Experimental Designs and Statistical Methods

Léo Belzile

Table of contents

Welcome

This book is a web complement to MATH 80667A *Experimental Designs and Statistical Methods*, a graduate course offered at HEC Montréal in the joint Ph.D. program in Management.

The objective of the course is to teach basic principles of experimental designs and statistical inference using the **R** programming language. We will pay particular attention to the correct reporting and interpretation of results and learn how to review critically scientific papers using experimental designs. The unifying theme of the book is that statistics are summary of evidence, and statistic as a field is the science of decision-making in the presence of uncertainty. We use examples drawn from published articles in management sciences to illustrate core concepts.

1 Introduction

The advancement of science is built on our ability to study and assess research hypotheses. This chapter covers the basic concepts of experiments, starting with vocabulary associated with the field. Emphasis is placed on the difference between experiments and observations.

This course covers experimental designs. In an experiment, the researcher manipulates one or more features (say the complexity of a text the person must read, or the type of advertisement campaign displayed, etc.) to study their impact. In general, however (Cox 1958)

> effects under investigation tend to be masked by fluctuations outside the experimenter's control.

The purpose of experiments is to arrange data collection so as to be capable of disentangling the differences due to treatment from those due to the (often large) intrinsic variation of the measurements. We typically expect differences between treatments (and thus the effect) to be *comparatively stable* relative to the measurement variation.

! Learning objectives:

- Learning the terminology associated to experiments.
- Assessing the generalizability of a study based on the consideration of the sample characteristics, sampling scheme and population.
- Distinguishing between observational and experimental studies.
- Understanding the rationale behind the requirements for good experimental studies.

1.1 Study type

There are two main categories of studies: observational and experimental. The main difference between the two is treatment assignment. In observational studies, a feature or potential cause is measured, but not assigned by the experimenter. By contrast, the

treatment assignment mechanism is fully determined by the experimenter in the latter case.

For example, an economist studying the impact of interest rates on the price of housing can only look at historical records of sales. Similarly, surveys studying the labour market are also observational: people cannot influence the type of job performed by employees or their social benefits to see what could have happened. Observational studies can lead to detection of association, but only an experiment in which the researcher controls the allocation mechanism through randomization can lead to *directly* establish existence of a causal relationship. Because everything else is the same in a well controlled experiment, any treatment effect should be in principle caused by the experimental manipulation.[1]

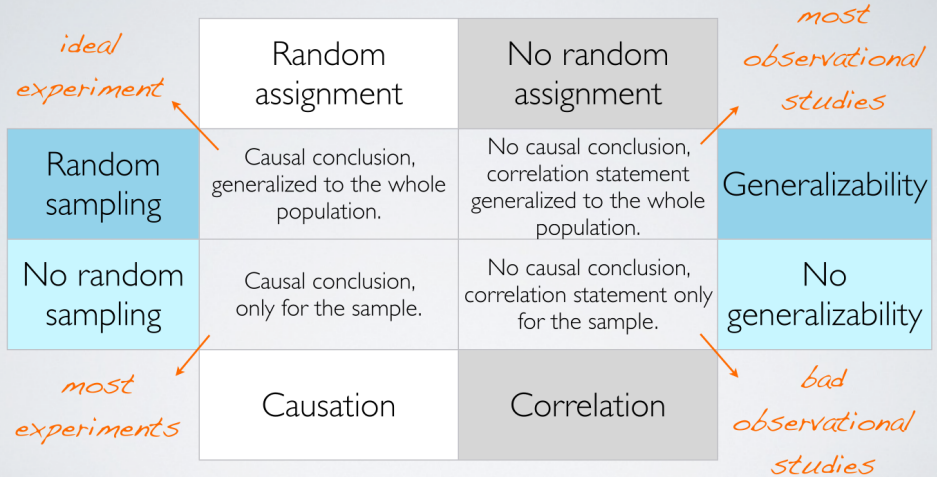


Figure 1.1: Two by two classification matrix for experiments based on sampling and study type. Source: Mine Çetinkaya-Rundel and OpenIntro, distributed under the CC BY-SA license.

Figure 1.1 summarizes the two preceding sections. Random allocation of observational units to treatment and random samples from the population lead to ideal studies, but may be impossible due to ethical considerations.

[1]The preceding paragraph shouldn't be taken to mean that one cannot get meaningful conclusions from observational studies. Rather, I wish to highlight that controlling for the non-random allocation and potential confounding is a much more complicated task, requires practitioners to make stronger (and sometimes unverifiable) assumptions and requires using a different toolbox (including, but not limited to differences in differences, propensity score weighting, instrumental variables). The book *The Effect: An Introduction to Research Design and Causality* by Nick Huntington-Klein gives a gentle nontechnical introduction to some of these methods.

1.2 Terminology

In its simplest form, an **experimental design** is a comparison of two or more treatments (experimental conditions):

- The subjects (or **experimental units**) in the different groups of treatment have similar characteristics and are treated exactly the same way in the experimentation except for the treatment they are receiving. Formally, an experimental unit is the smallest division such that any two units may receive different treatments.
- The **observational unit** is the smallest level (time point, individual) at which measurement are recorded.
- Explanatories (independent variables) are variables that impact the response. They can be continuous (dose) or categorical variables; in the latter case, they are termed **factors**.
- The **experimental treatments** or conditions are *manipulated and controlled* by the researcher. Oftentimes, there is a **control** or baseline treatment relative to which we measure improvement (e.g., a placebo for drugs).
- After the different treatments have been administered to subjects participating in a study, the researcher measures one or more outcomes (also called responses or dependent variables) on each subject.
- Observed differences in the outcome variable between the experimental conditions (treatments) are called treatment effects.

Example 1.1 (Pedagogical experience)**.** Suppose we want to study the effectiveness of different pedagogical approaches to learning. Evidence-based pedagogical researchs point out that active learning leads to higher retention of information. To corroborate this research hypothesis, we can design an experiment in which different sections of a course are assigned to different teaching methods. In this example, each student in a class group receives the same teaching assignment, so the experimental units are the sections and the observational units are the individual students. The treatment is the teaching method (traditional teaching versus flipped classroom).

Your turn

The marketing department of a company wants to know the value of its brand by determining how much more customers are willing to pay for their product relative to the cheaper generic product offered by the store. Economic theory suggests a substitution effect: while customers may prefer the brand product, they will switch to the generic version if the price tag is too high. To check this theory, one could design an experiment.

As a researcher, how would you conduct this study? Identify a specific product. For the latter, define

- an adequate response variable
- the experimental and observational units
- potential blocking factors

The main reason experiments should be preferred to collection of observational data is that they allow us, if they are conducted properly, to draw **causal conclusions** about the phenomenon of interest. If we take a random sample from the population of interest, split it randomly and manipulate only certain aspects, then all differences between groups must be due to those changes.

As Hariton and Locascio (2018) put it:

> Randomised controlled trials (RCTs) are the reference standard for studying causal relationships between interventions and outcomes as randomisation eliminates much of the bias inherent with other study designs

Quasi experiments

Sometimes, it is impossible or unethical to conduct an experiment. This seemingly precludes study many social phenomena, such as the effect on women and infantile mortality of strict bans on abortions. When changes in legislation occur (such as the Supreme court overturning Roe and Wade), this offers a window to compare neighbouring states.
Canadian economist David Card was co-awarded the 2021 Nobel Memorial Prize in Economic Sciences for his work in experimental economics. One of his most cited paper is Card and Krueger (1994), a study that looked at the impact of an increase in minimum wage on employment figures. Card and Krueger (1994) used a planned increase of the minimum wage of $0.80 USD in New Jersey to make comparisons with neighbouring Eastern Pennsylvania counties by studying 410 fast food outlets. The authors found no evidence of a negative impact on employment of this hike.

Point of terminology: internal and external validity

A study from which we can study causal relationships is said to have **internal validity**. By design, good experiments should have this desirable property because the random allocation of treatment guarantees, if randomization is well performed, that the effect of interest is causal. There are many other aspects, not covered in the class, that can

threaten internal validity.
External validity refers directly to generalizability of the conclusions of a study: Figure 1.1 shows that external validity is directly related to random sampling from the population

! Point of terminology: between-subjects and within-subjects designs

In between-subjects designs, subjects are randomly assigned to only one of the different experimental conditions. On the contrary, participants receive many or all of the experimental treatments in a within-subjects design, the order of assignment of the conditions typically being random.
While within-subject designs allow for a better use of available ressources (it is cheaper to have fewer participants perform multiple tasks), observations from within-design are correlated and more subject to missingness and learning effects, all of which require special statistical treatment.

1.3 Review of basic concepts

1.3.1 Variables

The choice of statistical model and test depends on the underlying type of the data collected. There are many choices: quantitative (discrete or continuous) if the variables are numeric, or qualitative (binary, nominal, ordinal) if they can be described using an adjective; I prefer the term categorical, which is more evocative. The choice of graphical representation for data is contingent on variable type. Specifically,

- a **variable** represents a characteristic of the population, for example the sex of an individual, the price of an item, etc.
- an **observation** is a set of measures (variables) collected under identical conditions for an individual or at a given time.

Most of the models we will deal with are so-called regression models, in which the mean of a quantitative variable is a function of other variables, termed explanatories. There are two types of numerical variables

- a discrete variable takes a countable number of values, prime examples being binary variables or count variables.

Figure 1.2: Illustration of continuous (left) and discrete variables (right). Artwork by Allison Horst shared under the CC BY 4.0 license.

- a continuous variable can take (in theory) an infinite possible number of values, even when measurements are rounded or measured with a limited precision (time, width, mass). In many case, we could also consider discrete variables as continuous if they take enough values (e.g., money).

Categorical variables take only a finite of values. They are regrouped in two groups, nominal if there is no ordering between levels (sex, colour, country of origin) or ordinal if they are ordered (Likert scale, salary scale) and this ordering should be reflected in graphs or tables. We will bundle every categorical variable using arbitrary encoding for the levels: for modelling, these variables taking K possible values (or levels) must be transformed into a set of $K-1$ binary variables $T_1, \ldots, T_K$, each of which corresponds to the logical group k (yes = 1, no = 0), the omitted level corresponding to a baseline when all of the $K-1$ indicators are zero. Failing to declare categorical variables in your software is a common mistake, especially when these are saved in the database using integers (1,2, . . .) rather than as text (Monday, Tuesday, . . .).

We can characterize the set of all potential values their measurements can take, together

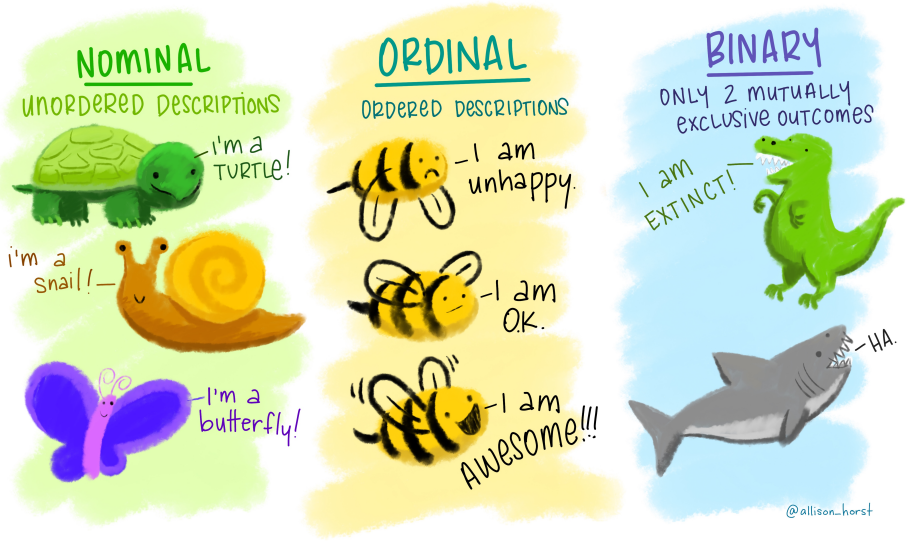


Figure 1.3: Examples of categorical nominal (left), ordinal (middle) and binary (right) variables. Artwork by Allison Horst shared under the CC BY 4.0 license.

with their frequency, via a **distribution**. The latter can be represented graphically using an histogram or a density plot[2] if the data are continuous, or a bar plot for discrete or categorical measurements.

The distribution of outcomes of a die toss is discrete and takes values $1, \ldots, 6$. Each outcome is equally likely with probability $1/6$.

Mathematical theory suggests that, under general conditions, the distribution of a sample average is approximately distributed according to a normal (aka Gaussian) distribution: this result is central to most of statistics. Normally distributed data are continuous; the distribution is characterized by a bell curve, light tails and it is symmetric around it's mean. The shape of the facade of Hallgrímskirkja church in Reykjavik, shown in Figure 1.4, closely resembles the density a normal distribution, which lead Khoa Vu to call it 'a normal church' (chuckles).

The normal distribution is fully characterized by two parameters: the average μ and the standard deviation σ. The left panel of Figure 1.5 shows an arbitrary continuous distribution and the values of a random sample of $n = 1000$ draws. The right panel shows the histogram of the sample mean value based on a very large number of random samples of size $n = 25$,

[2]Since continuous data can take any value in the interval, we can't talk about the probability of a specific value. Rather, the density curve encodes the probability for any given area: the higher the curve, the more likely the outcome.

Figure 1.4: Photography of the Hallgrímskirkja church in Reykjavik, Iceland by Dolf van der Haven, reproduced under a CC BY-ND-NC 2.0 license.

drawn from the same distribution. The superimposed black curve is a normal density curve whose parameters match those given by the central limit theorem: the approximation is seemingly quite accurate.

This fact explains the omnipresence of the normal distribution in introductory data science courses, as well as the prevalence of sample mean and sample variance as key summary statistics.[3]

Thinking outside the box

One key aspect, often neglected in studies, is the discussion of the metric used for measurement of the response. While previous research may have identified instruments (like questionnaires) and particular wording for studying a particular aspect of individuals, there is a lot of free room for researchers to choose from that may impact conclusions. For example, if one uses a Likert scale, what should be the range of the scale? Too coarse a choice may lead to limited capability to detect, but more truthfulness, while there may be larger intrinsic measurement with a finer scale.

[3] The parameters of most commonly used theoretical distributions do not directly relate to the mean and the variance, unlike the normal distribution.

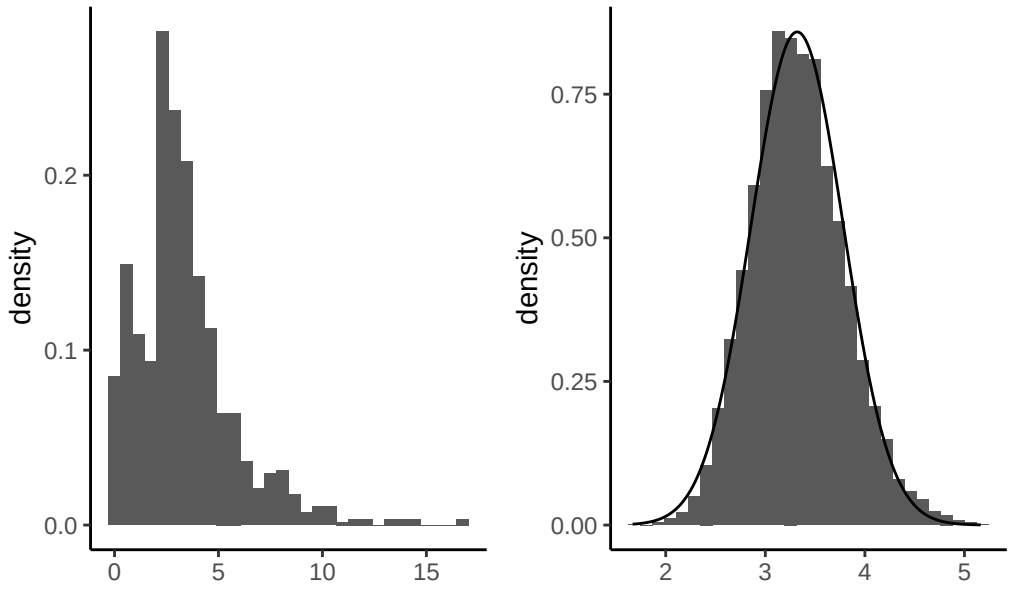


Figure 1.5: Graphical representation of the distribution of a continuous variable, with histogram of a sample of $n = 1000$ observations drawn from the distribution (left) and distribution of the sample mean, obtained by repeatedly drawing random sample of $n = 25$ observations and computing their average (right). The curve shows the normal distribution approximation based on the central limit theorem.

> Likewise, many continuous measures (say fMRI signal) can be discretized to provide a single numerical value. Choosing the average signal, range, etc. as outcome variable may lead to different conclusions.
> Choosing a particular instrument or metric could be in principle done by studying (apriori) the distribution of the values for the chosen metric using a pilot study: this will give researchers some grasp of the variability of those measures.

At the heart of most analysis are measurements. The data presented in the course have been cleaned and oftentimes the choice of explanatory variables and experimental factor[4] is evident from the context. In applications, however, this choice is not always trivial.

[4]A factor is a categorical variable, thus the experimental factor encodes the different groups to compare

1.3.2 Population and samples

Only for well-designed sampling schemes does results generalize beyond the group observed. It is thus of paramount importance to define the objective and the population of interest should we want to make conclusions.

Generally, we will seek to estimate characteristics of a population using only a sample (a sub-group of the population of smaller size). The **population of interest** is a collection of individuals which the study targets. For example, the Labour Force Survey (LFS) is a monthly study conducted by Statistics Canada, who define the target population as "all members of the selected household who are 15 years old and older, whether they work or not." Asking every Canadian meeting this definition would be costly and the process would be long: the characteristic of interest (employment) is also a snapshot in time and can vary when the person leaves a job, enters the job market or become unemployed. In this example, collecting a census would be impossible and too costly.

In general, we therefore consider only **samples** to gather the information we seek to obtain. The purpose of **statistical inference** is to draw conclusions about the population, but using only a share of the latter and accounting for sources of variability. The pollster George Gallup made this great analogy between sample and population:

> One spoonful can reflect the taste of the whole pot, if the soup is well-stirred

A **sample** is a sub-group of individuals drawn at random from the population. We won't focus on data collection, but keep in mind the following information: for a sample to be good, it must be representative of the population under study.

> **Your turn**
>
> The Parcours AGIR at HEC Montréal is a pilot project for Bachelor in Administration students that was initiated to study the impact of flipped classroom and active learning on performance.
> Do you think we can draw conclusions about the efficacy of this teaching method by comparing the results of the students with those of the rest of the bachelor program? List potential issues with this approach addressing the internal and external validity, generalizability, effect of lurking variables, etc.

Because the individuals are selected at **random** to be part of the sample, the measurement of the characteristic of interest will also be random and change from one sample to the next. While larger samples typically carry more information, sample size is not a guarantee of quality, as the following example demonstrates.

Example 1.2 (Polling for the 1936 USA Presidential Election)**.** *The Literary Digest* surveyed 10 millions people by mail to know voting preferences for the 1936 USA Presidential Election. A sizeable share, 2.4 millions answered, giving Alf Landon (57%) over incumbent President Franklin D. Roosevelt (43%). The latter nevertheless won in a landslide election with 62% of votes cast, a 19% forecast error. Biased sampling and differential non-response are mostly responsible for the error: the sampling frame was built using "phone number directories, drivers' registrations, club memberships, etc.' ', all of which skewed the sample towards rich upper class white people more susceptible to vote for the GOP.

In contrast, Gallup correctly predicted the outcome by polling (only) 50K inhabitants. Read the full story here.

> **i** Thinking outside the box
>
> What are the considerations that could guide you in determining the population of interest for your study?

1.3.3 Sampling

Because sampling is costly, we can only collect limited information about the variable of interest, drawing from the population through a sampling frame (phone books, population register, etc.) Good sampling frames can be purchased from sampling firms.

In general, *randomization* is necessary in order to obtain a **representative** sample[5], one that match the characteristics of the population. Failing to randomize leads to introduction of bias and generally the conclusions drawn from a study won't be generalizable.

Even when observational units are selected at random to participate, there may be **bias** introduced due to non-response. In the 1950s, conducting surveys was relatively easier because most people were listed in telephone books; nowadays, sampling firms rely on a mix of interactive voice response and live callers, with sampling frames mixing landlines, cellphones and online panels together with (heavy) weighting to correct for non-response. Sampling is a difficult problem with which we engage only cursorily, but readers are urged to exercise scrutiny when reading papers.

[5] Note this randomization is different from the one in assigning treatments to experimental units!

> **Thinking outside the box**
>
> Reflect on the choice of platform used to collect answers and think about how it could influence the composition of the sample returned or affect non-response in a systematic way.

Before examining problems related to sampling, we review the main random sampling methods. The simplest is simple random sampling, whereby n units are drawn completely at random (uniformly) from the N elements of the sampling frame. The second most common scheme is stratified sampling, whereby a certain numbers of units are drawn uniformly from strata, namely subgroups (e.g., gender). Finally, cluster sampling consists in sampling only from some of these subgroups.

Suppose we wish to look at student satisfaction regarding the material taught in an introductory statistics course offered to multiple sections. The population consists of all students enrolled in the course in a given semester and this list provides the sampling frame. We can define strata to consist of class group. A simple random sample would be obtaining by sampling randomly abstracting from class groups, a stratified sample by drawing randomly a number from each class group and a cluster sampling by drawing all students from selected class groups. Cluster sampling is mostly useful if all groups are similar and if the costs associated to sampling from multiple strata are expensive.

Figure 1.6 shows three sampling schemes: the middle corresponds to stratum (e.g., age bands) whereas the right contains clusters (e.g., villages or classrooms).

Stratified sampling is typically superior if we care about having similar proportions of sampled in each group and is useful for reweighting: in Figure 1.6, the true proportion of sampled is $1/3$, with the simple random sampling having a range of $[0.22, 0.39]$ among the strata, compared to $[0.31, 0.33]$ for the stratified sample.

> **Thinking outside the box**
>
> The credibility of a study relies in large part on the quality of the data collection. Why is it customary to report descriptive statistics of the sample and a description of the population?

There are other instances of sampling, most of which are non-random and to be avoided whenever possible. These include convenience samples, consisting of observational units that are easy to access or include (e.g., friends, students from a university, passerby in the street). Much like for anecdotal reports, these observational units need not be representative of the whole population and it is very difficult to understand how they relate to the latter.

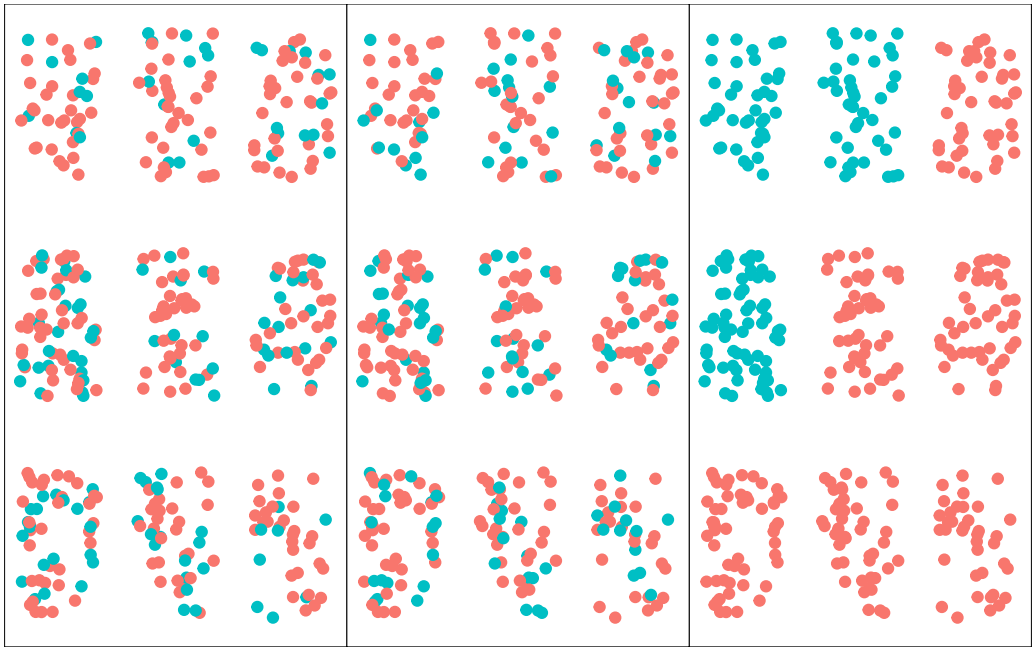


Figure 1.6: Illustration of three sampling schemes from nine groups: simple random sampling (left), stratified sampling (middle) and cluster sampling (right).

In recent years, there has been a proliferation of studies employing data obtained from web experimentation plateforms such as Amazon's Mechanical Turk (MTurk), to the point that the Journal of Management commissioned a review (Aguinis, Villamor, and Ramani 2021). These samples are subject to self-selection bias and articles using them should be read with a healthy dose of skepticism. Unless good manipulation checks are conducted (e.g., to ensure participants are faithful and answer in a reasonable amount of time), I would reserve these tools for paired samples (e.g., asking people to perform multiple tasks presented in random order) for which the composition of the population is less important. To make sure your sample matches the target population, you can use statistical tests and informal comparison and compare the distribution of individuals with the composition obtained from the census.

1.4 Examples of experimental designs

The field of experimental design has a long history, starting with agricultural field trials.

Figure 1.7: Sampling bias. Artwork by Jonathan Hey (Sketchplanations), licensed under CC-BY-SA 4.0

Example 1.3 (Agricultural field trials at the Rothamsted Research Station.)**.** The Rothamsted Research Station in the UK has been conducting experiments since 1843. Ronald A. Fisher, who worked 14 years at Rothamsted from 1919, developed much of the statistical theory underlying experimental design, inspired from his work there. Yates (1964) provides a recollection of his contribution to the field.

Experimental design revolves in large part in understanding how best to allocate our resources, determine the impact of policies or choosing the most effective "treatment" from a series of option.

Example 1.4 (Modern experiments and A/B testing)**.** Most modern experiments happen online, with tech companies running thousands of experiments on an ongoing basis in order to discover improvement to their interfaces that lead to increased profits. An Harvard Business Review article (Kohavi and Thomke 2017) details how small tweaks to the display of advertisements in the Microsoft Bing search engine landing page lead to a whooping 12% increase in revenues. Such randomized control trials, termed A/B experiments, involve splitting incoming traffic into separate groups; each group will see different views of the webpage that differ only ever slightly. The experimenters then compare traffic and click

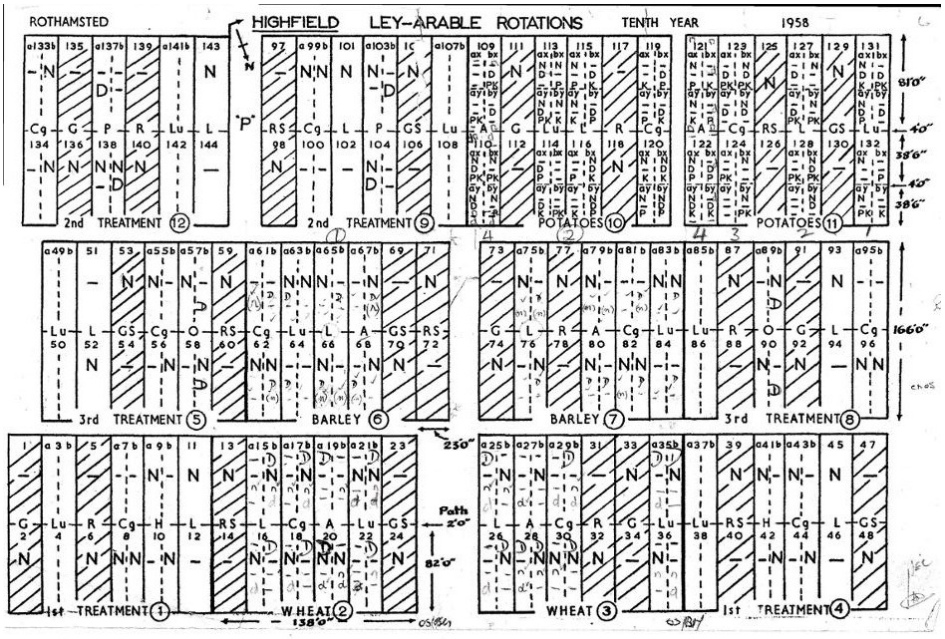


Figure 1.8: 1958 plan for the Highfield Ley–Arable Experiment. Source: Rothamsted Research Station, reproduced under the CC BY 4.0 license.

revenues. At large scale, even small effects can have major financial consequences and can be learned despite the large variability in customer behaviour.

There are also multiple examples of randomized control experiments used for policy making.

Example 1.5 (Experiments on wellness programs)**.** Song and Baicker (2019) conducted a large randomized trial over a period of 18 months to study the impact of wellness programs in US companies. The industry, worth more than 8 billions USD, has significantly increased following the passage of the Affordable Care Act, aka Obamacare. The findings are vulgarized in a press release by Jake Miller from Harvard News & Research: they show that, while there was seemingly an impact on physical activity and well-being, there were no evidence of changes in absenteeism, job tenure and job performance. Jones, Molitor, and Reif (2019) reach similar conclusion.

These findings are strikingly different from previous observational studies, which found increase participation in sportive activities, increased job duration, reduced medical spendings.

Example 1.6 (STAR)**.** The Tennessee's Student Teacher Achievement Ratio (STAR) project (Achilles et al. 2008) is another important example of large scale experiment with broad ramifications. The study suggested that smaller class sizes lead to better outcomes of pupils.

> Over 7,000 students in 79 schools were randomly assigned into one of 3 interventions: small class (13 to 17 students per teacher), regular class (22 to 25 students per teacher), and regular-with-aide class (22 to 25 students with a full-time teacher's aide). Classroom teachers were also randomly assigned to the classes they would teach. The interventions were initiated as the students entered school in kindergarten and continued through third grade.

Example 1.7 (RAND health care programs)**.** In a large-scale multiyear experiment conducted by the RAND Corporation (Brook et al. 2006), participants who paid for a share of their health care used fewer health services than a comparison group given free care. The study concluded that cost sharing reduced "inappropriate or unnecessary" medical care (overutilization), but also reduced "appropriate or needed" medical care.

> The HIE was a large-scale, randomized experiment conducted between 1971 and 1982. For the study, RAND recruited 2,750 families encompassing more than 7,700 individuals, all of whom were under the age of 65. They were chosen from six sites across the United States to provide a regional and urban/rural balance. Participants were randomly assigned to one of five types of health insurance plans created specifically for the experiment. There were four basic types of fee-for-service plans: One type offered free care; the other three types involved varying levels of cost sharing — 25 percent, 50 percent, or 95 percent coinsurance (the percentage of medical charges that the consumer must pay). The fifth type of health insurance plan was a nonprofit, HMO-style group cooperative. Those assigned to the HMO received their care free of charge. For poorer families in plans that involved cost sharing, the amount of cost sharing was income-adjusted to one of three levels: 5, 10, or 15 percent of income. Out-of-pocket spending was capped at these percentages of income or at $1,000 annually (roughly $3,000 annually if adjusted from 1977 to 2005 levels), whichever was lower.
>
> Families participated in the experiment for 3–5 years. The upper age limit for adults at the time of enrollment was 61, so that no participants would become eligible for Medicare before the experiment ended. To assess participant service use, costs, and quality of care, RAND served as the families' insurer and processed their claims. To assess participant health, RAND administered surveys at the beginning and end of the experiment and also conducted comprehensive physical exams. Sixty percent of participants were randomly chosen to receive

> exams at the beginning of the study, and all received physicals at the end. The random use of physicals at the beginning was intended to control for possible health effects that might be stimulated by the physical exam alone, independent of further participation in the experiment.

There are many other great examples in the dedicated section of Chapter 10 of *Telling stories with data* by Rohan Alexander (Alexander 2022). Section 1.4 of Berger, Maurer, and Celli (2018) also lists various applications of experimental designs in a variety of fields.

1.5 Requirements for good experiments

Section 1.2 of Cox (1958) describes the various requirements that are necessary for experiments to be useful. These are

1. absence of systematic error
2. precision
3. range of validity
4. simplicity

We review each in turn.

1.5.1 Absence of systematic error

This point requires careful planning and listing potential confounding variables that could affect the response.

Suppose we wish to consider the differences in student performance between two instructors. If the first teaches only morning classes, while the second only teaches in the evening, it will be impossible to disentangle the effect of timing with that of instructor performance. Such comparisons should only be undertaken if there is compelling prior evidence that timing does not impact the outcome of interest.

The first point raised by Cox is thus that we

> ensure that experimental units receiving one treatment differ in no systematic way from those receiving another treatment.

This point also motivates use of **double-blind** procedures (where both experimenters and participants are unaware of the treatment allocation) and use of placebo in control groups (to avoid psychological effects, etc. associated with receiving treatment or lack thereof).

Randomization[6] is at the core of achieving this goal, and ensuring measurements are independent of one another also comes out as corollary.

1.5.2 Variability

The second point listed by Cox (1958) is that of the variability of estimator. Much of the precision can be captured by the signal-to-noise ratio, in which the difference in mean treatment is divided by its standard error, a form of effect size. The intuition should be that it's easier to detect something when the signal is large and the background noise is low. The latter is a function of

(a) the accuracy of the experimental work and measurements apparatus and the intrinsic variability of the phenomenon under study,
(b) the number of experimental and observational units (the sample size).
(c) the choice of design and statistical procedures.

Point (a) typically cannot be influenced by the experimenter outside of choosing the response variable to obtain more reliable measurements. Point (c) related to the method of analysis, is usually standard unless there are robustness considerations. Point (b) is at the core of the planning, notably in choosing the number of units to use and the allocation of treatment to the different (sub)-units.

1.5.3 Generalizability

Most studies are done with an objective of generalizing the findings beyond the particular units analyzed. The range of validity thus crucially depends with the choice of population from which a sample is drawn and the particular sampling scheme. Non-random sampling severely limits the extrapolation of the results to more general settings. This leads Cox to advocate having

> not just empirical knowledge about what the treatment differences are, but also some understanding of the reasons for the differences.

[6] The percentage of participants need not be equiprobable, nor do we need to assign the same probability to each participant. However, going away from equal number of people per group has consequences and makes the statistical analysis more complicated.

Even if we believe a factor to have no effect, it may be wise to introduce it in the experiment to check this assumption: if it is not a source of variability, it shouldn't impact the findings and at the same time would provide some more robustness.

If we look at a continuous treatment, than it is probably only safe to draw conclusions within the range of doses administered. Comic in Figure 2.3 is absurd, but makes this point.

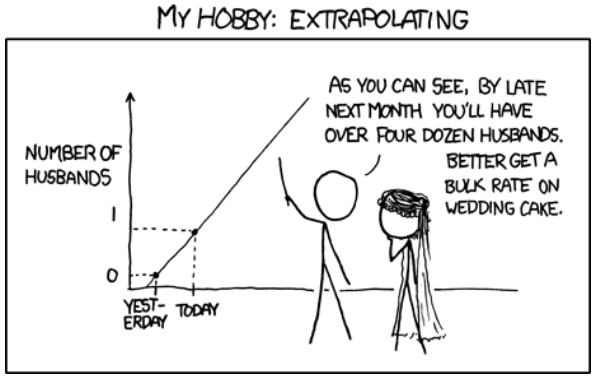


Figure 1.9: xkcd comic 605 (Extrapolating) by Randall Munroe. Alt text: By the third trimester, there will be thousands of babies inside you. Cartoon reprinted under the CC BY-NC 2.5 license.

Replication studies done in university often draw participants from students enrolled in the institutions. The findings are thus not necessarily robust if extrapolated to the whole population if there are characteristics for which they have strong (familiarity to technology, acquaintance with administrative system, political views, etc). These samples are often **convenience samples**.

Example 1.9 in Cox (1958) mentions recollections of "Student"[7] on Spratt-Archer barley, a new variety of barley that performed well in experiments and whose culture the Irish Department of Agriculture encouraged. Fuelled by a district skepticism with the new variety, the Department ran an experiment comparing the yield of the Spratt-Archer barley with that of the native race. Their findings surprised the experimenters: the native barley grew more quickly and was more resistant to weeds, leading to higher yields. It was concluded that the initial experiments were misleading because Spratt-Archer barley had been experimented in well-farmed areas, exempt of nuisance.

[7] William Sealy Gosset

1.5.4 Simplicity

The fourth requirement is one of simplicity of design, which almost invariably leads to simplicity of the statistical analysis. Randomized control-trials are often viewed as the golden rule for determining efficacy of policies or treatments because the set of assumptions they make is pretty minimalist due to randomization. Most researchers in management are not necessarily comfortable with advanced statistical techniques and this also minimizes the burden. Figure 1.10 shows an hypothetical graph on the efficacy of the Moderna MRNA vaccine for Covid: if the difference is clearly visible in a suitable experimental setting, then conclusions are easily drawn.

Randomization justifies the use of the statistical tools we will use under very weak assumptions, if units measurements are independent from one another. Drawing conclusions from observational studies, in contrast to experimental designs, requires making often unrealistic or unverifiable assumptions and the choice of techniques required to handle the lack of randomness is often beyond the toolbox of applied researchers.

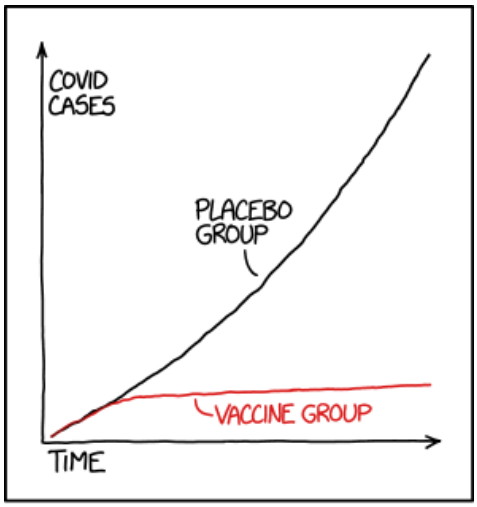


Figure 1.10: xkcd comic 2400 (Statistics) by Randall Munroe. Alt text: We reject the null hypothesis based on the 'hot damn, check out this chart' test. Cartoon reprinted under the CC BY-NC 2.5 license.

Your turn

- Define the following terms in your own word: experimental unit, factor, treatment.
- What is the main benefit of experimental studies over observational studies?

2 Hypothesis testing

In most applied domains, empirical evidences drive the advancement of the field and data from well designed experiments contribute to the built up of science. In order to draw conclusions in favour or against a theory, researchers turn (often unwillingly) to statistics to back up their claims. This has led to the prevalence of the use of the null hypothesis statistical testing (NHST) framework. One important aspect of the reproducibility crisis is the misuse of p-values in journal articles: falsification of a null hypothesis is not enough to provide substantive findings for a theory.

Because introductory statistics course typically present hypothesis tests without giving much thoughts to the underlying construction principles of such procedures, users often have a reductive view of statistics as a catalogue of pre-determined procedures. To make a culinary analogy, users focus on learning recipes rather than trying to understand the basics of cookery. This chapter focuses on understanding of key ideas related to testing.

! Key concept

Learning objectives:

- Understanding the role of uncertainty in decision making.
- Understanding the importance of signal-to-noise ratio as a measure of evidence.
- Knowing the basic ingredients of hypothesis testing and being capable of correctly formulating and identifying these components in a paper.
- Correctly interpreting p-values and confidence intervals for a parameter.

2.1 Hypothesis

The first step of a design is formulating a research question. Generally, this hypothesis will specify potential differences between population characteristics due to some intervention (a treatment) that the researcher wants to quantify. This is the step during which researchers decide on sample size, choice of response variable and metric for the measurement, write down the study plan, etc.

It is important to note that most research questions cannot be answered by simple tools. Researchers wishing to perform innovative methodological research should contact experts and consult with statisticians **before** they collect their data to get information on how best to proceed for what they have in mind so as to avoid the risk of making misleading and false claims based on incorrect analysis or data collection.

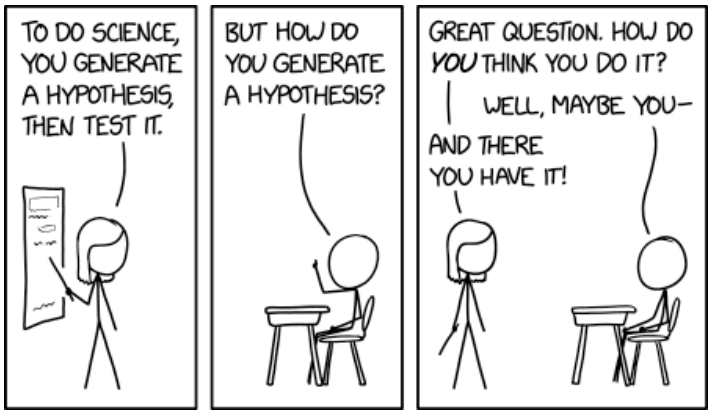


Figure 2.1: xkcd comic 2569 (Hypothesis generation) by Randall Munroe. Alt text: Frazzled scientists are requesting that everyone please stop generating hypotheses for a little bit while they work through the backlog. Cartoon reprinted under the CC BY-NC 2.5 license.

