

Week 4

Statistical Experiments and Significance Testing

Source: Bruce et al. (2026), Chapter 3 — Statistical Experiments and Significance Testing

Assoc. Prof. Dr. Onur Polat

Hacettepe University • Graduate School of Informatics
M.Sc. (Thesis) in Systems Engineering • Fall 2026, Lecture 4 / 14

OVERVIEW

This Week's Agenda

Chapter 3: Statistical Experiments

- A/B testing and control group design
- Hypothesis tests: null and alternative
- Resampling: permutation test
- Statistical significance and p-value
- t-Tests
- Multiple testing problem
- ANOVA and the F-statistic
- Chi-square test

Advanced Topics

- Degrees of freedom
- Two-way ANOVA
- Fisher's exact test
- Multi-arm bandit algorithm
- Power and sample size
- Role of p-value in data science
- Python implementations
- Epsilon-greedy and Thompson sampling

CLASSICAL STATISTICAL INFERENCE

Experimental Design Workflow



1. Formulate Hypothesis:

"Drug A is better than the current standard" or "Price B is more profitable"

2. Design Experiment:

Set up an A/B test or experiment to test the hypothesis

3. Collect Data:

Run the experiment and record the results

4. Inference:

Generalize results from a limited sample to the broader population

01

A/B Testing

Comparing two treatments to determine which is superior

A/B TESTING: KEY TERMS

A/B Testing Core Concepts

Treatment

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

Treatment Group

A group of subjects exposed to a specific treatment

Control Group

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

Randomization

The process of randomly assigning subjects to treatments

Subjects

The units (web visitors, patients, etc.) exposed to treatments

Test Statistic

A metric used to measure the effect of the treatment

A/B TEST EXAMPLES

Broad applications from web design to agriculture

- Test which of two soil treatments leads to better seed germination
- Determine which of two drugs more effectively suppresses cancer
- Test which of two prices yields more net profit
- Measure which of two web headlines gets more clicks
- Analyze which of two web ads generates more conversions

Table 3-1. E-commerce experiment results (2x2 table)

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

WHY A CONTROL GROUP?

Without a control group:

- No guarantee of "all other things being equal"
- Cannot tell if difference is due to treatment or chance

With a control group:

- Same conditions applied (except treatment)
- Any observed difference can come from only two sources:
 1. A real difference between treatments
 2. Chance in random assignment

Blinding: A study in which subjects do not know which treatment they are receiving. **Double blind:** Researchers also don't know. Blinding is not possible for treatments like cognitive therapy.

WHY JUST A/B? WHY NOT C, D?

A/B tests are popular, but not the only type of statistical experiment!

Data scientists are more interested in the question:

"Among multiple possible prices, which one is best?"

rather than the traditional question:

"Is the difference between A and B statistically significant?"

Solution: Multi-Arm Bandit Algorithm

- Tests multiple treatments simultaneously
- Steers toward the best performer during the experiment
- *Details later in this chapter*

02

Hypothesis Tests

Questioning whether an observed effect is due to chance

HYPOTHESIS TEST TERMINOLOGY

Core Concepts

Null Hypothesis (H_0)

The hypothesis that chance is responsible — "there is no difference"

Alternative Hypothesis (H_1)

The counterpart to the null — what you seek to prove

One-Way Test

A hypothesis test that counts chance effects in only one direction

Two-Way Test

A hypothesis test that counts chance effects in both directions

Why do we need hypothesis testing?

Humans tend to ascribe meaning to random behavior. Statistical hypothesis tests were developed to protect researchers from being deceived by random chance.

NULL AND ALTERNATIVE HYPOTHESIS

H_0 — Null Hypothesis

"There is no difference between treatments; any observed difference is due to chance."

Core assumption: Treatments are equivalent and any difference stems from chance.

Our goal is to disprove this hypothesis!

H_1 — Alternative Hypothesis

Examples:

- H_0 : "No difference between A and B"
 H_1 : "A is different from B" (two-way)
- H_0 : " $A \leq B$ "
 H_1 : " $A > B$ " (one-way)
- H_0 : "B is not X% better than A"
 H_1 : "B is X% better than A"

ONE-TAIL vs TWO-TAIL TEST

One-Tail Test

- New option (B) is tested against the existing default (A)
- Stick with A unless B is proven definitively better
- Directional alternative hypothesis: $H_1: B > A$
- Only extreme results in one direction count toward the p-value

Two-Tail Test

- Protection against being fooled by chance in either direction
- Two-sided alternative: $H_1: A \neq B$
- Extreme results in both directions count toward the p-value

Note: Python and R default to two-tail tests. In data science, the one-tail vs two-tail distinction is not critical; p-value precision at that level may not matter.

