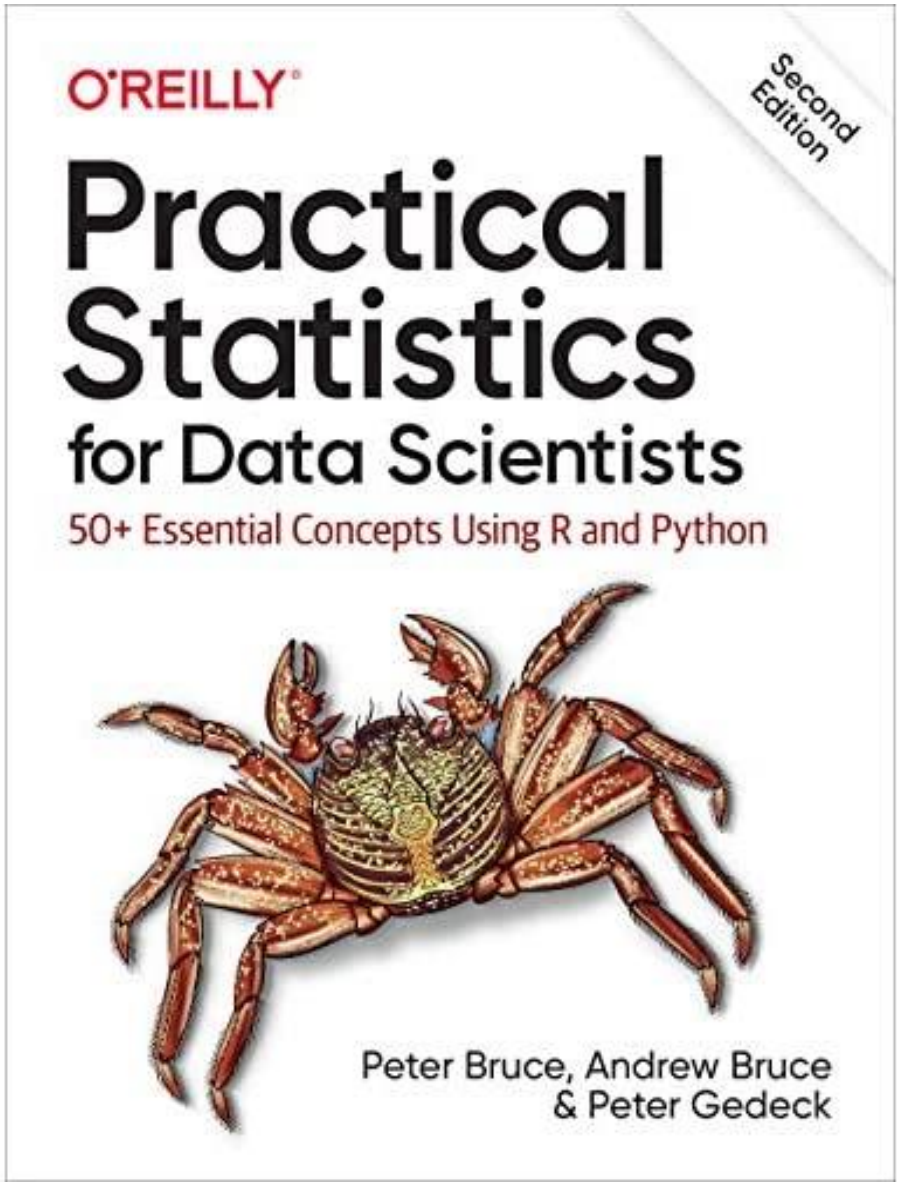
CHAPTER 3
Statistical Experiments and Significance Testing

Part 1

"Torture the data long enough, and it will confess."