2.2 Sampling variability

Given data, a researcher will be interested in estimating particular characteristics of the population. We can characterize the set of all potential values their measurements can take, together with their frequency, via a distribution.

The purpose of this section is to illustrate how we cannot simply use raw differences between groups to make meaningful comparisons: due to sampling variability, samples will be alike even if they are generated in the same way, but there will be always be differences between their summary statistics. Such differences tend to attenuate (or increase) as we collect more sample. Inherent to this is the fact that as we gather more data (and thus more information) about our target, the portrait becomes more precise. This is ultimately what allows us to draw meaningful conclusions but, in order to do so, we need first to determine what is likely or plausible and could be a stroke of luck, and what is not likely to occur solely due to randomness.

Example 2.1 (A/B testing)**.** Consider two webpage design: one is the current version (*status quo*) and the other implementation contains a clickable banner in a location where eye-tracker suggest that viewers eyes spend more time or attention. The number of clicks on those headlines are what generate longer viewing, and thus higher revenues from advertisement. The characteristic of interest here would be the average click conversation rate for each of the webpage design.

It is fairly simple to redirect traffic so that a random fraction gets assigned to the new design for study. After a suitable period of time, the data can be analyzed to see if the new webpage generates more clicks.

An hypothesis test will focus on one or multiple of these characteristics. Suppose for simplicity that we have only two groups, control and treatment, whose population averages are μ_C and μ_T we wish to compare. People commonly look at the difference in average, say $\delta = \mu_T - \mu_C$ as a measure of the effectiveness of the treatment.[1] If we properly randomized observations in each subgroup and nothing else changes, then this measures the impact of the treatment. Because we only have a sample at hand and not the whole population, we don't know for sure the values of μ_C and μ_T. These quantities exist, but are unknown to us so the best we can do is estimate them using our sample. If we have a random sample from the population, then the characteristics of the sample will be (noisy) proxys of those of the population.

We call numerical summaries of the data **statistics**. Its important to distinguish between procedures/formulas and their numerical values. An **estimator** is a rule or formula used to calculate an estimate of some parameter or quantity of interest based on observed data (like a recipe for cake). Once we have observed data we can actually compute the sample mean, that is, we have an estimate — an actual value (the cake), which is a single realization and not random. In other words,

- an estimand is our conceptual target, like the population characteristic of interest (population mean).
- an estimator is the procedure or formula telling us how to transform the sample data into a numerical summary that is a proxy of our target.
- an estimate is a number, the numerical value obtained once we apply the formula to observed data.

For example, we may use as estimand the population average of $Y_1, \ldots$, say μ. The estimator will be sample mean, i.e., the sum of the elements in the sample divided by the sample size, $\overline{Y} = (Y_1 + \cdots + Y_n)/n$. The estimate will be a numerical value, say 4.3.

[1] We could look at the ratio μ_T/μ_C instead.

(a) Estimand

(b) Estimator

(c) Estimate

Figure 2.2: Estimand (left), estimator (middle) and estimate (right) illustrated with cakes and based on an original idea of Simon Grund. Cake photos shared under CC BY-NC 2.0 license.

Because the inputs of the estimator are random, the output is also random and change from one sample to the next: even if you repeat a recipe, you won't get the exact same result every time, as in Figure 2.3.

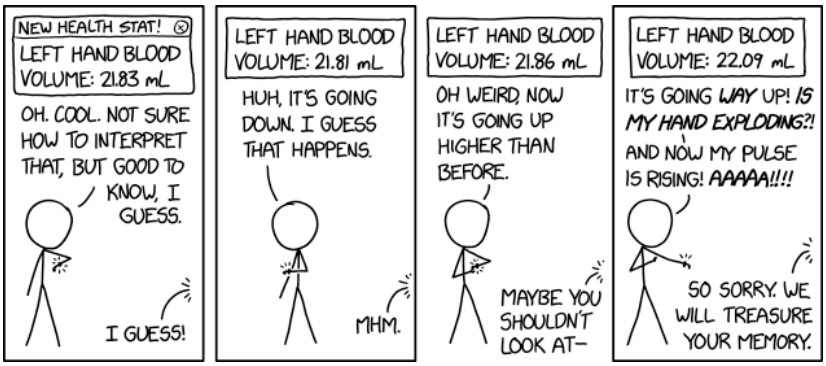


Figure 2.3: xkcd comic 2581 (Health Stats) by Randall Munroe. Alt text: You will live on forever in our hearts, pushing a little extra blood toward our left hands now and then to give them a squeeze. Cartoon reprinted under the CC BY-NC 2.5 license.

To illustrate this point, Figure 2.4 shows five simple random samples of size $n = 10$ drawn from an hypothetical population with mean μ and standard deviation σ, along with their sample mean $\overline{y}$. Thus, sampling variability implies that the sample means of the subgroups will always differ even if they share the same characteristics. You can view sampling variability as noise: our goal is to extract the signal (typically differences in means) but accounting

for spurious results due to the background noise.

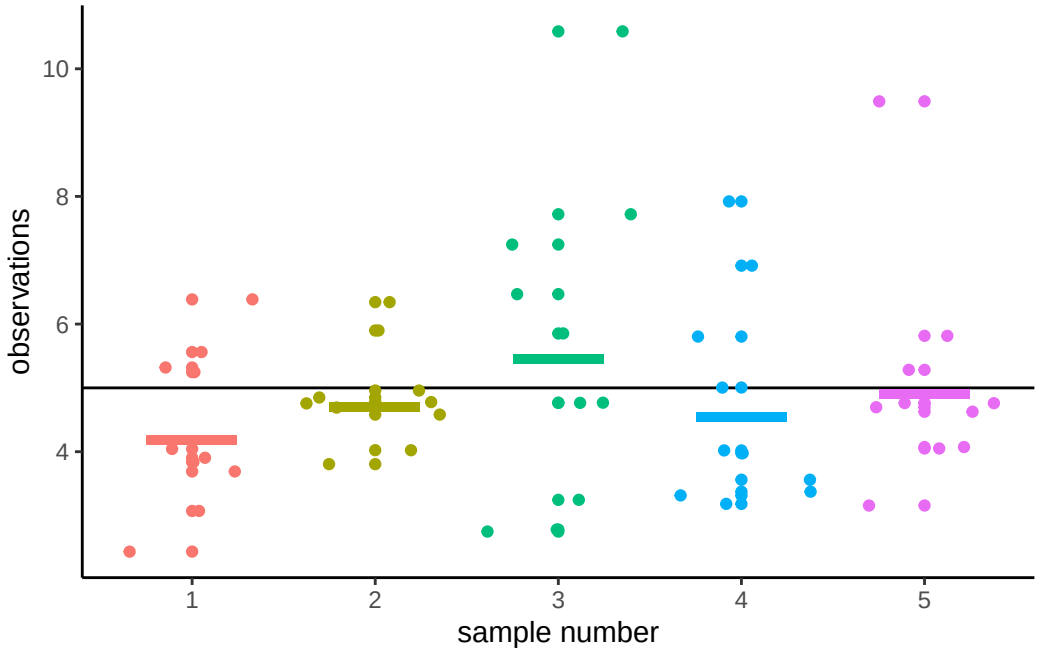


Figure 2.4: Five samples of size $n = 10$ drawn from a common population with mean μ (horizontal line). The colored segments show the sample means of each sample.

We can clearly see from Figure 2.4 that, even if each sample is drawn from the same population, the sample mean varies from one sample to the next as a result of the sampling variability. The astute eye might even notice that the sample means are less dispersed around the full black horizontal line representing the population average μ than are the individual measurements. This is a fundamental principle of statistics: information accumulates as you get more data.

Values of the sample mean don't tell the whole picture and studying differences in mean (between groups, or relative to a postulated reference value) is not enough to draw conclusions. In most settings, there is no guarantee that the sample mean will be equal to it's true value because it changes from one sample to the next: the only guarantee we have is that it will be on average equal to the population average in repeated samples. Depending on the choice of measurement and variability in the population, there may be considerable differences from one observation to the next and this means the observed difference could be a fluke.

To get an idea of how certain something is, we have to consider the variability of an observation Y_i. This variance of an observation drawn from the population is typically denoted

σ^2 and it's square root, the standard deviation, by σ.

The sample variance S_n is an estimator of the standard deviation σ, where

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

is the sum of squared difference between observations and the sample average, scaled by a factor proportional to the sample size.

The standard deviation *of a statistic* is termed **standard error**; it should not be confused with the standard deviation σ of the population from which the sample observations $Y_1, \ldots, Y_n$ are drawn. Both standard deviation and standard error are expressed in the same units as the measurements, so are easier to interpret than variance. Since the standard error is a function of the sample size, it is however good practice to report the estimated standard deviation in reports.

Example 2.2 (Sample proportion and uniform draws)**.** To illustrate the concept of sampling variability, we follow the lead of Matthew Crump and consider samples from a uniform distribution on $\{1, 2, \ldots, 10\}$ each number in this interval is equally likely to be sampled.

Even if they are drawn from the same population, the 10 samples in Figure 2.5 look quite different. The only thing at play here is the sample variability: since there are $n = 20$ observations in total, there should be on average 10% of the observations in each of the 10 bins, but some bins are empty and others have more counts than expected. This fluctuation is due to randomness, or chance.

How can we thus detect whether what we see is compatible with the model we think generated the data? The key is to collect more observations: the bar height is the sample proportion, an average of 0/1 values with ones indicating that the observation is in the bin and zero otherwise.

Consider now what happens as we increase the sample size: the top panel of Figure 2.6 shows uniform samples for increasing samples size. The histogram looks more and more like the true underlying distribution (flat, each bin with equal frequency) as the sample size increases. The sample distribution of points is nearly indistinguishable from the theoretical one (straight line) when $n = 10000$.[2] The bottom panel, on the other hand, isn't from a uniform distribution and larger samples come closer to the population distribution. We couldn't have spotted this difference in the first two plots, since the sampling variability is too important; there, the lack of data in some bins could have been attributed to chance, as they are comparable with the graph for data that are truly uniform. This is in line with most

[2]The formula shows that the standard error decreases by a tenfold every time the sample size increases by a factor 100.

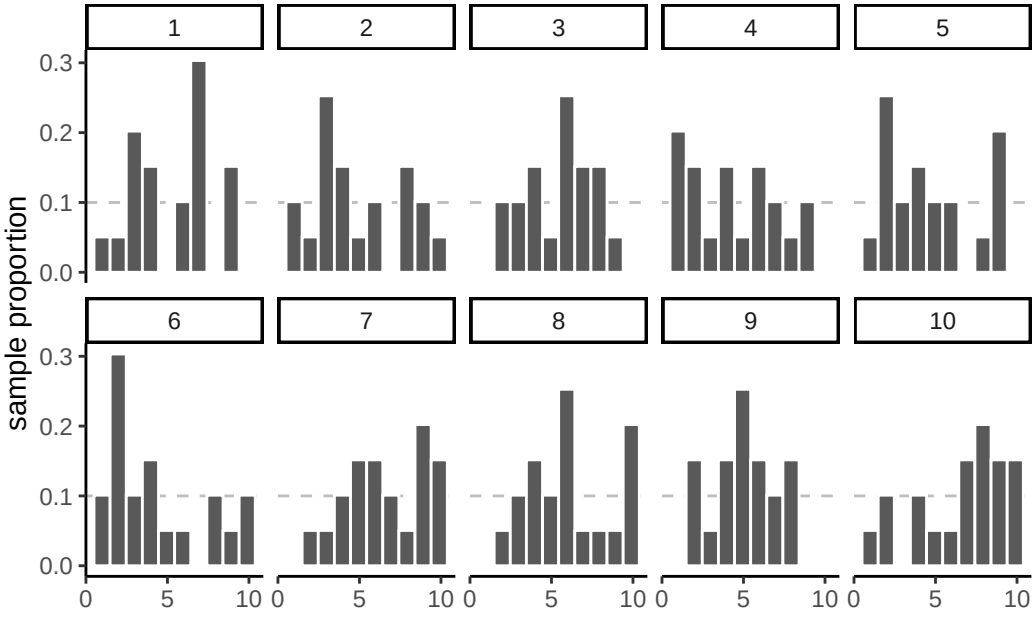


Figure 2.5: Histograms for 10 random samples of size $n = 20$ from a discrete uniform distribution.

practical applications, in which the limited sample size restricts our capacity to disentangle real differences from sampling variability. We must embrace this uncertainty: in the next section, we outline how hypothesis testing helps us disentangle the signal from the noise.

2.3 Hypothesis testing

An hypothesis test is a binary decision rule (yes/no) used to evaluate the statistical evidence provided by a sample to make a decision regarding the underlying population. The main steps involved are:

- define the model parameters
- formulate the alternative and null hypothesis
- choose and calculate the test statistic
- obtain the null distribution describing the behaviour of the test statistic under $\mathscr{H}_0$
- calculate the *p*-value
- conclude (reject or fail to reject $\mathscr{H}_0$) in the context of the problem.

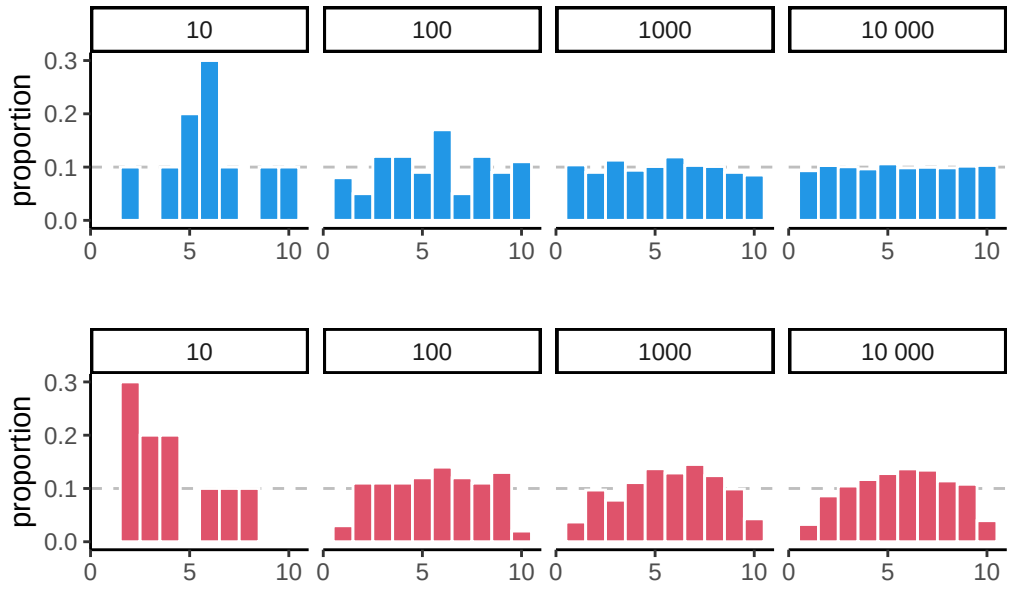


Figure 2.6: Histograms of data from a uniform distribution (top) and non-uniform (bottom) with increasing sample sizes of 10, 100, 1000 and 10 000 (from left to right).

A good analogy for hypothesis tests is a trial for murder on which you are appointed juror.

- The judge lets you choose between two mutually exclusive outcome, guilty or not guilty, based on the evidence presented in court.
- The presumption of innocence applies and evidences are judged under this optic: are evidence remotely plausible if the person was innocent? The burden of the proof lies with the prosecution to avoid as much as possible judicial errors. The null hypothesis $\mathscr{H}_0$ is *not guilty*, whereas the alternative $\mathscr{H}_a$ is *guilty*. If there is a reasonable doubt, the verdict of the trial will be not guilty.
- The test statistic (and the choice of test) represents the summary of the proof. The more overwhelming the evidence, the higher the chance the accused will be declared guilty. The prosecutor chooses the proof so as to best outline this: the choice of evidence (statistic) ultimately will maximize the evidence, which parallels the power of the test.
- The null distribution is the benchmark against which to judge the evidence (jurisprudence). Given the proof, what are the odds assuming the person is innocent? Since this is possibly different for every test, it is common to report instead a p-value, which gives the level of evidence on a uniform scale which is most easily interpreted.

- The final step is the verdict, a binary decision with outcomes: guilty or not guilty. For an hypothesis test performed at level α, one would reject (guilty) if the *p*-value is less than α. Even if we declare the person not guilty, this doesn't mean the defendant is innocent and vice-versa.

2.3.1 Hypothesis

In statistical tests we have two hypotheses: the null hypothesis ($\mathscr{H}_0$) and the alternative hypothesis ($\mathscr{H}_a$). Usually, the null hypothesis (the 'status quo') is a single numerical value. The alternative is what we're really interested in testing. In Figure 2.4, we could consider whether all five groups have the same mean $\mathscr{H}_0 : \mu_1 = \mu_2 = \cdots = \mu_5$ against the alternative that at least two of them are different. These two outcomes are mutually exclusive and cover all possible scenarios. A statistical hypothesis test allows us to decide whether or not our data provides enough evidence to reject $\mathscr{H}_0$ in favor of $\mathscr{H}_a$, subject to some pre-specified risk of error: while we know that the differences are just due to sampling variability in Figure 2.4 because the data is simulated, in practice we need to assess the evidence using a numerical summary.

Example 2.3 (A/B testing (continued))**.** We follow-up with our A/B test experiment. Given μ_1 the population average click conversation rate for the current webpage and μ_2, that of the redesign, we are interested in the *one-sided hypothesis* that $\mathscr{H}_0 : \mu_2 \leq \mu_1$ against the alternative (that we are trying to prove) $\mathscr{H}_a : \mu_2 > \mu_1$. In choosing as null hypothesis that the new design is no better or worst, we are putting all our weight to make sure the changes carry forward if there is overwhelming evidence that the new design is better and allow us to generate more revenues, given the costs associated to changes to the interface and the resulting disruption.

One-sided hypothesis are directional: we care only about a specific direction, and so here $\mathscr{H}_a : \mu_2 > \mu_1$. Indeed, if the experiment suggests that the conversion rate is worst with the new webpage design, we won't go forward.

Since neither of these population averages μ_1 and μ_2 are known to us, we can work instead with $\mathscr{H}_0 : \mu_2 - \mu_1 \geq 0$. We can use as estimator for the difference $\mu_2 - \mu_1$ the difference in sample average in each subgroup.

The null hypothesis here is an interval, but it suffices the consider the most beneficial scenario, which is $\mu_2 - \mu_1 = 0$. Indeed, if we can disprove that there is no difference and see an increase of the click rate with the updated version, all more extreme cases are automatically discarded in favour of the alternative that the new design is better.

One-sided tests for which the evidence runs contrary to the hypothesis (say the mean conversion rate is higher for the current design than for the new one) lead to *p*-values of 1,

since there is no proof against the null hypothesis that the old design (the status quo) is better.

The previous example illustrates the fact that, when writing down null and alternative hypotheses, what we are trying to prove is typically the alternative.

In pairwise comparisons or contrasts, we can assign a directionality. The benefit is that, if we are sure of the direction of the postulated effect, we only consider as extreme scenarios that run in the direction we postulated[3] However, if the empirical evidence runs contrary to our guess, then there is no support for the hypothesis.

In more general statistical models, it helps to view the null hypothesis as a simplification of a more complex model: the latter will fit the data better because it is more flexible, but we would fail to reject the null unless this improvement is drastic. For example, in an analysis of variance model, we compare different mean in each of K groups against a single common average.

2.3.2 Test statistic

A test statistic T is a function of the data which takes the data as input and outputs a summary of the information contained in the sample for a characteristic of interest, say the population mean. In order to assess whether the numerical value for T is unusual, we need to know what are the potential values taken by T and their relative probability if $\mathscr{H}_0$ is true. We need to know what values we should expect if, e.g., there was no difference in the averages of the different groups: this requires a benchmark.

Many statistics we will consider are of the form[4]

$$T = \frac{\text{estimated effect} - \text{postulated effect}}{\text{estimated effect variability}} = \frac{\widehat{\theta} - \theta_0}{\mathrm{se}(\widehat{\theta})}$$

where $\widehat{\theta}$ is an estimator of θ, θ_0 is the postulated value of the parameter and $\mathrm{se}(\widehat{\theta})$ is the standard error of the test statistic $\widehat{\theta}$. This quantity is designed so that, if our postulated value θ_0 is correct, T has approximately mean zero and variance one. This standardization makes comparison easier; in fact, the form of the test statistic is chosen so that it doesn't depend on the measurement units.

For example, if we are interested in mean differences between treatment group and control group, denoted μ_T and μ_C, then $\theta = \mu_T - \mu_C$ and $\mathscr{H}_0 : \mu_T = \mu_C$ corresponds to $\mathscr{H}_0 : \theta = 0$

[3] This implies that the level α is all on one side, rather than split equally between both tails of the distribution. In practice, this translates into increased power of detection provided the effect is in the postulated direction.

[4] This class of statistic, which includes t-tests, are called Wald statistics.

for no difference. The two-sample t-test would have numerator $\widehat{\theta} = \overline{Y}_T - \overline{Y}_C$, where $\overline{Y}_T$ is the sample average in treatment group and $\overline{Y}_C$ that of the control group. The postulated value for the mean difference is zero.

The numerator would thus consist of the difference in sample means and the denominator the standard error of that quantity, calculated using a software.[5]

2.3.3 Null distribution and *p*-value

The *p*-value allows us to decide whether the observed value of the test statistic T is plausible under $\mathscr{H}_0$. Specifically, the *p*-value is the probability that the test statistic is equal or more extreme to the estimate computed from the data, assuming $\mathscr{H}_0$ is true. Suppose that based on a random sample $Y_1, \ldots, Y_n$ we obtain a statistic whose value $T = t$. For a two-sided test $\mathscr{H}_0 : \theta = \theta_0$ vs. $\mathscr{H}_a : \theta \neq \theta_0$, the *p*-value is $\mathsf{Pr}_0(|T| \geq |t|)$.[6]

How do we determine the null distribution given that the true data generating mechanism is unknown to us? We ask a statistician! In simple cases, it might be possible to enumerate all possible outcomes and thus quantity the degree of outlyingness of our observed statistic. In more general settings, we can resort to simulations or to probability theory: the central limit theorem says that the sample mean behaves like a normal random variable with mean μ and standard deviation $\sigma/\sqrt{n}$ for n large enough. The central limit theorem has broader applications since most statistics can be viewed as some form of average or transformation thereof, a fact used to derive benchmarks for most commonly used tests. Most software use these approximations as proxy by default: the normal, Student's t, χ^2 and F distributions are the reference distributions that arise the most often.

Figure 2.7 shows the distribution of p-values for two scenarios: one in which there are no differences and the null is true, the other under an alternative. The probability of rejection is obtained by calculating the area under the density curve between zero and $\alpha = 0.1$, here 0.1 on the left. Under the null, the model is calibrated and the distribution of *p*-values is uniform (i.e., a flat rectangle of height 1), meaning all values in the unit interval are equally likely. Under the alternative (right), small *p*-values are more likely to be observed.

There are generally three ways of obtaining null distributions for assessing the degree of evidence against the null hypothesis

- exact calculations
- large sample theory (aka 'asymptotics' in statistical lingo)
- simulation

[5] Assuming equal variance, the denominator is estimated using the pooled variance estimator.
[6] If the distribution of T is symmetric around zero, the *p*-value reduces to $p = 2 \times \mathsf{Pr}_0(T \geq |t|)$.

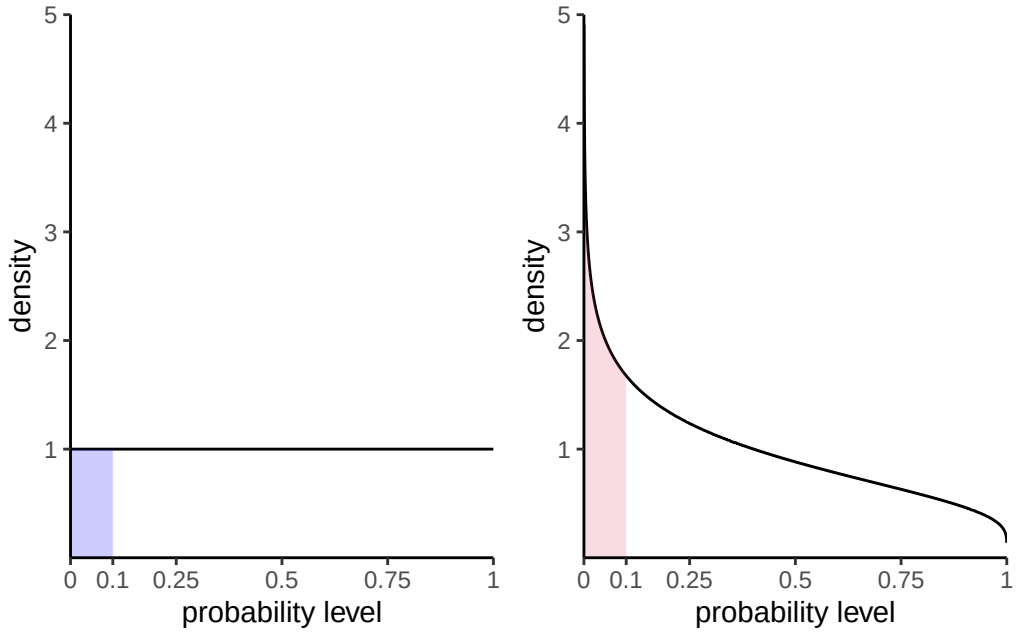


Figure 2.7: Density of p-values under the null hypothesis (left) and under an alternative with a signal-to-noise ratio of 0.5 (right).

While desirable, the first method is only applicable in simple cases (such as counting the probability of getting two six if you throw two fair die). The second method is most commonly used due to its generality and ease of use (particularly in older times where computing power was scarce), but fares poorly with small sample sizes (where 'too small' is context and test-dependent). The last approach can be used to approximate the null distribution in many scenarios, but adds a layer of randomness and the extra computations costs sometimes are not worth it.

2.3.4 Conclusion

The p-value allows us to make a decision about the null hypothesis. If $\mathscr{H}_0$ is true, the p-value follows a uniform distribution, as shown in Figure 2.7. Thus, if the p-value is small, this means observing an outcome more extreme than $T = t$ is unlikely, and so we're inclined to think that $\mathscr{H}_0$ is not true. There's always some underlying risk that we're making a mistake when we make a decision. In statistic, there are two type of errors:

- type I error: we reject the null hypothesis $\mathscr{H}_0$ when the null is true,

- type II error: we fail to reject the null hypothesis $\mathscr{H}_0$ when the alternative is true.

The two hypothesis are not judged equally: we seek to avoid error of type I (judicial errors, corresponding to condamning an innocent). To prevent this, we fix a the level of the test, α, which captures our tolerance to the risk of commiting a type I error: the higher the level of the test α, the more often we will reject the null hypothesis when the latter is true. The value of $\alpha \in (0,1)$ is the probability of rejecting $\mathscr{H}_0$ when $\mathscr{H}_0$ is in fact true,

$$\alpha = \mathsf{Pr}_0 (\text{ reject } \mathscr{H}_0) .$$

The level α is fixed beforehand, typically 1%, 5% or 10%. Keep in mind that the probability of type I error is α only if the null model for $\mathscr{H}_0$ is correct (sic) and correspond to the data generating mechanism.

The focus on type I error is best understood by thinking about costs of moving away from the status quo: a new website design or branding will be costly to implement, so you want to make sure there are enough evidence that the proposal is the better alternative and will lead to increased traffic or revenues.

Decision \ true model	$\mathscr{H}_0$	$\mathscr{H}_a$
fail to reject $\mathscr{H}_0$	✓	type II error
reject $\mathscr{H}_0$	type I error	✓

To make a decision, we compare our *p*-value P with the level of the test α:

- if $P < \alpha$, we reject $\mathscr{H}_0$;
- if $P \geq \alpha$, we fail to reject $\mathscr{H}_0$.

Do not mix up level of the test (a probability fixed beforehand by the researcher) and the *p*-value. If you do a test at level 5%, the probability of type I error (condemning an innocent by mistake) is by definition α and does not depend on the *p*-value. The latter is a conditional probability of observing a more extreme statistic given the null distribution $\mathscr{H}_0$ is true.

Pitfall

The American Statistical Association (ASA) published a list of principles guiding (mis)interpretation of *p*-values, some of which are reproduced below:

(2) *P*-values do not measure the probability that the studied hypothesis is true.

(3) Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold.

Table 2.2: Promotion recommandation to branch manager based on sex of the applicant.

	male	female
promote	32	19
hold file	12	30

(4) *P*-values and related analyses should not be reported selectively.

(5) *p*-value, or statistical significance, does not measure the size of an effect or the importance of a result.

Example 2.4 (Gender inequality and permutation tests)**.** We consider data from Rosen and Jerdee (1974), who look at sex role stereotypes and their impacts on promotion and opportunities for women candidates. The experiment took place in 1972 and the experimental units, which consisted of 95 male bank supervisors, were submitted to various memorandums and asked to provide ratings or decisions based on the information provided.

We are interested in Experiment 1 related to promotion of employees: managers were requested to decide on whether or not to promote an employee to become branch manager based on recommendations and ratings on potential for customer and employee relations.

The authors intervention focused on the description of the nature (complexity) of the manager's job (either simple or complex) and the sex of the candidate (male or female): all files were otherwise similar.

We consider for simplicity only sex as a factor and aggregate over job for the $n = 93$ replies. Table 2.2 shows the counts for each possibility.

The null hypothesis of interest here that sex has no impact, so the probability of promotion is the same for men and women. Let p_{m} and p_{w} denote these respective probabilities; we can thus write mathematically the null hypothesis as $\mathscr{H}_0 : p_{\text{m}} = p_{\text{w}}$ against the alternative $\mathscr{H}_a : p_{\text{m}} \neq p_{\text{w}}$.

The test statistic typically employed for contingency tables is a chi-square test[7], which compares the overall proportions of promoted to that in for each subgroup. The sample proportion for male is 32/42 = ~76%, compared to 19/49 or ~49% for female. While it seems that this difference of 16% is large, it could be spurious: the standard error for the sample proportions is roughly 3.2% for male and 3.4% for female.

[7] If you have taken advanced modelling courses, this is a score test obtained by fitting a Poisson regression with `sex` and `action` as covariates; the null hypothesis corresponding to lack of interaction term between the two.

Table 2.3: First five rows of the database in long format for experiment 1 of Rosen and Jerdee.

action	sex
hold file	female
promote	female
promote	male
hold file	female
hold file	female

If there was no discrimination based on sex, we would expect the proportion of people promoted to be the same overall; this is 51/93 =0.55 for the pooled sample. We could simply do a test for the mean difference, but rely instead on the Pearson contingency X_p^2 (aka chi-square) test, which compares the expected counts (based on equal promotion rates) to the observed counts, suitably standardized. If the discrepancy is large between expected and observed, than this casts doubt on the validity of the null hypothesis.

If the counts of each cell are large, the null distribution of the chi-square test is well approximated by a χ^2 distribution. The output of the test includes the value of the statistic, 10.79, the degrees of freedom of the χ^2 approximation and the *p*-value, which gives the probability that a random draw from a χ_1^2 distribution is larger than the observed test statistic **assuming the null hypothesis is true**. The *p*-value is very small, 0.001, which means such a result is quite unlikely to happen by chance if there was no sex-discrimination.

Another alternative to obtain a benchmark to assess the outlyingness of the observed odds ratio is to use simulations. Consider a database containing the raw data with 93 rows, one for each manager, with for each an indicator of `action` and the `sex` of the hypothetical employee presented in the task.

Under the null hypothesis, sex has no incidence on the action of the manager. This means we could get an idea of the "what-if" world by shuffling the sex labels repeatedly. Thus, we could obtain a benchmark by repeating the following steps multiple times:

1. permute the labels for `sex`,
2. recreate a contingency table by aggregating counts,
3. calculate a test statistic for the simulated table.

As test statistic, we use odds ratio: the odds of an event is the ratio of the number of success over failure: in our example, this would be the number of promoted over held files. The odds of promotion for male is $32/12$, whereas that of female is $19/30$. The odds ratio for male versus female is thus $\mathsf{OR} = (32/12)/(19/30) = 4.21$. Under the null hypothesis, $\mathscr{H}_0 : \mathsf{OR} = 1$ (same probability of being promoted) (why?)

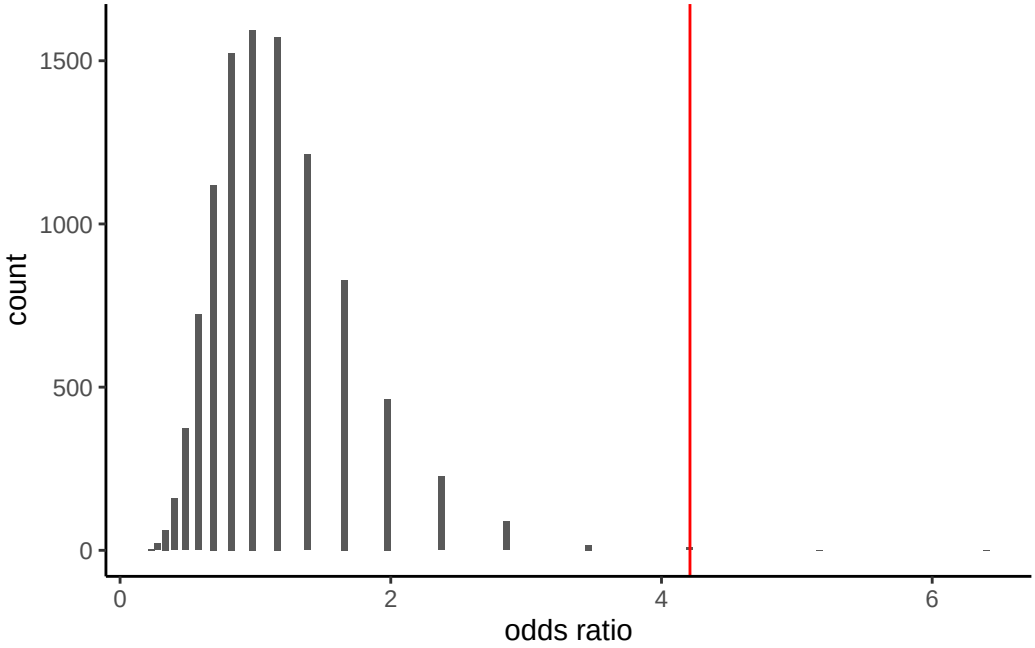


Figure 2.8: Histogram of the simulated null distribution of the odds ratio statistic obtained using a permutation test; the vertical red line indicates the sample odds ratio.

The histogram in Figure 2.8 shows the distribution of the odds ratio based on 10 000 permutations. Reassuringly, we again get roughly the same approximate p-value, here 0.002.[8]

The article concluded (in light of the above and further experiments)

> Results confirmed the hypothesis that male administrators tend to discriminate against female employees in personnel decisions involving promotion, development, and supervision.

Recap

- Model parameters: probability of promotion for men and women, respectively p_{m} and p_{w}.
- Hypotheses: no discrimination based on gender, meaning equal probability of promotion (null hypothesis $\mathscr{H}_0 : p_{\text{m}} = p_{\text{w}}$, versus alternative hypothesis $\mathscr{H}_a : p_{\text{m}} \neq p_{\text{w}}$).
- Test statistic: (1) chi-square test for contingency tables and (2) odds ratio.

[8] The p-value obtained for the permutation test would change from one run to the next since it's input is random. However, the precision of the proportion statistic is sufficient for decision making purposes.

- p-value: (1) $.0010$ and (2) $.0024$ based on permutation test.
- Conclusion: reject null hypothesis, as there is evidence of a gender-discrimination with different probability of promotion for men and women.

Following the APA guidelines, the χ^2 statistic would be reported as $\chi^2(1, n = 93) = 10.79$, $p = .001$ along with counts and sample proportions.

Pitfall

In the first experiment, managers were also asked to rank applications on their potential for both employee and customer relations using a Likert scale of six items ranging from (1) extremely unfavorable to (6) extremely favorable. However, only the averages are reported in Table 1 along with (Rosen and Jerdee 1974)

> Mean rating for the male candidate was 4.73 compared to a mean rating of 4.25 for the female candidate ($F = 4.76$, df $= 1/80$, $p < .05$)

The degrees of freedom (80) are much too few compared to the number of observations, implying non-response that isn't discussed.
Partial or selective reporting of statistical procedures hinders reproducibility. In general, the presentation should explicitly state the name of the test statistic employed, the sample size, mean and variance estimates, the null distribution used to assess significance and its parameters, if any. Without these, we are left to speculate.

2.4 Confidence intervals

A **confidence interval** is an alternative way to present the conclusions of an hypothesis test performed at significance level α by giving a range of all values for which the null isn't rejected at the chosen level. It is often combined with a point estimator $\hat{\theta}$ to give an indication of the variability of the estimation procedure. Wald-based $(1 - \alpha)$ confidence intervals for a parameter θ are of the form

$$\widehat{\theta} + \text{critical value } \mathrm{se}(\widehat{\theta})$$

based on the Wald statistic W,

$$W = \frac{\widehat{\theta} - \theta}{\mathrm{se}(\widehat{\theta})},$$

and where θ represents the postulated value for the fixed, but unknown value of the parameter. The critical values are quantile of the null distribution and are chosen so that the probability of being more extreme is α.

The bounds of the confidence intervals are random variables, since both estimators of the parameter and its standard error, $\widehat{\theta}$ and $\mathrm{se}(\widehat{\theta})$, are random: their values will vary from one sample to the next.

For generic random samples, there is a $1-\alpha$ probability that θ is contained in the **random** confidence interval computed. Once we obtain a sample and calculate the confidence interval, there is no more notion of probability: the true value of the parameter θ is either inside the confidence interval or not. We can interpret confidence interval's as follows: if we were to repeat the experiment multiple times, and calculate a $1-\alpha$ confidence interval each time, then roughly $1-\alpha$ of the calculated confidence intervals would contain the true value of θ in repeated samples (in the same way, if you flip a coin, there is roughly a 50-50 chance of getting heads or tails, but any outcome will be either). Our confidence is in the *procedure* we use to calculate confidence intervals and not in the actual values we obtain from a sample.

If we are only interested in the binary decision rule reject/fail to reject $\mathscr{H}_0$, the confidence interval is equivalent to a *p*-value since it leads to the same conclusion. Whereas the $1-\alpha$ confidence interval gives the set of all values for which the test statistic doesn't provide enough evidence to reject $\mathscr{H}_0$ at level α, the *p*-value gives the probability under the null of obtaining a result more extreme than the postulated value and so is more precise for this particular value. If the *p*-value is smaller than α, our null value θ will be outside of the confidence interval and vice-versa.

Example 2.5 ("The Surprise of Reaching Out")**.** Liu et al. (2022+) studies social interactions and the impact of surprise on people reaching out if this contact is unexpected. Experiment 1 focuses on questionnaires where the experimental condition is the perceived appreciation of reaching out to someone (vs being reached to). The study used a questionnaire administered to 200 American adults recruited on the Prolific Academic platform. The response index consists of the average of four questions measured on a Likert scale ranging from 1 to 7, with higher values indicating higher appreciation.

We can begin by inspecting summary statistics for the sociodemographic variables (gender and age) to assess whether the sample is representative of the general population as a whole. The proportion of `other` (including non-binary people) is much higher than that of the general census, and the population skews quite young according to Table 2.4.

Since there are only two groups, initiator and responder, we are dealing with a pairwise comparison. The logical test one could use is a two sample t-test, or a variant thereof.