03

Resampling

Exploring the role of random variation through the permutation test

RESAMPLING TERMINOLOGY

Core Concepts

Permutation Test

Combining two or more samples and randomly reallocating them into new subsamples

Synonyms

Randomization test, random permutation test, exact test

Resampling

Drawing additional samples from an observed data set

With / Without Replacement

Whether an item is returned to the pool after being drawn

Permutation Test Procedure:

Combine groups → Shuffle and re-split → Compute statistic → Repeat R times → Compare

PERMUTATION TEST STEPS

- 1. Combine the results from the different groups into a single data set
- 2. Shuffle the combined data and randomly draw (without replacement) a sample of size n_A
- 3. Set aside the remaining data as Group B
- 4. If there are groups C, D, etc., repeat accordingly
- 5. Compute the statistic for the resampled groups as for the original samples — this is one permutation iteration
- 6. Repeat steps R times to obtain the permutation distribution of the test statistic

Conclusion: If the observed difference exceeds most of the permuted differences → chance is not responsible → *statistically significant*

EXAMPLE: WEB STICKINESS

Page A vs Page B — Session time comparison

Scenario: A company selling high-value services wants to test which of two web presentations is better. Since sales are rare and expensive, session time is used as a proxy variable.

n_A = 21

Page A
page views

n_B = 15

Page B
page views

**35.67
sec**

B – A mean diff
(B is longer)

12.9%

p-value
(NOT significant)

Interpretation: After 1000 iterations of the permutation test, 12.9% of random differences exceeded the observed difference. This difference is within the range of chance variation → not statistically significant.

PYTHON: PERMUTATION TEST

Web stickiness example code

```
import numpy as np
import pandas as pd

session_times = pd.read_csv("web_page_data.csv")
session_times["Time"] *= 100

mean_a = session_times[session_times.Page == "Page A"].Time.mean()
mean_b = session_times[session_times.Page == "Page B"].Time.mean()

def perm_fun(x, n_a):
    x_resampled = np.random.permutation(x)
    x_a = x_resampled[:n_a]
    x_b = x_resampled[n_a:]
    return x_b.mean() - x_a.mean()

n_a = session_times[session_times.Page == "Page A"].shape[0]
perm_diffs = [perm_fun(session_times.Time, n_a)
               for _ in range(1000)]

p_value = np.mean(perm_diffs > mean_b - mean_a)
print(f"p-value: {p_value:4f}")
```

TYPES OF PERMUTATION TESTS

Random Permutation

Randomly shuffle data and repeat R times. The most common approach.

Exhaustive Permutation

Enumerate all possible splits. Practical for small samples. Also known as the exact test.

Bootstrap Permutation

Sampling with replacement. Models randomness in both treatment assignment and population selection.

For Data Science: Permutation tests are easy to code, easy to interpret, and provide a practical alternative to formula-based methods. Data can be numerical or binary; no normal distribution assumption is needed.

04

Statistical Significance and the p-Value

Measuring whether a result could be due to chance

p-VALUE AND ALPHA

Core Concepts

p-value

The probability of obtaining a result as extreme as or more extreme than the observed, given the null hypothesis is true

Alpha (α)

The threshold of statistical significance. Typical values: 5% and 1%

Type 1 Error

Concluding a real effect when there is none (false positive)

Type 2 Error

Failing to detect a real effect (false negative)

Caution: The p-value is NOT "the probability the result is due to chance." Correct: "Given a chance model, the probability of results as extreme as the observed" — a subtle but important distinction.

EXAMPLE: PRICE TEST

E-commerce conversion rate comparison

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

0.8425

%

Price A
conversion rate

0.8057

%

Price B
conversion rate

0.0368

Percentage point diff
(in favor of A)

$p \approx 0.31$

p-value
(NOT significant)

Interpretation: Despite over 45,000 data points, the actual conversion counts are only in the hundreds. With $p \approx 0.31$, the difference is within chance variation → not statistically significant.

TYPE 1 AND TYPE 2 ERRORS

	H_0 Actually True	H_0 Actually False
H_0 Not Rejected	Correct decision (True negative)	TYPE 2 ERROR (False negative)
H_0 Rejected	TYPE 1 ERROR (False positive)	Correct decision (True positive)

The primary goal of hypothesis testing is to protect against Type 1 errors. Type 2 error is not truly an error — it indicates the sample size was insufficient to detect the effect.