Ronald Coase



Presentation Outline

- A/B Testing
- Hypothesis testing
- Resampling
- Statistical Significance
- P-value
- T- Tests
- Multiple Testing

Chapter 3 Goals

- Reviews traditional experimental design
- Discusses some common challenges in data science.
- Cover some oft-cited concepts in statistical inference and explains their meaning and relevance (or lack of relevance) to data science.

Types of statistical studies

Survey study

Aim :

Estimate the value of a parameter for a population.

Example :

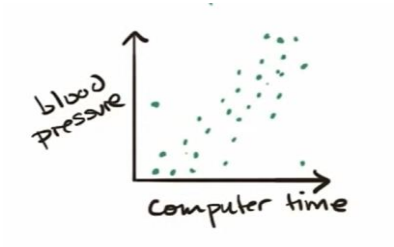
How much time does a population (randomly sampled) spend on a computer ?

→ Parameter: average daily time on computer

Observational study

Aim: **Attempts** to find a relationship between variables (correlation ?)

Example:
How does Average daily time spent on computer relate to people's blood pressure ?



Intrepretation: more computer time correlates with high BP
Correlation
Causality
→ counfounding variable
(lack of activity ..)

Experiments

Are the basis of the scientific method
Aim: attempts to **establish** a cause-and-effect (causality) relationship between variables.
(Does one variable affect the other ?)

In order to avoid Counfouonding variable :
Randomly assign **Subjects** to two groups :
→ **Treatement** group
→ **Control** group

Measure (**Metric**) BP before & after being in front of computer

Quizz: For each of the following study descriptions, identify whether the study is a survey, an observational study, or an experiment, and give a reason for your answer.

A study determines whether taking aspirin regularly helps to prevent heart attacks. A large group of male physicians of comparable health were randomly assigned equally to taking an aspirin every second day or to taking a placebo. After several years, the proportion of the study participants who had suffered heart attacks in each group was compared.

Quizz: For each of the following study descriptions, identify whether the study is a survey, an observational study, or an experiment, and give a reason for your answer.

A study determines whether taking aspirin regularly helps to prevent heart attacks. A large group of male physicians of comparable health were randomly assigned equally to taking an aspirin every second day or to taking a placebo. After several years, the proportion of the study participants who had suffered heart attacks in each group was compared.

Experiment

→The male physicians were randomly assigned to one of two treatments.
→The treatment variable: is aspirin or no aspirin administered.
→ The response variable is whether or not the subject suffered a heart attack.
(Note that this is a well-known experiment called the Physicians' Health Study)

Quizz: For each of the following study descriptions, identify whether the study is a survey, an observational study, or an experiment, and give a reason for your answer.

A study investigated whether boys are quicker at learning video games than girls. Twenty randomly selected boys and twenty randomly selected girls played a video game that they had never played before. The time it took them to reach a certain level of expertise was recorded.

Quizz: For each of the following study descriptions, identify whether the study is a survey, an observational study, or an experiment, and give a reason for your answer.

A study investigated whether boys are quicker at learning video games than girls. Twenty randomly selected boys and twenty randomly selected girls played a video game that they had never played before. The time it took them to reach a certain level of expertise was recorded.

Observational study.

→ The children were observed, and no treatment was administered to them.

→The population of interest: boys and girls who would play the video game that they had never played before.

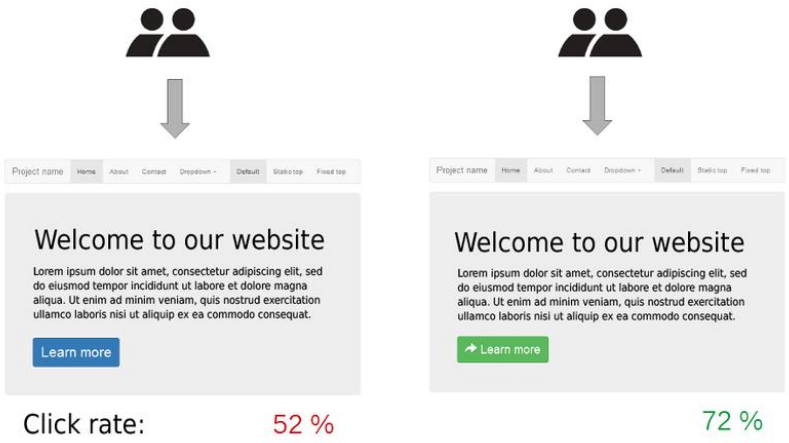
→The study was to see who is quicker at achieving a certain level of expertise.

