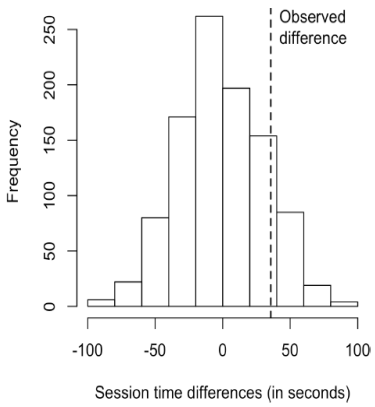


Figure 3-4. Frequency distribution for session time differences between pages A and B; the vertical line shows the observed difference

The histogram, in Figure 3-4 shows that mean difference of random permutations often exceeds the observed difference in session times (the vertical line). For our results, this happens in 12.1% of the cases:

```
np.mean(perm_diffs > mean_b - mean_a)
---
0.121 = 12.1%
```

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*

· A *bootstrap permutation test*

- **Exhaustive permutation test (exact test):**

Instead of randomly shuffling, it considers *all possible rearrangements* of the data into groups. This guarantees exact results but is only feasible for small datasets (too many combinations for large samples).

- **Bootstrap permutation test:**

Similar to the random permutation test, but the resampling is done *with replacement* (instead of without). This not only captures random assignment of treatments but also simulates the randomness of sampling subjects from a population.

In practice, random permutation (with many shuffles) gives results close to exhaustive, and the bootstrap adds an extra layer of randomness modeling.

Permutation Tests: The Bottom Line for Data Science

- They help us **see how much random variation** might explain observed differences.
- They are **simple to implement and interpret**, without heavy math or strict formulas.
- Unlike traditional formula-based methods (e.g., t-tests), they avoid giving a **false sense of certainty** from precise-looking equations.
- **Resampling methods** (like permutation tests) are flexible — they can handle different data types (numeric, binary), unequal sample sizes, and don't require assumptions like normality.

In short: permutation tests provide a practical, assumption-light, and broadly applicable way to test hypotheses.

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.

3.4 Statistical Significance and p-Values

- **Statistical significance** tells us whether an observed result is likely due to real effects rather than random chance.
- If the outcome is so unusual that random variation alone is very unlikely to produce it, we call it **statistically significant**.
- In practice, this is usually judged using a **p-value** (probability of observing such an extreme result under the null hypothesis). If the p-value is smaller than a chosen threshold (like 0.05), the result is considered significant.

Note: Statistical significance means the result is unlikely to be explained just by randomness.

p-Value

- Just looking at a graph isn't precise for deciding significance.
- Instead, we calculate the **p-value**, which is the probability of seeing a result as extreme (or more extreme) than the observed one if the null hypothesis (chance model) is true.
- In a permutation test, this is estimated by checking how often the simulated (permuted) differences are greater than or equal to the observed difference.

- In R or Python, True is treated as 1 and False as 0, so taking the mean of these comparisons gives the proportion of times the permuted difference exceeds the observed one.

```
mean(perm_diffs > obs_pct_diff)
[1] 0.308

np.mean([diff > obs_pct_diff for diff in perm_diffs])
```

- In this example, the p-value is **0.308**, meaning that such a result could occur by random chance about **30% of the time**.
- Since 0.308 is much higher than the usual threshold (like 0.05), the observed difference is **not statistically significant**.

In R (`prop.test`):

- `x` = number of successes in each group (`200` and `182`)
- `n` = number of trials (`23739` and `22588`)
- The test compares proportions:
  - Group 1 proportion = 200 / 23739 ≈ 0.00842
  - Group 2 proportion = 182 / 22588 ≈ 0.00806
- It tests if **Group 1 > Group 2** (one-sided test).
- The p-value = **0.3498**, meaning there's ~35% chance of seeing such a difference (or larger) if the true proportions are equal.

- In the earlier example, a **permutation test** was used to compute a p-value.
- But since the data come from a **binomial setting** (successes out of trials), we can also use a **statistical test based on the binomial/normal approximation**.
- A chi-square test gives a two-sided p-value; dividing by 2 gives a one-sided result.
- The p-value is again ≈ **0.35**, close to the permutation test result (0.308).

Note: Both the permutation test and the normal/chi-square approximation lead to the same conclusion: the difference in proportions is **not statistically significant**.