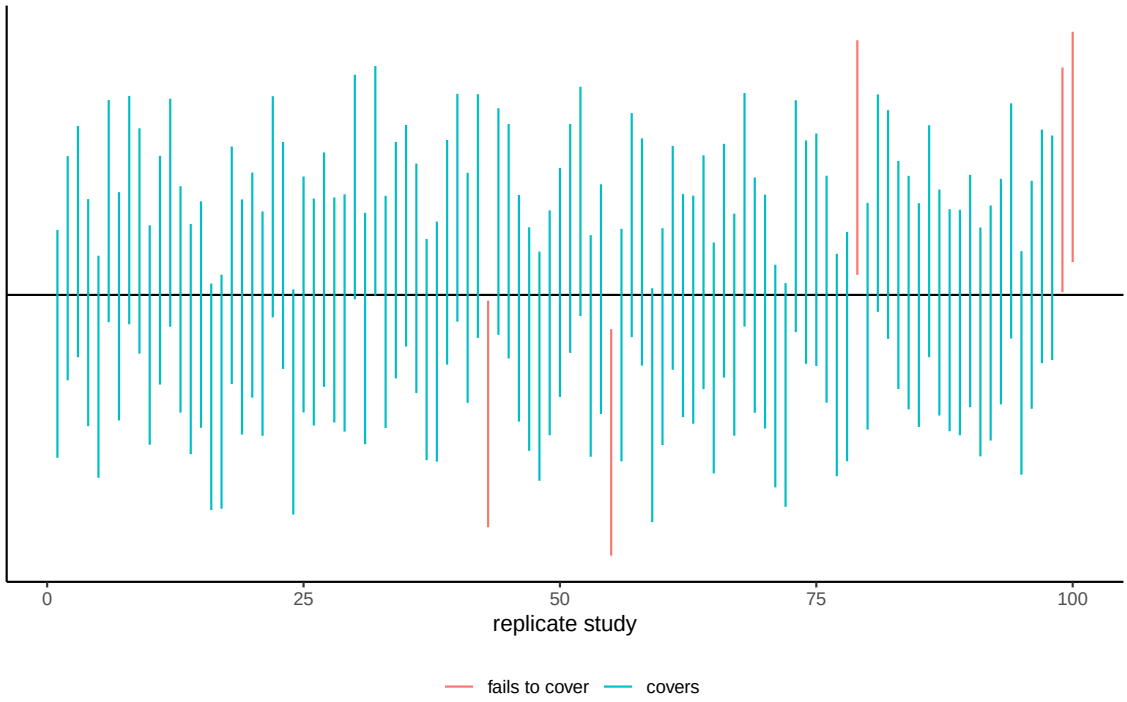


Figure 2.9: 95% confidence intervals for the mean of a standard normal population for 100 random samples. On average, 5% of these intervals fail to include the true mean value of zero (in red).

Table 2.4: Summary statistics of the age of participants, and counts per gender

gender	min	max	mean	n
male	18	78	32.03	105
female	19	68	36.50	92
other	24	30	27.67	3

Using Welch two sample t-test statistic, both group average and standard deviation are estimated using the data provided and the latter are used to build a statistic. This explains the non-integer degrees of freedom.

The software returns $t(197.52) = -2.05$, $p = .041$, which leads to the rejection of the null hypothesis of no difference in appreciation depending on the role of the individual (initiator or responder). The estimated mean difference is $\Delta M = -0.37$, 95% CI $[-0.73, -0.01]$; since 0 is not included in the confidence interval, we also reject the null hypothesis at level 5%.

Table 2.5: Mean ratings, standard deviation and number of participants per experimental condition.

role	mean	sd	n
initiator	5.50	1.28	103
responder	5.87	1.27	97

The estimate suggests that initiators underestimate the appreciation of reaching out.[9]

Recap

- Model parameters: average expected appreciation score μ_{i} and μ_{r} of initiators and responder, respectively
- Hypothesis: expected appreciation score is the same for initiator and responders, $\mathscr{H}_0 : \mu_{\mathrm{i}} = \mu_{\mathrm{r}}$ against alternative $\mathscr{H}_0 : \mu_{\mathrm{i}} \neq \mu_{\mathrm{r}}$ that they are different.
- Test statistic: Welch two sample t-test
- p-value: 0.041
- Conclusion: reject the null hypothesis, average appreciation score differs depending on the role

Example 2.6 (Virtual communication curbs creative idea generation)**.** A Nature study performed an experiment to see how virtual communications teamwork by comparing the output both in terms of ideas generated during a brainstorming session by pairs and of the quality of ideas, as measured by external referees. The sample consisted of 301 pairs of participants who interacted via either videoconference or face-to-face.

The authors compared the number of creative ideas, a subset of the ideas generated with creativity score above average. The mean number of the number of creative ideas for face-to-face 7.92 ideas (sd 3.40) relative to videoconferencing 6.73 ideas (sd 3.27).

Brucks and Levav (2022) used a negative binomial regression model: in their model, the expected number creative ideas generated is

$$\mathsf{E}(\texttt{ncreative}) = \exp(\beta_0 + \beta_1 \texttt{video})$$

where $\texttt{video} = 0$ if the pair are in the same room and $\texttt{video} = 1$ if they interact instead via videoconferencing.

The mean number of ideas for videoconferencing is thus $\exp(\beta_1)$ times that of the face-to-face: the estimate of the multiplicative factor is $\exp(\beta_1)$ is 0.85 95% CI $[0.77, 0.94]$.

[9] Assuming that the variance of each subgroup were equal, we could have used a two-sample t-test instead. The difference in the conclusion is immaterial, with a nearly equal *p*-value.

No difference between experimental conditions translates into the null hypothesis as $\mathscr{H}_0 : \beta_1 = 0$ vs $\mathscr{H}_0 : \beta_1 \neq 0$ or equivalently $\mathscr{H}_0 : \exp(\beta_1) = 1$. The likelihood ratio test comparing the regression model with and without `video` the statistic is $R = 9.89$ (p-value based on χ^2_1 of $.002$). We conclude the average number of ideas is different, with summary statistics suggesting that virtual pairs generate fewer ideas.

If we had resorted to a two sample t-test, we would have found a mean difference in number of creative idea of $\Delta M = 1.19$, 95% CI $[0.43, 1.95]$, $t(299) = 3.09$, $p = .002$.

Both tests come with slightly different sets of assumptions, but yield similar conclusions: there is evidence of a smaller number of creative ideas when people interact via videoconferencing.

2.5 Conclusion

This chapter has focused on presenting the tools of the trade and some examples outlining the key ingredients that are common to any statistical procedure and the reporting of the latter. The reader is not expected to know which test statistic to adopt, but rather should understand at this stage how our ability to do (scientific) discoveries depends on a number of factors.

Richard McElreath in the first chapter of his book (McElreath 2020) draws a parallel between statistical tests and golems (i.e., robots): neither

> discern when the context is inapropriate for its answers. It just knows its own procedure [...] It just does as it's told.

The responsibility therefore lies with the user to correctly use statistical procedures and be aware of their limitations. A *p*-value does not indicate whether the hypothesis is reasonable, whether the design is proper, whether the choice of measurement is adequate, etc.

Your turn

Pick a journal paper (e.g., one of the dataset documented in the course webpage) and a particular study.
Look up for the ingredients of the testing procedure (parameters, hypotheses, test statistic name and value, summary statistics, *p*-value, conclusion).
You may encounter other measures, such as effect size, that will be discussed later.

3 Completely randomized designs

This chapter focuses on experiments where potentially multiple factors of interest are manipulated by the experimenter to study their impact. If the allocation of observational units to each treatment combination is completely random, the resulting experiment is a completely randomized design.

The one-way analysis of variance describes the most simple experimental setup one can consider: completely randomized experiments with one factor, in which we are solely interested in the effect of a single treatment variable with multiple levels.

3.1 One-way analysis of variance

The focus is on comparisons of the average of a single outcome variable with K different treatments levels, each defining a sub-population differing only in the experimental condition they received. A **one-way analysis of variance** compares the sample averages of each treatment group $T_1, \ldots, T_K$ to try and determine if the population averages could be the same. Since we have K groups, there will be K averages (one per group) to estimate.

Let $\mu_1, \ldots, \mu_K$ denote the theoretical (unknown) mean (aka expectation) of each of the K sub-populations defined by the different treatments. Lack of difference between treatments is equivalent to equality of means, which translates into the hypotheses

$$\begin{aligned}\mathscr{H}_0 &: \mu_1 = \cdots = \mu_K \\ \mathscr{H}_a &: \text{at least two treatments have different averages,}\end{aligned}$$

The null hypothesis is, as usual, a single numerical value, μ. The alternative consists of all potential scenarios for which not all expectations are equal. Going from K averages to one requires imposing $K-1$ restrictions (the number of equality signs), as the value of the global mean μ is left unspecified.

3.1.1 Parametrizations and contrasts

This section can be skipped on first reading. It focuses on the interpretation of the coefficients obtained from a linear model or analysis of variance model.

The most natural parametrization is in terms of group averages: the (theoretical unknown) average for treatment T_j is μ_j, so we obtain K parameters $\mu_1, \ldots, \mu_K$ whose estimates are the sample averages $\widehat{\mu}_1, \ldots, \widehat{\mu}_K$. One slight complication arising from the above is that the values of the population average are unknown, so this formulation is ill-suited for hypothesis testing because none of the μ_i values are known in practice and we need to make comparisons in terms of a known numerical value.

The most common parametrization for the linear model is in terms of **differences to a baseline**, say T_1. The theoretical average of each group is written as $\mu_1 + a_i$ for treatment T_i, where $a_1 = 0$ for T_1 and $a_i = \mu_i - \mu_1$ otherwise. The parameters are $\mu_1, a_2, \ldots, a_K$.

An equivalent formulation writes for each treatment group the average of subpopulation j as $\mu_j = \mu + \delta_j$, where δ_j is the difference between the treatment average μ_j and the global average of all groups. Imposing the constraint $\delta_1 + \cdots + \delta_K = 0$ ensures that the average of effects equals μ. Thus, if we know any $K-1$ of $\{\delta_1, \ldots, \delta_K\}$, we automatically can deduce the last one.

In **R**, the `lm` function fits a linear model based on a formula of the form `response ~ explanatory`. If the explanatory is categorical (i.e., a factor), the parameters of this model are the intercept, which is the sample average of the baseline group and the other parameters are simply contrasts, i.e., the a_i's.

The sum-to-zero parametrization is obtained with `contrasts = list(... = contr.sum)`, where the ellipsis is replaced by the name of the categorical variable; an easier alternative is `aov`, which enforces this parametrization by default. With the sum-to-zero parametrization, the intercept is the average of each treatment average, $(\widehat{\mu}_1 + \cdots + \widehat{\mu}_5)/5$; this need not coincide with the (overall) mean of the response $\widehat{\mu} = \overline{y}$ unless the sample the number of observations in each group is the same.[1] The other coefficients of the sum-to-zero parametrization are the differences between this intercept and the group means.

We show the function call to fit a one-way ANOVA in the different parametrizations along with the sample average of each arithmetic group (the two controls who were taught separately and the groups that were praised, reproved and ignored in the third class). Note that the omitted category changes depending on the parametrization.

[1] We say a sample is balanced if each (sub)group contains the same number of observations.

Table 3.1: Coefficients of the analysis of variance model for the arithmetic scores using different parametrizations.

group	mean	contrasts	sum-to-zero
intercept		19.67	21.00
control 1	19.67		-1.33
control 2	18.33	-1.33	-2.67
praise	27.44	7.78	6.44
reprove	23.44	3.78	2.44
ignore	16.11	-3.56	

```
mod_contrast <- lm(score ~ group,
                   data = arithmetic)
mod_sum2zero <- lm(score ~ group,
                   data = arithmetic,
                   contrasts = list(group = contr.sum))
```

We can still assess the hypothesis by comparing the sample means in each group, which are noisy estimates of the population mean: their inherent variability will limit our ability to detect differences in averages if the signal-to-noise ratio is small.

3.1.2 Sum of squares decomposition

The following section can be safely skipped on first reading: it attempts to shed some light into how the F-test statistic works as a summary of evidence, as it isn't straightforward in the way it appears.

The usual notation for the sum of squares decomposition is as follows: suppose y_{ik} represents the ith person in the kth treatment group ($k = 1, \ldots, K$) and the sample size n can be split between groups as $n_1, \ldots, n_K$; in the case of a balanced sample, $n_1 = \cdots = n_K = n/K$ and the number of observations in each group is the same. We denote by $\widehat{\mu}_k$ the sample average in group k and $\widehat{\mu}$ the overall average $(y_{11} + \cdots + y_{n_K K})/n = \sum_k \sum_i y_{ik}/n$, where $\sum_i$ denotes the sum over all individuals in the group.

Under the null model, all groups have the same mean, so the natural estimator for the latter is the sample average of the pooled sample $\widehat{\mu}$ and likewise the group averages $\widehat{\mu}_1, \ldots, \widehat{\mu}_K$ are the best estimators for the group averages if each group has a (potentially) different mean. The more complex model, which has more parameters, will always fit better because it has more possibility to accommodate differences observed in a group, even if these are

spurious. The sum of squares measures the (squared) distance between the observation and the fitted values, with the terminology total, within and between sum of squares linked to the decomposition

$$\underbrace{\sum_i \sum_k (y_{ik} - \widehat{\mu})^2}_{\text{total sum of squares}} = \underbrace{\sum_i \sum_k (y_{ik} - \widehat{\mu}_k)^2}_{\text{within sum of squares}} + \underbrace{\sum_k n_i (\widehat{\mu}_k - \widehat{\mu})^2}_{\text{between sum of squares}} \quad .$$

The term on the left is a measure of the variability for the null model ($\mu_1 = \cdots = \mu_K$) under which all observations are predicted by the overall average $\widehat{\mu}$. The within sum of squares measures the distance between observations and their group mean, which describes the alternative model in which each group has (potentially) a different average, but the same variability.

We can measure how much worst we do with the alternative model (different average per group) relative to the null by calculating the between sum of square. This quantity in itself varies with the sample size (the more observations, the larger it is) so we must standardize as usual this quantity so that we have a suitable benchmark.

The F-statistic is

$$\begin{aligned} F &= \frac{\text{between-group variability}}{\text{within-group variability}} \\ &= \frac{\text{between sum of squares}/(K-1)}{\text{within sum of squares}/(n-K)} \end{aligned} \tag{3.1}$$

If there is no mean difference (null hypothesis), the numerator is an estimator of the population variance, and so is the denominator of eq. 3.1 and the ratio of the two is approximately 1 on average. However, the between sum of square is more variable and this induces skewness: for large enough sample, the null distribution of the F-statistic is approximately an F-distribution, whose shape is governed by two parameters named degrees of freedom which appear in Equation 3.1 as scaling factors to ensure proper standardization. The first degree of freedom is the number of restrictions imposed by the null hypothesis ($K-1$, the number of groups minus one for the one-way analysis of variance), and the second degree of freedom is the number of observations minus the number of *parameters estimates* for the mean ($n-K$, where n is the overall sample size and K is the number of groups).[2]

Figure 3.1 shows how the difference between these distances can encompass information that the null is wrong. The sum of squares is obtained by computing the squared length of these vectors and adding them up. The left panel shows strong signal-to-noise ratio, so that, on average, the black segments are much longer than the colored ones. This indicates that

[2] There are only K parameter estimates for the mean, since the overall mean is full determined by the other averages with $n\widehat{\mu} = n_1\widehat{\mu}_1 + \cdots + n_K\widehat{\mu}_K$.

the model obtained by letting each group have its own mean is much better than the other. The picture in the right panel is not as clear: on average, the colored arrows are shorter, but the difference in length is much smaller relative to the colored arrows.

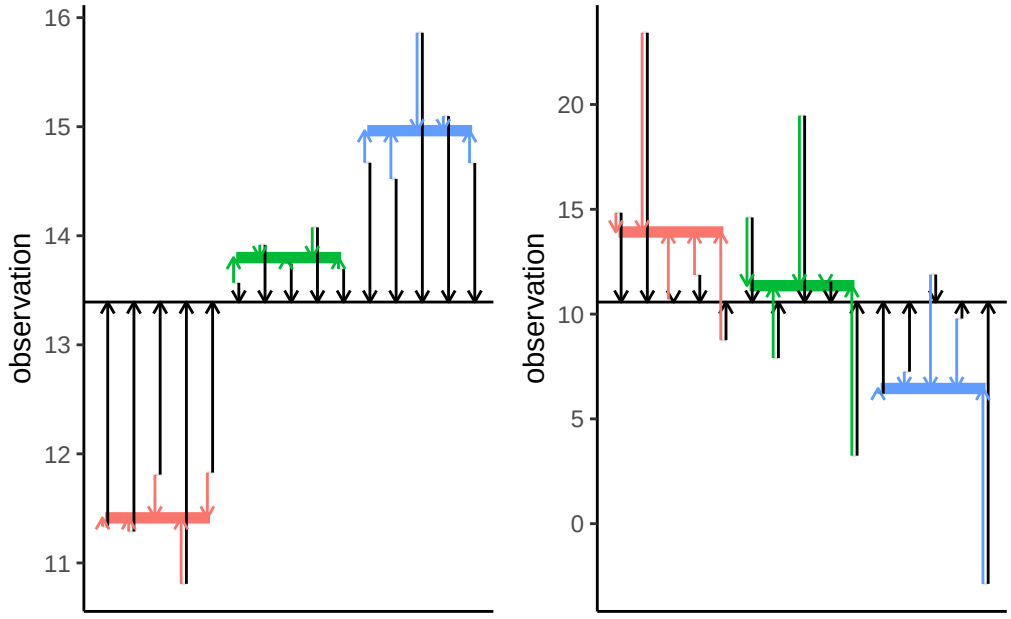


Figure 3.1: Observations drawn from three groups from a model with a strong (left) and weak (right) signal-to-noise ratio, along with their sample mean (colored horizontal segments) and the overall average (horizontal line). Arrows indicate the magnitude of the difference between the observation and the (group/average) mean.

The F-distribution is what we call a **large sample approximation** to the behaviour of the statistic if there is truly no difference between group averages (and if model assumptions are satisfied): it tells us what to expect if there is nothing going on. The quality of the approximation depends on the sample size in each group: it is more accurate when there are more observations in each group, as average estimation becomes more reliable[3].

As was alluded to in the last chapter, large sample approximations are not the only option for assessing the null, but they are cheap and easy to obtain. If the distributions are the same under the null and alternative except for a location shift, we could instead resort to a permutation-based approach to generate those alternative samples by simply shuffling the labels. We see in Figure 3.2 that the histogram of the F-statistic values obtained from 1000

[3] Mostly because the central limit theorem kicks in

permutations closely matches that of the large-sample F-distribution when there are on average 20 observations per group (right), so the computational burden associated with running this simulation outweights the benefits. However, with smaller samples (left), the large sample approximation appears underdispersed relative to the permutation-based distribution, with more extreme outcomes; the latter should be viewed as more accurate in this setting.

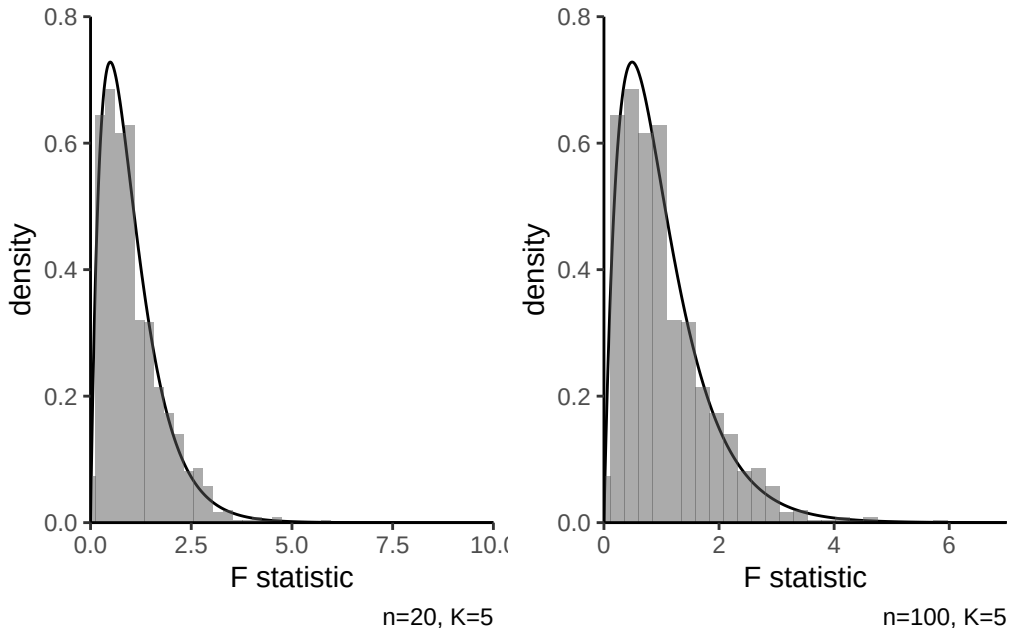


Figure 3.2: One-way analysis of variance for a sample of size 20 (left) and 100 (right), split in five groups. The histograms shows the computed test values based on 1000 permutations, which are compared to the density of the large-sample F-distribution.

More interestingly perhaps is what happens to the values taken by the statistic when not all of the averages are the same. We can see in Figure 3.3 that, when there are some difference between group means, the values taken by the statistic for a random sample are more to the right than the null distribution: the larger those differences, the more the curve will shift to the right and the more often we will obtain a value in the rejection region (in red).

If there are only two groups, then one can show that the F-statistic is mathematically equivalent to squaring the t-statistic: the null distributions are $\mathsf{St}(n-K)$ and $\mathsf{F}(1, n-K)$ and lead to the same p-values and thus same statistical inference and conclusions.

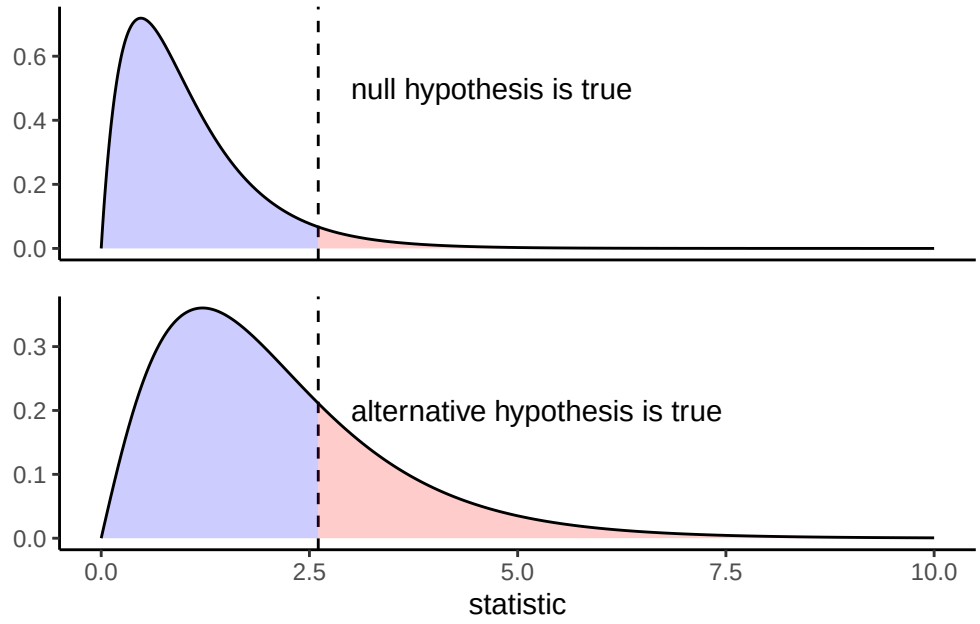


Figure 3.3: Distribution of the F-test statistic for the one-way analysis of variance when the true group means are equal (top) and under a specific alternative when they are not (bottom). Any value falling within the red region leads to rejection of the null hypothesis at level $\alpha = 0.05$.

3.2 Graphical representation

How to represent data for a one-way analysis in a publication? The purpose of the visualization is to provide intuition that extends beyond the reported descriptive statistics and to check the model assumptions. Most of the time, we will be interested in averages and dispersion, but plotting the raw data can be insightful. It is also important to keep in mind that summary statistics are estimators of population quantities that are perhaps unreliable (much too variable) in small samples to be meaningful quantities. Since the mean estimates will likely be reported in the text, the graphics should be used to convey additional information about the data. If the samples are extremely large, then graphics will be typically be used to present salient features of the distributions.

In a one-way analysis of variance, the outcome is a continuous numerical variable, whereas the treatment or explanatory is a categorical variable. Basic graphics include dot plots, histograms and density plots, or rugs for the raw data.

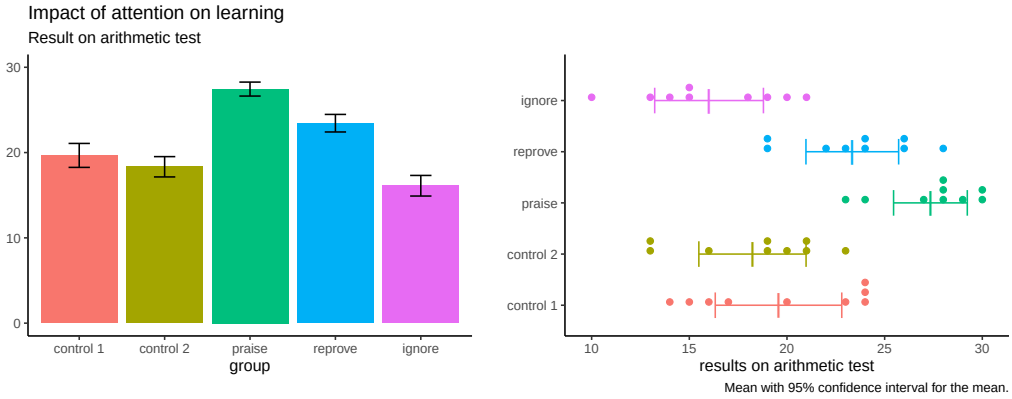


Figure 3.4: Two graphical representations of the arithmetic data: dynamite plot (left) showing the sample average with one standard error above and below, and dot plot with the sample mean (right).

Typically, scatterplots are not a good option because observations get overlaid. There are multiple workarounds, involving transparency, bubble plots for discrete data with ties, adding noise (jitter) to every observation or drawing values using a thin line (rugs) if the data are continuous and take on few distinct values.

Journals are plagued with poor vizualisations, a prime example of which is the infamous dynamite plot: it consists of a bar plot with one standard error interval. The problem with this (or with other summary statistics) is that they hide precious information about the spread and values taken by the data, as many different data could give rise to the same average while being quite different in nature. The height of the bar is the sample average and the bars extend beyond one standard error: this makes little sense as we end up comparing areas, whereas the mean is a single number. The right panel of Figure 3.4 shows instead a dot plot for the data, i.e., sample values with ties stacked for clarity, along with the sample average and a 95% confidence interval for the latter as a line underneath. In this example, there are not enough observations per group to produce histograms, and a five number summary of nine observations isn't really necessary so boxplot are useless. Weissgerber et al. (2015) discusses alternative solutions and can be referenced when fighting reviewers who insist on bad visualizations.

If we have a lot of data, it sometimes help to represent selected summary statistics or group data. A box-and-whiskers plot (or boxplot) is a commonly used graphic representing the whole data distribution using five numbers

- The box gives the quartiles, say q_1, q_2 (median) and q_3 of the distribution: 50% of the observations are smaller or larger than q_2, 25% are smaller than q_1 and 75% are smaller

than q_3 for the sample.
- The whiskers extend up to 1.5 times the box width $(q_3 - q_1)$ (so the largest observation that is smaller than $q_3 + 1.5(q_3 - q_1)$, etc.)

Observations beyond the whiskers are represented by dots or circles, sometimes termed outliers. However, beware of this terminology: the larger the sample size, the more values will fall outside the whiskers (about 0.7% for normal data). This is a drawback of boxplots, which were conceived at a time where big data didn't exist. If you want to combine boxplots with the raw data, remove the display of outliers to avoid artefacts.

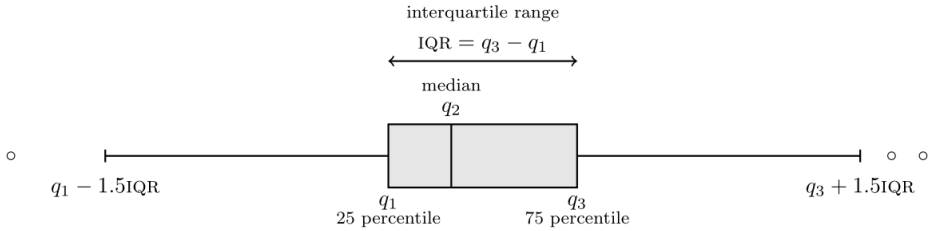


Figure 3.5: Box-and-whiskers plot

Weissgerber et al. (2019) contains many examples of how to build effective visualizations, including highlighting particular aspects using color, jittering, transparency and how to adequately select the display zone.

3.3 Pairwise tests

If the global test of equality of mean for the one-way ANOVA leads to rejection of the null, the conclusion is that one of the group has a different mean. However, the test does not indicate which of the groups differ from the rest nor does it say how many are different. There are different options: one is custom contrasts, a special instance of which is pairwise comparisons.

We are interested in looking at the difference between the (population) average of group i and j, say. The null hypothesis of no difference translate into $\mu_i - \mu_j = 0$, so the numerator of our statistic will be the estimator $\widehat{\mu}_i - \widehat{\mu}_j$ of the difference in sample mean, minus zero.

Assuming equal variances, the two-sample t-test statistic is

$$t_{ij} = \frac{(\widehat{\mu}_i - \widehat{\mu}_j) - 0}{\mathsf{se}(\widehat{\mu}_i - \widehat{\mu}_j)} = \frac{\widehat{\mu}_i - \widehat{\mu}_j}{\widehat{\sigma}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)^{1/2}},$$

where $\widehat{\mu}_i$ and n_i are respectively the sample average and the number of observations of group i, and $\widehat{\sigma}$ is the estimator of the standard deviation derived using the whole sample (assuming equal variance). As usual, the denominator of t_{ij} is the standard error of the $\widehat{\mu}_i - \widehat{\mu}_j$, whose postulated difference is zero. We can compare the value of the observed statistic to a Student-t distribution with $n-K$ degrees of freedom, denoted $\mathsf{St}(n-K)$. For a two-sided alternative, we reject if $|t_{ij}| > \mathrm{t}_{1-\alpha/2}$, for $\mathrm{t}_{1-\alpha/2}$ the $1-\alpha/2$ quantile of $\mathsf{St}(n-K)$.

Figure 3.6 shows the density of the benchmark distribution for pairwise comparisons in mean for the `arithmetic` data. The blue area under the curve defines the set of values for which we fail to reject the null hypothesis, whereas all values of the test statistic falling in the red area lead to rejection at level 5%.

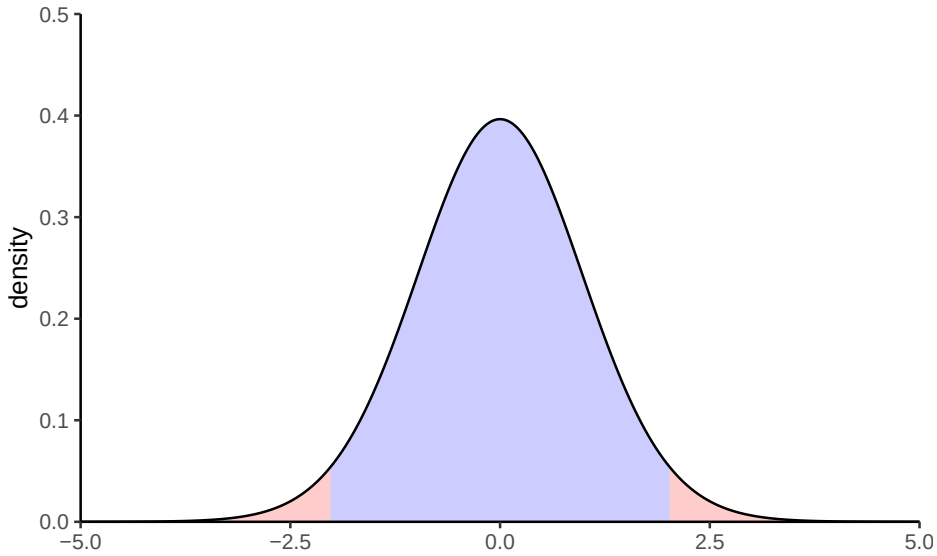


Figure 3.6: Student-t null distribution and rejection region for a t-test.

We fail to reject $\mathscr{H}_0$ as $\mathrm{t}_{\alpha/2} \leq t_{ij} \leq \mathrm{t}_{1-\alpha/2}$[4]: this gives us another way of presenting the same conclusion in terms of the set of mean differences $\delta_{ij} = \mu_i - \mu_j$ for which

$$\mathrm{t}_{\alpha/2} \leq \frac{\widehat{\delta}_{ij} - \delta_{ij}}{\mathsf{se}\left(\widehat{\delta}_{ij}\right)} \leq \mathrm{t}_{1-\alpha/2}$$

which is equivalent upon rearranging to the $(1-\alpha)$ confidence interval for δ_{ij},

$$\mathsf{CI} = \left[\widehat{\delta}_{ij} - \mathrm{t}_{1-\alpha/2}\mathsf{se}\left(\widehat{\delta}_{ij}\right), \widehat{\delta}_{ij} - \mathrm{t}_{\alpha/2}\mathsf{se}\left(\widehat{\delta}_{ij}\right)\right].$$

[4]Note that the Student-t distribution is symmetric, so $\mathrm{t}_{1-\alpha/2} = -\mathrm{t}_{\alpha/2}$.

We consider the pairwise average difference in scores between the praised (group C) and the reproved (group D) of the `arithmetic` study. The sample averages are respectively $\widehat{\mu}_C = 27.4$ and $\widehat{\mu}_D = 23.4$ and the estimated pooled standard deviation for the five groups is 1.15. Thus, the estimated average difference between groups C and D is $\widehat{\delta}_{CD} = 4$ and the standard error for the difference is $\mathsf{se}(\widehat{\delta}_{CD}) = 1.6216$; all of these are calculated by software.

If we take as null hypothesis $\mathscr{H}_0 : \delta_{CD} = 0$, the t statistic is

$$t = \frac{\widehat{\delta}_{CD} - 0}{\mathsf{se}(\widehat{\delta}_{CD})} = \frac{4}{1.6216} = 2.467$$

and the p-value is $p = 0.018$. We therefore reject the null hypothesis at level $\alpha = 0.05$ to conclude that there is a significant difference (at level $\alpha = 0.05$) between the average scores of students praised and reproved.

3.4 Model assumptions

So far, we have brushed all of the model assumptions under the carpet. These are necessary requirements for the inference to be valid: any statement related to *p*-values, etc. will approximately hold only if a set of assumptions is met in the first place. This section is devoted to the discussion of these assumptions, showcasing examples of where things can go wrong.

It is customary to write the ith observation of the kth group in the one-way analysis of variance model as

$$\underset{\text{observation}}{Y_{ik}} = \underset{\text{mean of group } k}{\mu_k} + \underset{\text{error term}}{\varepsilon_{ik}}, \tag{3.2}$$

where the error terms ε_{ik}, which account for unexplained variability and individual differences, are independent from one with mean zero and variance σ^2.

3.4.1 Additivity

The basic assumption of most designs is that we can decompose the outcome into two components (Cox 1958)

$$\begin{pmatrix}\text{quantity depending}\\ \text{on the treatment used}\end{pmatrix} + \begin{pmatrix}\text{quantity depending only}\\ \text{on the particular unit}\end{pmatrix} \tag{3.3}$$

This **additive** decomposition further assumes that each unit is unaffected by the treatment of the other units and that the average effect of the treatment is constant. Thus, it is justified

to use difference in sample mean to estimate the treatment effect since on average, the individual effect is zero.

The decomposition of observations in terms of group average and mean-zero noise in Equation 3.2 suggests that we could plot the error term ε_{ik} against observations, or against other factors or explanatories, to see if there is any unusual structure unexplained by the model and indicating problems with the randomization or additivity. However, we do not have access to ε_{ik} since both the true group mean μ_k and the error ε_{ik} are unknown. However, a good proxy is the **ordinary residual** $e_{ik} = y_{ik} - \widehat{\mu}_k$ where $\widehat{\mu}_k$ is the sample mean of all observations in experimental group k. By construction, the sample mean of the residuals will be zero, but local deviations may indicate violations of the analysis (for example, plotting residuals against time could show a learning effect).

Many graphical diagnostics use residuals, i.e., some variant of the observations minus the group mean $y_{ik} - \widehat{\mu}_k$, to look for violation of the assumptions.

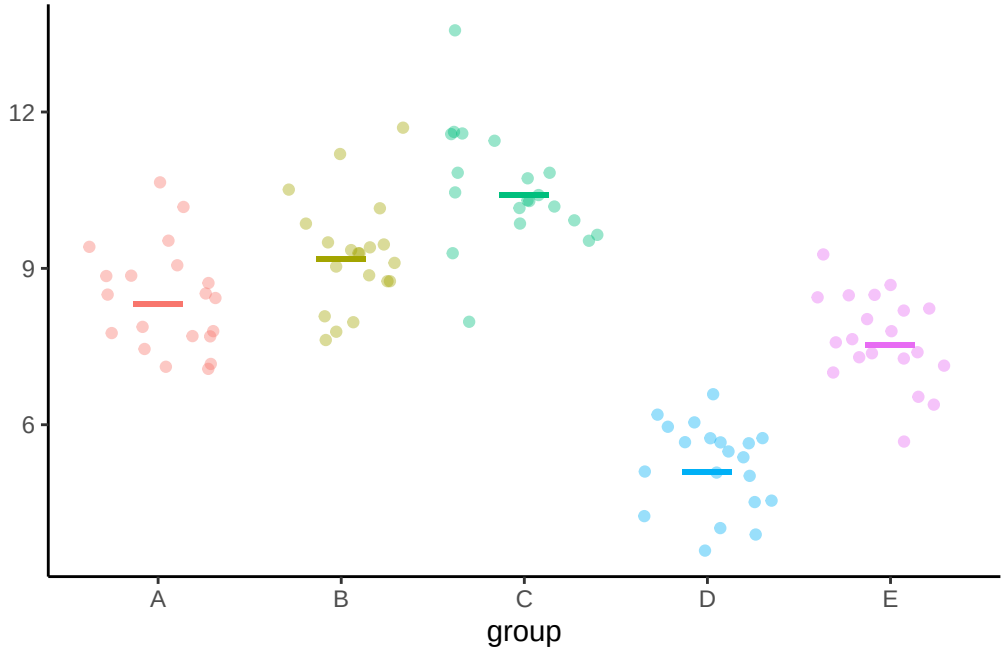


Figure 3.7: Data satisfying the assumptions of the one-way analysis of variance model, with additive effects, independent observations and common variance.

More generally, the test statistic may make further assumptions. The F-test of the global null $\mu_1 = \cdots \mu_K$ assumes that the ith observation of group k, say y_{ik}, has average $\mathsf{E}(Y_{ik}) = \mu_k$ and variance $\mathsf{Va}(Y_{ik}) = \sigma^2$. The latter is estimated using all of the residuals, with $\widehat{\sigma}^2 = \sum_k \sum_i (y_{ik} - \widehat{\mu}_k)^2/(n-K)$. Under these assumptions, the F-test statistic for the global null

$\mu_1 = \cdots = \mu_K$ is the most powerful because it uses all of the data to get a more precise estimation of the variability. Generally, there may be other considerations than power that may guide the choice of test statistic, including robustness (sensitivity to extremes and outliers). For unequal variance, other statistics than the F-test statistic may be more powerful.

Chapter 2 of Cox (1958) discusses the assumption of additivity and provides useful examples showing when it cannot be taken for granted. One of them, Example 2.3, is a scenario in which the experimental units are participants and they are asked to provide a ranking of different kindergarten students on their capacity to interact with others in games, ranked on a scale of 0 to 100. A random group of students receives additional orthopedagogical support, while the balance is in the business-as-usual setting (control group). Since there are intrinsic differences at the student level, one could consider a **paired experiment** and take as outcome the difference in sociability scores at the beginning and at the end of the school year.

One can expect the treatment to have more impact on people with low sociability skills who were struggling to make contacts: a student who scored 50 initially might see an improvement of 20 points with support relative to 10 in the business-as-usual scenario, whereas another who is well integrated and scored high initially may see an improvement of only 5 more had (s)he been assigned to the support group. This implies that the treatment effects are not constant over the scale, a violation of the additivity assumption. One way to deal with this is via transformations: Cox (1958) discusses the transformation $\log\{(x + 0.5)/(100.5 - x)\}$ to reduce the warping due to scale.

Another example is in experiments where the effect of treatment is multiplicative, so that the output is of the form

$$\begin{pmatrix}\text{quantity depending only}\\ \text{on the particular unit}\end{pmatrix} \times \begin{pmatrix}\text{quantity depending}\\ \text{on the treatment used}\end{pmatrix}$$

Usually, this arises for positive responses and treatments, in which case taking natural logarithms on both sides, with $\log(xy) = \log x + \log y$ yielding again an additive decomposition.

This example is adapted from Cox (1958), Example 2.2. Children suffering from attention deficit hyperactivity disorder (ADHD) may receive medication to increase their attention span, measured on a scale of 0 to 100, with 0 indicating normal attention span. An experiment can be designed to assess the impact of a standardized dose in a laboratory by comparing performances of students on a series of task before and after, when to a placebo. To make a case, suppose that students with ADHD fall into two categories: low symptoms and strong symptoms. In the low symptom group, the average attention is 8 per cent with the drug and 12 per cent with the placebo, whereas for people with strong symptoms, the

average is 40 per cent among treated and 60 per cent with the placebo. If these two categories are equally represented in the experiment and the population, we would estimate an average reduction of 12 percent in the score (thus higher attention span among treated). Yet, this quantity is artificial, and a better measure would be that symptoms are for the treatment are 2/3 of those of the control (the ratio of proportions).