A/B Testing

A/B testing is a way to compare **two versions** of a single variable, typically by testing a subject's response to variant A against variant B, and determining which of the two variants is more effective.

A/B tests are common in web design and marketing.

Example:



Metric: number of clicks
control group: Version A (current version)
Treatment group: Version B (modified version)
Subjects: real-time visitors (taken **randomly**)

A/B testing refers to the experiments where **two or more variations** of the same webpage are compared against each other by displaying them to real-time **visitors** to determine which one performs better for a given goal.

❑ An A/B test is typically constructed with a hypothesis in mind.

Hypothesis Testing

Hypothesis testing is a formal procedure for investigating our ideas about the world using statistics. It is most often used by scientists to test specific predictions, called hypotheses, that arise from theories.

Their purpose is to help you learn whether random chance might be responsible for an observed effect.

our experiments we will require proof that the difference between groups is more extreme than what chance might reasonably produce.

Why do we need a hypothesis?

- the tendency of the human mind to underestimate the scope of natural random behavior.
- the tendency to misinterpret random events as having patterns of some significance.

Hypothesis Testing

The Null Hypothesis & Alternative Hypothesis

There are 5 main steps in hypothesis testing:

1. State your research hypothesis as a null hypothesis (H_o) and alternate hypothesis (Ha or H1).
2. Collect data in a way designed to test the hypothesis.
3. Perform an appropriate statistical test.
4. Decide whether to reject or fail to reject your null hypothesis.
5. Present the findings in your results and discussion section.

Example 1:

You want to test whether there is a relationship between gender and height. Based on your knowledge of human physiology, you formulate a hypothesis that men are, on average, taller than women. To test this hypothesis, you restate it as:

H_o: Men are, on average, not taller than women.
H_a: Men are, on average, taller than women.

Step 1: State your null and alternate hypothesis

- The **alternate hypothesis** is usually your initial hypothesis that predicts a relationship between variables.
- The **null hypothesis** is a prediction of **no** relationship between the variables you are interested in.

Step 2: Collect data

For a statistical test to be valid, it is important to perform sampling and collect data in a way that is designed to test your hypothesis. If your data are not representative, then you cannot make statistical inferences about the population you are interested in.

Hypothesis Testing

Step 3: Perform a statistical test

There are a variety of statistical tests available, but they are all based on the comparison of **within-group variance** (how spread out the data is within a category) versus **between-group variance** (how different the categories are from one another).
If the between-group variance is large enough that there is little or no overlap between groups, then your statistical test will reflect that by showing a low *p*-value. This means it is unlikely that the differences between these groups came about by chance.
Alternatively, if there is high within-group variance and low between-group variance, then your statistical test will reflect that with a high *p*-value. This means it is likely that any difference you measure between groups is due to chance.
Your choice of statistical test will be based on the type of data you collected.

Hypothesis Testing

***Step 4*:** Decide whether to reject or fail to reject your null hypothesis

Based on the outcome of your statistical test, you will have to decide whether to reject or fail to reject your null hypothesis.

In most cases you will use the *p*-value generated by your statistical test to guide your decision. And in most cases, your predetermined level of significance for rejecting the null hypothesis will be 0.05 – that is, when there is a less than 5% chance that you would see these results if the null hypothesis were true.

In some cases, researchers choose a more conservative level of significance, such as 0.01 (1%). This minimizes the risk of incorrectly rejecting the null hypothesis (Type I error).