Example scenario:

Suppose we are classifying images from the **Caltech-101 dataset** (101 object categories like airplanes, cars, faces, etc.).

We train **two models**:

- **Model A**: CNN baseline
- **Model B**: CNN + data augmentation

After evaluation:

- Model A correctly classifies **200 out of 23,739 images**.

- Model B correctly classifies **182 out of 22,588 images**.

So we want to test: *Is Model A significantly better than Model B?*

Step 1: Permutation Test

- Combine all predictions into one pool.
- Randomly shuffle the labels to simulate "no real difference" between models.
- Recompute the accuracy difference many times.
- See how often the permuted difference is ≥ the observed difference.
- Suppose this gives a **p-value ≈ 0.308** → about 30% of the time such a gap appears by chance.
- ⇒ Not significant.

Step 2: Binomial / Normal Approximation (`prop.test` in R, `chi2_contingency` in Python)

- Treat the number of **correct classifications** as "successes" out of total trials (images).
- Compare proportions:
  - Model A accuracy = 200 / 23739 ≈ 0.84%
  - Model B accuracy = 182 / 22588 ≈ 0.81%
- Use **chi-square / z-test** to test if Model A > Model B.
- Result: **p-value ≈ 0.35** (very close to permutation test).
- ⇒ Again, no significant difference.

- Even though Model A's accuracy looks slightly higher, the difference is **tiny relative to dataset size**.
- The **p-value (~0.3–0.35)** means this small difference could easily happen by chance.
- Thus, we cannot confidently claim Model A is truly better.

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

Alpha

Alpha and Statistical Significance

- Statisticians dislike letting researchers decide after the fact what counts as "too unusual" — the threshold is set **before** the experiment.
- This threshold is called **alpha** (commonly 5% or 1%).
- Alpha is arbitrary; it does **not** guarantee correct decisions a certain percentage of the time.
- The key question answered is:
  "Given a chance model (null hypothesis), how likely is a result this extreme?"
  Not: "What's the probability this happened by chance?"

p-Value Controversy

- Many researchers **misinterpret p-values** as the probability that results are due to chance.
- In reality, a p-value tells us:

"If the null hypothesis is true, how likely are results as extreme as observed?"

- Misuse leads to "p-hacking" — searching data until a significant p-value is found, producing questionable research.

ASA Guidelines (2016)

The American Statistical Association clarified that:

1. P-values show **incompatibility** with a statistical model.

2. They **do not** measure the truth of a hypothesis.
3. Decisions **should not** rely solely on p-values.
4. Full **reporting and transparency** are essential.
5. P-values **do not measure effect size or importance**.
6. They alone are **not strong evidence** for a hypothesis.

Practical vs. Statistical Significance

- A statistically significant result isn't always **practically meaningful**.
- Large samples can make trivial differences statistically significant.
- Statistical significance ≠ importance.

Type 1 and Type 2 Errors

1. **Type 1 Error (False Positive):**
   - Concluding that an effect is real when it's actually due to chance.
   - Significance tests are designed primarily to **minimize this error**.
2. **Type 2 Error (False Negative):**
   - Concluding that an effect is not real when it actually exists.
   - Often occurs because the **sample size is too small** to detect the effect.
   - If a p-value is not significant (e.g., >5%), it means **"effect not proven,"** not necessarily **"effect absent."**

Significance tests mainly guard against claiming false effects (Type 1), but failing to detect real effects (Type 2) can happen if the study is underpowered.

Data Science and p-Values

For **data scientists**, the p-value is less about "proof" and more about **guidance**:

- It helps check whether a result or model feature is **likely due to chance**.
- It should **inform decisions**, not dictate them.
- In practice, p-values can be used as **intermediate inputs** in statistical or machine learning models—for example, to decide whether to include a feature in the model.

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).

3.5 t-Tests

There are many types of significance tests, chosen based on the type of data (count or measured), the number of samples, and the effect being studied. One of the most common is the **t-test**, based on *Student's t-distribution*, developed by W. S. Gosset to approximate the distribution of a single 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.

All significance tests require a test statistic to measure the effect of interest and determine whether the observed effect could be due to chance. In resampling tests, the null hypothesis distribution is created directly from the data, making the scale irrelevant. Historically, resampling was impractical, so statisticians used the *t-test*—developed from Gosset's t-distribution—as an approximation, especially for two-sample comparisons (A/B tests).