Equation 3.3 also implies that the effect of the treatment is constant for all individuals. This often isn't the case: in an experimental study on the impact of teaching delivery type (online, hybrid, in person), it may be that the response to the choice of delivery mode depends on the different preferences of learning types (auditory, visual, kinestetic, etc.) Thus, recording additional measurements that are susceptible to interact may be useful; likewise, treatment allotment must factor in this variability should we wish to make it detectable. The solution to this would be to setup a more complex model (two-way analysis of variance, general linear model) or stratify by the explanatory variable (for example, compute the difference within each level).

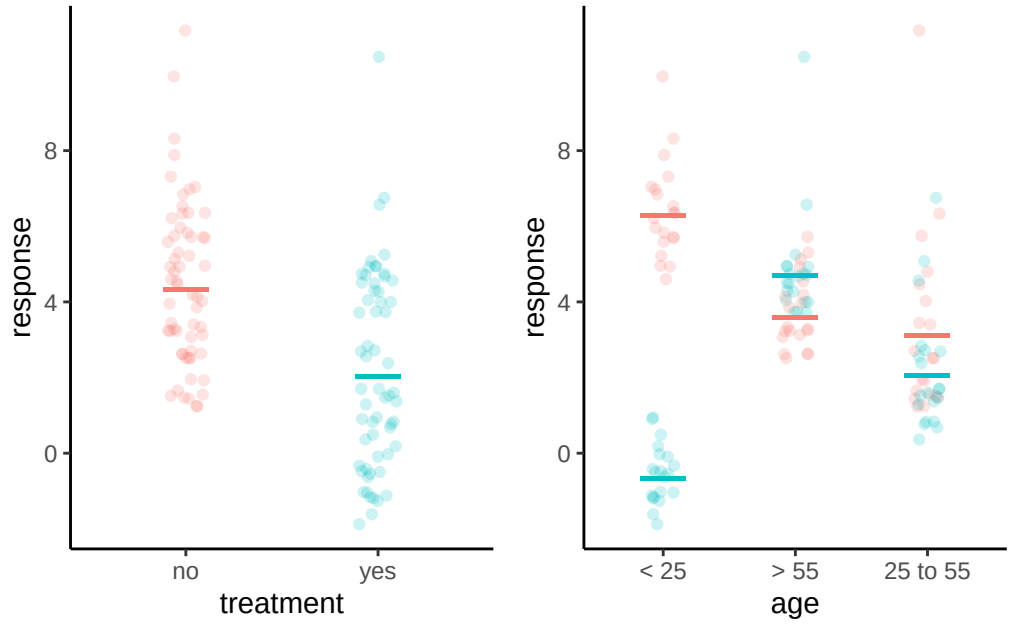


Figure 3.8: Difference in average response; while the treatment seems to lead to a decrease in the response variable, a stratification by age group reveals this only occurs in the younger population aged less than 25 years, with a seemingly reversed effect for the older adults. Thus, the marginal model implied by the one-way analysis of variance is misleading.

3.4.2 Heterogeneity

The one-way ANOVA builds on the fact that the variance in each group is equal, so that upon recentering, we can estimate it from the variance of the residuals $y_{ik} - \widehat{\mu}_k$. Specifically, the unbiased variance estimator is the denominator of the F-statistic formula, i.e., the within sum of squares divided by $n - K$ with n the total number of observations and K the number of groups under comparison.

For the time being, we consider hypothesis tests for the homogeneity (equal) variance assumption. The most commonly used tests are Bartlett's test[5] and Levene's test (a more robust alternative, less sensitive to outliers). For both tests, the null distribution is $\mathscr{H}_0 : \sigma_1^2 = \cdots = \sigma_K^2$ against the alternative that at least two differ. The Bartlett test statistic has a χ^2 null distribution with $K-1$ degrees of freedom, whereas Levene's test has an F-distribution with $(K-1, n-K)$ degrees of freedom: it is equivalent to computing the one-way ANOVA F-statistic with the absolute value of the centered residuals, $|y_{ik} - \widehat{\mu}_k|$, as observations.

```
bartlett.test(score ~ group,
              data = arithmetic)
```

```
    Bartlett test of homogeneity of variances

data:  score by group
Bartlett's K-squared = 2.3515, df = 4, p-value = 0.6714
```

```
car::leveneTest(score ~ group,
                data = arithmetic,
                center = mean)
```

```
Levene's Test for Homogeneity of Variance (center = mean)
      Df F value Pr(>F)
group  4   1.569 0.2013
      40
```

```
# compare with one-way ANOVA
mod <- lm(score ~ group, data = arithmetic)
```

[5] For the connoisseur, this is a likelihood ratio test under the assumption of normally distributed data, with a Bartlett correction to improve the χ^2 approximation to the null distribution.

```
arithmetic$absresid <- abs(resid(mod)) #|y_{ik}-mean_k|
anova(aov(absresid ~ group, data = arithmetic))
```

```
Analysis of Variance Table

Response: absresid
          Df  Sum Sq Mean Sq F value Pr(>F)
group      4  17.354  4.3385   1.569 0.2013
Residuals 40 110.606  2.7652
```

We can see in both cases that the p-values are large enough to dismiss any concern about the inequality of variance. However, should the latter be a problem, we can proceed with a test statistic that does not require variances to be equal. The most common choice is a modification due to Satterthwaite called Welch's ANOVA. It is most commonly encountered in the case of two groups ($K = 2$) and is the default option in **R** with `t.test` or `oneway.test`.

What happens with the example of the arithmetic data when we use this instead of the usual F statistic? Here, the evidence is overwhelming so no changes to the conclusion. Generally, the only drawback of using Welch's ANOVA over the usual F statistic is the need to have enough observations in each of the group to reliably estimate a separate variance[6]. For Welch's ANOVA, we have to estimate $2K$ parameters (one mean and one variance per group), rather than $K + 1$ parameters for the one-way ANOVA (one mean per group, one overall variance).

```
# Welch ANOVA
oneway.test(score ~ group, data = arithmetic,
            var.equal = FALSE)
```

```
    One-way analysis of means (not assuming equal variances)

data:  score and group
F = 18.537, num df = 4.000, denom df = 19.807, p-value = 1.776e-06
```

```
# Usual F-test statistic
oneway.test(score ~ group, data = arithmetic,
            var.equal = TRUE)
```

[6] Coupled with a slight loss of power if the variance are truly equal, more on this later.

```
    One-way analysis of means

data:  score and group
F = 15.268, num df = 4, denom df = 40, p-value = 1.163e-07
```

Notice how the degrees of freedom of the denominator have decreased. If we use `pairwise.t.test` with argument `pool.sd=FALSE`, this amounts to running Welch t-tests separately for each pair of variable.

What are the impacts of unequal variance if we use the F-test instead? For one, the pooled variance will be based on a weighted average of the variance in each group, where the weight is a function of the sample size. This can lead to size distortion (meaning that the proportion of type I error is not the nominal level α as claimed) and potential loss of power. The following toy example illustrates this.

We consider for simplicity a problem with $K = 2$ groups, which is the two-sample t-test. We simulated 50 observations from a $\mathsf{No}(0, 1)$ distribution and 10 observations from $\mathsf{No}(0, 9)$, comparing the distribution of the p-values for the Welch and the F-test statistics. Figure 3.9 shows the results. The percentage of p-values less than $\alpha = 0.05$ based on 10 000 replicates is estimated to be 4.76% for the Welch statistic, not far from the level. By contrast, we reject 28.95% of the time with the one-way ANOVA global F-test: this is a large share of innocents sentenced to jail based on false premises! While the size distortion is not always as striking, heterogeneity should be accounted in the design by requiring sufficient sample sizes (whenever costs permits) in each group to be able to estimate the variance reliably and using an adequate statistic.

There are alternative graphical ways of checking the assumption of equal variance, many including the standardized residuals $r_{ik} = (y_{ik} - \widehat{\mu}_k)/\widehat{\sigma}$ against the fitted values $\widehat{\mu}_k$. We will cover these in later sections.

Oftentimes, unequal variance occurs because the model is not additive. You could use variance-stabilizing transformations (e.g., log for multiplicative effects) to ensure approximately equal variance in each group. Another option is to use a model that is suitable for the type of response you have (including count and binary data). Lastly, it may be necessary to explicitly model the variance in more complex design (including repeated measures) where there is a learning effect over time and variability decreases as a result. Consult an expert if needed.

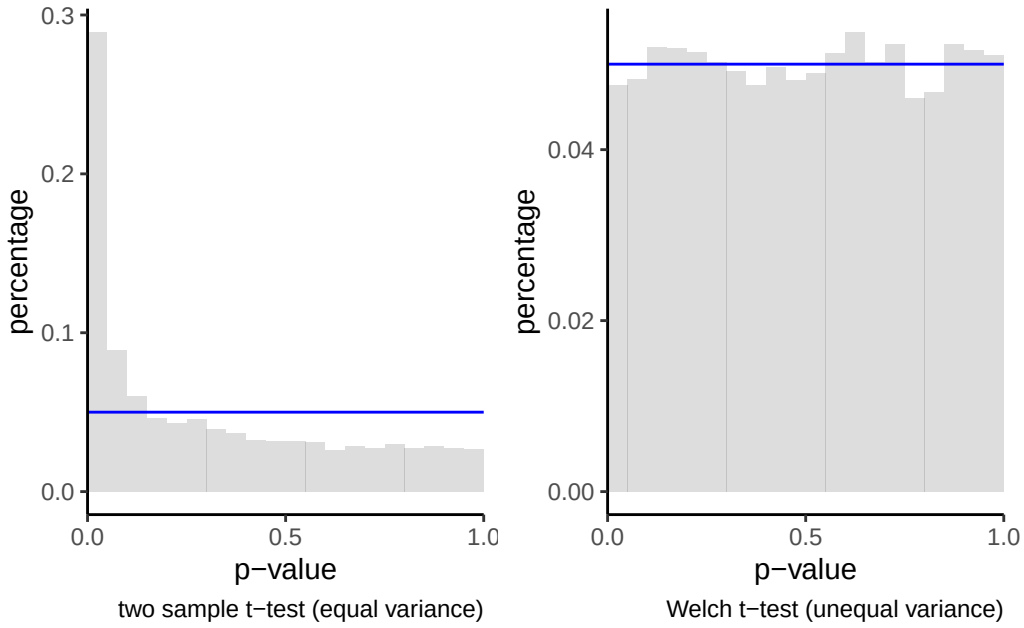


Figure 3.9: Histogram of the null distribution of p-values obtained through simulation using the classical analysis of variance F-test (left) and Welch's unequal variance alternative (right), based on 10 000 simulations. Each simulated sample consist of 50 observations from a $\mathsf{No}(0,1)$ distribution and 10 observations from $\mathsf{No}(0,9)$. The uniform distribution would have 5% in each of the 20 bins used for the display.

3.4.3 Normality

There is a persistent yet incorrect claim in the literature that the data (either response, explanatory or both) must be normal in order to use (so-called parametric) models like the one-way analysis of variance. With normal data and equal variances, the eponymous distributions of the F and t tests are exact: knowing the exact distribution does no harm and is convenient for mathematical derivations. However, it should be stressed that this condition is **unnecessary**: the results hold approximately for large samples by virtue of the central limit theorem. This probability results dictates that, under general conditions nearly universally met, the sample mean behaves like a normal distribution in large samples. This applet lets you explore the impact of the underlying population from which the data are drawn and the interplay with the sample size before the central limit theorem kicks in. You can view this in Figure 3.2, where the simulated and theoretical large-sample distributions are undistinguishable with approximately 20 observations per group.

While many authors may advocate rules of thumbs (sample size of $n > 20$ or $n > 30$ per group, say), these rules are arbitrary: the approximation is not much worst at $n = 19$ than at $n = 20$. How large must the sample size be for the approximation to hold? It largely depends on the distribution in the population: the more extremes, skewness, etc. you have, the larger the number of observation must be in order for the approximation to be valid. Figure 3.10 shows a skewed to the right bimodal distribution and the distribution of the sample mean under repeated sampling. Even with $n = 5$ observations (bottom left), the approximation is not bad but it may still be very far off with $n = 50$ for heavy-tailed data.

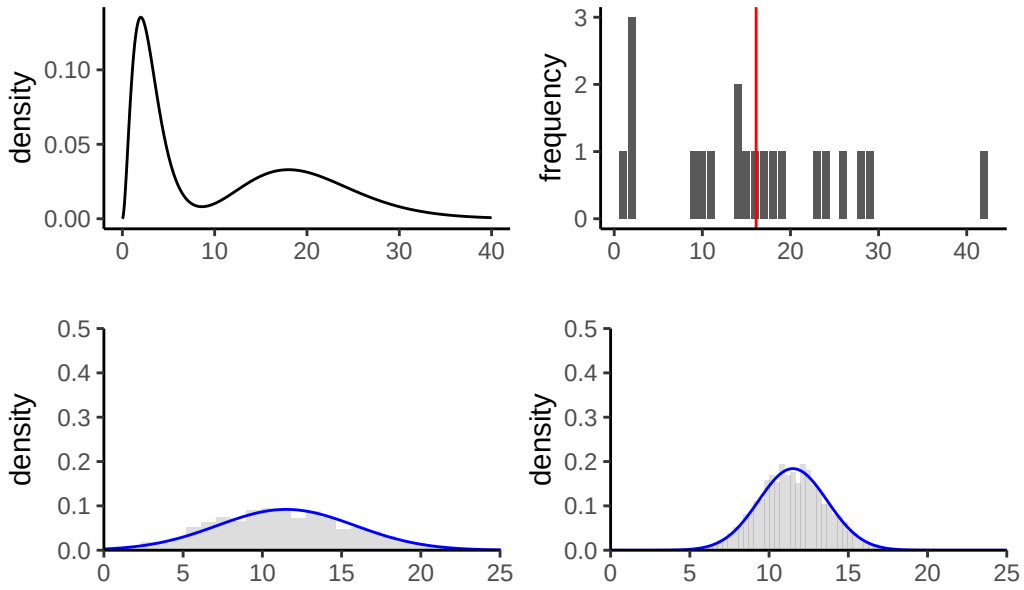


Figure 3.10: Graphical representation of the central limit theorem. Top left: density of the underlying population from which samples are drawn. Top right: a sample of 20 observations with its sample mean (vertical red). Bottom panels: histogram of sample averages for samples of size 5 (left) and 20 (right) with normal approximation superimposed. As the sample size increases, the normal approximation for the mean is more accurate and the standard error decreases.

It is important to keep in mind that all statistical statements are typically approximate and their reliability depends on the sample size: too small a sample may hampers the strength of your conclusions. The default graphic for checking whether a sample matches a postulated distribution is the quantile-quantile plot.

3.4.4 Independence

While I am not allowed to talk of independence as a Quebecer[7], this simply means that knowing the value of one observation tells us nothing about the value of any other in the sample. Independence may fail to hold in case of group structure (family dyads, cluster sampling) which have common characteristics or more simply in the case of repeated measurements. Random assignment to treatment is thus key to ensure that the measure holds, and ensuring at the measurement phase that there is no spillover.

There are many hidden ways in which measurements can impact the response. Physical devices that need to be calibrated before use (scales, microscope) require tuning: if measurements are done by different experimenters or on different days, it may impact and add systematic shift in means for the whole batch.

Special care must be taken whenever group testing is used, and blocking for potential impacts can salvage an analysis.

What is the impact of dependence between measurements? Heuristically, correlated measurements carry less information than independent ones. In the most extreme case, there is no additional information and measurements are identical. The reason why this makes a difference is the following: the denominator of the F-test is the sample variance, which is based on the within sum of squares divided by $n-K$. If each observation is counted 10 times, say, then the real number of measurements is n but the F statistic gets multiplied by a factor 10.[8]

The lack of independence can also have drastic consequences on inference and lead to false conclusions: Figure 3.11 shows an example with correlated samples within group (or equivalently repeated measurements from individuals) with 25 observations per group. The y-axis shows the proportion of times the null is rejected when it shouldn't be. Here, since the data are generated from the null model (equal mean) with equal variance, the inflation in the number of spurious discoveries, false alarm or type I error is alarming and the inflation is substantial even with very limited correlation between measurements.

[7] All credits for this pun are due to C. Genest

[8] The null distribution also changes with the sample size, but for n large the impact is less than that of the scaling since the $\mathsf{F}(\nu_1, \nu_2)$ distribution is approximately $\chi^2(\nu_1)$ when ν_2 is large.

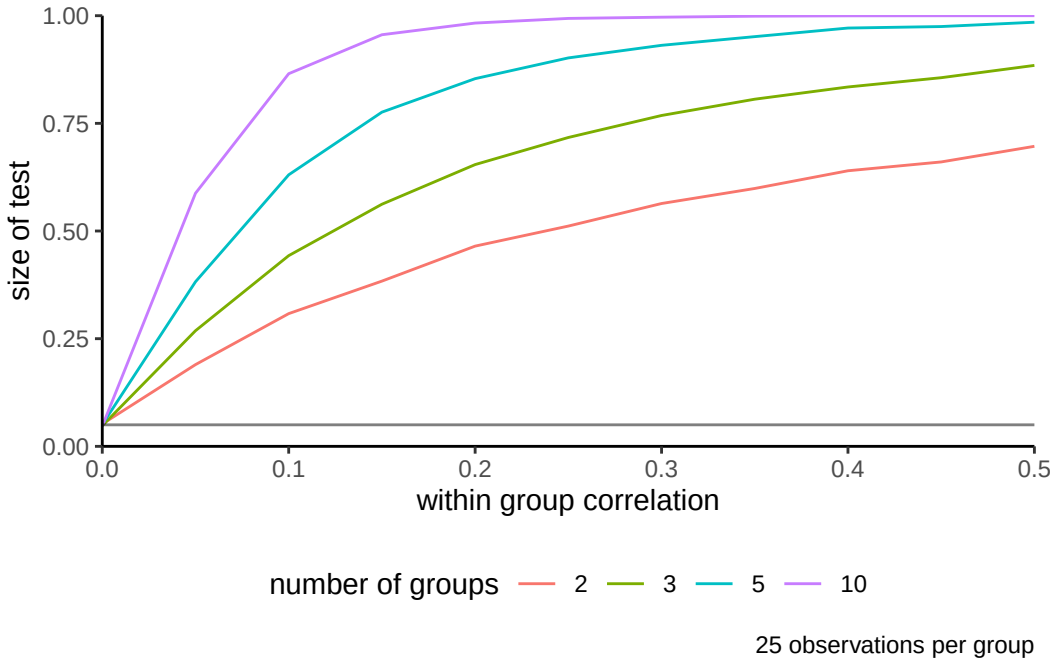


Figure 3.11: Percentage of rejection of the null hypothesis for the F-test of equality of means for the one way ANOVA with data generated with equal mean and variance from an equicorrelation model (within group observations are correlated, between group observations are independent). The nominal level of the test is 5%.

4 Contrasts and multiple testing

> ! Key concept
>
> **Learning objectives**:
>
> - Specifying contrast weights to test for differences between groups
> - Understanding the problem of multiple testing and the danger for selective reporting
> - Adjusting for multiple testing to control the global error rate.

4.1 Contrasts

Suppose we perform an analysis of variance and the F-test for the (global) null hypothesis that the averages of all groups are equal is very large: we reject the null hypothesis in favour of the alternative, which states that at least one of the group average is different. The follow-up question will be where these differences lie. Indeed, in an experimental context, this implies one or more of the manipulation has a different effect from the others on the mean response. Oftentimes, this isn't interesting in itself: we could be interested in comparing different options relative to a *status quo* (e.g., for new drugs or medical treatment), or determine whether specific combinations work better than separately, or find the best treatment by comparing all pairs.

The scientific question of interest that warranted the experiment may lead to a specific set of hypotheses, which can be formulated by researchers as comparisons between means of different subgroups. We can normally express these as **contrasts**. As Dr. Lukas Meier puts it, if the global F-test for equality of means is equivalent to a dimly lit room, contrasts are akin to spotlight that let one focus on particular aspects of differences in treatments.

Formally speaking, a contrast is a linear combination of averages: in plain English, this means we assign a weight to each group average and add them up, and then compare that summary to a postulated value a, typically zero. Contrasts encode research question of interest: if c_i denotes the weight of group average μ_i $(i = 1, \ldots, K)$, then we can write

the contrast as $C = c_1\mu_1 + \cdots + c_K\mu_K$ with the null hypothesis $\mathscr{H}_0 : C = a$ for a two-sided alternative. The sample estimate of the linear contrast is obtained by replacing the unknown population average μ_i by the sample average of that group, $\widehat{\mu}_i = \overline{y}_i$. We can easily obtain the standard error of the linear combination C.[1] We can then build a t statistic as usual by looking at the difference between our postulated value and the observed weighted mean, suitably standardized. If the global F-test leads to rejection of the null, there exists a contrast which is significant at the same level.

4.1.1 Orthogonal contrasts

Sometimes, linear contrasts encode disjoint bits of information about the sample: for example, one contrast that compares groups the first two groups versus one that compares the third and fourth is in effect using data from two disjoint samples, as contrasts are based on sample averages. Whenever the contrasts vectors are orthogonal, the tests will be uncorrelated. Mathematically, if we let c_i and c_i^* denote weights attached to the mean of group i comprising n_i observations, contrasts are orthogonal if $c_1c_1^*/n_1 + \cdots + c_Kc_K^*/n_K = 0$; if the sample is balanced with the same number of observations in each group, $n/K = n_1 = \cdots = n_K$, we can consider the dot product of the two contrast vectors and neglect the subsample sizes.

If we have K groups, there are $K-1$ contrasts for pairwise differences, the last one being captured by the sample mean for the overall effect[2]. If we care only about difference between groups (as opposed to the overall effect of all treatments), we impose a sum-to-zero constraint on the weights so $c_1 + \cdots + c_K = 0$. Keep in mind that, although independent tests are nice mathematically, contrasts should encode the hypothesis of interest to the researchers: we choose contrasts because they are meaningful, not because they are orthogonal.

The `arithmetic` data example considered five different treatment groups with 9 individuals in each. Two of them were control groups, one received praise, another was reproved and the last was ignored.

Suppose that researchers were interested in assessing whether the experimental manipulation had an effect, and whether the impact of positive and negative feedback is the same on

[1] Should you ever need the formula, the standard error assuming subsample size of $n_1, \ldots, n_K$ and a common variance σ^2 is $\sqrt{\mathsf{Va}(\widehat{C})}$, where

$$\mathsf{Va}(\widehat{C}) = \widehat{\sigma}^2\left(\frac{c_1^2}{n_1} + \cdots + \frac{c_K^2}{n_K}\right).$$

[2] The constraint $c_1 + \cdots + c_K = 0$ ensures that linear contrasts are orthogonal to the mean, which has weight $c_i = n_i/n$ and for balanced samples $c_i = 1/n$.

Table 4.1: Contrasts estimates for the arithmetic data

contrast	estimate	std. error	df	lower (CI)	upper (conf. limit)CI)
control vs manip	-3.33	1.05	40	-5.45	-1.22
praised vs reproved	4.00	1.62	40	0.72	7.28

students.[3]

Suppose we have five groups in the order (control 1, control 2, praised, reproved, ignored). We can express these hypothesis as

- $\mathscr{H}_{01}$: $\mu_{\text{praise}} = \mu_{\text{reproved}}$
- $\mathscr{H}_{02}$:

$$\frac{1}{2}(\mu_{\text{control}_1} + \mu_{\text{control}_2}) = \frac{1}{3}\mu_{\text{praised}} + \frac{1}{3}\mu_{\text{reproved}} + \frac{1}{3}\mu_{\text{ignored}}$$

Note that, for the hypothesis of control vs experimental manipulation, we look at average of the different groups associated with each item. Using the ordering, the weights of the contrast vector are $(1/2, 1/2, -1/3, -1/3, -1/3)$ and $(0, 0, 1, -1, 0)$. There are many equivalent formulation: we could multiply the weights by any number (different from zero) and we would get the same test statistic, as the latter is standardized.

```
library(emmeans)
data(arithmetic, package = "hecedsm")
linmod <- aov(score ~ group, data = arithmetic)
linmod_emm <- emmeans(linmod, specs = 'group')
contrast_specif <- list(
  controlvsmanip = c(0.5, 0.5, -1/3, -1/3, -1/3),
  praisedvsreproved = c(0, 0, 1, -1, 0)
)
contrasts_res <-
  contrast(object = linmod_emm,
                   method = contrast_specif)
# Obtain confidence intervals instead of p-values
confint(contrasts_res)
```

Example 4.1 (Teaching to read)**.** We consider data from Baumann, Seifert-Kessell, and Jones (1992). The abstract of the paper provides a brief description of the study

[3]These would be formulated *at registration time*, but for the sake of the argument we proceed as if they were.

Table 4.2: Estimated group averages with standard errors and 95% confidence intervals for post-test 1.

terms	marg. mean	std. err.	dof	lower (CI)	upper (CI)
DR	6.68	0.68	63	5.32	8.04
DRTA	9.77	0.68	63	8.41	11.13
TA	7.77	0.68	63	6.41	9.13

> This study investigated the effectiveness of explicit instruction in think aloud as a means to promote elementary students' comprehension monitoring abilities. Sixty-six fourth-grade students were randomly assigned to one of three experimental groups: (a) a Think-Aloud (TA) group, in which students were taught various comprehension monitoring strategies for reading stories (e.g., self-questioning, prediction, retelling, rereading) through the medium of thinking aloud; (b) a Directed reading-Thinking Activity (DRTA) group, in which students were taught a predict-verify strategy for reading and responding to stories; or (c) a Directed reading Activity (DRA) group, an instructed control, in which students engaged in a noninteractive, guided reading of stories.

Looking at Table 4.2, we can see that `DRTA` has the highest average, followed by `TA` and directed reading (`DR`).

```
library(emmeans) #load package
data(BSJ92, package = "hecedsm")
mod_post <- aov(posttest1 ~ group, data = BSJ92)
emmeans_post <- emmeans(object = mod_post,
                        specs = "group")
```

The purpose of Baumann, Seifert-Kessell, and Jones (1992) was to make a particular comparison between treatment groups. From the abstract:

> The primary quantitative analyses involved two planned orthogonal contrasts—effect of instruction (TA + DRTA vs. 2 x DRA) and intensity of instruction (TA vs. DRTA)—for three whole-sample dependent measures: (a) an error detection test, (b) a comprehension monitoring questionnaire, and (c) a modified cloze test.

The hypothesis of Baumann, Seifert-Kessell, and Jones (1992) is $\mathscr{H}_0 : \mu_{\text{TA}} + \mu_{\text{DRTA}} = 2\mu_{\text{DRA}}$ or, rewritten slightly,

$$\mathscr{H}_0 : -2\mu_{\text{DR}} + \mu_{\text{DRTA}} + \mu_{\text{TA}} = 0.$$

Table 4.3: Estimated contrasts for post-test 1.

contrast	estimate	std. err.	dof	stat	p-value
C1: DRTA+TA vs 2DR	4.18	1.67	63	2.51	0.01
C1: average (DRTA+TA) vs DR	2.09	0.83	63	2.51	0.01
C2: DRTA vs TA	2.00	0.96	63	2.08	0.04
C2: TA vs DRTA	-2.00	0.96	63	-2.08	0.04

with weights $(-2, 1, 1)$; the order of the levels for the treatment are (DRA, DRTA, TA) and it must match that of the coefficients. An equivalent formulation is $(2, -1, -1)$ or $(1, -1/2, -1/2)$: in either case, the estimated differences will be different (up to a constant multiple or a sign change). The vector of weights for $\mathscr{H}_0 : \mu_{\text{TA}} = \mu_{\text{DRTA}}$ is (0, −1, 1): the zero appears because the first component, DRA doesn't appear. The two contrasts are orthogonal since $(-2 \times 0) + (1 \times -1) + (1 \times 1) = 0$.

```
# Identify the order of the level of the variables
with(BSJ92, levels(group))
```

```
[1] "DR"   "DRTA" "TA"
```

```
# DR, DRTA, TA (alphabetical)
contrasts_list <- list(
  "C1: DRTA+TA vs 2DR" = c(-2, 1, 1),
  # Contrasts: linear combination of means, coefficients sum to zero
  # 2xDR = DRTA + TA => -2*DR + 1*DRTA + 1*TA = 0 and -2+1+1 = 0
  "C1: average (DRTA+TA) vs DR" = c(-1, 0.5, 0.5),
  #same thing, but halved so in terms of average
  "C2: DRTA vs TA" = c(0, 1, -1),
  "C2: TA vs DRTA" = c(0, -1, 1)
  # same, but sign flipped
)
contrasts_post <-
  contrast(object = emmeans_post,
           method = contrasts_list)
contrasts_summary_post <- summary(contrasts_post)
```

We can look at these differences; since `DRTA` versus `TA` is a pairwise difference, we could have obtained the t-statistic directly from the pairwise contrasts using `pairs(emmeans_post)`.

Note that the two different ways of writing the comparison between DR and the average of the other two methods yield different point estimates, but same inference (i.e., the same p-values). For contrast C_{1b}, we get half the estimate (but the standard error is also halved) and likewise for the second contrasts we get an estimate of $\mu_{\text{DRTA}} - \mu_{\text{TA}}$ in the first case (C_2) and $\mu_{\text{TA}} - \mu_{\text{DRTA}}$: the difference in group averages is the same up to sign.

What is the conclusion of our analysis of contrasts? It looks like the methods involving teaching aloud have a strong impact on reading comprehension relative to only directed reading. The evidence is not as strong when we compare the method that combines directed reading-thinking activity and thinking aloud.

Example 4.2 (Paper or plastic)**.** Sokolova, Krishna, and Döring (2023) consider consumer bias when assessing how eco-friendly packages are. Items such as cereal are packaged in plastic bags, which themselves are covered in a box. They conjecture (and find) that consumers tend to view the packaging as being more eco-friendly when the amount of cardboard or paper surrounding the box is large, relative to the sole plastic package. We consider the data Study 2A, which measures the perceived environmental friendliness (PEF) as a function of the proportion of paper wrapping (either none, half of the area of the plastic, equal or twice). The authors are interested in comparing none with other choices.

If $\mu_0, \mu_{0.5}, \mu_1, \mu_2$ denote the true mean of the PEF score as a function of the proportion of paper, we are interested in pairwise differences, but only relative to the reference μ_0:

$$\begin{aligned}
\mu_0 = \mu_{0.5} &\iff 1\mu_0 - 1\mu_{0.5} + 0\mu_1 + 0\mu_2 = 0\\
\mu_0 = \mu_1 &\iff 1\mu_0 + 0\mu_{0.5} - 1\mu_1 + 0\mu_2 = 0\\
\mu_0 = \mu_2 &\iff 1\mu_0 + 0\mu_{0.5} + 0\mu_1 - 1\mu_2 = 0
\end{aligned}$$

so contrast vectors $(1,-1,0,0)$, $(1,0,-1,0)$ and $(1,0,0,-1)$ would allow one to test the hypothesis.

```
data(SKD23_S2A, package = "hecedsm") # load data
linmod <- lm(pef ~ proportion, data = SKD23_S2A) # fit simple linear regression
anova(linmod) # check for significance of slope
coef(linmod) # extract intercept and slope
anovamod <- lm(pef ~ factor(proportion), data = SKD23_S2A) # one-way ANOVA
margmean <- anovamod |>  emmeans::emmeans(specs = "proportion") # group means
contrastlist <- list( # specify contrast vectors
   refvshalf = c(1, -1, 0, 0),
   refvsone =  c(1, 0, -1, 0),
   refvstwo =  c(1, 0, 0, -1))
# compute contrasts relative to reference
margmean |> emmeans::contrast(method = contrastlist)
```

Table 4.4: Estimated group averages of PEF per proportion with standard errors

proportion	marg. mean	std. err.	dof	lower (CI)	upper (CI)
0.0	2.16	0.093	798	1.98	2.34
0.5	2.91	0.093	798	2.73	3.09
1.0	3.06	0.092	798	2.88	3.24
2.0	3.34	0.089	798	3.17	3.52

Table 4.5: Estimated contrasts for differences of PEF to no paper.

contrast	estimate	std. err.	dof	stat	p-value
refvshalf	-0.75	0.13	798	-5.71	0
refvsone	-0.90	0.13	798	-6.89	0
refvstwo	-1.18	0.13	798	-9.20	0

The group averages are reported in Table 4.4, match those reported by the authors in the paper. They suggest an increased perceived environmental friendliness as the amount of paper used in the wrapping increases. We could fit a simple regression model to assess the average change, treating the proportion as a continuous explanatory variable. The estimated slope for the change in PEF score, which ranges from 1 to 7 in increments of 0.25, is 0.53 per area of paper. There is however strong evidence, given the data, that the change isn't quite linear, as the fit of the linear regression model is significantly worse than the corresponding linear model.

All differences reported in Table 4.5 are significant and positive, in line with the researcher's hypothesis.

4.2 Multiple testing

Beyond looking at the global null, we will be interested in a set of contrast statistics and typically this number can be large-ish. There is however a catch in starting to test multiple hypothesis at once.

If you do a **single** hypothesis test and the testing procedure is well calibrated (meaning that the model model assumptions hold), p-values are generated uniformly on the interval $[0, 1]$ and there is a probability of α of making a type I error (i.e., concluding in favour of the alternative and rejecting the null incorrectly) if the null is true. The problem of the above approach is that the more tests you perform, the higher the chance of finding (incorrectly)

something: with 20 independent tests, we expect that, on average, one of them will yield a p-value less than 5% even if this is a fluke. The problem with multiple testing is not so much that it occurs, but more than researchers tend to report selectively findings and only give the results of tests for which $p \leq \alpha$, even if these are typically the product of chance. This makes most findings will not replicate: if we rerun the experiment, we will typically not find the same result.

There is an infinite potential number of contrasts with more than two factos. Not all tests are of interest: standard software will report all possible pairwise comparisons, but this may not be of interest as showcased in Example 4.3. If there are K groups to compare and any comparison is of interest, than we could performs $\binom{K}{2}$ pairwise comparisons with $\mathscr{H}_0 : \mu_i = \mu_j$ for $i \neq j$. For $K = 3$, there are three such comparisons, 10 pairwise comparisons if $K = 5$ and 45 pairwise comparisons if $K = 10$. The number of pairwise comparisons grows quickly.

The number of tests performed in the course of an analysis can be very large. Y. Benjamini investigated the number of tests performed in each study of the Psychology replication project (Nosek et al. 2015): this number ranged from 4 to 700, with an average of 72 — most studies did not account for the fact they were performing multiple tests or selected the model and thus some 'discoveries' are bound to be spurious. It is natural to ask then how many results are spurious findings that correspond to type I errors. The paramount (absurd) illustration is the cartoon presented in Figure 4.1: note how there is little scientific backing for the theory (thus such test shouldn't be of interest to begin with) and likewise the selective reporting made of the conclusions, despite nuanced conclusions.

We can also assess mathematically the problem. Assume for simplicity that all tests are independent[4], then the probability of any rejecting the null incorrectly is α, but larger over the collection (with tests A and B, we could reject by mistake if A is a type I error and B isn't, or vice-versa, or if both are incorrect rejections.

The probability of making at least one type I error if each test is conducted at level α, say $\alpha^\star$, is[5]

$$\begin{aligned} \alpha^\star &= 1 - \text{probability of making no type I error} && (4.1)\\ &= 1 - (1-\alpha)^m && (4.2)\\ &\leq m\alpha && (4.3) \end{aligned}$$

[4]This is the case if tests are based on different data, or if the contrasts considered are orthogonal under normality.

[5]The second line holds with independent observations, the second follows from the use of Boole's inequality and does not require independent tests.

With $\alpha = 5\%$ and $m = 4$ tests, $\alpha^\star \approx 0.185$ whereas for $m = 72$ tests, $\alpha^\star \approx 0.975$: this means we are almost guaranteed even when nothing is going on to find "statistically significant" yet meaningless results.

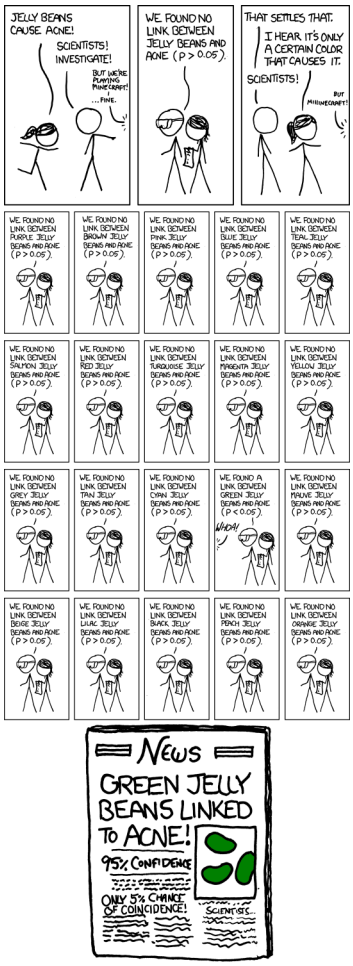


Figure 4.1: xkcd 882: Significant. The alt text is 'So, uh, we did the green study again and got no link. It was probably a–' 'RESEARCH CONFLICTED ON GREEN JELLY BEAN/ACNE LINK; MORE STUDY RECOMMENDED!'

It is sensible to try and reduce or bound the number of false positive or control the probability of getting spurious findings. We consider a **family** of m null hypothesis $\mathscr{H}_{01}, \ldots, \mathscr{H}_{0m}$, i.e. a collection of m hypothesis tests. The exact set depends on the context, but this comprises all hypothesis that are scientifically relevant and could be reported. These comparisons are called **pre-planned comparisons**: they should be chosen before the experiment takes place and pre-registered to avoid data dredging and selective reporting. The

number of planned comparisons should be kept small relative to the number of parameters: for a one-way ANOVA, a general rule of thumb is to make no more comparisons than the number of groups, K.

Suppose that we perform m hypothesis tests in a study and define binary indicators

$$R_i = \begin{cases} 1 & \text{if we reject the null hypothesis } \mathscr{H}_{0i} \\ 0 & \text{if we fail to reject } \mathscr{H}_{0i} \end{cases} \tag{4.4}$$

$$V_i = \begin{cases} 1 & \text{type I error for } \mathscr{H}_{0i} \quad (R_i = 1 \text{ and } \mathscr{H}_{0i} \text{ is true}) \\ 0 & \text{otherwise.} \end{cases} \tag{4.5}$$

With this notation, $R = R_1 + \cdots + R_m$ simply encodes the total number of rejections ($0 \leq R \leq m$), and $V = V_1 + \cdots + V_m$ is the number of null hypothesis rejected by mistake ($0 \leq V \leq R$).

The **familywise error rate** is the probability of making at least one type I error for the whole collection or test, in other words per family, is

$$\mathsf{FWER} = \Pr(V \geq 1).$$

To control the familywise error rate, one must be more stringent in rejecting the null and perform each test with a smaller level α so that the overall or simultaneous probability is less than FWER.

4.2.1 Bonferroni's procedure

The easiest way to control for multiple testing is to perform each test at level α/m, thereby ensuring that the family-wise error is controlled at level α. This is a good option if m is small and the Bonferroni adjustment also controls the **per-family error rate**, which is the expected (theoretical average) number of false positive $\mathsf{PFER} = \mathsf{E}(V)$. The latter is a more stringent criterion than the familywise error rate because $\Pr(V \geq 1) \leq \mathsf{E}(V)$: the familywise error rate does not make a distinction between having one or multiple type I errors.[6]

Why is Bonferroni's procedure popular? It is conceptually easy to understand and simple, and it applies to any design and regardless of the dependence between the tests. However, the number of tests to adjust for, m, must be prespecified and the procedure leads to low power when the size of the family is large, meaning it makes detection of non-null effects more difficult. Moreover, if our sole objective is to control for the familywise error rate,

[6] By definition, the expected number of false positive (PFER) is $\mathsf{E}(V) = \sum_{i=1}^{m} i \Pr(V = i) \geq \sum_{i=1}^{m} \Pr(V = i) = \Pr(V \geq 1)$, so larger than the probability of making at least type 1 error. Thus, any procedure that controls the per-family error rate (e.g., Bonferroni) also automatically bounds the familywise error rate.

then there are other procedures that are always better in the sense that they still control the FWER while leading to increased capacity of detection when the null is false.

If the raw (i.e., unadjusted) p-values are reported, we reject hypothesis $\mathscr{H}_{0i}$ if $m \times p_i \geq \alpha$: operationally, we multiply each p-value by m and reject if the result exceeds α.

