# Experimental Design

## Johns Hopkins University

July 3, 2024

**Contents**

# 1 What does experimental design mean?

- Experimental design is organising an experiment so that we have plentiful and correct data to enable us to clearly and effectively answer the data science question.

- The process involves clearly formulating the question before data collection, followed by designing the best possible setup for gathering data, identifying possible sources of error in the design, and only then collecting the appropriate data.

# 2 Why does the study design matter?

- It is important to plan your analysis in advance to avoid incorrect conclusions. An incorrect analysis can lead to incorrect conclusions.

- There is a website called Retraction Watch which identifies research papers that have been retracted due to poor scientific practices.

- Poor practices often result from poor experimental design and analysis.

- Erroneous conclusions can have significant effects, especially in human health.

- Example: A retracted paper attempted to predict chemotherapy response based on genome analysis; it was cited nearly 450 times before retraction. Erroneous data then influenced clinical trials for cancer treatment plans.

- Proper experimental design is crucial when stakes are high.

# 3 Principles of experimental design

There are many concepts and terms inherent to experiment design:

- **Independent variable**: the variable that the experimenter manipulates. The independent variable does not depend on other variables being measured; it is often displayed on the x-axis.

- **Dependent variable**: the variable that is expected to change as a result of changes in the independent variable. The dependent variable is often displayed on the y-axis, so that changes in the independent variable effect changes in the dependent variable.

- **Hypothesis**: an educated guess as to the relationship between the variables and the outcome of the experiment.

## 3.1 Relationship between variables and hypotheses

Consider the following example experiment:

- Suppose that we hypothesise a positive linear relationship between shoe size and literacy, i.e., we hypothesise that literacy level *depends* on shoe size.

- In this experiment, we will take a measure of literacy such as reading fluency as the dependent variable and shoe size as the independent variable.

- We design an experiment in which we measure the shoe size and literacy level of 100 individuals.

- **Sample size** is the number of experimental subjects that are included in the experiment. There are many methods to determine the optimal sample size.

Before we collect any data, we need to consider if there are problems with this experiment that might cause an erroneous result. In our case, the experiment may be fatally flawed by a **confounder**:

- A **confounder** is an extraneous variable that may affect the relationship between the dependent and independent variables.

- In our example, age also affects foot size, and literacy is affected by age. Thus, observation of a relationship between the two variables may in fact be due to the presence of a confounding variable - age - which "confounds" our experimental design.

We can control for this situation by taking the following actions:

- Measure the age of each individual to account for the effects of age on literacy levels;

- Fix the age of all participants, effectively removing the effect of age on literacy.

In other experimental design paradigms, a **control group** may be appropriate:

- A control group is a group of experimental subjects that are *not* manipulated.

- Example: If we studied the effect of a drug on survival, one group would receive the drug (treatment group) and one group that did not (control group).

- This allows us to compare the effects of the drug in the treatment versus control group.

- The control group still has their dependent variables measured.

In these study designs, there are other strategies we can use to control for confounding effects. For one, we could **blind** the subjects to their assigned treatment group by giving them a **placebo**:

- A **placebo** is used in randomised clinical trials to test the efficacy of treatments and is most often used in drug studies. It is a substance or treatment which is designed to have no therapeutic value.

- The treatment group receives the actual drug, while the control group receives the placebo.

- Participants in both groups do not know if they have received the actual drug or the placebo. This way, researchers can measure if the drug works by comparing how both groups react.

- If both groups have the same reaction - improvement or not - the drug is deemed not to work.

The above strategy essentially spreads any possible confounding effects equally across the groups being compared. If we assume that age is a possible confounding effect, ensuring that both groups have similar ages and age ranges will help to mitigate any effect that age may cause on the dependent variable, i.e., the effect of age is equal between the two groups.