To apply the t-distribution regardless of data scale, the test statistic must be standardized, enabling comparisons to the standard t-distribution.

The function `scipy.stats.ttest_ind` can be used in *Python*:

```
res = stats.ttest_ind(session_times[session_times.Page == 'Page A'].Time,
                      session_times[session_times.Page == 'Page B'].Time,
                      equal_var=False)
print(f'p-value for single sided test: {res.pvalue / 2:.4f}')
```

The alternative hypothesis in testing compares means (e.g., session time of page A < page B). P-values from formula-based tests (e.g., 0.1408) are often close to those from resampling tests (e.g., 0.121, 0.126).

In resampling, the focus is on the observed data and hypothesis, without worrying about data type, sample balance, or variance. While statisticians study detailed formula-based methods, data scientists mainly use these tests pragmatically rather than for formal publication.

Key Ideas

- Before the advent of computers, resampling tests were not practical, and statisticians used standard reference distributions.
- A test statistic could then be standardized and compared to the reference distribution.
- One such widely used standardized statistic is the t-statistic.

3.6 Multiple Testing

As mentioned previously, there is a saying in statistics: "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.

Significance testing risks false positives when many tests are run, a problem called **alpha inflation** or **multiplicity**. For example, with 20 tests at $\alpha = 0.05$, there's a 64% chance of at least one false positive (Type I error). This is linked to **overfitting**, where models fit noise instead of true patterns.

Statistical adjustments (e.g., **Bonferroni correction**, **Tukey's HSD**) ***lower α*** thresholds to control false positives in structured settings, but they are limited in scope. In data science, where repeated testing and data dredging are common, the risks are greater.

Key takeaways for data scientists:

- Cross-validation and holdout set help avoid overfitting in predictive modeling.
- Without labelled holdouts, one must rely on awareness of multiplicity risks, and use **resampling or simulation** to gauge whether results could arise by chance.

Key Terms for Multiple Testing

Type 1 error
Mistakenly concluding that an effect is statistically significant.

False discovery rate
Across multiple tests, the rate of making a Type 1 error.

Alpha inflation
The multiple testing phenomenon, in which *alpha*, the probability of making a Type 1 error, increases as you conduct more tests.

Adjustment of p-values
Accounting for doing multiple tests on the same data.

Overfitting
Fitting the noise.

In today's world, we have huge amounts of data and many studies being published, which increases the chances of finding spurious results. Examples of multiple comparisons include:

- Comparing many groups at once (pairwise differences).
- Checking effects in many subgroups (like a specific age or gender group).
- Trying many different statistical models.
- Including lots of variables in your analysis.
- Asking multiple questions about the data.

For data scientists, traditional statistical adjustments are **too rigid** for general use, so here's the practical advice:

- **Predictive modeling:** Avoid "illusory models" (models that only seem good due to random chance) by using **cross-validation** and **holdout samples**.
- **Other analyses (without holdout data):**
  - Be aware that the more you explore or manipulate data, the more chance affects results.
  - Use **resampling or simulation methods** to create random benchmarks, helping you see if your results are truly meaningful or just luck.

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.

3.7 Degrees of Freedom

Degrees of freedom (d.f.) measure the number of values in a calculation that are free to vary.

- **Example:** In a sample of 10 values, if you know the mean, only 9 values can vary freely—the 10th is fixed. So, the sample has **9 degrees of freedom**.
- **Why it matters:** In calculations like variance or standard deviation, using **n – 1** instead of n corrects a small downward bias in the estimate.

In traditional statistics:

- Degrees of freedom are used in formulas for t-tests, F-tests, and other hypothesis tests to make sure your statistic matches the right reference distribution.

In data science:

- Usually **not critical**, because:
  1. Formal hypothesis tests are used less often.
  2. Large datasets make the difference between n and n – 1 negligible.
- **Important exception:** When using categorical variables in regression (like "day of the week"), degrees of freedom matter.
  - If you encode 7 days with 7 binary variables, you get **multicollinearity**, because knowing 6 variables automatically determines the 7th.
  - Solution: Only include 6 indicators (d.f. = 6), leaving one as a reference.

Key takeaway: Degrees of freedom are about **how many values can vary freely**, and they mostly matter in small samples or when creating dummy variables for regression.