4.2.2 Holm–Bonferroni's procedure

The idea of Holm's procedure is to use a sharper inequality bound and amounts to performing tests at different levels, with more stringent for smaller p-values. To perform Holm–Bonferroni,

1. order the p-values of the family of m tests from smallest to largest, $p_{(1)} \leq \cdots \leq p_{(m)}$
2. test sequentially the hypotheses: coupling Holm's method with Bonferroni's procedure, we compare $p_{(1)}$ to $\alpha_{(1)} = \alpha/m$, $p_{(2)}$ to $\alpha_{(2)} = \alpha/(m-1)$, etc. If $p_{(j)} \geq \alpha_{(j)}$ but $p_{(i)} \leq \alpha_{(i)}$ for $i = 1, \ldots, j-1$ (all smaller p-values), we reject the associated hypothesis $\mathscr{H}_{0(1)}, \ldots, \mathscr{H}_{0(j-1)}$ but fail to reject $\mathscr{H}_{0(j)}, \ldots, \mathscr{H}_{0(m)}$.

If all of the p-values are less than their respective levels, than we still reject each null hypothesis. Otherwise, we reject all the tests whose p-values exceeds the smallest nonsignificant one. This procedure doesn't control the per-family error rate, but is uniformly more powerful (lingo to say that it's universally better for control) and thus leads to increased detection than Bonferroni's method. To see this, consider a family of $m = 3$ p-values with values 0.01, 0.04 and 0.02. Bonferroni's adjustment would lead us to reject the second and third hypotheses at level $\alpha = 0.05$, but not Holm-Bonferroni.

4.2.3 Multiple testing methods for analysis of variance

There are specialized procedures for the analysis of variance problem that leverages some of the assumptions (equal variance, large sample approximation for the distribution of means). There are three scenarios

1. Dunnett's method for comparison to a reference or control group, controlling only for $K-1$ pairwise differences
2. Tukey's range procedure, also termed *honestly significant difference* (HSD), for **all** pairwise differences. We can obtain control on the type I error by looking at what happens between the minimum and maximum group averages under the null.
3. Scheffe's method for contrasts. This is useful when the number of contrasts of interest is not specified apriori.

If the global F-test does not find differences at level α, then Scheffe's method will also find no significant contrast α but nothing can be said about other methods. Generally, the more tests we control the type error for, the more conservative the procedures are.

In **R**, we can use the `multcomp` or `emmeans` packages for the tests to adjust, or compute results manually. The test statistics do not change, only the benchmark null distribution is different. Figure 4.2 shows what the p-value would be depending on how we control for contrasts. For reasonable values, we get larger p-values for the methods that provide control.

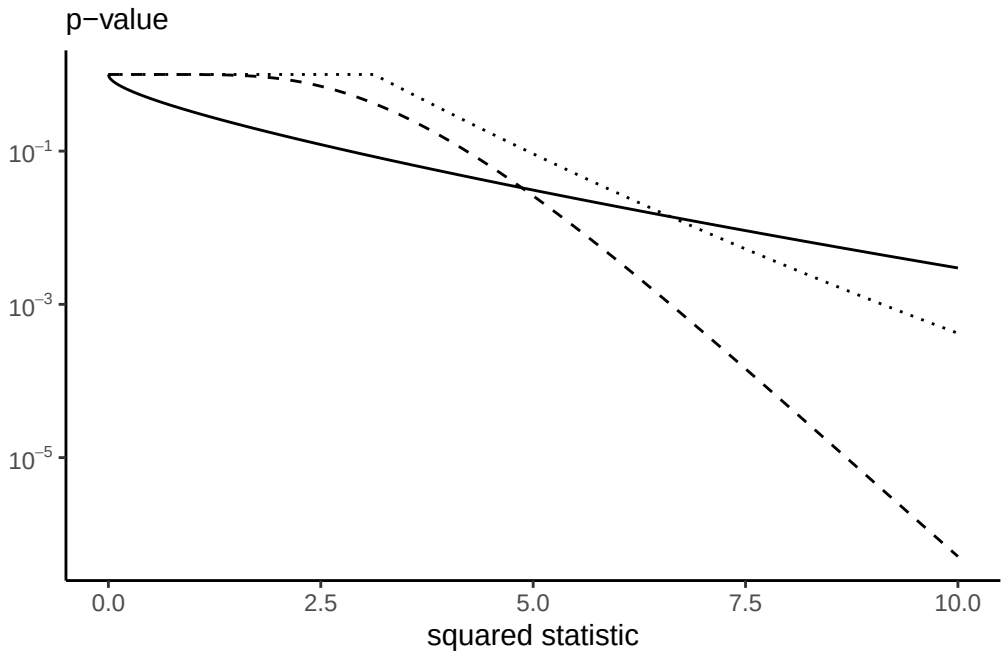


Figure 4.2: P-value as a function of the squared t-statistic for a contrast for no adjustment (full line), Tukey's HSD (dashed line) and Scheffe's adjustment (dotted).

Example 4.3 (Multiple testing for paper or plastic)**.** Sokolova, Krishna, and Döring (2023) considered pairwise difference relative to the control where only plastic wrapping is used. We could use either Bonferroni, Holm–Bonferoni or Dunnett's method. Since the p-values are tiny (less than 10^{-4}), this has no impact on the conclusions whatsoever. To better appreciate the impact in small samples, we subsample 20 observation per group to inflate p-values. We can also see differences by inspecting the width of the confidence intervals for the pairwise differences to the reference group: more conservative references lead to wider intervals.

```r
data(SKD23_S2A, package = "hecedsm") # load data
set.seed(80667) # Set seed for reproducibility
SKD23_S2A_sub <- SKD23_S2A |>
  # Create a categorical variable (factor) and ensure reference is 0
  # By default, it would be (first alphanumerical value of labels)
  dplyr::mutate(propfact = relevel(factor(proportion), ref = "0")) |>
  # Sample only fourty observations by group -
  # for illustration purposes only, otherwise p-values are too small
  dplyr::slice_sample(n = 20, by = propfact)
anovamod <- lm(pef ~ propfact, data = SKD23_S2A_sub)
library(emmeans)
margmean <- emmeans(
  anovamod, # fitted model
  # 'specs': vector with names of factors to adjust for
  specs = "propfact")
contrastlist <- list( # specify contrast vectors
  refvshalf = c(1, -1, 0, 0),
  refvsone =  c(1, 0, -1, 0),
  refvstwo =  c(1, 0, 0, -1))
contrasts <- margmean |> contrast(method = contrastlist)
# Bonferroni and Holm-Bonferroni adjustments
summary(contrasts, adjust = "bonferroni")
```

```
 contrast  estimate    SE df t.ratio p.value
 refvshalf   -0.450 0.331 76  -1.359  0.5346
 refvsone    -1.150 0.331 76  -3.473  0.0026
 refvstwo    -0.662 0.331 76  -2.001  0.1470

P value adjustment: bonferroni method for 3 tests
```

```r
summary(contrasts, adjust = "holm")
```

```
 contrast  estimate    SE df t.ratio p.value
 refvshalf   -0.450 0.331 76  -1.359  0.1782
 refvsone    -1.150 0.331 76  -3.473  0.0026
 refvstwo    -0.662 0.331 76  -2.001  0.0980

P value adjustment: holm method for 3 tests
```

```
# Note that the p-values for the latter are equal or smaller

# Adjustments for ANOVA to get simultaneous statements
# Number of groups minus 1 for Scheffe (correct here)
# This 'rank' often needs to be manually specified in multi-way ANOVA
summary(contrasts, adjust = "scheffe", scheffe.rank = 3)
```

```
 contrast  estimate    SE df t.ratio p.value
 refvshalf   -0.450 0.331 76  -1.359  0.6070
 refvsone    -1.150 0.331 76  -3.473  0.0104
 refvstwo    -0.662 0.331 76  -2.001  0.2696

P value adjustment: scheffe method with rank 3
```

```
# This would be the better option here
summary(contrasts, adjust = "dunnett")
```

```
 contrast  estimate    SE df t.ratio p.value
 refvshalf   -0.450 0.331 76  -1.359  0.3934
 refvsone    -1.150 0.331 76  -3.473  0.0025
 refvstwo    -0.662 0.331 76  -2.001  0.1260

P value adjustment: dunnettx method for 3 tests
```

```
# The less you adjust for, the smaller the p-values
# For Tukey, use 'contrast(method = "pairwise")' instead

# Since we have a small number of pairwise comparisons
# We could use the less stringent of Holm-Bonferroni and Dunnett's
# The latter provides shorter intervals here.
contrasts |> confint(adjust = "dunnett")
```

```
 contrast  estimate    SE df lower.CL upper.CL
 refvshalf   -0.450 0.331 76    -1.25    0.348
 refvsone    -1.150 0.331 76    -1.95   -0.352
 refvstwo    -0.662 0.331 76    -1.46    0.135
```

```
Confidence level used: 0.95
Conf-level adjustment: dunnettx method for 3 estimates
```

```
contrasts |> confint(adjust = "holm")
```

```
 contrast  estimate    SE df lower.CL upper.CL
 refvshalf   -0.450 0.331 76    -1.26    0.361
 refvsone    -1.150 0.331 76    -1.96   -0.339
 refvstwo    -0.662 0.331 76    -1.47    0.148

Confidence level used: 0.95
Conf-level adjustment: bonferroni method for 3 estimates
```

We can see that more stringent adjustments lead to higher p-values and wider intervals.

If we wanted to perform tests for multiple variables, or for subgroups, we can obtain overall control by using a procedure in each subset with a lower α, and combining the overall errors afterwards. If the data arise from different independent samples, the tests are indeed independent.

5 Complete factorial designs

We next consider experiments and designs in which there are multiple factors being manipulated by the experimenter simultaneously. Before jumping into the statistical analysis, let us discuss briefly some examples that will be covered in the sequel.

Supplemental Study 5 from Sharma, Tully, and Cryder (2021) checks the psychological perception of borrowing money depending on the label. The authors conducted a 2 by 2 between-subject comparison (two-way ANOVA) varying the type of debt (whether the money was advertised as credit or loan) and the type of purchase the latter would be used for (discretionary spending or need). The response is the average of the likelihood and interest in the product, both measured using a 9 point Likert scale from 1 to 9.

Maglio and Polman (2014) measured the subjective distance on travel based on the direction of travel. They conducted an experiment in the Toronto subway green line, asking commuters from Bay station to answer the question "How far away does the [name] station feel to you?" using a 7 point Likert scale ranging from very close (1) to very far (7). The stations name were one of Spadina, St. George, Bloor–Yonge and Sherbourne (from West to East).

As there are four stations and two directions of travel (a 4 by 2 design), the scientific question of interest for subjective measures of distance would consist of perceiving differently the distance depending on the direction of travel. We could also wonder whether destinations that are two stations away from Bay (Spadina and Sherbourne) would be considered equidistant, and similarly for the other two.

5.1 Efficiency of multiway analysis of variance.

Consider the setting of Sharma, Tully, and Cryder (2021) and suppose we want to check the impact of debt and collect a certain number of observations in each group. If we suspected the label had an influence, we could run a one-way analysis of variance for each spending type separately (thus, two one-way ANOVA each with two groups). We could do likewise if we wanted instead to focus on whether the spending was discretionary in nature or not, for each label: together, this would give a total of eight sets of observations. Combining the two factors allows us to halve the number of groups/samples we collect in this simple setting:

this highlights the efficiency of running an experiment modifying all of these instances at once, over a series of one-way analysis of variance. This concept extends to higher dimension when we manipulate two or more factors. Factorial designs allow us to study the impact of multiple variables simultaneously with **fewer overall observations**.

The drawback is that as we increase the number of factors, the total number of subgroups increases: with a complete design[1] and with factors A, B, C, etc. with n_a, n_b, n_c, ... levels, we have a total of $n_a \times n_b \times n_c \times \cdots$ combinations and the number of observations needed to efficiently measure the group means increases quickly. This is the **curse of dimensionality**: the larger the number of experimental treatments manipulated together, the larger the sample size needed. A more efficient approach, which we will cover in later section, relies on measuring multiple observations from the same experimental units, for example by giving multiple tasks (randomly ordered) to participants.

Intrinsically, the multiway factorial design model description does not change relative to a one-way design: the analysis of variance describes the sample mean for the response in each subgroup,

Consider a two-way analysis of variance model. This is a linear model with two factors, A and B, with respectively n_a and n_b levels. The response Y_{ijk} of the kth measurement in group (a_i, b_j) is

$$\underset{\text{response}}{Y_{ijk}} = \underset{\text{subgroup mean}}{\mu_{ij}} + \underset{\text{error term}}{\varepsilon_{ijk}} \tag{5.1}$$

where

- Y_{ijk} is the kth replicate for ith level of factor A and jth level of factor B
- μ_{ij} is the average response of measurements in group (a_i, b_j)
- ε_{ijk} are independent error terms with mean zero and standard deviation σ.

This, it turns out, is a special case of linear regression model. We could build contrasts for comparing group averages, but it will more convenient to reparametrize the model so that hypotheses of interest are directly expressed in terms of the parameters.

For example, in the Maglio and Polman (2014) study, we could gather observations for each factor combination in a table, where `direction` is the row and `station` the column.

[1] By complete design, it is meant that we gather observations for each subcategory.

Table 5.1: Conceptual depiction of cell average for the two by two design of Maglio and Polman (2014)

A `station`	B `direction`	b_1 (`east`)	b_2 (`west`)	row mean
a_1 (`Spadina`)		μ_{11}	μ_{12}	$\mu_{1.}$
a_2 (`St. George`)		μ_{21}	μ_{22}	$\mu_{2.}$
a_3 (`Bloor-Yonge`)		μ_{31}	μ_{32}	$\mu_{3.}$
a_4 (`Sherbourne`)		μ_{41}	μ_{42}	$\mu_{4.}$
column mean		$\mu_{.1}$	$\mu_{.2}$	μ

The ith row mean represents the average response across all levels of B, $\mu_{i.} = (\mu_{i1} + \cdots + \mu_{in_b})/n_b$ and similarly for the average of the jth column, $\mu_{.j} = (\mu_{1j} + \cdots + \mu_{n_aj})/n_a$. Finally, the overall average is

$$\mu = \frac{\sum_{i=1}^{n_a}\sum_{j=1}^{n_b}\mu_{ij}}{n_a n_b}.$$

Each subgroup average μ_{ij} will be estimated as the sample mean of observations in their group and we would use the above formulae to obtain estimates of the row, column and overall means $\widehat{\mu}_{i.}$, $\widehat{\mu}_{.j}$ and $\widehat{\mu}$. If the sample is balanced, meaning the number of observations is the same, these will be the same as summing over all observations in a row, column or table and then averaging. In general setup, however, we will give equal weight to each subgroup average.

5.2 Interactions

Table Table 5.1 shows the individual mean of each subgroup. From these, we may be interested in looking at the experiment as a single one-way analysis of variance model with eight subgroups, or as a series of one-way analysis of variance with either `direction` or `station` as sole factor.

We will use particular terminology to refer to these:

- **simple effects**: difference between levels of one in a fixed combination of others. Simple effects are comparing cell averages within a given row or column.
- **main effects**: differences relative to average for each condition of a factor. Main effects are row/column averages.
- **interaction effects**: when simple effects differ depending on levels of another factor. Interactions effects are difference relative to the row or column average.

In other words, an interaction occurs when some experimental factors, when coupled together, have different impacts than the superposition of each. An interaction between two factors occurs when the average effect of one independent variable depends on the level of the other.

If there is a significant interaction, the main effects are **not** of interest since they are misleading. Rather, we will compute the simple effects by making the comparison one at level at the time.

In our example of Maglio and Polman (2014), a simple effect would be comparing the distance between Spadina and Sherbourne for east. The main effect for the direction would be the average perceived distance for east and for west. Finally, the interaction would measure how much these differ by station depending on direction.

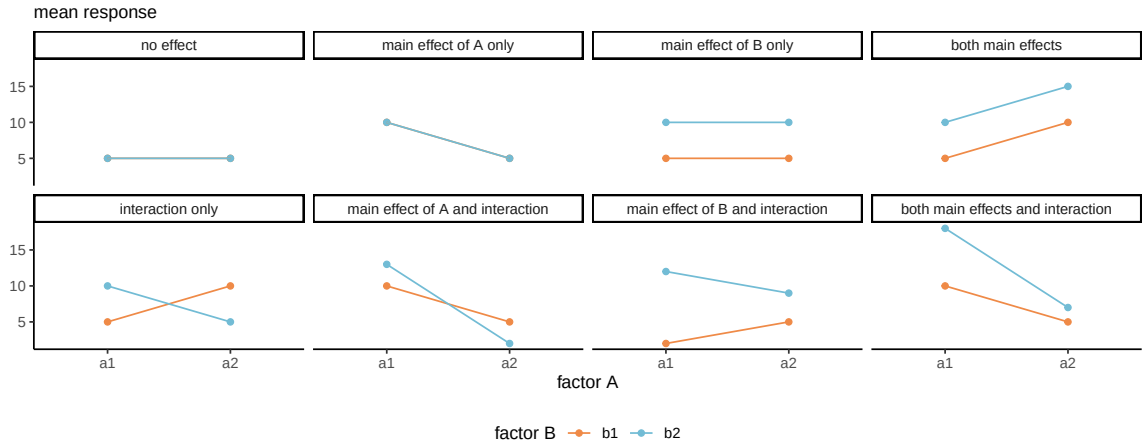


Figure 5.1: Interaction plots (line graphs) for example patterns for means for each of the possible kinds of general outcomes in a 2 by 2 design. Illustration adapted from Figure 10.2 of Crump, Navarro, and Suzuki (2019) by Matthew Crump (CC BY-SA 4.0 license).

To better understand, we consider the average response and suppose we have access to the true population average for each sub-treatment. We can then represent the population using a line graph with the two factors, one being mapped to color and another to the x-axis. Figure 5.1 shows what happens under all possible scenarios with a 2 by 2 design. When there is no overall effect, the mean is constant. If there isn't a main effect of A, the average of the two mean response for a_1 and a_2 are the same, etc. Interactions are depicted by **non-parallel lines**.

It's clear from Figure 5.1 that looking only at the average of A alone (the main effect) isn't instructive when we are in the presence of an interaction: rather, we should be comparing

the values of A for b_1 separately than those for b_2, and vice-versa using simple effects, otherwise our conclusions may be misleading.

The hypothesis of interest is the interaction; for the time being, we can simply plot the average per group. Since the summary statistics can hide important information such as the uncertainty, we add 95% confidence intervals for the subgroup averages and superimpose jittered observations to show the spread of the data. Based on Figure 5.2, there appears to be at least an interaction between station and direction of travel, in addition to a main effect for station. Formal hypothesis testing can help check this intuition.

```
`summarise()` has grouped output by 'direction'. You can override using the
`.groups` argument.
```

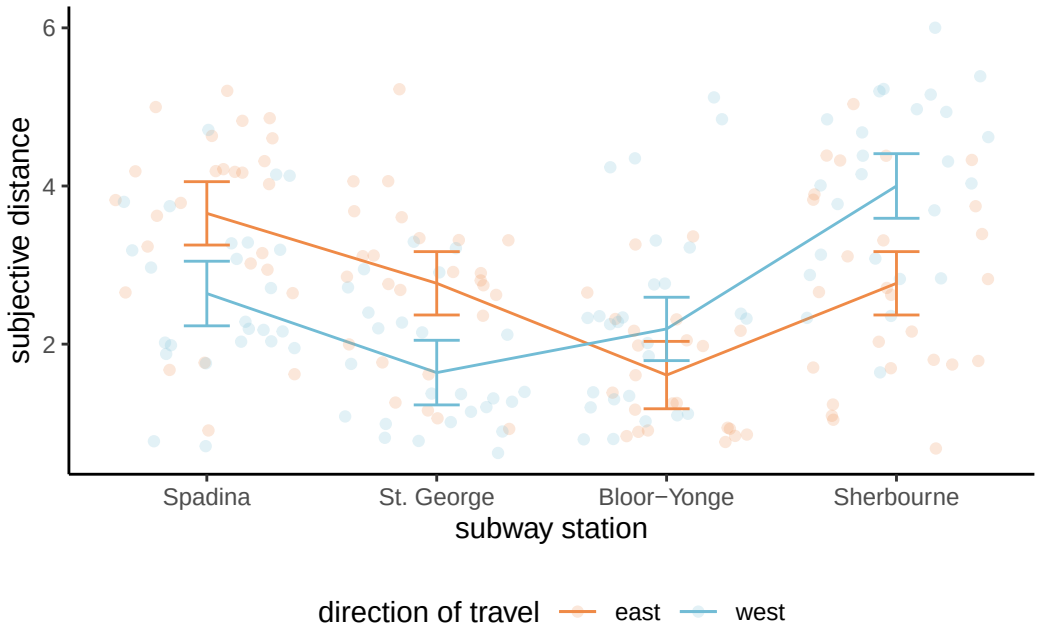


Figure 5.2: Interaction plot for Study 1 of Maglio and Polman (2014), showing group averages and 95% confidence intervals for the means. Observations are overlaid on the graph.

5.3 Model parametrization

The model parametrized in terms of subgroup or cell average is okay in Equation 5.1, but it doesn't help us if we want to check for the presence of main effects and interaction, even if it would be possible to specify the contrasts required to test these hypotheses. We can however express the model in terms of main effects and interactions.

We consider the alternative formulation

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk},$$

where

- μ is the average of all subgroup averages, termed overall mean.
- $\alpha_i = \mu_{i.} - \mu$ is the mean of level A_i minus the overall mean.
- $\beta_j = \mu_{.j} - \mu$ is the mean of level B_j minus overall mean.
- $(\alpha\beta)_{ij} = \mu_{ij} - \mu_{i.} - \mu_{.j} + \mu$ is the interaction term for A_i and B_j which encodes the effect of both variable not already captured by the main effects.

A rapid calculation shows that there are more coefficients than the number of cells and subgroups ($n_a n_b$ cells overall) in our table. The model is **overparametrized**: to get away with this, we impose constraints to remove redundancies. The idea is that if we know $n_a - 1$ of the mean for factor A and the global average is a combination of these, we can deduce the value for the last row mean. The model formulation in terms of difference from the global average or main effect ensures that we can test for main effects for factor A by setting $\mathscr{H}_0 : \alpha_1 = \cdots = \alpha_{n_a-1} = 0$. The $1 + n_a + n_b$ **sum to zero** constraints,

$$\sum_{i=1}^{n_a} \alpha_i = 0, \quad \sum_{j=1}^{n_b} \beta_j = 0, \quad \sum_{j=1}^{n_b} (\alpha\beta)_{ij} = 0, \quad \sum_{i=1}^{n_a} (\alpha\beta)_{ij} = 0,$$

restore identifiability.

The redundancy in information, due to the fact main effects are expressible as row and column averages, and the overall mean as the average of all observations, will arise again when we consider degrees of freedom for tests.

To be continued...

Sharma, Tully, and Cryder (2021) first proceeded with the test for the interaction. Since there are one global average and two main effect (additional difference in average for both factors `debttype` and `purchase`), the interaction involves one degree of freedom since we go from a model with three parameters describing the mean to one that has a different average for each of the four subgroups.

The reason why this is first test to carry out is that if the effect of one factor depends on the level of the other, as shown in Figure 5.1, then we need to compare the label of debt type separately for each type of purchase and vice-versa using simple effects. If the interaction on the contrary isn't significant, then we could pool observations instead and average across either of the two factors, resulting in the marginal comparisons with the main effects.

Fitting the model including the interaction between factors ensures that we keep the additivity assumption and that our conclusions aren't misleading: the price to pay is additional mean parameters to be estimated, which isn't an issue if you collect enough data, but can be critical when data collection is extremely costly and only a few runs are allowed.

In **R**, we include both factors in a formula as `response ~ factorA * factorB`, the `*` symbol indicating that both are allowed to interact; in the main effect model, we would use instead `+` to reflect that the effects of both factors add up.

```
# Analysing Supplementary Study 5
# of Sharma, Tully, and Cryder (2021)
data(STC21_SS5, package = "hecedsm")
mod <- aov(likelihood ~ purchase*debttype,
           data = STC21_SS5)
model.tables(mod, type = "means")
```

```
Tables of means
Grand mean

4.879747

 purchase
    discretionary    need
            4.182   5.579
rep       751.000 750.000

 debttype
     credit    loan
      5.127   4.631
rep 753.000 748.000

 purchase:debttype
               debttype
purchase        credit loan
  discretionary   4.5    3.8
```

```
  rep           392.0  359.0
  need            5.7    5.4
  rep           361.0  389.0
```

```
# Analysis of variance reveals
# non-significant interaction
# of purchase and type
car::Anova(mod, type = 3)
```

```
Anova Table (Type III tests)

Response: likelihood
                   Sum Sq   Df   F value    Pr(>F)
(Intercept)        7974.0    1 1040.9610 < 2.2e-16 ***
purchase            282.8    1   36.9137 1.563e-09 ***
debttype             88.5    1   11.5483 0.0006959 ***
purchase:debttype    13.7    1    1.7852 0.1817132
Residuals         11467.4 1497
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Main effects
emmeans::emmeans(mod,
                 specs = "debttype",
                 contr = "pairwise")
```

```
NOTE: Results may be misleading due to involvement in interactions

$emmeans
 debttype emmean    SE   df lower.CL upper.CL
 credit     5.12 0.101 1497     4.93     5.32
 loan       4.63 0.101 1497     4.43     4.83

Results are averaged over the levels of: purchase
Confidence level used: 0.95

$contrasts
 contrast      estimate    SE   df t.ratio p.value
```

```
 credit - loan    0.496 0.143 1497   3.469  0.0005

Results are averaged over the levels of: purchase
```

```
# Pairwise comparisons within levels of purchase
# Simple effect
emmeans::emmeans(mod,
                 specs = c("purchase", "debttype"),
                 by = "purchase",
                 contr = "pairwise")
```

```
$emmeans
purchase = discretionary:
 debttype emmean    SE   df lower.CL upper.CL
 credit     4.51 0.140 1497     4.24     4.78
 loan       3.82 0.146 1497     3.54     4.11

purchase = need:
 debttype emmean    SE   df lower.CL upper.CL
 credit     5.74 0.146 1497     5.45     6.02
 loan       5.43 0.140 1497     5.16     5.71

Confidence level used: 0.95

$contrasts
purchase = discretionary:
 contrast      estimate    SE   df t.ratio p.value
 credit - loan    0.687 0.202 1497   3.398  0.0007

purchase = need:
 contrast      estimate    SE   df t.ratio p.value
 credit - loan    0.305 0.202 1497   1.508  0.1318
```

In the analysis of variance table, we only focus on the last line with the sum of squares for `purchase:debttype`. The F statistic is 1.7852164; using the F (1, 1497) distribution as benchmark, we obtain a p-value of 0.18 so there is no evidence the effect of purchase depends on debt type.

We can thus pool data and look at the effect of debt type (`loan` or `credit`) overall by combining the results for all purchase types, one of the planned comparison reported in the

Supplementary material. To do this in **R** with the `emmeans` package, we use the `emmeans` function and we quote the factor of interest (i.e., the one we want to keep) in `specs`. By default, this will compute the estimate marginal means: the `contr = "pairwise"` indicates that we want the difference between the two, which gives us the contrasts.

To get the simple effects, we give both variables in `specs` as factors for which to compute subgroup means, then set additionally the `by` command to specify which variable we want separate results for. We get the difference in average between `credit` and `loan` labels for each purchase type along with the t statistics for the marginal contrast and the p-value. The simple effects suggest that the label has an impact on perception only for discretionary expenses rather than needed ones, which runs counter-intuitively with the lack of interaction.

! Summary

- Complete factorial designs consist of experiments in which we manipulate multiple experimental factors at once and collect observations for each subgroup.
- Factorial designs are more efficient than running repeatedly one-way analysis of variance with the same sample size per group.
- Interactions occur when the effect of a variable depends on the levels of the others.
- Interaction plots (group average per group) can help capture this difference, but beware of overinterpretation in small samples.
- If there is an interaction, we consider differences and contrasts for each level of the other factor (**simple effects**).
- If there is no interaction, we can pool observations and look at **main effects**.
- A multiway analysis of variance can be treated as a one-way analysis of variance by collapsing categories; however, only specific contrasts will be of interest.
- The number of observations increases quickly with the dimension as we increase the number of factors considered.

6 Designs to reduce the error

The previous chapter dealt with factorial experiments in which all experimental factors are of interest. In many instances, some of the characteristics of observational units are not of interest: for example, EEG measurements of participants in a lab may differ due to time of the day, to the lab technician, etc. These are instances of **blocking factors**: variables that impact the measurements's variability, but that are not of direct interest. By filtering their effect out and looking at the residual variability that is unexplained by the blocking factors. Block designs reduce the error term, at the cost of including and estimating additional parameters (group average or slope).

We will analyse block designs in the same as we did for multi-way analysis of variance model, with one notable exception. Typically, we will assume that there is **no interaction** between experimental factor and blocking factors.[1] Thus, we will be interested mostly in marginal effects.

A related design includes a continuous covariate to the analysis of variance, whose slope governs the relationship with the response. The strict inclusion isn't necessary to draw valid causal conclusion, but adding the term helps again reduce the residual variability. Such a design was historically called **analysis of covariance**, an instance of a linear model.

Including blocking factor or covariates should in principle increase power and our ability to detect real differences due to experimental manipulations, provided the variables used as control are correlated with the response. Generally, they are not needed for valid inference, which is guaranteed by randomization, and shouldn't be used to assign treatment.

6.1 Analysis of covariance

In an analysis of covariance, we include a linear component for a (continuous) covariate, with the purpose again to reduce residual error. A prime example is prior/post experiment measurements, whereby we monitor the change in outcome due to the manipulation.

[1] We can always check for this assumption.

In such setting, it may seem logical to take the difference in post and prior score as response: this is showcased in Example 6.1 and Baumann, Seifert-Kessell, and Jones (1992), an analysis of which is presented on the course website.

When we add a covariate, we need the latter to have a strong linear correlation for the inclusion to make sense. We can assess graphically whether the relationship is linear, and whether the slopes for each experimental condition are the same.[2]

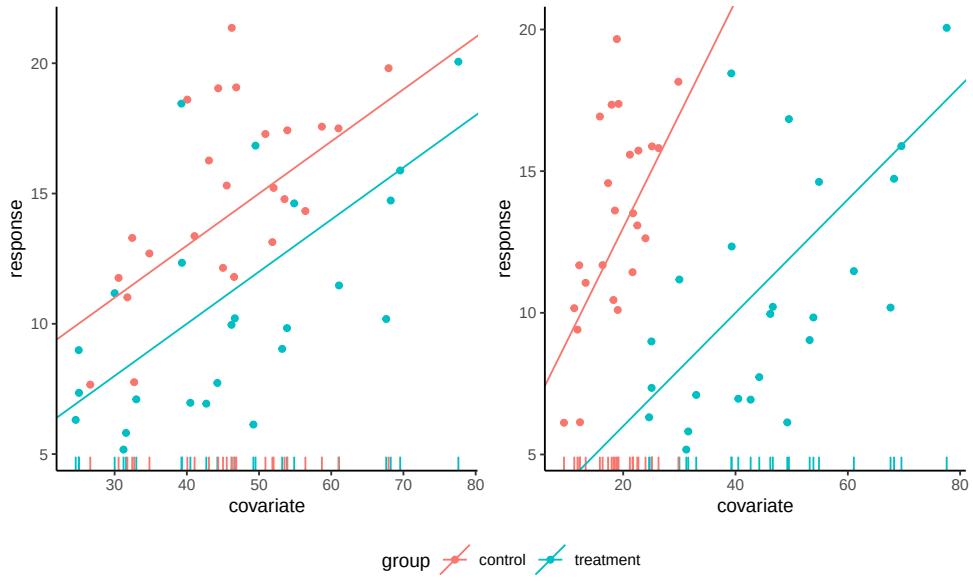


Figure 6.1: Simulated data from two groups with an analysis of covariance model.

The left panel of Figure 6.1 shows the ideal situation for an analysis of covariate: the relationship between response and covariate is linear with strong correlation, with the same slope and overlapping support. Since the slopes are the same, we can compare the difference in average (the vertical difference between slopes at any level of the covariate) because the latter is constant, so this depiction is useful. By contrast, the right-hand panel of Figure 6.1 shows an interaction between the covariate and the experimental groups, different slopes: there, the effect of the experimental condition increases with the level of the covariate. One may also note that the lack of overlap in the support, the set of values taken by the covariate, for the two experimental conditions, makes comparison hazardous at best in the right-hand panel.

Figure 6.2 shows that, due to the strong correlation, the variability of the measurements is smaller on the right-hand panel (corresponding to the analysis of covariance model) than

[2]If not, this implies that the covariate interacts with the experimental condition.

for the centred response on the left-hand panel; note that the y-axes have different scales.

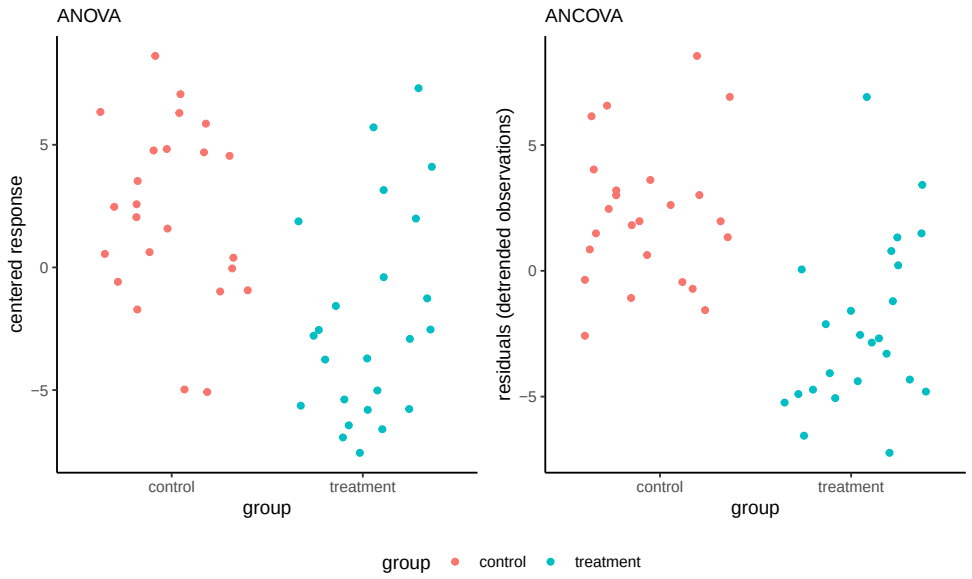


Figure 6.2: Response after subtracting mean (left) and after detrending (right).

We present two examples of analysis of covariance, showing how the inclusion of covariates helps disentangle differences between experimental conditions.

Lee and Choi (2019) measured the impact of discrepancies between descriptions and visual depiction of items in online retail. They performed an experiment in which participants were presented with descriptions of a product (a set of six toothbrushes) that was either consistent or inconsistent with the description. The authors postulated that a discrepancy could lead to lower appreciation score, measured using three Likert scales. They also suspected that the familiarity with the product brand should impact ratings, and controlled for the latter using another question.

One way to account for familiarity when comparing the mean is to use a linear regression with familiarity as another explanatory variable. The expected value of the product evaluation is

$$\mathsf{E}(\texttt{prodeval}) = \beta_0 + \beta_1 \texttt{familiarity} + \beta_2 \texttt{consistency}, \tag{6.1}$$

where `familiarity` is the score from 1 to 7 and `consistency` is a binary indicator equal to one if the output is inconsistent and zero otherwise. The coefficient β_2 thus measures the difference between product evaluation rating for consistent vs inconsistent displays, for the same familiarity score.

Table 6.1: Analysis of variance tables

(a) model without familiarity

term	sum. sq.	df	stat	p-value
consistency	7.04	1	2.55	.113
Residuals	259.18	94		

(b) model with familiarity

term	sum. sq.	df	stat	p-value
familiarity	55.9	1	25.60	< .001
consistency	9.8	1	4.49	.037
Residuals	203.2	93		

We can look at coefficient (standard error) estimates $\widehat{\beta}_2 = -0.64(0.302)$. No difference between groups would mean $\beta_2 = 0$ and we can build a test statistic by looking at the standardized regression coefficient $t = \widehat{\beta}_2/\mathsf{se}(\widehat{\beta}_2)$. The result output is $b = -0.64$, 95% CI $[-1.24, -0.04]$, $t(93) = -2.12$, $p = .037$. We reject the null hypothesis of equal product evaluation for both display at level 5%: there is evidence that there is a small difference, with people giving on average a score that is 0.64 points smaller (on a scale of 1 to 9) when presented with conflicting descriptions and images.

We can compare the analysis of variance table obtained by fitting the model with and without `familiarity`. Table 6.1 shows that the effect of consistency is small and not significant and a two-sample t-test shows no evidence of difference between the average familiarity score in both experimental conditions (p-value of $.532$). However, we can explain roughly one fifth of the residual variability by the familiarity with the brand (see the sum of squares in Table 6.1): removing the latter leads to a higher signal-to-noise ratio for the impact of consistency, at the expense of a loss of one degree of freedom. Thus, it appears that the manipulation was successful.

Figure 6.3 shows that people more familiar with the product or brand tend to have a more positive product evaluation, as postulated by the authors. The graph also shows two straight lines corresponding to the fit of a linear model with different intercept and slope for each display group: there is little material difference, one needs to assess formally whether a single linear relationship as the one postulated in Equation 6.1 can adequately characterize the relation in both groups.

To this effect, we fit a linear model with different slopes in each group, and compare the

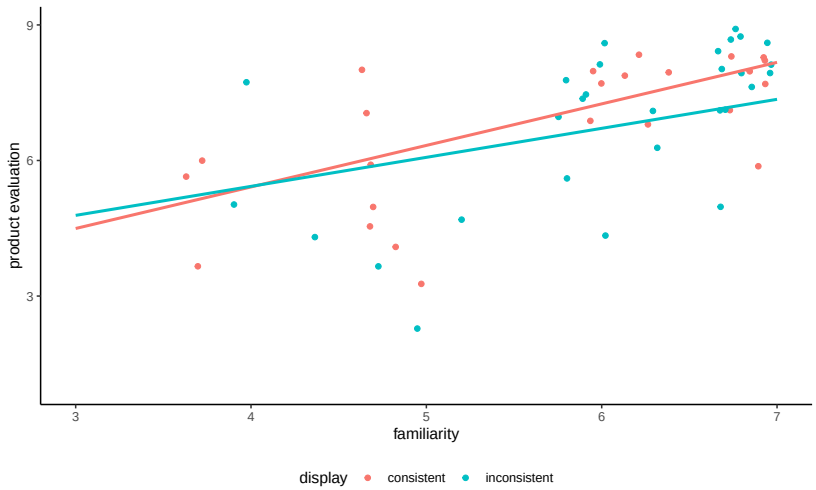


Figure 6.3: Scatterplot of product evaluation as a function of the familiarity score, split by experimental manipulation.

fit of the latter with the analysis of covariance model that includes a single slope for both groups: we can then test if the slopes are the same, or alternatively if the difference between the slopes is zero. The t-statistic indicates no difference in slope (p-value of .379), thus the assumption is reasonable. Levene's test for homogeneity of variance indicates no discernible difference between groups. Thus, it appears there is a difference in perception of product quality due to the manipulation.

Example 6.1 (Effect of scientific consensus on false beliefs)**.** We consider Study 3 of Stekelenburg et al. (2021), who studied changes in perception of people holding false beliefs or denying (to some extent) the scientific consensus by presenting them with news article showcasing information about various phenomena. The experimental manipulation consisted in presenting boosting, a form of training to help readers identify and establish whether scientifists were truly expert in the domain of interest, how strong was the consensus, etc.[3]

The third and final experiment of the paper focused on genetically modified organisms: it is a replication of Study 2, but with a control group (since there were no detectable difference between experimental conditions `Boost` and `BoostPlus`) and a larger sample size (because Study 2 was underpowered).

[3]The article is interesting because lack of planning/changes led them to adapt the design from experiment 1 to 3 until they found something. Without preregistration, it is unlikely such findings would have been publishable.

The data include 854 observations with `prior`, the negative of the prior belief score of the participant, the `post` experiment score for the veracity of the claim. Both were measured using a visual scale ranging from -100 (I am 100% certain this is false) to 100 (I am 100% certain this is true), with 0 (I don't know) in the middle. Only people with negative prior beliefs were recruited to the study. The three experimental conditions were `BoostPlus`, `consensus` and a `control` group. Note that the scores in the data have been negated, meaning that negative posterior scores indicate agreement with the consensus on GMO.