THE p-VALUE IN DATA SCIENCE

The p-Value Controversy

- Some psychology journals have banned p-value-based publication decisions
- In 2016, the ASA published 6 principles: p-value alone is not sufficient evidence
- Statistical significance $\neq$ practical significance

A Pragmatic View for Data Scientists

- The p-value is a useful metric for determining whether a model result is within the range of normal chance variation
- Should be used as a tool that informs the decision, not one that controls it
- In ML: can be used as an intermediate input to decide whether to include a variable in a model

t-TESTS

Core Concepts

Test Statistic

A metric for the difference or effect of interest

t-statistic

A standardized form of common test statistics such as differences in means

t-distribution

A reference distribution derived from the null hypothesis; the observed t-statistic is compared against it

```
from scipy import stats

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 (one-tail): {res.pvalue / 2:.4f}")
# Output: p-value (one-tail): 0.1408
```

THE MULTIPLE TESTING PROBLEM

Alpha inflation and false discovery rate

Core Concepts

Type 1 Error

Mistakenly concluding that an effect is statistically significant

False Discovery Rate (FDR)

The rate of committing Type 1 errors across multiple tests

Alpha Inflation

The increase in Type 1 error probability as the number of tests increases

p-value Adjustment

Accounting for performing multiple tests on the same data

Overfitting

Fitting noise rather than signal

Example: 20 random variables + 1 outcome $\rightarrow$ at $\alpha=0.05$, probability of at least one false significant result: $1 - 0.95^{20} = 64\%$

05

ANOVA

Analyzing differences among multiple groups

ANOVA: KEY CONCEPTS

Analysis of Variance Terminology

Pairwise Comparison

Hypothesis testing between pairs among multiple groups

Omnibus Test

A single test of the overall variance among multiple group means

Variance Decomposition

Separating the components contributing to an individual value

F-statistic

A standardized statistic measuring whether differences among group means exceed what is expected under a chance model

SS (Sum of Squares)

Sum of squared deviations from a mean value

ANOVA = the generalization of the A/B test to multiple groups. A single omnibus test replaces pairwise comparisons.

EXAMPLE: FOUR WEB PAGES

A/B/C/D test — stickiness (seconds)

	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
Mean	172	185	176	162

Grand Mean: 173.75

4 groups → 6 pairwise comparisons. **Multiple comparison problem!** ANOVA tests all groups simultaneously with a single

```
import statsmodels.api as sm
import statsmodels.formula.api as smf

model = smf.ols("Time ~ Page", data=four_sessions).fit()
```

F-STATISTIC AND VARIANCE DECOMPOSITION

Decomposition of Variance

Each observation = Grand Mean + Treatment Effect + Residual Error

Example (top-left cell = 164):

1. Grand mean: 173.75
2. Treatment effect: -1.75 ($172 - 173.75$)
3. Residual: -8 ($164 - 172$)
4. Total: 164 ✓

	df	SS	MS	F	p
Page	3	831.4	277.1	2.74	0.0776
Residual	16	1618.4	101.2		

$F = MS(\text{Treatment}) / MS(\text{Error}) = 277.1 / 101.2 = 2.74$

TWO-WAY ANOVA

One-Way ANOVA: Only one factor varies (e.g., page design)

Two-Way ANOVA: Two factors vary simultaneously (e.g., page design + weekend/weekday)

Two-way ANOVA tests 3 effects:

1. Main effect — Factor A (page design)
2. Main effect — Factor B (weekend/weekday)
3. Interaction effect ($A \times B$)

ANOVA is the first step toward regression and logistic regression — multiple factors and their effects can be modeled.

06

Chi-Square Test

Measuring the significance of deviations from expected counts

CHI-SQUARE: HEADLINE TEST EXAMPLE

Click-through rate comparison across three headlines

	Headline A	Headline B	Headline C
Click	14	8	12
No Click	986	992	988

Expected (H_0 : equal click rates): $34/3 = 11.33$ clicks / headline

Pearson Residual (R) = (Observed – Expected) / $\sqrt{\text{Expected}}$

Chi-square statistic = $\sum R^2 = 1.666$

Degrees of freedom = $(r-1) \times (c-1) = 1 \times 2 = 2$

p-value $\approx 0.4348 \rightarrow$ **NOT significant — difference may be due to chance**

```
from scipy import stats
survivors = np.array([[14, 8, 12], [986, 992, 988]])
res = stats.chi2_contingency(survivors)
print(f"Chi-square: {res.statistic:.4f}, p: {res.pvalue:.4f}")
```

FISHER'S EXACT TEST

When to use it?