Resampling

Resampling in statistics means to repeatedly sample values from observed data, with a general goal of assessing random variability in a statistic.

Resampling Method	Application	Sampling procedure used
Bootstrap	Standard Deviation, Confidence Interval, Hypothesis testing, bias	Samples drawn at random, **with** replacement
Permutation	Hypothesis testing;	Samples drawn at random, **without** replacement

Resampling

The permutation procedure is as follows:

1. Combine the results from the different groups into a single data set.
2. Shuffle the combined data and then randomly draw (**without replacement**) a resample of the same size as group A (clearly it will contain some data from the other groups).
3. From the remaining data, randomly draw (without replacement) a resample of the same size as group B.
4. Do the same for groups C, D, and so on. You have now collected one set of resamples that mirror the sizes of the original samples.
5. Whatever statistic or estimate was calculated for the original samples (e.g., difference in group proportions), calculate it now for the resamples, and record; this constitutes one permutation iteration.
6. Repeat the previous steps *R* times to yield a permutation distribution of the test statistic.

Statistical significance

A result is statistically significant when it has a very low chance of occurring if there were no true effect in a research study.

The ***p*-value**, or probability value, tells you how likely it is that your data could have occurred under the null hypothesis.

- A **null hypothesis** (H_0) always predicts no true effect, no relationship between variables, or no difference between groups.
- An **alternative hypothesis** (H_a or H_1) states your main prediction of a true effect, a relationship between variables, or a difference between groups.

The smaller the *p*-value, the more likely you are to reject the null hypothesis.

Statistical significance

What is a significance level?

The **significance level**, or alpha (α), is a value that the researcher sets in advance as the threshold for statistical significance. It is the maximum risk of making a false positive conclusion (Type I error) that you are willing to accept.
In a hypothesis test, the *p* value is compared to the significance level to decide whether to reject the null hypothesis.

- If the *p* value is **higher** than the significance level, the null hypothesis is not refuted, and the results are **not statistically significant**.

- If the *p* value is **lower** than the significance level, the results are interpreted as refuting the null hypothesis and reported as **statistically significant**.

Usually, the significance level is set to 0.05 or 5%. That means your results must have a 5% or lower chance of occurring under the null hypothesis to be considered statistically significant.
The significance level can be lowered for a more conservative test. That means an effect has to be larger to be considered statistically significant.

T-Tests

A t-test is a statistical test that is used to compare the **means** of two groups.
To determine whether a process or treatment actually **has an effect** on the population of interest, or whether **two groups are different from one another.**

T-test formula

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{{S_1}^2}{N_1} + \frac{{S_2}^2}{N_2}}}$$

In this formula, t is the t-value, x_1 and x_2 are the means of the two groups being compared, s_2 is the pooled standard error of the two groups, and n_1 and n_2 are the number of observations in each of the groups.

A larger t-value shows that the difference between group means is greater than the pooled standard error, indicating a more significant difference between the groups.

Type I & Type II Errors

- **Type I error (false positive):** the test result says you have coronavirus, but you actually don't.

A Type I error means rejecting the null hypothesis when it's actually true. It means concluding that results are **statistically significant** when, in reality, they came about purely by chance or because of unrelated factors.

- **Type II error (false negative):** the test result says you don't have coronavirus, but you actually do.

a Type II error means failing to conclude there was an effect when there actually was. In reality, your study may not have had enough statistical power to detect an effect of a certain size.

Resources

- https://www.youtube.com/watch?v=SaP1O0i1bdc (KhanAcademy)
- https://www.invespcro.com/blog/ab-testing-statistics-made-simple/
- https://www.scribbr.com/category/statistics/
- https://www.youtube.com/watch?v=cn4S3QqEBRg (hypothesis testing)
- https://www.youtube.com/watch?v=0zZYBALbZgg
- https://www.scribbr.com/statistics/t-test/