Preliminary checks suggest that, although the slopes for prior beliefs could plausibly be the same in each group and the data are properly randomized, there is evidence of unequal variance for the changes in score. As such, we fit a model with mean

$$\mathsf{E}(\texttt{post}) = \begin{cases} \beta_0 + \beta_1 \texttt{prior} + \alpha_1 & \texttt{condition} = \texttt{BoostPlus} \\ \beta_0 + \beta_1 \texttt{prior} + \alpha_2 & \texttt{condition} = \texttt{consensus} \\ \beta_0 + \beta_1 \texttt{prior} + \alpha_3 & \texttt{condition} = \texttt{control} \end{cases}$$

with $\alpha_1 + \alpha_2 + \alpha_3 = 0$, using the sum-to-zero parametrization, and with different variance for each experimental condition,

$$\mathsf{Va}(\texttt{post}) = \begin{cases} \sigma_1^2, & \texttt{condition} = \texttt{BoostPlus}, \\ \sigma_2^2, & \texttt{condition} = \texttt{consensus}, \\ \sigma_3^2, & \texttt{condition} = \texttt{control}. \end{cases}$$

Because of the unequal variances, we cannot use multiple testing procedures reserved for analysis of variance and resort instead to Holm–Bonferroni to control the familywise error rate. We here look only at pairwise differences between conditions.[4]

Repeating the exercise of comparing the amount of evidence for comparison with and without inclusion of a covariate shows that the value of the test statistic is larger (Table 6.2), indicative of stronger evidence with the analysis of covariance model: the conclusion would be unaffected with such large sample sizes. We of course care very little for the global F test of equality of mean, as the previous study had shown large differences. What is more interesting here is quantifying the change between conditions.

Table 6.3 shows the pairwise contrasts, which measure two different things: the analysis of variance model compares the average in group, whereas the analysis of covariance (the linear model with `prior`) uses detrended values and focuses on the change in perception. Because the data are unbalanced and we estimate group mean and variance separately, the degrees of freedom change from one pairwise comparison to the next. Again, using the covariate `prior`, which is somewhat strongly correlated with `post` as seen in Figure 6.4, helps decrease background noise.

[4] In Study 2, the interest was comparing manipulation vs control and the Boost vs BoostPlus conditions, two orthogonal contrasts.

Table 6.2: Analysis of variance tables

(a) ANOVA model (without prior belief)

term	df	stat	p-value
condition	2	42.5	< .001

(b) ANCOVA model (with prior belief)

term	df	stat	p-value
prior	1	289	< .001
condition	2	57	< .001

Table 6.3: Pairwise contrasts with _p_-values adjusted using Holm--Bonferroni

(a) ANOVA model (without prior belief score).

contrast	estimate	std.error	df	statistic	p.value
consensus vs control	-12.0	4.0	558	-3.01	.003
consensus vs BoostPlus	16.3	4.7	546	3.49	< .001
BoostPlus vs control	-28.3	4.4	505	-6.49	< .001

(b) ANCOVA model (with prior belief score).

contrast	estimate	std.error	df	statistic	p.value
consensus vs control	-11.8	3.3	543	-3.54	< .001
consensus vs BoostPlus	17.5	4.3	524	4.11	< .001
BoostPlus vs control	-29.3	3.9	459	-7.45	< .001

Pitfall

Stekelenburg et al. (2021) split their data to do pairwise comparisons two at the time (thus taking roughly two-third of the data to perform a two sample t-test with each pair). Although it does not impact their conclusion, this approach is conceptually incorrect: if the variance was equal, we would want to use all observations to estimate it (so their approach would be suboptimal, since we would estimate the variance three

Table 6.4: Summary statistics of belief as a function of time of measurement and experimental condition.

time	condition	mean	se
prior	BoostPlus	57.65	1.69
prior	consensus	56.32	1.67
prior	control	56.49	1.68
post	BoostPlus	2.62	3.53
post	consensus	18.93	3.06
post	control	30.91	2.56

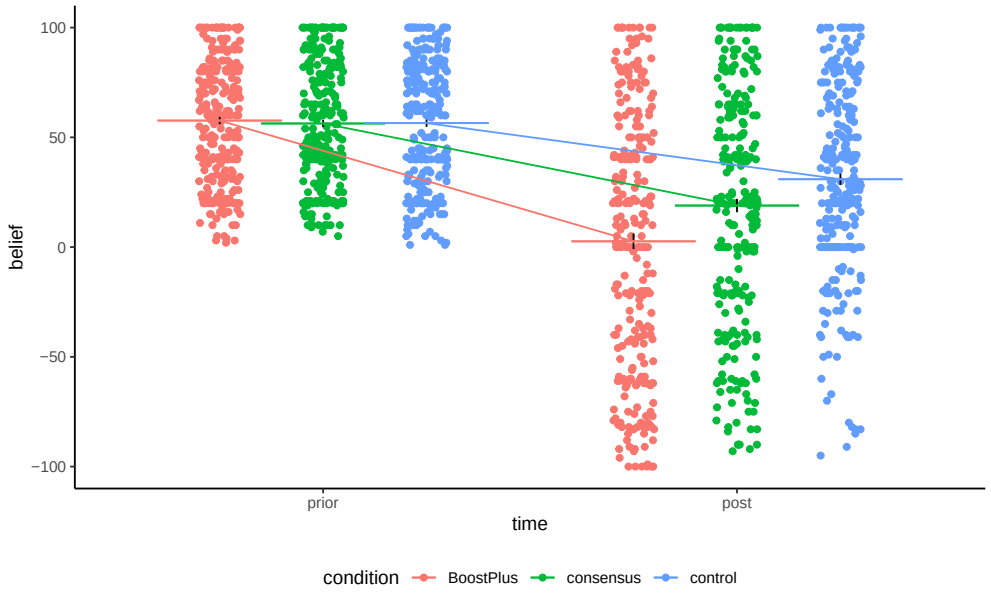


Figure 6.4: Difference between prior and post experiment beliefs on genetically engineered food.

times with smaller samples).
On the contrary, using a model that assumes equal variance when it is not the case leads to distortion: the variance we estimate will be some sort of average of the variability σ_i and σ_j in experimental condition i and j, again potentially leading to distortions. With large samples, this may be unconsequential, but illustrates caveats of subsample analyses.

Pitfall

Figure 6.5 shows the relationship between prior and posterior score. The data show clear difference between individuals: many start from completely disbelieving of genetically engineered food and change their mind (sometimes drastically), there are many people who do not change idea at all and have similar scores, and many who give a posterior score of zero. This heterogeneity in the data illustrates the danger of only looking at the summary statistics and comparing averages. It does not tell the whole picture! One could investigate whether the strength of religious or political beliefs, or how much participants trust scientists, could explain some of the observed differences.

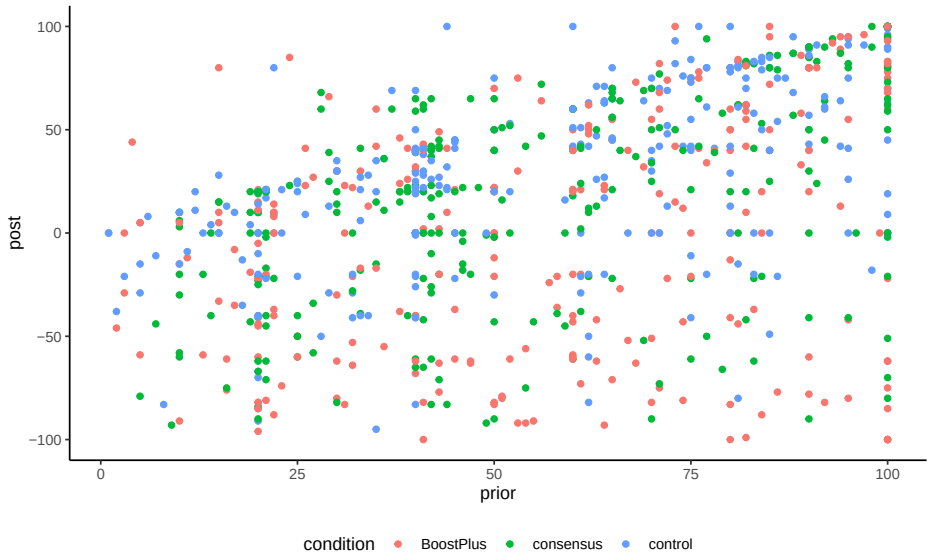


Figure 6.5: Scatterplot of negated prior and posterior belief score.

Summary

- Inclusion of blocking factor and continuous covariates may help filtering out unwanted variability.
- These are typically concomitant variables (measured alongside the response variable).
- These designs reduce the residual error, leading to an increase in power (more ability to detect differences in average between experimental conditions).
- We are only interested in differences due to experimental condition (marginal

effects).
- In general, there should be no interaction between covariates/blocking factors and experimental conditions.
- This hypothesis can be assessed by comparing the models with and without interaction, if there are enough units (e.g., equality of slope for ANCOVA).

Your turn

- Box, Hunter, and Hunter (1978) write on page 103 the following motto:

  Block what you can and randomize what you cannot.

Explain the main benefit of blocking for confounding variables (when possible) over randomization.

7 Effect sizes and power

In social studies, it is common to write a paper containing multiple studies on a similar topic. These may use different designs, with varying sample size. If the studies uses different questionnaires, or change the Likert scale, the results and the mean difference between groups are not directly comparable between experiments.

We may also wish replicate a study by using the same material and re-run an experiment. For the replication to be somewhat successful (or at least reliable), one needs to determine beforehand how many participants should be recruited in the study.

We could think for an example of comparing statistics or p-values, which are by construction standardized unitless measures, making them comparable across study. Test statistics show how outlying observed differences between experimental conditions relative to a null hypothesis, typically that of no effect (equal mean in each subgroup). However, statistics are usually a function of both the sample size (the number of observations in each experimental condition) and the effect size (how large the standardized differences between groups are), making them unsuitable for describing differences.

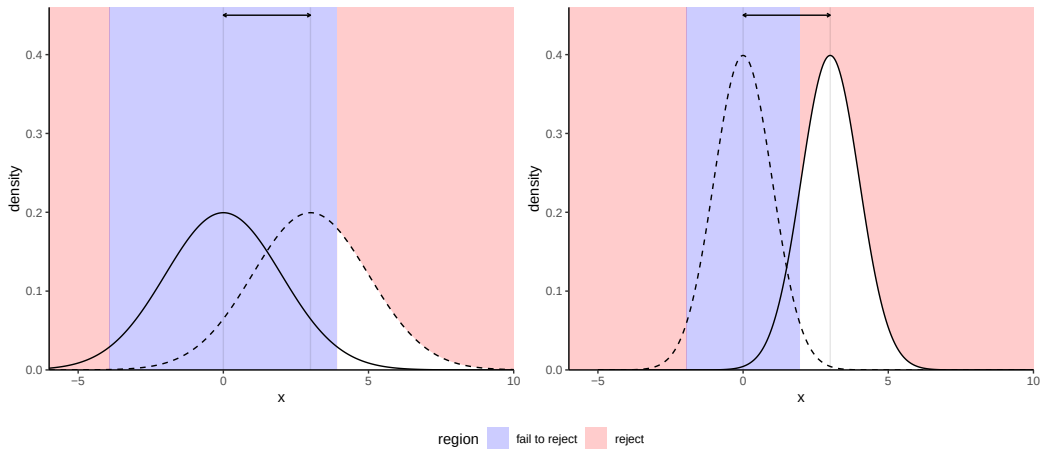


Figure 7.1: True sampling distribution for a two-sample t-test under the alternative (rightmost curve) and null distribution (leftmost curve) with small (left panel) and large (right panel) sample sizes.

Figure 7.1 shows an example with the sampling distributions of the difference in mean under the null (curve centered at zero) and the true alternative (mean difference of two). The area in white under the curve represents the power, which is larger with larger sample size and coincides with smaller average p-values for the testing procedure.

One could argue that, on the surface, every null hypothesis is wrong and that, with a sufficiently large number of observation, all observed differences eventually become "statistically significant". This has to do with the fact that we become more and more certain of the estimated means of each experimental sub-condition. Statistical significance of a testing procedure does not translate into practical relevance, which itself depends on the scientific question at hand. For example, consider the development of a new drug for commercialization by Health Canada: what is the minimum difference between two treatments that would be large enough to justify commercialization of the new drug? If the effect is small but it leads to many lives saved, would it still be relevant? Such decision involve a trade-off between efficacy of new treatment relative to the status quo, the cost of the drug, the magnitude of the improvement, etc.

Effect size are summaries to inform about the standardized magnitude of these differences; they are used to combine results of multiple experiments using meta-analysis, or to calculate sample size requirements to replicate an effect in power studies.

7.1 Effect sizes

There are two main classes of effect size: standardized mean differences and ratio (percentages) of explained variance. The latter are used in analysis of variance when there are multiple groups to compare.

Unfortunately, the literature on effect size is quite large. Researchers often fail to distinguish between estimand (unknown target) and the estimator that is being used, with frequent notational confusion arising due to conflicting standards and definitions. Terms are also overloaded: the same notation may be used to denote an effect size, but it will be calculated differently depending on whether the design is between-subject or within-subject (with repeated correlated measures per participant), or whether there are blocking factors.

7.1.1 Standardized mean differences

To gather intuition, we begin with the task of comparing the means of two groups using a two-sample t-test, with the null hypothesis of equality in means or $\mathscr{H}_0 : \mu_1 = \mu_2$. The test

statistic is

$$T = \frac{\widehat{\mu}_2 - \widehat{\mu}_1}{\widehat{\sigma}} \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1/2}$$

where $\widehat{\sigma}$ is the pooled sample size estimator. The first term, $\widehat{d}_s = (\widehat{\mu}_2 - \widehat{\mu}_1)/\widehat{\sigma}$, is termed Cohen's d (Cohen 1988) and it measures the standardized difference between groups, a form of signal-to-noise ratio. As the sample size gets larger and larger, the sample mean and pooled sample variance become closer and closer to the true population values μ_1, μ_2 and σ; at the same time, the statistic T becomes bigger as n becomes larger because of the second term.[1]

The difference $d = (\mu_1 - \mu_2)/\sigma$ has an obvious interpretation: a distance of a indicates that the means of the two groups are a standard deviation apart. Cohen's d is sometimes loosely categorized in terms of weak ($d = 0.2$), medium ($d = 0.5$) and large ($d = 0.8$) effect size; these, much like arbitrary p-value cutoffs, are rules of thumbs. Alongside d, there are many commonly reported metrics that are simple transformations of d describing the observed difference. This interactive applet by Kristoffer Magnusson (Magnusson 2021) shows the visual impact of changing the value of d along. There are different estimators of d depending on whether or not the pooled variance estimator is used. Cohen's d, is upward biased, meaning it gives values that are on average larger than the truth. Hedge's g (Hedges 1981) offers a bias-correction and should always be preferred as an estimator.

For these different estimators, it is possible to obtain (asymmetric) confidence intervals or tolerance intervals.[2]

We consider a two-sample t-test for the study of Liu et al. (2022+) discussed in Example 2.5. The difference in average response index is 0.371, indicating that the responder have a higher score. The p-value is 0.041, showing a small effect.

If we consider the standardized difference d, the group means are -0.289 standard deviations apart based on Hedge's g, with an associated 95% confidence interval of [-0.567, -0.011]: thus, the difference found is small (using Cohen (1988)'s convention) and there is a large uncertainty surrounding it.

There is a 42% probability that an observation drawn at random from the responder condition will exceed an observation drawn at random of the initiator group (probability of superiority) and 38.6% of the responder observations will exceed the median of the initiator (Cohen's U_3).

[1] If we consider a balanced sample, $n_1 = n_2 = n/2$ we can rewrite the statistic as $T = \sqrt{n}\widehat{d}_s/2$ and the statement that T increases with n on average becomes more obvious.

[2] By using the pivot method, e.g., Steiger (2004), and relating the effect size to the noncentrality parameter of the null distribution, whether St, F or χ^2.

```
data(LRMM22_S1, package = "hecedsm")
ttest <- t.test(
  appreciation ~ role,
  data = LRMM22_S1,
  var.equal = TRUE)
effect <- effectsize::hedges_g(
  appreciation ~ role,
  data = LRMM22_S1,
  pooled_sd = TRUE)
```

7.1.2 Ratio and proportion of variance

Another class of effect sizes are obtained by considering either the ratio of the variance due to an effect (say differences in means relative to the overall mean) relative to the background level of noise as measured by the variance.

One common measure employed in software is Cohen's f (Cohen 1988), which for a one-way ANOVA (equal variance σ^2) with more than two groups,

$$f^2 = \frac{1}{\sigma^2}\sum_{j=1}^{k}\frac{n_j}{n}(\mu_j - \mu)^2 = \frac{\sigma^2_{\text{effect}}}{\sigma^2},$$

a weighted sum of squared difference relative to the overall mean μ. σ^2_{effect} is a measure of the variability that is due to the difference in mean, so standardizing it by the measurement variance gives us a ratio of variance with values higher than one indicating that more variability is explainable, leading to higher effect sizes. If the means of every subgroup is the same, then $f = 0$. For $k = 2$ groups, Cohen's f and Cohen's d are related via $f = d/2$.

Cohen's f can be directly related to the behaviour of the F statistic under an alternative, as explained in Section 7.2.1. However, since the interpretation isn't straightforward, we typically consider proportions of variance (rather than ratios of variance).

To build such an effect size, we break down the variability that is explained by our experimental manipulation (σ^2_{effect}), here denoted by effect, from the leftover unexplained part, or residual (σ^2_{resid}). In a one-way analysis of variance,

$$\sigma^2_{\text{total}} = \sigma^2_{\text{resid}} + \sigma^2_{\text{effect}}$$

and the percentage of variability explained by the effect.

$$\eta^2 = \frac{\text{explained variability}}{\text{total variability}} = \frac{\sigma^2_{\text{effect}}}{\sigma^2_{\text{resid}} + \sigma^2_{\text{effect}}} = \frac{\sigma^2_{\text{effect}}}{\sigma^2_{\text{total}}}.$$

Simple arithmetic manipulations reveal that $f^2 = \eta^2/(1-\eta^2)$, so we can relate any proportion of variance in terms of ratio and vice-versa.

Such an effect size depends on unknown population quantities (the true means of each subgroup, the overall mean and the variance). There are multiple alternative estimators to estimate η^2, and researchers are often carefree when reporting as to which is used. To disambiguate, I will put $\hat{\eta}^2$ to denote an estimator. To make an analogy, there are many different recipes (estimators) that can lead to a particular cake, but some may lead to a mixing that is on average too wet if they are not well calibrated.

The default estimator for η^2 is the coefficient of determination of the linear regression, denoted $\widehat{R}^2$ or $\widehat{\eta}^2$. The latter can be reconstructed from the analysis of variance table using the formula

$$\widehat{R}^2 = \frac{F\nu_1}{F\nu_1 + \nu_2}$$

where for the one-way ANOVA $\nu_1 = K - 1$ and $\nu_2 = n - K$ are the degrees of freedom of a design with n observations and K experimental conditions.

Unfortunately, $\widehat{R}^2$ is an upward biased estimator (too large on average), leading to optimistic measures. Another estimator of η^2 that is recommended in Keppel and Wickens (2004) for power calculations is $\widehat{\omega}^2$, which is

$$\widehat{\omega}^2 = \frac{\nu_1(F-1)}{\nu_1(F-1)+n}.$$

Since the F statistic is approximately 1 on average, this measure removes the mode. Both $\widehat{\omega}^2$ and $\widehat{\epsilon}^2$ have been reported to be less biased and thus preferable as estimators of the true proportion of variance (Lakens 2013).

7.1.3 Partial effects and variance decomposition

In a multiway design with several factors, we may want to estimate the effect of separate factors or interactions. In such cases, we can break down the variability explained by manipulations per effect. The effect size for such models are build by comparing the variance explained by the effect σ^2_{effect}.

For example, say we have a completely randomized balanced design with two factors A, B and their interaction AB. We can decompose the total variance as

$$\sigma^2_{\text{total}} = \sigma^2_A + \sigma^2_B + \sigma^2_{AB} + \sigma^2_{\text{resid}}.$$

When the design is balanced, these variance terms can be estimated using the mean squared error from the analysis of variance table output. If the design is unbalanced, the sum of

square decomposition is not unique and we will get different estimates when using Type II and Type III sum of squares.

We can get formula similar to the one-sample case with now what are termed **partial** effect sizes, e.g.,

$$\widehat{\omega}^2_{\langle \text{effect} \rangle} = \frac{\text{df}_{\text{effect}}(F_{\text{effect}} - 1)}{\text{df}_{\text{effect}}(F_{\text{effect}} - 1) + n},$$

where n is the overall sample size and F_{effect} and the corresponding degrees of freedom could be the statistic associated to the main effects A and B, or the interaction term AB. In **R**, the `effectsize` package reports these estimates with one-sided confidence intervals derived using the pivot method (Steiger 2004).[3]

Software will typically return estimates of effect size alongside with the designs, but there are small things to keep in mind. One is that the decomposition of the variance is not unique with unbalanced data. The second is that, when using repeated measures and mixed models, the same notation is used to denote different quantities.

Lastly, it is customary to report effect sizes that include the variability of blocking factors and random effects, leading to so-called **generalized** effect sizes. Include the variance of all blocking factors and interactions (only with the effect!) in the denominator.[4]

For example, if A is the experimental factor whose main effect is of interest, B is a blocking factor and C is another experimental factor, use

$$\eta^2_{\langle A \rangle} = \frac{\sigma^2_A}{\sigma^2_A + \sigma^2_B + \sigma^2_{AB} + \sigma^2_{\text{resid}}}.$$

as generalized partial effect. In **R**, most effect sizes for variance proportion have a `generalized` argument to which the vector of names of blocking factor can be passed. The reason for including blocking factors and random effects is that they would not necessarily be available in a replication. The correct effect size measure to calculate and to report depends on the design, and there are numerous estimators that can be utilized. Since they are related to one another, it is oftentimes possible to compute them directly from the output or convert. The formula highlight the importance of reporting (with enough precision) exactly the values of the test statistic.

[3] The confidence intervals are based on the F distribution, by changing the non-centrality parameter and inverting the distribution function (pivot method). This yields asymmetric intervals.

[4] Typically, there won't be any interaction with blocking factors, but it there was for some reason, it should be included in the total.

7.2 Power

There are typically two uses to hypothesis test: either we want to show it is not unreasonable to assume the null hypothesis (for example, assuming equal variance), or else we want to show beyond reasonable doubt that a difference or effect is significative: for example, one could wish to demonstrate that a new website design (alternative hypothesis) leads to a significant increase in sales relative to the status quo (null hypothesis).

Our ability make discoveries depends on the power of the test: the larger the power, the greater our ability to reject the null hypothesis $\mathscr{H}_0$ when the latter is false.

The **power of a test** is the probability of **correctly** rejecting the null hypothesis $\mathscr{H}_0$ when $\mathscr{H}_0$ is false, i.e.,

$$\mathsf{Pr}_a(\text{reject}\mathscr{H}_0)$$

Whereas the null alternative corresponds to a single value (equality in mean), there are infinitely many alternatives... Depending on the alternative models, it is more or less easy to detect that the null hypothesis is false and reject in favour of an alternative. Power is thus a measure of our ability to detect real effects. Different test statistics can give broadly similar conclusions despite being based on different benchmark. Generally, however, there will be a tradeoff between the number of assumptions we make about our data or model (the fewer, the better) and the ability to draw conclusions when there is truly something going on when the null hypothesis is false.

We want to choose an experimental design and a test statistic that leads to high power, so that this power is as close as possible to one. Under various assumptions about the distribution of the original data, we can derive optimal tests that are most powerful, but some of the power comes from imposing more structure and these assumptions need not be satisfied in practice.

Minimally, the power of the test should be α because we reject the null hypothesis α fraction of the time even when $\mathscr{H}_0$ is true. Power depends on many criteria, notably

- the effect size: the bigger the difference between the postulated value for θ_0 under $\mathscr{H}_0$ and the observed behaviour, the easier it is to detect departures from θ_0. (Figure 7.4); it's easier to spot an elephant in a room than a mouse.
- variability: the less noisy your data, the easier it is to assess that the observed differences are genuine, as Figure 7.3 shows;

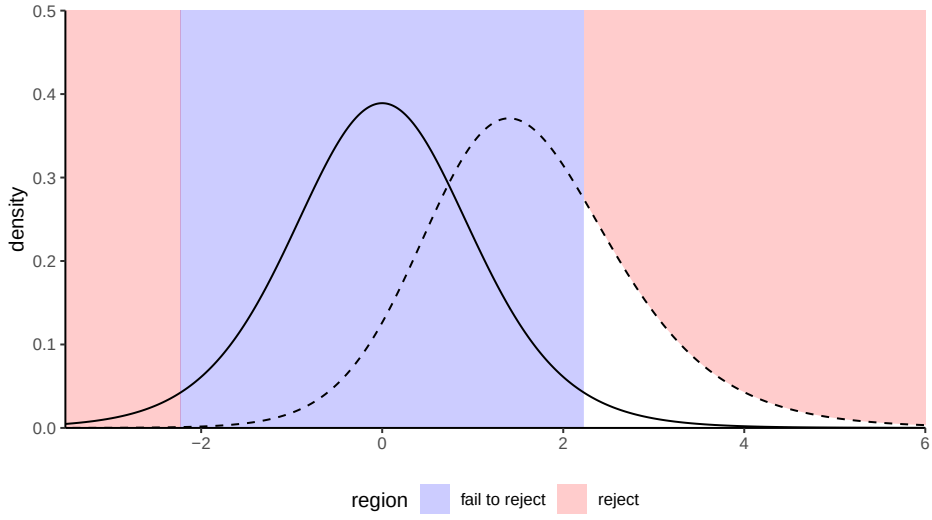


Figure 7.2: Comparison between null distribution (full curve) and a specific alternative for a t-test (dashed line). The power corresponds to the area under the curve of the density of the alternative distribution which is in the rejection area (in white).

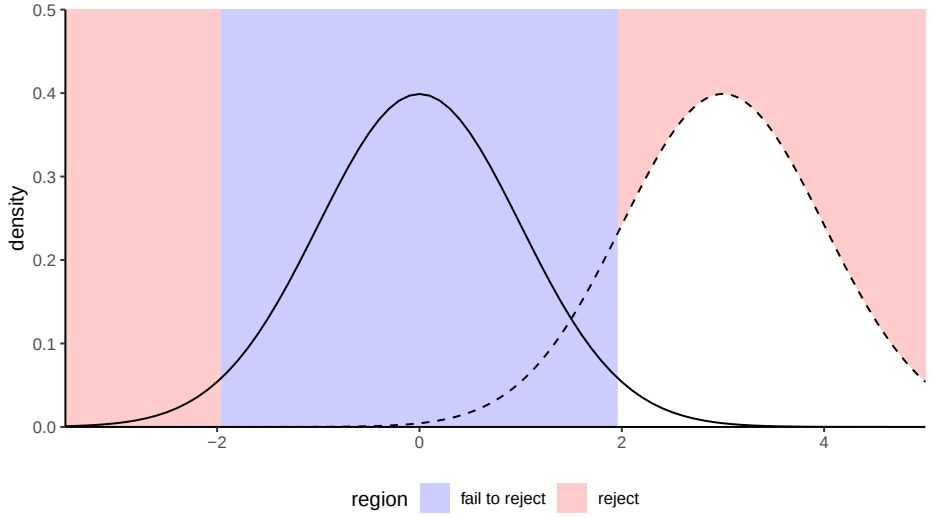


Figure 7.3: Increase in power due to an increase in the mean difference between the null and alternative hypothesis. Power is the area in the rejection region (in white) under the alternative distribution (dashed): the latter is more shifted to the right relative to the null distribution (full line).

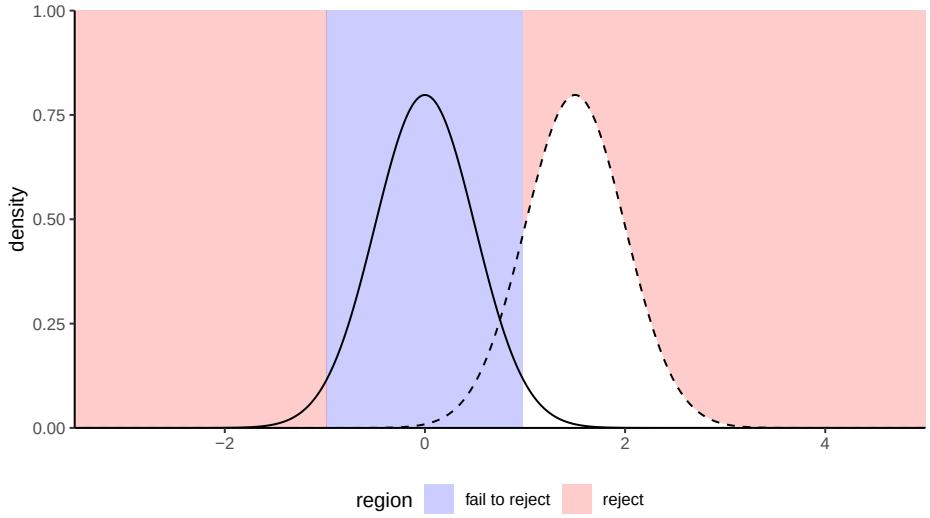


Figure 7.4: Increase of power due to an increase in the sample size or a decrease of standard deviation of the population: the null distribution (full line) is more concentrated. Power is given by the area (white) under the curve of the alternative distribution (dashed). In general, the null distribution changes with the sample size.

- the sample size: the more observation, the higher our ability to detect significative differences because the amount of evidence increases as we gather more observations.[5] In experimental designs, the power also depends on how many observations are allocated to each group.[6]
- the choice of test statistic: there is a plethora of possible statistics to choose from as a summary of the evidence against the null hypothesis. Choosing and designing statistics is usually best left out to statisticians, as there may be tradeoffs. For example, rank-based statistics discard information about the observed values of the response, focusing instead on their relative ranking. The resulting tests are typically less powerful, but they are also less sensible to model assumptions, model misspecification and outliers.

Changing the value of α also has an impact on the power, since larger values of α move the cutoff towards the bulk of the distribution. However, it entails a higher percentage of rejection also when the alternative is false. Since the value of α is fixed beforehand to control the type I error (avoid judicial mistakes), it's not a parameter we consider.

[5]Specifically, the standard error decreases with sample size n at a rate (typically) of $n^{-1/2}$. The null distribution also becomes more concentrated as the sample size increase.

[6]While the default is to assign an equal number to each subgroup, power may be maximized by specifying different sample size in each group if the variability of the measurement differ in these groups.

There is an intricate relation between effect size, power and sample size. Journals and grant agencies oftentimes require an estimate of the latter before funding a study, so one needs to ensure that the sample size is large enough to pick-up effects of scientific interest (good signal-to-noise), but also not overly large as to minimize time and money and make an efficient allocation of resources. This is Goldilock's principle, but having more never hurts.

If we run a pilot study to estimate the background level of noise and the estimated effect, or if we wish to perform a replication study, we will come up with a similar question in both cases: how many participants are needed to reliably detect such a difference? Setting a minimum value for the power (at least 80%, but typically 90% or 95% when feasible) ensures that the study is more reliable and ensures a high chance of success of finding an effect of at least the size specified. A power of 80% ensures that, on average, 4 in 5 experiments in which we study a phenomenon with the specified non-null effect size should lead to rejecting the null hypothesis.

In order to better understand the interplay between power, effect size and sample size, we consider a theoretical example. The purpose of displaying the formula is to (hopefully) more transparently confirm some of our intuitions about what leads to higher power. There are many things that can influence the power:

- the experimental design: a blocking design or repeated measures tend to filter out some of the unwanted variability in the population, thus increasing power relative to a completely randomized design
- the background variability σ: the noise level is oftentimes intrinsic to the measurement. It depends on the phenomenon under study, but instrumentation and the choice of scale, etc. can have an impact. Running experiments in a controlled environment helps reduce this, but researchers typically have limited control on the variability inherent to each observation.
- the sample size: as more data are gathered, information accumulates. The precision of measurements (e.g., differences in mean) is normally determined by the group with the smallest sample size, so (approximate) balancing increases power if the variance in each group is the same.
- the size of the effect: the bigger the effect, the easier it is to accurately detect (it's easier to spot an elephant than a mouse hiding in a classroom).
- the level of the test, α: if we increase the rejection region, we technically increase power when we run an experiment under an alternative regime. However, the level is oftentimes prespecified to avoid type I errors. We may consider multiplicity correction within the power function, such as Bonferonni's method, which is equivalent to reducing α.

7.2.1 Power for one-way ANOVA

To fix ideas, we consider the one-way analysis of variance model. In the usual setup, we consider K experimental conditions with n_k observations in group k, whose population average we denote by μ_k. We can parametrize the model in terms of the overall sample average,

$$\mu = \frac{1}{n}\sum_{j=1}^{K}\sum_{i=1}^{n_j}\mu_j = \frac{1}{n}\sum_{j=1}^{K} n_j\mu_j,$$

where $n = n_1 + \cdots + n_K$ is the total sample size. The F-statistic of the one-way ANOVA is

$$F = \frac{\text{between sum of squares}/(K-1)}{\text{within sum of squares}/(n-K)}$$

The null distribution is $F(K-1, n-K)$. Our interest is in understanding how the F-statistic behaves under an alternative.

During the construction, we stressed out that the denominator is an estimator of σ^2 under both the null and alternative. What happens to the numerator? We can write the population average for the between sum of square as

$$\mathsf{E}(\text{between sum of squares}) = \sigma^2\{(K-1) + \Delta\}.$$

where

$$\Delta = \frac{\sum_{j=1}^{K} n_j(\mu_j - \mu)^2}{\sigma^2} = nf^2.$$

and where f^2 is the square of Cohen's f. Under the null hypothesis, all group means are equal and $\mu_j = \mu$ for $j = 1, \ldots, K$ and $\Delta = 0$, but if some groups have different average the displacement will be non-zero. The greater Δ, the further the mode (peak of the distribution) is from unity and the greater the power.

Closer examination reveals that Δ increases with n_j (sample size) and with the true squared mean difference $(\mu_j - \mu)^2$ increases effect size represented by the difference in mean, but decreases as the observation variance increases.

Under the alternative, the distribution of the F statistic is a noncentral Fisher distribution, denoted $\mathsf{F}(\nu_1, \nu_2, \Delta)$ with degrees of freedom ν_1 and ν_2 and noncentrality parameter Δ.[7] To calculate the power of a test, we need to single out a specific alternative hypothesis.

[7] Note that the $F(\nu_1, \nu_2)$ distribution is indistinguishable from $\chi^2(\nu_1)$ for ν_2 large. A similar result holds for tests with χ^2 null distributions.

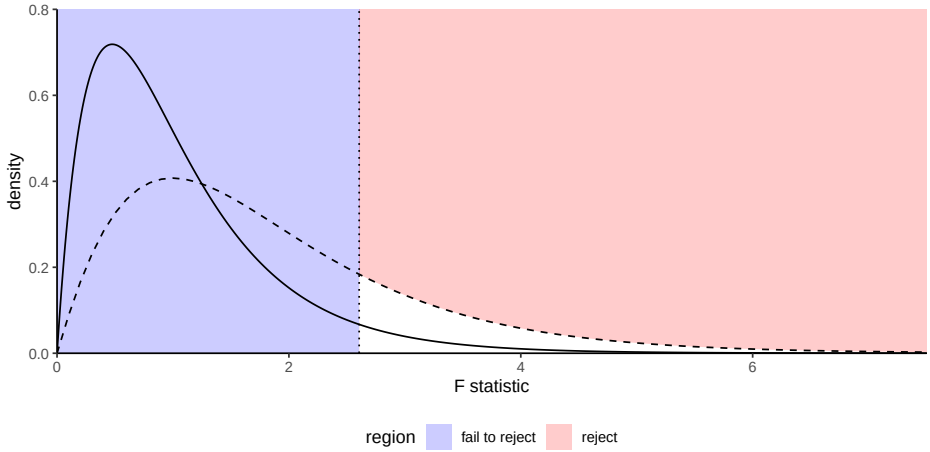


Figure 7.5: Density curves for the null distribution (full line) and true distribution (dashed line) under noncentrality parameter $\Delta = 3$. The area in white under the curve denotes the power under this alternative.

The plot in Figure 7.5 shows the null (full line) distribution and the true distribution (dashed line) for a particular alternative. The noncentral F is shifted to the right and right skewed, so the mode (peak) is further away from 1.

Given a value of $\Delta = nf^2$ and information about the effect of interest (degrees of freedom of the effect and the residuals), we can compute the tail probability as follows

1. Compute the cutoff point: the value under $\mathscr{H}_0$ that leads to rejection at level α
2. Compute probability below the alternative curve, from the cutoff onwards.

```
cutoff <- qf(p = 1-alpha, df1 = df1, df2 = df2)
pf(q = cutoff,  df1 = df1, df2 = df2,
    ncp = Delta, lower.tail = FALSE)
```

In practice, a software will return these quantities and inform us about the power. Note that these results are trustworthy provided the model assumptions are met, otherwise they may be misleading.

The most difficult question when trying to estimate sample size for a study is determining which value to use for the effect size. One could opt for a value reported elsewhere for a similar scale to estimate the variability and provide educated guesses for the mean differences. Another option is to run a pilot study and use the resulting estimates to inform about sensible values, perhaps using confidence intervals to see the range of plausible effect sizes.

Reliance on estimated effect sizes reported in the literature is debatable: many such effects are inflated as a result of the file-drawer problem and, as such, can lead to unreasonably high expectations about power.

The `WebPower` package in **R** offers a comprehensive solution for conducting power studies, as is the free software G*Power.

7.2.2 Power in complex designs

In cases where an analytic derivations isn't possible, we can resort to simulations to approximate the power. For a given alternative, we

- simulate repeatedly samples from the model from the hypothetical alternative world
- we compute the test statistic for each of these new samples
- we transform these to the associated p-values based on the postulated null hypothesis.

At the end, we calculate the proportion of tests that lead to a rejection of the null hypothesis at level α, namely the percentage of p-values smaller than α. We can vary the sample size and see how many observations we need per group to achieve the desired level of power.

! Summary

- Effect sizes are used to provide a standardized measure of the strength of a result, independent of the design and the sample size.
- There are two classes: standardized differences and proportions of variance.
- Multiple estimators exists: report the latter along with the software used to compute confidence intervals.
- The adequate measure of variability to use for the effect size depends on the design: we normally include the variability of blocking factors and residual variance.
- Given a design, we can deduce either the sample size, the power or the effect size from the other two metrics. This allows us to compute sample size for a study or replication.

8 Replication crisis

In recent years, many team efforts have performed so-called replications of existing methodological papers to assess the robustness of their findings. Perhaps unsurprisingly, many replications failed to yield anything like what authors used to claim, or found much weaker findings. This chapter examines some of the causes of this lack of replicability.

> **!** Learning objectives
>
> - Defining replicability and reproducibility.
> - Understanding the scale of the replication crisis.
> - Recognizing common statistical fallacies.
> - Listing strategies for enhancing reproducibility.

We adopt the terminology of Claerbout and Karrenbach (1992): a study is said to be **reproducible** if an external person with the same data and enough indications about the procedure (for example, the code and software versions, etc.) can obtain consistent results that match those of a paper. A related scientific matter is **replicability**, which is the process by which new data are collected to test the same hypothesis, potentially using different methodology. Reproducibility is important because it enhances the credibility of one's work. Extensions that deal with different analyses leading to the same conclusion are described in The Turing Way and presented in Figure 8.1.

Why is reproducibility and replicability important? In a thought provoking paper, Ioannidis (2005) claimed that most research findings are wrong. The abstract of his paper stated

> There is increasing concern that most current published research findings are false. [...] In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance.