The balancing of confounders is often achieved by **randomisation**. Since we do not know what will be a confounder beforehand, we would not want to accidentally risk biasing one group to be enriched for a confounder. To avoid this, we can randomly assign individuals to each of our groups. This means that any potential confounding variables should be distributed between each group roughly equally, to help eliminate/reduce systematic errors.

The type of study described above is called a single-blinded or double-blinded randomised control trial (RCT); single-blind is when the participants of both groups are blinded to the treatment, whereas double-blind is where neither the researchers nor the patients know the treatments being given.

## 3.2 Replication

**Replication** is one of the key ways in which scientists build confidence in the scientific merit of results. Replication involves repeating the experiment with different experimental subjects. A single experiment's results may have occurred by chance due to factors such as:

- A confounder was unevenly distributed across both groups;

- There was a systematic error in the data collection;

- There were some outliers in the data

When the result from one study is found to be consistent by another study, it is more likely to represent a reliable claim to new knowledge. Replication allows you to measure the **variability** of the data more accurately, which enables us to better assess whether any differences you see in your data are significant.

Replication studies are a great way to bolster your experimental results and calculate the variability in the data. However, a successful replication does not guarantee that the original scientific results of a study were correct, nor does a single failed replication conclusively refute the original claims. A failure to replicate previous results can be due to any number of factors, including:

- the discovery of an unknown effect;

- inherent variability in the system;

- inability to control complex variables;

- substandard research practices;

- chance/probability.

## 4 Sharing data

Once we have collected and analysed our data, one of the next steps is to share the data and code for analysis. GitHub is a great place to share that code, using version controlled data and analyses.

## 5 Beware $p$-hacking

The **$p$-value** is a commonly reported statistical metric that tells us the probability that the results of the experiment were observed by chance. Something to consider is when you manipulate $p$-values towards your own end. When $p < 0.05$, there is less than 5% probability that the differences we observed were by chance, implying that the result is **significant**. However, if we do 20 tests, by chance, we would expect one of the 20 (5%) to be significant.

In the age of big data, testing 20 hypotheses is a very easy proposition. This is where the term "**$p$-hacking**" originates from. $p$-hacking is when we exhaustively search a dataset to find patterns and correlations that *appear* statistically

significant by virtue of the sheer number of tests you have performed. **These spurious correlations can be reported as significant and if you perform enough tests, you can find a dataset and analysis that will show you what you wanted to see**.

## 6 Summary

We have covered the following topics:

- What experimental design is
- Why good experimental design matters
- Principles of experimental design
- Definitions of common terms in experimental design
- Importance of sharing data and code for analysis
- Dangers of $p$-hacking and manipulating data to achieve significance.

## 7 Quiz

1. In a study measuring the effect of diet on BMI, cholesterol, lipid levels, triglyceride levels, and glycemic index, which is an independent variable:
   - Diet
   - BMI
   - Lipid levels

2. Which of the following is <u>not</u> a method to control your experiments?
   - Control group
   - Placebo effect
   - Blinding

3. What might a confounder be in an experiment looking at the relationship between the prevalence of white hair in a population and wrinkles?
   - Socioeconomic status
   - Smoking status
   - Age

4. According to Leek group recommendations, what data do you need to share with a collaborating statistician?
   - The raw data
   - A tidy data set
   - A code book describing each variable and its values in the tidy data set.
   - An explicit and exact recipe of how you went from the raw data to the tidy data and the code book.
   - All of the above.

5. If you set your significance level at $p \leq 0.01$, how many significant tests would you expect to see by chance if you carry out 1000 tests?
   - 10
   - 50
   - 100

6. What is an experimental design tool that can be used to address variables that may be confounders at the design phase of an experiment?
   - Stratifying variables
   - Data cleaning
   - Using all the data you have access to

7. Which of the following describes a descriptive analysis?
   - Use your sample data distribution to make predictions for the future

- Draw conclusions from your sample data distribution and infer for the larger population

- Generate a table summarising the number of observations in your dataset as well as the central tendencies and variances of each variable.