- When counts are very small (especially single digits)
- When the chi-square approximation is inadequate
- Enumerates all possible permutations to calculate an exact p-value

How does it work?

- Computes all possible arrangements the data could take
- Tabulates the probability of each arrangement to yield the exact p-value
- Headline example: Fisher $p \approx 0.4824$ (resampling $p \approx 0.4853$)

For Data Science: The chi-square test is useful in sample size calculations and web experiments. In automated feature selection, it can test whether class prevalence deviates from random variation.

07

Multi-Arm Bandit Algorithm

Optimal balance between exploration and exploitation

MULTI-ARM BANDIT ALGORITHM

A modern alternative to traditional A/B testing

Traditional A/B Test

- Fixed sample allocation
- Result: "effect not proven"
- Excessive exposure to underperforming treatment
- Rigid and inflexible process
- Complex for 2+ treatments

Multi-Arm Bandit

- Dynamic sample allocation
- Steers toward the best during the experiment
- Less exposure to underperformers
- Flexible and adaptive
- Naturally supports 3+ treatments

Slot machine analogy: Each arm pays out at a different rate. The goal: maximize total reward — find the most rewarding arm as early as possible.

EPSILON-GREEDY ALGORITHM

A simple epsilon-greedy algorithm for A/B testing:

1. Generate a uniform random number between $[0, 1]$
2. If the number is in $[0, \epsilon]$ $\rightarrow$ explore mode, flip a coin:
 - a. Heads $\rightarrow$ show Offer A, b. Tails $\rightarrow$ show Offer B
3. If the number $\geq \epsilon \rightarrow$ exploit mode, show the offer with the highest response rate so far

$$\epsilon = 1$$

Fully random
(standard A/B)

$$\epsilon = 0$$

Fully greedy
(no exploration)

$$0 < \epsilon < 1$$

Hybrid approach
(optimal)

Advanced: Thompson sampling (Bayesian approach) — updates priors using the beta distribution, maximizing the probability of selecting the best arm at each draw.

POWER AND SAMPLE SIZE

Core Concepts

Effect Size

The minimum effect you hope to detect (e.g., 20% increase in click-through rate)

Power

The probability of detecting a given effect size with a given sample size

Significance Level (α)

The level of statistical significance at which the test is conducted

4 interrelated components in power analysis:

n

Sample
Size

urth

d

Effect
Size

Signific

α

Significance
Level

$1-\beta$

Power

SAMPLE SIZE CALCULATION

Ad click-through rate example

Scenario: Current click rate is 1.1%, and we want the new ad to perform 10% better (1.21%).

- $n=300$, 10% diff → **cannot detect**
- $n=2000$, 50% diff → **still cannot detect (thousands of impressions needed)**

```
import statsmodels.api as sm

# Compute effect size
effect_size = sm.stats.proportion_effectsize(0.0121, 0.011)

# Compute required sample size
analysis = sm.stats.TTestIndPower()
result = analysis.solve_power(
    effect_size=effect_size,
    alpha=0.05, power=0.8, alternative="larger")
print(f"Required sample size: {result:.0f}")
# Output: 116,602 (~ 120K impressions per group)
```

DEGREES OF FREEDOM

A key parameter of statistical tests and distributions

What is degrees of freedom (d.f.)?

A concept applied to statistics computed from sample data, referring to the number of values that are free to vary.

Example: If you know the mean of 10 values, you can freely choose 9 of them — the 10th is automatically determined → d.f. = 9

Why does it matter?

- Reason for using $n-1$ instead of n when computing variance and std (bias correction)
- Determines the shapes of t-, F-, and chi-square distributions
- In regression: encoding categorical variables with $n-1$ indicators (dummies) to avoid multicollinearity

For data science: With large data sets, the difference between n and $n-1$ is negligible. It matters more in regression and indicator coding.

CHAPTER SUMMARY

Experimental Design

Splitting into control and treatment groups via randomization — the foundation of valid inference

Hypothesis Tests

H_0 (chance) vs H_1 (real effect) — the statistical inference framework

Resampling

Permutation tests: easy to code, easy to interpret, no formulas required

p-Value

A pragmatic tool — a metric that informs the decision, not one that controls it

ANOVA & Chi-Square

Multi-group and count data comparisons

Multi-Arm Bandit

Exploration/exploitation balance — a flexible alternative to traditional A/B testing