Since its publication, collaborative efforts have tried to assess the scale of the problem by reanalysing data and trying to replicate the findings of published research. For example,

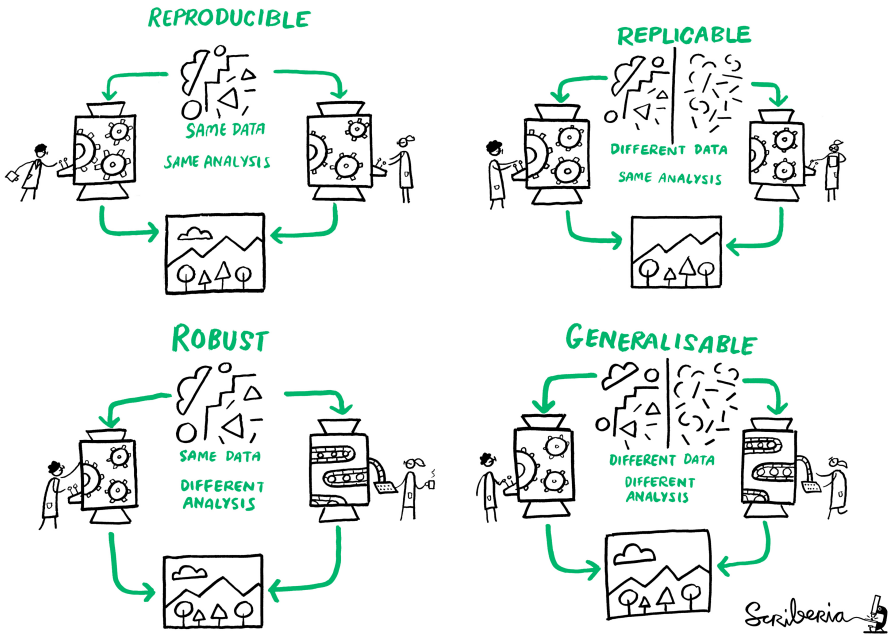


Figure 8.1: Definition of different dimensions of reproducible research (from The Turing Way project, illustration by Scriberia).

the "Reproducibility [sic] Project: Psychology" (Nosek et al. 2015)

> conducted replications of 100 experimental and correlational studies published in three psychology journals using high powered designs and original materials when available. Replication effects were half the magnitude of original effects, representing a substantial decline. Ninety seven percent of original studies had significant results. Thirty six percent of replications had significant results; 47% of original effect sizes were in the 95% confidence interval of the replication effect size; 39% of effects were subjectively rated to have replicated the original result; and, if no bias in original results is assumed, combining original and replication results left 68% with significant effects. [...]

A large share of findings in the review were not replicable or the effects were much smaller than claimed, as shown by Figure 2 from the study. Such findings show that the peer-review procedure is not foolproof: the "publish-or-perish" mindset in academia is leading many researchers to try and achieve statistical significance at all costs to meet the 5% level criterion, whether involuntarily or not. This problem has many names: p-hacking, harking or to paraphrase a story of Jorge Luis Borges, the garden of forking paths. There are many

degrees of freedom in the analysis for researchers to refine their hypothesis after viewing the data, conducting many unplanned comparisons and reporting selected results.

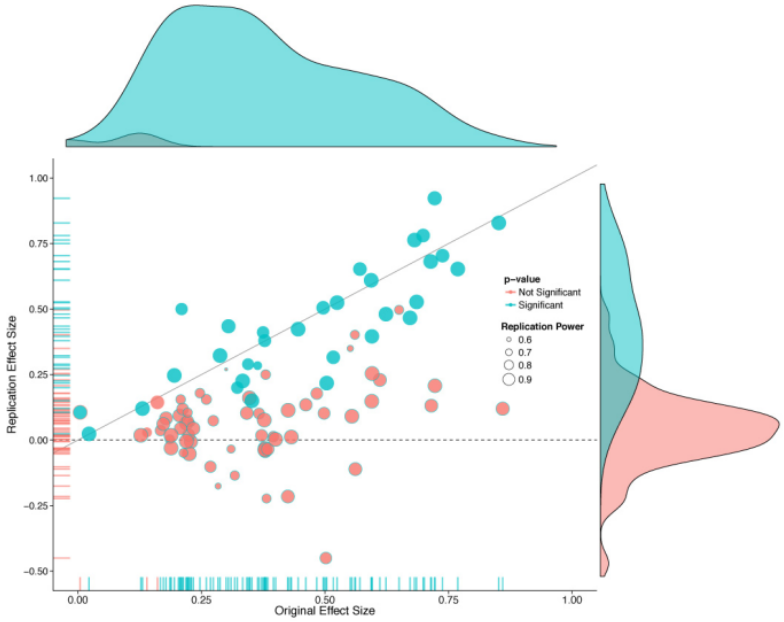


Figure 8.2: Figure 2 from Nosek et al. (2015), showing scatterplot of effect sizes for the original and the replication study by power, with rugs and density plots by significance at the 5% level.

Another problem is selective reporting. Because a large emphasis is placed on statistical significance, many studies that find small effects are never published, resulting in a gap. Figure 8.3 from Zwet and Cator (2021) shows z-scores obtained by transforming confidence intervals reported in Barnett and Wren (2019). The authors used data mining techniques to extract confidence intervals from abstracts of nearly one million publication in Medline published between 1976 and 2019. If most experiments yielded no effect and were due to natural variability, the z-scores should be normally distributed, but Figure 8.3 shows a big gap in the bell curve between approximately -2 and 2, indicative of selective reporting. The fact that results that do not lead to $p < 0.05$ are not published is called the **file-drawer** problem.

The ongoing debate surrounding the reproducibility crisis has sparked dramatic changes in the academic landscape: to enhance the quality of studies published, many journal now require authors to provide their code and data, to pre-register their studies, etc. Teams lead effort (e.g., the Experimental Economics Replication Project) try to replicate studies, with mitigated success so far. This inside recollection by a graduate student shows the extent of

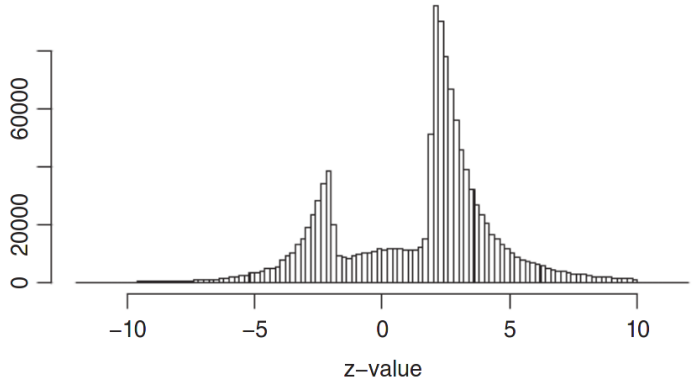


Figure 8.3: Figure from Zwet and Cator (2021) based on results of Barnett and Wren (2019); histogram of z-scores from one million studies from Medline.

the problem.

This course will place a strong emphasis on identifying and avoiding statistical fallacies and showcasing methods than enhance reproducibility. How can reproducible research enhance your work? For one thing, this workflow facilitates the publication of negative research, forces researchers to think ahead of time (and receive feedback). Reproducible research and data availability also leads to additional citations and increased credibility as a scientist.

Among good practices are

- pre-registration of experiments and use of a logbook.
- clear reporting of key aspects of the experiment (choice of metric, number of items in a Likert scale, etc.)
- version control systems (e.g., Git) that track changes to files and records.
- archival of raw data in a proper format with accompanying documentation.

Keeping a logbook and documenting your progress helps your collaborators, reviewers and your future-self understand decisions which may seem unclear and arbitrary in the future, even if they were the result of a careful thought process at the time you made them. Given the pervasiveness of the garden of forking paths, pre-registration helps you prevents harking because it limits selective reporting and unplanned tests, but it is not a panacea. Critics often object to pre-registration claiming that it binds people. This is a misleading claim in my view: pre-registration doesn't mean that you must stick with the plan exactly, but merely requires you to explain what did not go as planned if anything.

Version control keeps records of changes to your file and can help you retrieve former versions if you make mistakes at some point.

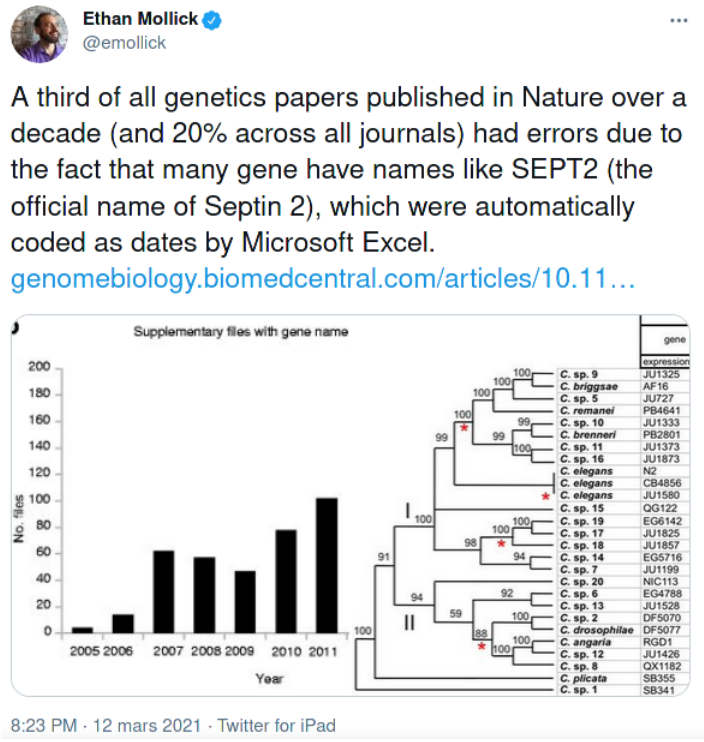


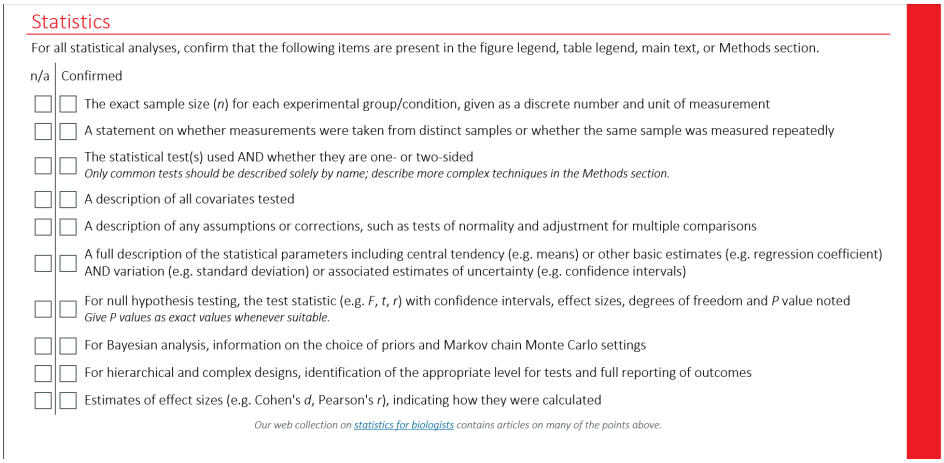
Figure 8.4: Tweet showing widespread problems related to unintentional changes to raw data by software.

Archival of data helps to avoid unintentional and irreversible manipulations of the original data, examples of which can have large scale consequences as illustrated in Figure 8.4, who report flaws in genetic journals due to the automatic conversion of gene names to dates in Excel. These problems are far from unique. While sensitive data cannot be shared "as is" because of confidentiality issues, in many instances the data can and should be made available with a licence and a DOI to allow people to reuse it, cite and credit your work.

To enforce reproducibility, many journals now have policy regarding data, material and code availability. Some journals encourage such, while the trend in recent years has been to enforce. For example, Nature require the following to be reported in all published papers:

8.1 Causes of the replication crisis

Below are multiple (non-exclusive) explanations for the lack of replication of study findings.



Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☐ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☐ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☐ A description of all covariates tested
☐	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Figure 8.5: Screenshot of the Nature Reporting summary for statistics, reproduced under the CC BY 4.0 license.

8.1.1 The garden of forking paths

The garden of forking paths, named after a novel of Borges, is a term coined by Andrew Gelman to refer to researchers' degrees of freedom. With vague hypothesis and data collection rules, it is easy for the researcher to adapt and interpret the conclusions in a way that fits his or her chosen narratives. In the words of Gelman and Loken (2014)

> Given a particular data set, it can seem entirely appropriate to look at the data and construct reasonable rules for data exclusion, coding, and analysis that can lead to statistical significance. In such a case, researchers need to perform only one test, but that test is conditional on the data.

This user case is not accomodated by classical testing theory. Research hypothesis are often formulated in a vague way, such that different analysis methods, tests may be compatible. Abel et al. (2022) recent preprint found that preregistration alone did not solve this problem, but that publication bias in randomized control trial was alleviated by publication of pre-analysis plans. This is directly related to the garden of forking path.

8.1.2 Selective reporting

Also known as the file-drawer problem, selective reporting occurs because publication of results that fail to reach statistical significance (sic) are harder to publish. In much the same way as multiple testing, if 20 researchers perform a study but only one of them writes

a paper and the result is a fluke, then this indicates. There are widespread indications publication bias, as evidence by the distribution of p-values reported in papers. A recent preprint of a study found the prevalance to be higher in online experiments such as Amazon MTurks.

P-hacking and the replication crisis has lead many leading statisticians to advocate much more stringent cutoff criterion such as $p < 0.001$ instead of the usual $p < 0.05$ criterion as level for the test. The level $\alpha = 5\%$ is essentially arbitrary and dates back to Fisher (1926), who wrote

> If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty or one in a hundred. Personally, the writer prefers to set a low standard of significance at the 5 per cent point, and ignore entirely all results which fails to reach this level.

i Thinking outside the box

Methods that pool together results, such as meta-analysis, are sensitive to selective reporting. Why does it matter?

8.1.3 Non-representative samples

Many researchers opt for convenience samples by using online panels such as Qualtrics, Amazon MTurks, etc. The quality of those observations is at best dubious: ask yourself whether you would answer such as survey for a small amount. Manipulation checks to ensure participants are following, information is not completed by bots, a threshold for the minimal time required to complete the study, etc. are necessary (but not sufficient) conditions to ensure that the data are not rubbish.

A more important criticism is that the people who answer those surveys are not representative of the population as a whole: sampling bias thus plays an important role in the conclusions and, even if the summary statistics are not too different from the general population, they may exhibit different opinions, levels of skills, etc. than most.

The same can be said of panels of students recruited in universities classes, who are more young, educated and perhaps may infer through backward induction the purpose of the study and answer accordingly.

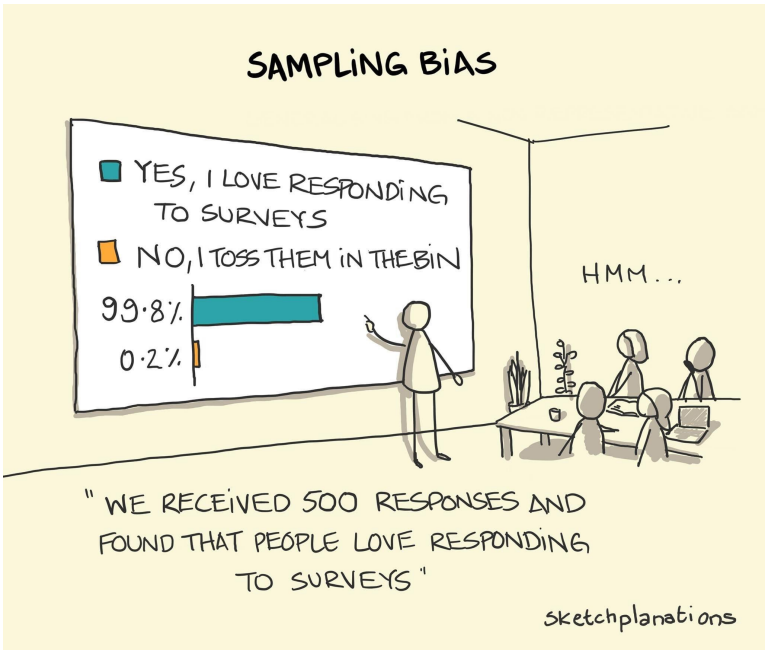


Figure 8.6: Sampling bias. Artwork by Jonathan Hey (Sketchplanations) shared under the CC BY-NC 4.0 license.

8.2 Summary

Operating in an open-science environment should be seen as an opportunity to make better science, offer more opportunities to increase your impact and increase the likelihood that your work gets published regardless of whether the results turn out to be negative. It is the *right thing* to do and it increases the quality of research produced, with collateral benefits because it forces researchers to validate their methodology before, to double-check their data and their analysis and to adopt good practice.

There are many platforms for preregistering studies and sharing preanalysis plans, scripts and data, with different level of formality. One such is the Research Box.

> **Your turn**
>
> Reflect on your workflow as applied researcher when designing and undertaking experiments. Which practical aspects could you improve upon to improve the reproducibility of your study?

9 Repeated measures and multivariate models

So far, all experiments we have considered can be classified as between-subject designs, meaning that each experimental unit was assigned to a single experimental (sub)-condition. In many instances, it may be possible to randomly assign multiple conditions to each experimental unit. For example, an individual coming to a lab to perform tasks in a virtual reality environment may be assigned to all treatments, the latter being presented in random order to avoid confounding. There is an obvious benefit to doing so, as the participants can act as their own control group, leading to greater comparability among treatment conditions.

For example, consider a study performed at Tech3Lab that looks at the reaction time for people texting or talking on a cellphone while walking. We may wish to determine whether disengagement is slower for people texting, yet we may also postulate that some elderly people have slower reflexes.

In a between-subjects design, subjects are **nested** within experimental condition, as a subject can only be assigned a single treatment. In a within-subjects designs, experimental factors and subjects are **crossed**: it is possible to observed all combination of subject and experimental conditions.

By including multiple conditions, we can filter out effect due to subject, much like with blocking: this leads to increased precision of effect sizes and increased power (as we will see, hypothesis tests are based on within-subject variability). Together, this translates into the need to gather fewer observations or participants to detect a given effect in the population and thus experiments are cheaper to run.

There are of course drawbacks to gathering repeated measures from individuals. Because subjects are confronted with multiple tasks, there may be carryover effects (when one task influences the response of the subsequent ones, for example becoming more fluent as manipulations go on), period effects (practice of fatigue, e.g., leading to a decrease in acuity), permanent changes in the subject condition after a treatment or attrition (loss of subjects over time).

To minimize potential biases, there are multiple strategies one can use. While can randomize the order of treatment conditions among subjects to reduce confounding, or use a balanced crossover design and include the period and carryover effect in the statistical

model via control variables so as to better isolate the treatment effect. The experimenter should also allow enough time between treatment conditions to reduce or eliminate period or carryover effects and plan tasks accordingly.

Due to fatigue or learning effects, randomization of the order of the within-subject experimental conditions. If each is assigned a single time, one good way to do this is via **counterbalancing**. We proceed as follows: first, enumerate all possible orders of the condition and then assign participants as equally as possible between conditions. For example, with a single within-factor design with three conditions A, B, C, we have six possible orderings (either ABC, ACB, BAC, BCA, CAB or CBA). Much like other forms of randomization, this helps us remove confounding effects and let's us estimate what is the average effect of task ordering on the response.

There are multiple approaches to handling repeated measures. The first option is to take averages over experimental condition per subject and treat them as additional blocking factors, but it may be necessary to adjust the resulting statistics. The second approach consists in fitting a multivariate model for the response and explicitly account for the correlation, otherwise the null distribution commonly used are off and so are the conclusions, as illustrated with the absurd comic displayed in Figure 11.1.

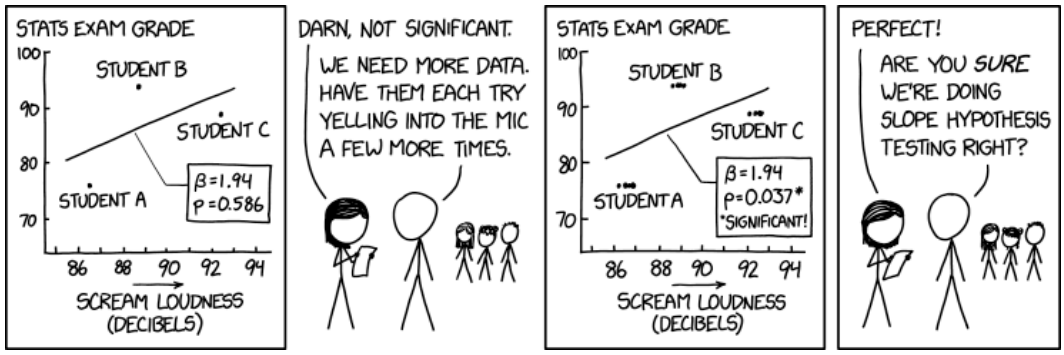


Figure 9.1: xkcd comic 2533 (Slope Hypothesis Testing) by Randall Munroe. Alt text: What? I can't hear– I said, are you sure–; CAN YOU PLEASE SPEAK–. Cartoon reprinted under the CC BY-NC 2.5 license.

Multivariate analysis of variance (MANOVA) leads to procedures that are analogous to univariate analysis of variance, but we now need to estimate correlation and variance parameters for each measurement separately and there are multiple potential statistics that can be defined for testing effects. While we can benefit from the correlation and find differences that wouldn't be detected from univariate models, the additional parameters to estimate lead to a loss of power. Finally, the most popular method nowadays for handling repeated measures is to fit a mixed model, with random effects accounting to subject-specific characteristics. By doing so, we assume that the levels of a factor (here the subject

identifiers) form a random sample from a large population. These models can be difficult to fit and one needs to take great care in specifying the model.

9.1 Repeated measures

We introduce the concept of repeated measure and within-subject ANOVA with an example.

We consider an experiment conducted in a graduate course at HEC, *Information Technologies and Neuroscience*, in which PhD students gathered electroencephalography (EEG) data. The project focused on human perception of deepfake image created by a generative adversarial network: Amirabdolahian and Ali-Adeeb (2021) expected the attitude towards real and computer generated image of people smiling to change.

The response variable is the amplitude of a brain signal measured at 170 ms after the participant has been exposed to different faces. Repeated measures were collected on 9 participants given in the database `AA21`, who were expected to look at 120 faces. Not all participants completed the full trial, as can be checked by looking at the cross-tabs of the counts

```
data(AA21, package = "hecedsm")
xtabs(~stimulus + id, data = AA21)
```

```
        id
stimulus  1  2  3  4  5  6  7  8  9 10 11 12
    real 30 32 34 32 38 29 36 36 40 30 39 33
    GAN1 32 31 40 33 38 29 39 31 39 28 35 34
    GAN2 31 33 37 34 38 29 34 36 40 33 35 32
```

The experimental manipulation is encoded in the `stimuli`, with levels control (`real`) for real facial images, whereas the others were generated using a generative adversarial network (GAN) with be slightly smiling (`GAN1`) or extremely smiling (`GAN2`); the latter looks more fake. While the presentation order was randomized, the order of presentation of the faces within each type is recorded using the `epoch` variable: this allows us to measure the fatigue effect.

Since our research question is whether images generated from generative adversarial networks trigger different reactions, we will be looking at pairwise differences with the control.

(a) real

(b) slightly modified

(c) extremely modified

Figure 9.2: Example of faces presented in Amirabdolahian and Ali-Adeeb (2021).

We could first grouping the data and compute the average for each experimental condition `stimulus` per participant and set `id` as blocking factor. The analysis of variance table obtained from `aov` would be correct, but would fail to account for correlation.

The one-way analysis of variance with n_s subjects, each of which was exposed to the n_a experimental conditions, can be written

$$\underset{\text{response}}{Y_{ij}} = \underset{\text{global mean}}{\mu} + \underset{\text{mean difference}}{\alpha_j} + \underset{\text{subject difference}}{s_i} + \underset{\text{error}}{\varepsilon_{ij}}$$

```
# Compute mean for each subject +
# experimental condition subgroup
AA21_m <- AA21 |>
  dplyr::group_by(id, stimulus) |>
  dplyr::summarize(latency = mean(latency))
# Use aov for balanced sample
fixedmod <- aov(
  latency ~ stimulus + Error(id/stimulus),
  data = AA21_m)
# Print ANOVA table
summary(fixedmod)
```

```
Error: id
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals 11  187.8   17.07

Error: id:stimulus
          Df Sum Sq Mean Sq F value Pr(>F)
stimulus   2   1.94  0.9704   0.496  0.615
Residuals 22  43.03  1.9557
```

Since the design is balanced after averaging, we can use `aov` in **R**: we need to specify the subject identifier within `Error` term. This approach has a drawback, as variance components can be negative if the variability due to subject is negligible. While `aov` is fast, it only works for simple balanced designs.

9.1.1 Contrasts

With balanced data, the estimated marginal means coincide with the row averages. If we have a single replication or the average for each subject/condition, we could create a new column with the contrast and then fit a model with an intercept-only (global mean) to check whether the latter is zero. With 12 participants, we should thus expect our test statistic to have 11 degrees of freedom, since one unit is spent on estimating the mean parameter and we have 12 participants.

Unfortunately, the `emmeans` package analysis for object fitted using `aov` will be incorrect: this can be seen by passing a contrast vector and inspecting the degrees of freedom. The `afex` package includes functionalities that are tailored for within-subject and between-subjects and has an interface with `emmeans`.

```
afexmod <- afex::aov_ez(
  id = "id",           # subject id
  dv = "latency",      # response variable
  within = "stimulus", # within-subject factor
  data = AA21,
  fun_aggregate = mean)
```

The `afex` package has different functions for computing the within-subjects design and the `aov_ez` specification, which allow people to list within and between-subjects factor separately with subject identifiers may be easier to understand. It also has an argument, `fun_aggregate`, to automatically average replications.

```
# Set up contrast vector
cont_vec <- list(
  "real vs GAN" = c(1, -0.5, -0.5))
library(emmeans)
# Correct output
afexmod |>
  emmeans::emmeans(
    spec = "stimulus",
    contr = cont_vec)
```

```
$emmeans
 stimulus emmean    SE df lower.CL upper.CL
 real      -10.8 0.942 11    -12.8    -8.70
 GAN1      -10.8 0.651 11    -12.3    -9.40
 GAN2      -10.3 0.662 11    -11.8    -8.85

Confidence level used: 0.95

$contrasts
 contrast    estimate    SE df t.ratio p.value
 real vs GAN   -0.202 0.552 11  -0.366  0.7213
```

```
# Incorrect output -
# note the wrong degrees of freedom
fixedmod |>
  emmeans::emmeans(
    spec = "stimulus",
    contr = cont_vec)
```

```
Note: re-fitting model with sum-to-zero contrasts


$emmeans
 stimulus emmean    SE   df lower.CL upper.CL
 real      -10.8 0.763 16.2    -12.4    -9.15
 GAN1      -10.8 0.763 16.2    -12.4    -9.21
 GAN2      -10.3 0.763 16.2    -11.9    -8.69

Warning: EMMs are biased unless design is perfectly balanced
Confidence level used: 0.95

$contrasts
 contrast    estimate    SE df t.ratio p.value
 real vs GAN   -0.202 0.494 22  -0.409  0.6867
```

9.1.2 Sphericity assumption

The validity of the F statistic null distribution relies on the model having the correct structure.

In repeated-measure analysis of variance, we assume again that each measurement has the same variance. We equally require the correlation between measurements of the same subject to be the same, an assumption that corresponds to the so-called compound symmetry model.[1]

What if the within-subject measurements have unequal variance or the correlation between those responses differs?

Since we care only about differences in treatment, can get away with a weaker assumption than compound symmetry (equicorrelation) by relying instead on *sphericity*, which holds if the variance of the difference between treatment is constant. Sphericity is not a relevant concept when there is only two measurements (as there is a single correlation); we could check this by comparing the fit of a model with an unstructured covariance (difference variances for each and correlations for each pair of variable)

The most popular approach to handling correlation in tests is a two-stage approach: first, check for sphericity (using, e.g., Mauchly's test of sphericity). If the null hypothesis of sphericity is rejected, one can use a correction for the F statistic by modifying the parameters of the Fisher F null distribution used as benchmark.

An idea due to Box is to correct the degrees of freedom of the $\mathsf{F}(\nu_1, \nu_2)$ distribution by multiplying them by a common factor $\epsilon < 1$ and use $\mathsf{F}(\epsilon\nu_1, \epsilon\nu_2)$ as null distribution instead to benchmark our statistics and determine how extreme our observed one is. Since the F statistic is a ratio of variances, the ϵ terms would cancel. Using the scaled F distribution leads to larger p-values, thus accounting for the correlation.

There are three widely used corrections: Greenhouse–Geisser, Huynh–Feldt and Box correction, which divides by ν_1 both degrees of freedom and gives a very conservative option. The Huynh–Feldt method is reported to be more powerful so should be preferred, but the estimated value of ϵ can be larger than 1.

Using the `afex` functions, we get the result for Mauchly's test of sphericity and the p values from using either correction method

```
summary(afexmod)
```

```
Univariate Type III Repeated-Measures ANOVA Assuming Sphericity

            Sum Sq num Df Error SS den Df  F value    Pr(>F)
(Intercept) 4073.1      1  187.814     11 238.5554 8.373e-09 ***
```

[1] Note that, with two measurements, there is a single correlation parameter to estimate and this assumption is irrelevant.

```
stimulus       1.9      2   43.026     22   0.4962    0.6155
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Mauchly Tests for Sphericity

         Test statistic p-value
stimulus        0.67814 0.14341


Greenhouse-Geisser and Huynh-Feldt Corrections
 for Departure from Sphericity

          GG eps Pr(>F[GG])
stimulus 0.75651     0.5667

            HF eps Pr(>F[HF])
stimulus 0.8514944  0.5872648
```

We consider a model with both within-subject and between-subject factors. Data for a study on visual acuity of participants. The data represent the number of words correctly detected at different font size; interest is in effect of illusory contraction on detection. The mixed analysis of variance includes the experimental factors `adaptation` (2 levels, within), `fontsize` (4 levels, within), `position` (5 levels, within) and visual `acuity` (2 levels, between). There are a total of 1760 measurements for 44 participants in `LBJ17_S1A`, balanced. The within-subject factors give a total of 40 measurements ($2 \times 4 \times 5$) per participant; all of these factors are crossed and we can estimate interactions for them. The subjects are nested within visual acuity groups, The participants were dichotomized in two groups based on their visual acuity, obtained from preliminary checks, using a median split.

To fit the model, we rely on the `aov_ez` function from `afex`. By default, the latter includes all interactions.

```
LBJ_mod <- afex::aov_ez(
  id = "id",     # subject id
  dv = "nerror", # response
  between = "acuity",
  within = c("adaptation",
             "fontsize",
```

Table 9.1: Analysis of variance for the four-way model with partial effect sizes (partial eta-square)

	df1	df2	F	pes	p-value
acuity	1	42	30.8	0.42	<0.001
adaptation	1	42	7.8	0.16	0.008
acuity:adaptation	1	42	12.7	0.23	<0.001
fontsize	3	126	1705.7	0.98	<0.001
acuity:fontsize	3	126	10.0	0.19	<0.001
position	4	168	9.4	0.18	<0.001
acuity:position	4	168	4.2	0.09	0.003
adaptation:fontsize	3	126	3.3	0.07	0.023
acuity:adaptation:fontsize	3	126	7.0	0.14	<0.001
adaptation:position	4	168	0.6	0.01	0.662
acuity:adaptation:position	4	168	0.9	0.02	0.464
fontsize:position	12	504	9.1	0.18	<0.001
acuity:fontsize:position	12	504	2.7	0.06	0.002
adaptation:fontsize:position	12	504	0.5	0.01	0.907
acuity:adaptation:fontsize:position	12	504	1.2	0.03	0.295

```
            "position"),
  data = hecedsm::LBJ17_S1A)
anova_tbl <- anova(LBJ_mod,  # model
      correction = "none", # no correction for sphericity
      es = "pes")
#partial eta-square for effect sizes (es)
```

This is the most complicated model we tested so far: there are four experimental factor being manipulated at once, and all interactions of order two, three and four are included!

The fourth order interaction isn't statistically significant: this means that we can legitimately marginalize over and look at each of the four three-way ANOVA designs in turn. We can also see that the third order interaction `adaptation:fontsize:position` and `acuity:adaptation:position` are not really meaningful.

The following paragraph is technical and can be skipped. One difficult bit with designs including both within-subject and between-subject factors is the degrees of freedom and the correct sum of square terms to use to calculate the F statistics for each hypothesis of interest. The correct setup is to use the next sum of square (and the associated degrees of

freedom) from this. For any main effect or interaction, we count the number of instances of this particular (e.g., 10 for the interaction between position and adaptation). We subtract the number of mean parameter used to estimate means and differences in mean (1 global mean, 4 means for position, 1 for adaptation), which gives $4 = 10 - 6$ degrees of freedom. Next, this term is compared to the mean square which contains only subject (here via acuity levels, since subjects are nested within acuity) and the corresponding variables; the correct mean square is for `acuity:adaptation:position`. In the balanced design setting, this can be formalized using Hasse diagram (Oehlert 2000).

We can produce an interaction plot to see what comes out: since we can't draw in four dimensions, we map visual acuity and adaptation level to panels with different colours for the position. The figure looks different from the paper, seemingly because their y-axis is flipped.

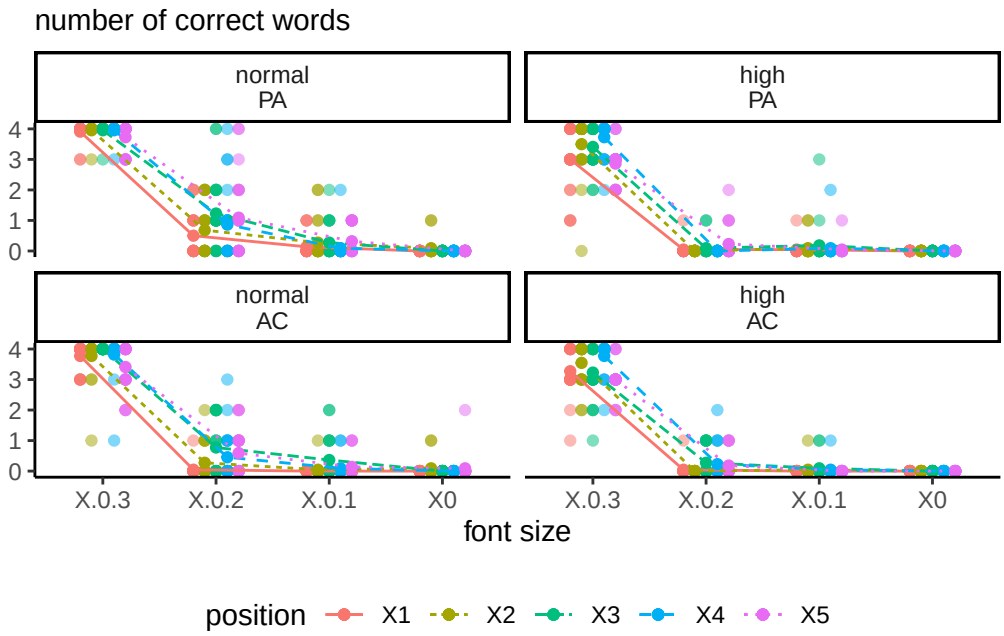


Figure 9.3: Interaction plot for visual acuity levels.

9.2 Multivariate analysis of variance

The second paradigm for modelling is to specify that the response from each subject is in fact a multivariate object: we can combine all measurements from a given individual in a

vector $\boldsymbol{Y}$. In the example with the happy fakes, this would be the tuple of measurements for (`real`, `GAN1`, `GAN2`).

The multivariate analysis of variance model is designed by assuming observations follow a (multivariate) normal distribution with mean vector $\boldsymbol{\mu}_j$ in group j and common covariance matrix $\boldsymbol{\Sigma}$ and comparing means between groups. As in univariate analysis of variance, the multivariate normal assumption holds approximately by virtue of the central limit theorem in large samples, but the convergence is slower and larger numbers are needed to ensure this is valid.

The difference with the univariate approach is now that we will compare a global mean vector $\boldsymbol{\mu}$ between comparisons. In the one-way analysis of variance model with an experimental factor having K levels and a balanced sample n_g observations per group and $n = n_g K$ total observations, we assume that each group has average $\boldsymbol{\mu}_k$ $(k = 1, \ldots, K)$, which we can estimate using only the observations from that group. Under the null hypothesis, all groups have the same mean, so the estimator is the overall mean $\boldsymbol{\mu}$ combining all n observations.

The statistic is obtained by decomposing the total variance around the global mean into components due to the different factors and the leftover variability. Because these equivalent to the sum of square decomposition results in multiple matrices, there are multiple ways of constructing test statistics. Wilk's Λ is the most popular choice. Another common choice, which leads to a statistic giving lower power but which is also more robust to departure from model assumptions is Pillai's trace.

The MANOVA model assumes that the covariance matrices are the same within each experimental condition. We can use Box's M statistic to test the normality hypothesis.

9.2.1 Data format

With repeated measures, it is sometimes convenient to store measurements associated to each experimental condition in different columns of a data frame or spreadsheet, with lines containing participants identifiers. Such data are said to be in **wide format**, since there are multiple measurements in each row. While this format is suitable for storate, many statistical routines will instead expect data to be in **long format**, for which there is a single measurement per line. Figure 9.4 illustrates the difference between the two formats.

Ideally, a data base in long format with repeated measures would also include a column giving the order in which the treatments were assigned to participants. This is necessary in order to test whether there are fatigue or crossover effects, for example by plotting the residuals after accounting for treatment subject by subject, ordered over time. We could

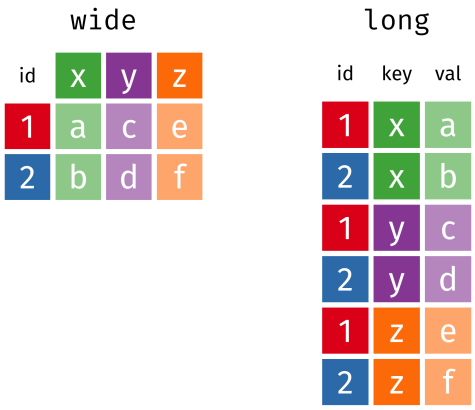


Figure 9.4: Long versus wide-format for data tables (illustration by Garrick Aden-Buie).

also perform formal tests by including time trends in the model and checking whether the slope is significant.

Overall, the biggest difference with within-subject designs is that observations are correlated whereas we assumed measurements were independent until now. This needs to be explicitly accounted for, as correlation has an important impact on testing as discussed Section 3.4.4: failing to account for correlation leads to p-values that are much too low. To see why, think about a stupid setting under which we duplicate every observation in the database: the estimated marginal means will be the same, but the variance will be halved despite the fact there is no additional information. Intuitively, correlation reduces the amount of information provided by each individual: if we have repeated measures from participants, we expect the effective sample size to be anywhere between the total number of subjects and the total number of observations.

9.2.2 Mathematical complement

This section is technical and can be omitted. Analogous to the univariate case, we can decompose the variance estimator in terms of within, between and total variance. Let $\boldsymbol{Y}_{ik}$ denote the response vector for the ith observation of group k; then, we can decompose the

variance as

$$\underbrace{\sum_{k=1}^{K}\sum_{i=1}^{n_g}(\boldsymbol{Y}_{ik}-\widehat{\boldsymbol{\mu}})(\boldsymbol{Y}_{ik}-\widehat{\boldsymbol{\mu}})^\top}_{\text{total variance}}$$
$$=\underbrace{\sum_{k=1}^{K}\sum_{i=1}^{n_g}(\boldsymbol{Y}_{ik}-\widehat{\boldsymbol{\mu}}_k)(\boldsymbol{Y}_{ik}-\widehat{\boldsymbol{\mu}}_k)^\top}_{\text{within variance}}+\underbrace{\sum_{k=1}^{K}n_g(\boldsymbol{\mu}_k-\widehat{\boldsymbol{\mu}})(\widehat{\boldsymbol{\mu}}_k-\widehat{\boldsymbol{\mu}})^\top}_{\text{between variance}}$$

defining covariance matrix estimators. If we write $\widehat{\boldsymbol{\Sigma}}_T$, $\widehat{\boldsymbol{\Sigma}}_W$, and $\widehat{\boldsymbol{\Sigma}}_B$ for respectively the total, within and between variance estimators, we can build a statistic from these ingredients to see how much variability is induced by centering using a common vector. When $K > 2$, there are multiple statistics that be constructed, including

- Wilk's Λ: $|\widehat{\boldsymbol{\Sigma}}_W|/|\widehat{\boldsymbol{\Sigma}}_W + \widehat{\boldsymbol{\Sigma}}_B|$
- Roy's maximum root: the largest eigenvalue of $\widehat{\boldsymbol{\Sigma}}_W^{-1}\widehat{\boldsymbol{\Sigma}}_B$
- Lawley–Hotelling trace: $\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_W^{-1}\widehat{\boldsymbol{\Sigma}}_B)$
- Pillai's trace: $\mathrm{tr}\left\{\widehat{\boldsymbol{\Sigma}}_B(\widehat{\boldsymbol{\Sigma}}_W + \widehat{\boldsymbol{\Sigma}}_B)^{-1}\right\}$.

All four criteria lead to equivalent statistics and the same p-values if $K = 2$.

With a two-way balanced MANOVA, we can perform a similar decomposition for each factor or interaction, with

$$\widehat{\boldsymbol{\Sigma}}_T = \widehat{\boldsymbol{\Sigma}}_A + \widehat{\boldsymbol{\Sigma}}_B + \widehat{\boldsymbol{\Sigma}}_{AB} + \widehat{\boldsymbol{\Sigma}}_W.$$

Wilk's Λ is based on taking the ratio of the determinant of the within-variance and that of the sum of effect-variance plus within-variance, e.g., $|\widehat{\boldsymbol{\Sigma}}_{AB} + \widehat{\boldsymbol{\Sigma}}_W|$ for the interaction term.

9.2.3 Model fitting

We can treat the within-subject responses as a vector of observations and estimate the model using using multivariate linear regression. Contrary to the univariate counterpart, the model explicitly models the correlation between observations from the same subject.

In order to fit a model with a multivariate response, we first need to pivot the data into wider format so as to have a matrix with rows for the number of subjects and M columns for the number of response variables.

Once the data are in a suitable format, we fit the multivariate model with the `lm` function using the sum-to-zero constraints, here imposed globally by changing the `contrasts` option.

Syntax-wise, the only difference with the univariate case is that the response on the left of the tilde sign (`~`) is now a matrix composed by binding together the vectors with the different responses.

We use the data from Amirabdolahian and Ali-Adeeb (2021), but this time treating the averaged repeated measures for the different stimulus as a multivariate response. We first pivot the data to wide format, then fit the multivariate linear model.

```
# Pivot to wide format
AA21_mw <- AA21_m |>
  tidyr::pivot_wider(names_from = stimulus, # within-subject factor labels
                     values_from = latency) # response measurements
# Model with each variable with a different mean
# Specify all columns with column bind
# left of the ~, following
options(contrasts = c("contr.sum", "contr.poly"))
mlm <- lm(cbind(real, GAN1, GAN2) ~ 1,
          data = AA21_mw)
```

Since the within-subject factor `stimulus` disappeared when we consider the multivariate response, we only specify a global mean vector $\boldsymbol{\mu}$ via `~1`. In general, we would add the between-subject factors to the right-hand side of the equation. Our hypothesis of equal mean translates into the hypothesis $\boldsymbol{\mu} = \mu \mathbf{1}_3$, which can be imposed using a call to `anova`. The output returns the statistic and p-values including corrections for sphericity.

We can also use `emmeans` to set up post-hoc contrasts. Since we have no variable, we need to set in `specs` the repeated measure variable appearing on the left hand side of the formula; the latter is labelled `rep.meas` by default.

```
# Test the multivariate model against
# equal mean (X = ~1)
anova(mlm, X = ~1, test = "Spherical")
```

```
Analysis of Variance Table


Contrasts orthogonal to
~1

Greenhouse-Geisser epsilon: 0.7565
Huynh-Feldt epsilon:        0.8515
```

```
            Df      F num Df den Df  Pr(>F)  G-G Pr  H-F Pr
(Intercept)  1 0.4962      2     22 0.61549 0.56666 0.58726
Residuals   11
```

```
# Follow-up contrast comparisons
library(emmeans)
emm_mlm <- emmeans(mlm, specs = "rep.meas")
emm_mlm |> contrast(method = list(c(1,-0.5,-0.5)))
```

```
 contrast          estimate    SE df t.ratio p.value
 c(1, -0.5, -0.5)   -0.202 0.552 11  -0.366  0.7213
```

We can check that the output is the same in this case as the within-subject analysis of variance model fitted previously with the `afex` package.

We consider a between-subject repeated measure multivariate analysis of variance model with the Baumann, Seifert-Kessell, and Jones (1992). The data are balanced by experimental condition and they include the results of three tests performed after the intervention: an error detection task, an expanded comprehension monitoring questionnaire and a cloze test. Note that the scale of the tests are different (16, 18 and 56).

We could obtain the estimated covariance matrix of the fitted model by extracting the residuals $Y_{ik} - \widehat{\mu}_k$ and computing the empirical covariance. The results shows a strong dependence between tests 1 and 3 (correlation of 0.39), but much weaker dependence with test 2.

Let us compute the multivariate analysis of variance model

```
data(BSJ92, package = "hecedsm")
# Force sum-to-zero parametrization
options(contrasts = c("contr.sum", "contr.poly"))
# Fit MANOVA model
mmod <- lm(
  cbind(posttest1, posttest2, posttest3) ~ group,
   data = BSJ92)
# Calculate multivariate test
mtest <- car::Anova(mmod, test = "Wilks")
# mtest
# Get all statistics and univariate tests
summary(car::Anova(mmod), univariate = TRUE)
```

```
Type II MANOVA Tests:

Sum of squares and products for error:
          posttest1  posttest2 posttest3
posttest1 640.50000   30.77273  498.3182
posttest2  30.77273  356.40909 -104.3636
posttest3 498.31818 -104.36364 2511.6818

----------------------------------------

Term: group

Sum of squares and products for the hypothesis:
           posttest1 posttest2 posttest3
posttest1 108.121212  6.666667 190.60606
posttest2   6.666667 95.121212  56.65152
posttest3 190.606061 56.651515 357.30303

Multivariate Tests: group
                 Df test stat approx F num Df den Df     Pr(>F)
Pillai            2 0.4082468 5.300509      6    124 6.7654e-05 ***
Wilks             2 0.6320200 5.243287      6    122 7.7744e-05 ***
Hotelling-Lawley  2 0.5185169 5.185169      6    120 8.9490e-05 ***
Roy               2 0.3184494 6.581288      3     62 0.00062058 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

 Type II Sums of Squares
          df posttest1 posttest2 posttest3
group      2    108.12    95.121     357.3
residuals 63    640.50   356.409    2511.7

 F-tests
      posttest1 posttest2 posttest3
group      5.32      8.41      4.48

 p-values
      posttest1  posttest2  posttest3
group 0.00734676 0.00058043 0.01515115
```

By default, we get Pillai's trace statistic. Here, there is clear evidence of differences between groups of observations regardless of the statistic being used.

We can compute effect size as before by passing the table, for example using `eta_squared(mtest)` to get the effect size of the multivariate test, or simple the model to get the individual variable effect sizes.

Having found a difference, one could in principle investigate for which component of the response they are by performing univariate analysis of variance and accounting for multiple testing using, e.g., Bonferroni's correction. A more fruitful avenue if you are trying to discriminate is to use descriptive discriminant analysis as a follow-up, which computes the best fitting hyperplanes that separate groups.

```
MASS::lda(group ~ posttest1 + posttest2 + posttest3,
          data = BSJ92)
```

This amounts to compute the weights $\boldsymbol{w}$ such, that, computing $\boldsymbol{w}^\top \boldsymbol{Y}$ creating a composite score by adding up weighted components that leads to maximal separation between groups. Figure 9.5 shows the new coordinates.

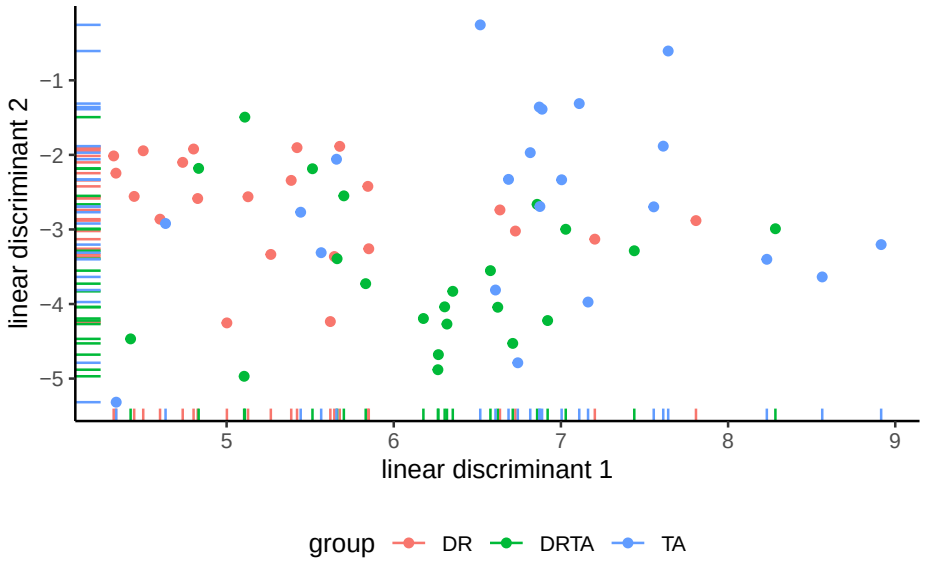


Figure 9.5: Scatterplot of observations projected onto the linear discriminants for the post-experiment tests, by group.

Linear discriminant analysis is a topic on it's own that is beyond the scope of the course.

9.2.4 Model assumptions

In addition to the usual model assumptions (independence of measurements from different subjects, equal variance, additivity, etc.), the MANOVA model adds two hypothesis that altogether determine how reliable our p-values and conclusions are.

The first assumption is that of multivariate normality of the response. The central limit theorem can be applied to a multivariate response, but the sample size needed overall to reliably estimate the correlation and variance is larger than in the univariate setting. This hypothesis can be tested using the Shapiro-Wilk normality test (null hypothesis is that of normality) by passing the residuals of the multivariate model. Such a test can lead to rejection of the null hypothesis when specific variables are far from normal, or when the dependence structure isn't the one exhibited by a multivariate normal model. With decent sample sizes (say $n = 50$ per group), this assumption isn't as important as others.

```
# Shapiro-Wilk normality test
# Must transpose the residuals
# to get a 3 by n matrix
mvnormtest::mshapiro.test(U = t(resid(mmod)))
```

```
    Shapiro-Wilk normality test

data:  Z
W = 0.96464, p-value = 0.05678
```

The second assumption is that the covariance matrix is the same for all individuals, regardless of their experimental group assignment. We could try checking whether a covariance model in each group: under multivariate normal assumption, this leads to a test statistic called Box's M test. Unfortunately, this test is quite sensitive to departures from the multivariate normal assumption and, if the p-value is small, it may have to do more with the normality than the heterogeneity.

```
with(BSJ92,
     biotools::boxM(
       data = cbind(posttest1, posttest2, posttest3),
       grouping = group))
```

```
    Box's M-test for Homogeneity of Covariance Matrices

data:  cbind(posttest1, posttest2, posttest3)
Chi-Sq (approx.) = 15.325, df = 12, p-value = 0.2241
```

In our example, there is limited evidence against any of those model assumptions. We should of course also check the assumptions of the analysis of variance model for each of `postest1`, `posttest2` and `posttest3` in turn; such a check is left as an exercice to the reader.

9.2.5 Power and effect size

Since all of the multivariate statistics can be transformed for a comparison with a univariate F distribution, we can estimate partial effect size as before. The package `effectsize` offers a measure of partial $\widehat{\eta}^2$ for the multivariate tests.[2]

Power calculations are beyond the reach of ordinary software as one needs to specify the variance of each observation, their correlation and their mean. Simulation is an obvious way for this kind of design to obtain answers, but the free **G*Power** software (Faul et al. 2007) also offers some tools. See also Läuter (1978) for pairwise comparisons.

[2] I must confess I haven't checked whether the output is sensical.

10 Introduction to mixed models

This chapter considers tools for models with repeated measures from a modern perspective, using random effects for modelling. This class of model, called hierarchical models, multilevel models or mixed models in simple scenarios, give us more flexibility to account for complex scenarios in which there may be different sources of variability.

For example, consider a large-scale replication study about teaching methods. We may have multiple labs partaking in a research program and each has unique characteristics. Because of these, we can expect that measurements collected within a lab will be correlated. At the same time, we can have repeated mesures for participants in the study. One can view this setup as a hierarchy, with within-subject factor within subject within lab. In such settings, the old-school approach to analysis of variance becomes difficult, if not impossible; it doesn't easily account for the heterogeneity in the lab sample size and does not let us estimate the variability within labs.

We begin our journey with the same setup as for repeated measures ANOVA by considering one-way within-subject ANOVA model. We assign each participant (subject) in the study to all of the experimental treatments, in random order. If we have one experimental factor A with n_a levels, the model is

$$\underset{\text{response}}{Y_{ij}} = \underset{\text{global mean}}{\mu} + \underset{\text{mean difference}}{\alpha_j} + \underset{\text{random effect for subject}}{S_i} + \underset{\text{error}}{\varepsilon_{ij}}\,.$$

In a random effect model, we assume that the subject effect S_i is a random variable; we take $S_i \sim \mathsf{No}(0, \sigma_s^2)$ and the latter is assumed to be independent of the noise $\varepsilon_{ij} \sim \mathsf{No}(0, \sigma_e^2)$. The model parameters that we need to estimate are the global mean μ, the mean differences $\alpha_1, \ldots, \alpha_{n_a}$, the subject-specific variability σ_s^2 and the residual variability σ_e^2, with the sum-to-zero constraint $\alpha_1 + \cdots + \alpha_{n_a} = 0$.

Inclusion of random effects introduces positive correlation between measurements: specifically, the correlation between two observations from the same subject will be $\rho = \sigma_s^2/(\sigma_s^2 + \sigma_e^2)$ and zero otherwise. This correlation structure is termed compound symmetry, since the correlation between measurements, ρ, is the same regardless of the order of the observations. If there are multiple random effects, the dependence structure will be more complicated.

In the repeated measure models, we need to first reduce measurements to a single average per within-subject factor, then fit the model by including the subject as a blocking factor. We are therefore considering subjects as fixed effects by including them as blocking factors, and estimate the mean effect for each subject: the value of σ_s^2 is estimated from the mean squared error of the subject term, but this empirical estimate can be negative. By contrast, the mixed model machinery will directly estimate the variance term, which will be constrained to be strictly positive.

10.1 Fixed vs random effects

Mixed models include, by definition, both **random** and **fixed** effects. Fixed effects are model parameters corresponding to overall average or difference in means for the experimental conditions. These are the terms for which we want to perform hypothesis tests and compute contrasts. So far, we have only considered models with fixed effects.

Random effects, on the other hand, assumes that the treatments are random samples from a population of interest. If we gathered another sample, we would be looking at a new set of treatments. Random effects model the variability arising from the sampling of that population and focuses on variance and correlation parameters. Addition of random effects does not impact the population mean, but induces variability and correlation within subject. There is no consensual definition, but Gelman (2005) lists a handful:

> When a sample exhausts the population, the corresponding variable is fixed; when the sample is a small (i.e., negligible) part of the population the corresponding variable is random [Green and Tukey (1960)].
>
> Effects are fixed if they are interesting in themselves or random if there is interest in the underlying population (e.g., Searle, Casella and McCulloch [(1992), Section 1.4])

In terms of estimation, fixed effect terms are mean parameters, while all random effects will be obtained from variance and correlation parameters. In the repeated measure approach with fixed effects and blocking, we would estimate the average for each subject despite the fact that this quantity is of no interest. Estimating a mean with only a handful of measurements is a risky business and the estimated effects are sensitive to outliers.

Random effects would proceed to directly estimate the variability arising from different subjects. We can still get predictions for the subject-specific effect, but this prediction will be shrunk toward the global mean for that particular treatment category. As we gather more data about the subjects, the predictions will become closer to the fixed effect estimates

when the number of observations per subject or group increases, but these prediction can deviate from mean estimates in the case where there are few measurements per subject.

Oehlert (2000) identifies the following step to perform a mixed model

1. Identify sources of variation
2. Identify whether factors are crossed are nested
3. Determine whether factors should be fixed or random
4. Figure out which interactions can exist and whether they can be fitted.

To fit the model, identifiers of subjects must be declared as factors (categorical variables).

We say to factors are nested (A within B) when one can only coexist within the levels of the other: this has implications, for we cannot have interaction between the two. In between-subject experiments, factors are crossed, meaning we can assign an experimental unit or a subject to each factor combination and thus interactions can occur. For example, subjects are nested in between-level factors.

The next step will be in determining whether we have enough observations to support the inclusion of a random term. In a pure within-subject design, we could not include an interaction between subject and the within-factor unless we have multiple replications for each subject, as is the case here. We can therefore include both subject identifier and experimental factor, as well as their interaction. Note that in `lme4` package, the random effects are specified inside parenthesis.

We consider again the experiment of Amirabdolahian and Ali-Adeeb (2021) on smiling fakes and the emotion, this time from a pure mixed model perspective. This means we can simply keep all observations and model them accordingly.

Figure 10.1 shows the raw measurements, including what are notable outliers that may be due to data acquisition problems or instrumental manipulations. Since the experiment was performed in a non-controlled setting (pandemic) with different apparatus and everyone acting as their own technician, it is unsurprising that the signal-to-noise ratio is quite small. We will exclude here (rather arbitrarily) measurements below a latency of minus 40.

```
library(lmerTest)
# fit and tests for mixed models
options(contrasts = c("contr.sum", "contr.poly"))
mixedmod <- lmer(
  latency ~ stimulus +
    (1 | id) + # random effect for subject
    (1 | id:stimulus),
  # random effect for interaction
```

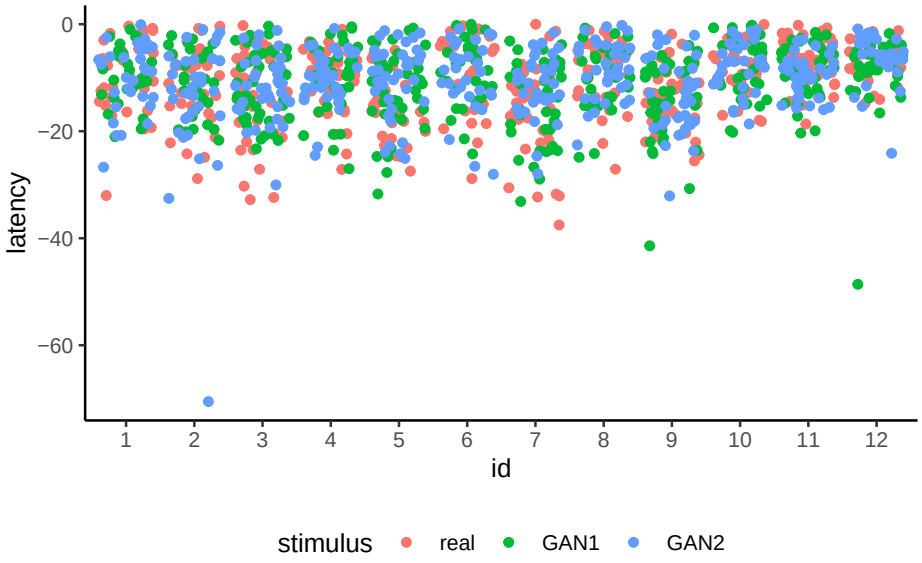


Figure 10.1: Jittered scatterplot of individual measurements per participant and stimulus type.

```
  data = hecedsm::AA21 |> #remove outliers
    dplyr::filter(latency > -40))
# Output parameter estimates
print(mixedmod)
```

```
Linear mixed model fit by REML ['lmerModLmerTest']
Formula: latency ~ stimulus + (1 | id) + (1 | id:stimulus)
   Data: dplyr::filter(hecedsm::AA21, latency > -40)
REML criterion at convergence: 8007.913
Random effects:
 Groups      Name        Std.Dev.
 id:stimulus (Intercept) 0.7371
 id          (Intercept) 2.2679
 Residual                6.2235
Number of obs: 1227, groups:  id:stimulus, 36; id, 12
Fixed Effects:
(Intercept)    stimulus1    stimulus2
   -10.5374      -0.2529      -0.1394
```

We see that there is quite a bit of heterogeneity between participants and per stimulus participant pair, albeit less so for the interaction. All estimated variance terms are rather large.

We can also look globally at the statistical evidence for the

```
Type III Analysis of Variance Table with Satterthwaite's method
         Sum Sq Mean Sq NumDF  DenDF F value Pr(>F)
stimulus  65.62   32.81     2 23.337  0.8471 0.4414
```

The global F test of significance for stimulus is based on an approximation; the denominator degrees of freedom for the approximate F statistic are based on Satterthwaite's method, which provides a correction. There is again no evidence of differences between experimental conditions. This is rather unsurprising if we look at the raw data in Figure 10.1.

We consider a replication study of Elliott et al. (2021), which studied verbalization and verbalization of kids aged 5, 6, 7 and 10. The replication was performed in 17 different school labs, adapting a protocol of Flavell, Beach, and Chinsky (1966), with an overall sample of 977 child partaking in the experiment.

Each participant was assigned to three tasks (`timing`): `delayed` recall with 15 seconds wait, or `immediate`, and finally a naming task (`point-and-name`). The `taskorder` variable records the order in which these were presented: the order of `delayed` and `immediate` was counterbalanced across individuals, with the naming task always occurring last. The response variable is the number of words correctly recalled out of five. The experimenters also recorded the frequency at which students spontaneously verbalized during the task (except the naming task, where they were instructed to do so).

The timing is a within-subject factor, whereas task order and age are between-subject factors: we are particularly interested in the speech frequency and the improvement over time (pairwise differences and trend).

To fit the linear mixed model with a random effect for both children `id` and `lab`: since children are nested in lab, we must specify the random effects via `(1 | id:lab) + (1 | lab)` if `id` are not unique.

We modify the data to keep only 5 and 6 years old students, since most older kids verbalized during the task and we would have large disbalance (14 ten years old out of 235, and 19 out of 269 seven years old). We also exclude the point-and-name task, since verbalization was part of the instruction. This leaves us with 1419 observations and we can check that there are indeed enough children in each condition to get estimates.

```
data(MULTI21_D2, package = "hecedsm")
MULTI21_D2_sub <- MULTI21_D2 |>
  dplyr::filter(
    age %in% c("5yo", "6yo"),
    timing != "point-and-name") |>
  dplyr::mutate(
    verbalization = factor(frequency != "never",
                           labels = c("no", "yes")),
    age = factor(age)) # drop unused age levels
xtabs(~ age + verbalization, data = MULTI21_D2_sub)
```

```
     verbalization
age    no yes
  5yo 106 334
  6yo  56 450
```

Given that we have multiple students of every age group, we can include two-way and three-way interactions in the 2^3 design. We also include random effects for the student and the lab.

```
library(lmerTest)
library(emmeans)
hmod <- lmer(
  mcorrect ~ age*timing*verbalization + (1 | id:lab) + (1 | lab),
  data = MULTI21_D2_sub)
# Parameter estimates
#summary(hmod)
```

We focus here on selected part of the output from `summary()` giving the estimated variance terms.

```
#> Random effects:
#>  Groups   Name        Variance Std.Dev.
#>  id:lab   (Intercept) 0.3587   0.599
#>  lab      (Intercept) 0.0625   0.250
#>  Residual             0.6823   0.826
#> Number of obs: 946, groups:  id:lab, 473; lab, 17
```

We can interpret the results as follows: the total variance is the sum of the `id`, `lab` and `residual` variances components give us an all but negligible effect of lab with 7 percent

of the total variance, versus 40.5 percent for the children-specific variability. Since there are only 17 labs, and most of the individual specific variability is at the children level, the random effect for lab doesn't add much to the correlation.

```
anova(hmod, ddf = "Kenward-Roger")
```

```
Type III Analysis of Variance Table with Kenward-Roger's method
                          Sum Sq Mean Sq NumDF  DenDF F value    Pr(>F)
age                      20.5045 20.5045     1 459.37 30.0507 6.955e-08 ***
timing                    3.2464  3.2464     1 469.00  4.7579   0.02966 *
verbalization            13.6053 13.6053     1 459.45 19.9395 1.007e-05 ***
age:timing                0.2434  0.2434     1 469.00  0.3567   0.55062
age:verbalization         0.0849  0.0849     1 462.08  0.1245   0.72439
timing:verbalization      2.6443  2.6443     1 469.00  3.8754   0.04959 *
age:timing:verbalization  0.2699  0.2699     1 469.00  0.3955   0.52972
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Check estimated marginal means for age
emm <- emmeans(hmod, specs = "age")
```

```
NOTE: Results may be misleading due to involvement in interactions
```

```
emm
```

```
 age emmean     SE   df lower.CL upper.CL
 5yo   1.85 0.0914 35.8     1.67     2.04
 6yo   2.44 0.1053 60.9     2.23     2.66

Results are averaged over the levels of: timing, verbalization
Degrees-of-freedom method: kenward-roger
Confidence level used: 0.95
```

```
# Pairwise differences
pairdiff <- emm |> pairs()
pairdiff
```

```
 contrast  estimate    SE  df t.ratio p.value
 5yo - 6yo    -0.59 0.108 459  -5.482  <.0001

Results are averaged over the levels of: timing, verbalization
Degrees-of-freedom method: kenward-roger
```

The type III ANOVA table shows that there is no evidence of interaction between task order, age and verbalization (no three-interaction) and a very small difference for timing and verbalization. Thus, we could compute the estimated marginal means (95% confidence interval) for age with an estimated correct number of words of 1.8543233 (1.6689094, 2.0397372) words out of 5 for the 5 years olds and 2.4446996 (2.2340654, 2.6553337) words for six years old. Note that, despite the very large number children in the experiment, the degrees of freedom from the Kenward–Roger method are much fewer, respectively 35.8271155 and 60.943758 for five and six years old.

The t-test for the pairwise difference of the marginal effect is 0.5903763 words with standard error 0.1076964. Judging from the output, the degrees of freedom calculation for the pairwise t-test are erroneous — they seem to be some average between the number of entries for the five years old (440) and six years old (506), but this fails to account for the fact that each kid is featured twice. Given the large magnitude of the ratio, this still amounts to strong result provided the standard error is correct.

We can easily see the limited interaction and strong main effects from the interaction plot in Figure 10.2. The confidence intervals are of different width because of the sample inbalance.

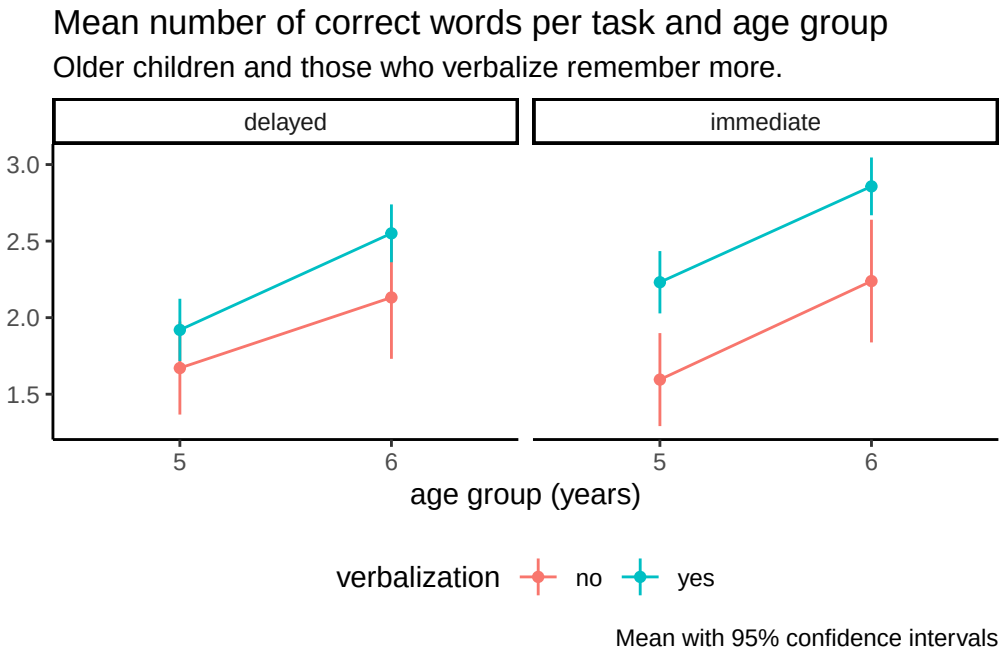


Figure 10.2: Interaction plot for the recall task for younger children.

11 Causal inference

A pet peeve of statisticians is to state that correlation (or association) between two phenomena is not the same as causation. For example, weather forecasts of rain and the number of people carrying umbrellas in the streets are positively correlated, but the relationship is directed: if I intervene as an experimenter and force everyone around to carry umbrellas, it won't impact weather forecasts nor the weather itself. The website Spurious correlations by Tyler Vigen shows multiple graphs of absurd non-causal relations, many of which are simply artefact of population growth.

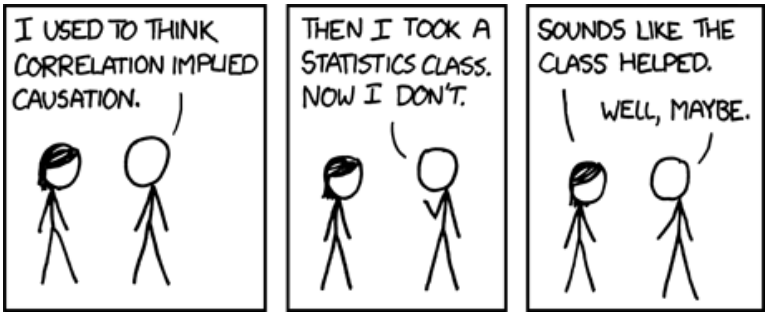


Figure 11.1: xkcd comic 552 (Correlation) by Randall Munroe. Alt text: Correlation doesn't imply causation, but it does waggle its eyebrows suggestively and gesture furtively while mouthing 'look over there'. Cartoon reprinted under the CC BY-NC 2.5 license.

Causal inference is concerned with inferring the effect of an action or manipulation (intervention, policy, or treatment) applied to an observational unit and identifying and quantifying the effect of one variable on other variables. Such action may be conceptual: we can imagine for example looking at student's success (as measured by their grades) by comparing two policies: giving them timely feedback and encouragement, versus no feedback. In reality, only one of these two scenarios can be realized even if both can conceptually be envisioned as **potential outcomes**.

The content of this chapter is divided as follows. First, we discuss the logical interrelation between variables using directed acyclic graphs and focus on relations between triples defining confounders, colliders and mediators. We then describe how we can retrieve causal

effects (in an abstract setting). Finally, we present the linear mediation model popularized by Baron and Kenny (1986). We focus on the hidden assumptions that can undermine causal claims of mediation.

11.1 Basics of causal inference

In an experiment, we can manipulate assignment to treatment a and randomize units to each value of the treatment to avoid undue effects from other variables. The potential outcomes applies in between-subject design because experimental units are assigned to a single treatment, whereas in within-subject designs a single ordering is presented. If we denote the outcome by Y, then we are effectively comparing $(Y \mid X)$ for different values of the action set X (for example, X could be an experimental factor. We talk about causation when for treatment $(X = j)$ and control $(X = 0)$, the distributions of $(Y \mid X = j)$ differs from that of $(Y \mid X = 0)$. The fundamental problem of causal inference is that, while we would like to study the impact of every action on our response, we can only observe the outcome for treatment j, written $(Y_i \mid X = j)$.[1]

Rather than look at the individual treatment effect, we must focus on the population effects. The most common measure of causation is the **average treatment effect**, which is the difference between the population averages of treatment group j and the control group,

$$\mathsf{ATE}_j = \mathsf{E}(Y \mid X = j) - \mathsf{E}(Y \mid X = 0)$$

In experimental designs, we can target the average treatment effect if our subjects comply with their treatment assignment and if we randomize the effect and use a random sample which is representative of the population.

In many fields, the unconditional effect is not interesting enough to warrant publication free of other explanatory variables. It may be that the effect of the treatment is not the same for everyone: for example, a study on gender discrimination may reveal different perceptions depending on gender, in which case the average treatment effect might not be a sensible measure and we could look at conditional effects. We may also be interested in seeing how different mechanisms and pathways are impacted by treatment and how this affects the response. VanderWeele (2015) provides an excellent non-technical overview of mediation and interaction.

Causal inference requires a logical conceptual model of the interrelation between the variables. We will look at directed acyclic graphs to explain concepts of confounding, collision and mediation and how they can influence our conclusions.

[1] In a within-subject design, the analog is that a single ordering can be presented.

To illustrate the relationship between variables, we use diagrams consisting of directed acyclic graph (DAG). A DAG is a graph with no cycle: each node represents a variable of interest and these are linked with directed edges indicating the nature of the relation (if X causes Y, then $X \to Y$). Directed acyclic graphs are used to represent the data generating process that causes interdependencies, while abstracting from the statistical description of the model. This depiction of the conceptual model helps to formalize and identify the assumptions of the model. To identify a causal effect of a variable X on some response Y, we need to isolate the effect from that of other potential causes. Figure 11.2 shows an example of DAG in a real study; the latter is a simplification or abstraction of a world view, but is already rather complicated.

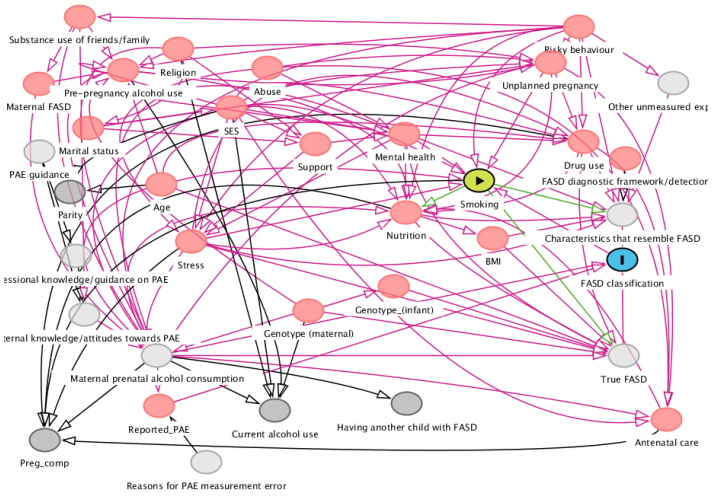


Figure 11.2: Directed acyclic graph of McQuire et al. (2020) reproduction by Andrew Heiss, licensed under CC BY-NC 4.0.

At a theoretical level, the DAG will help identify which paths and relations to control through conditioning arguments to strip the relation to that of interest. Judea Pearl (e.g., Pearl, Glymour, and Jewell 2016) identifies three potential relations between triples of (sets of) variables:

- chains ($X \to Z \to Y$),
- forks ($Z \leftarrow X \to Y$) and
- reverse forks ($Z \to X \leftarrow Y$).

These are represented in Figure 11.3. In the graph, X represents an explanatory variable, typically the experimental factor, Y is the response and Z is another variable whose role depends on the logical flow between variables (collider, confounder or mediator).

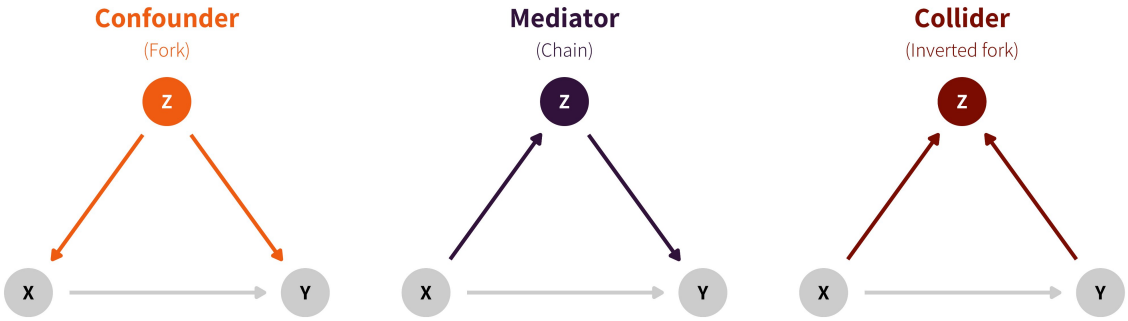


Figure 11.3: Type of causal relations by Andrew Heiss, licensed under CC BY-NC 4.0.

In an experimental design, confounding effects from the experimental treatment X to the response Y are controlled by randomization or sample selection: all incoming arrows inside X from other variables are removed. If we include additional variables in the model which happen to be colliders, then we won't recover the causal effect of interest. Addition of mediators will let us filter the effect due to Z from the direct effect of X.

It is *essential* to determine via logic or otherwise (experiments can help!) the direction of the relationship, lest we run into trouble. Many statistical models commonly used, including regression models, cannot provide an answer to a problem that is philosophical or conceptual in nature. Indeed, correlation is symmetric and insufficient to infer the direction of the arrows in the directed acyclic graph.

The conclusions we will draw from statistical models depend on the nature of the relation. For example, in an observational setting, we could eliminate the effect of a confounding variable by controlling in a regression model or by stratifying for different values of the confounders in order to extract the causal estimate of X on Y. However, the same strategy with a collider would backfire and we would get erroneous conclusions: Kowal (2021) reportedly found out married couples with more children were less happy. As Richard McElreath hinted, controlling for marriage (through sample selection) is incorrect since unhappy couples tend to divorce, but families with many children are less likely to divorce!

In a randomized experiment, we can check the average outcome of a manipulation by comparing groups: assuming random sampling, these conclusions can be broadly generalized to the population of interest from which the sample is drawn. However, it may be that the effect of the treatment depends on other variables: cultural differences, gender or education may change. In the statistical model, inclusion of interaction terms (typically product of the moderator variable with the factor encoding the experimental sub-condition) will allow us to estimate those differences.

11.2 Mediation

In order to do inference and tests relations, we need to add to our logical causal model represented by a directed acyclic graph a data generating mechanisms that prescribes how variables how interrelated. Our ability to establish mediation will depend on the model and a set of assumptions, some of which won't be verifiable.

To define in full generality the treatment and mediation effect, we need to consider the potential outcome framework. Following Pearl (2014), we use $Y_i(x, m)$ to denote the potential outcome for individual i with explanatory or experimental covariate/factor x and mediator m. Likewise, $M_i(x)$ is the potential mediator when applying treatment level x.[2]

The **total effect** measures the overall impact of changes in outcome Y (both through M and directly) when experimentally manipulating X,

$$\mathsf{TE}(x, x^*) = \mathsf{E}[Y \mid \mathrm{do}(X = x)] - \mathsf{E}[Y \mid \mathrm{do}(X = x^*)],$$

where x^* is the factor level of the reference or control and x is another treatment value. This definition generalizes when X is a continuous variable.

The **average controlled directed effect** measures the flow along $X \to Y$, disabling the pathway $X \to M \to Y$ by fixing the mediatior: it is

$$\begin{aligned}\mathsf{CDE}(m, x, x^*) &= \mathsf{E}[Y \mid \mathrm{do}(X = x, m = m)] - \mathsf{E}[Y \mid \mathrm{do}(X = x^*, m = m)] \\ &= \mathsf{E}\{Y(x, m) - Y(x^*, m)\}\end{aligned}$$

This measures the expected change in response when the experimental factor changes from x to x^* and the mediator is set to a fixed value m uniformly over the population.

The **natural direct effect**,

$$\mathsf{NDE}(x, x^*) = \mathsf{E}[Y\{x, M(x^*)\} - Y\{x^*, M(x^*)\}],$$

is the expected change in Y under treatment x if M is set to whatever value it would take under control x^*.

The **natural indirect effect** (NIE) is the expected change in the response Y if we set our experimental value X to its control value x^* and change the mediator value which it would attain under x,

$$\mathsf{NIE}(x, x^*) = \mathsf{E}[Y\{x^*, M(x)\} - Y\{x^*, M(x^*)\}]$$

[2] The notation is important to distinguish between association $Y \mid X$ when observing X from manipulations or interventions, $Y \mid \mathrm{do}(X)$ and counterfactuals $Y(X)$.

Armed with these definitions, we can consider the **sequential ignorability assumption**, which is decomposed into two components.

The first component is: given pre-treatment covariates W, treatment assignment is independent of potential outcomes for mediation and outcome,

$$Y_i(x', m), M_i(x) \perp\!\!\!\perp X_i \mid W_i = w.$$

In order words, the values taken by the mediator and by the response exist independently of the treatment assignment and don't change.[3]

The second component of the sequential ignorability assumption is as follows: given pre-treatment covariates and observed treatment, mediation is independent of potential outcomes,

$$Y_i(x', m) \perp\!\!\!\perp M_i(x) \mid X_i = x, W_i = w$$

The set of assumptions from Imai, Keele, and Tingley (2010) and Pearl (2014) are equivalent under randomization of treatment assignment, as we consider thereafter.

The total effect can be written

$$\mathsf{TE}(x, x^*) = \mathsf{NDE}(x, x^*) - \mathsf{NIE}(x^*, x).$$

In the linear mediation model, the reversal of argument amounts to changing the sign of the coefficient, giving an additive decomposition of the total effect as $\mathsf{TE} = \mathsf{NDE} + \mathsf{NIE}$ (Pearl 2014).

When measuring effects in psychology and marketing, it will often be the case that the conceptual causal model includes variables that cannot be directly measured. The proxy, as in Figure 11.4, add an additional layer of complexity and potential sources of confounding.

11.3 Linear mediation model

One of the most popular model in social sciences is the linear mediation model, popularized by Baron and Kenny (1986) although the method predates this publication. Another inferential approach, suggested by Preacher and Hayes (2004), uses the same model with different test statistics and is extremely popular because it comes with software; Hayes' PROCESS macros for SAS, SPSS and **R** have lead to the widespread adoption by researchers.

[3]The dependence on W is used for situations where we can perform randomization based on pre-treatment assignment (i.e., we specify a mechanisms that is not equal based, but the probability of assignment is known from each individual).

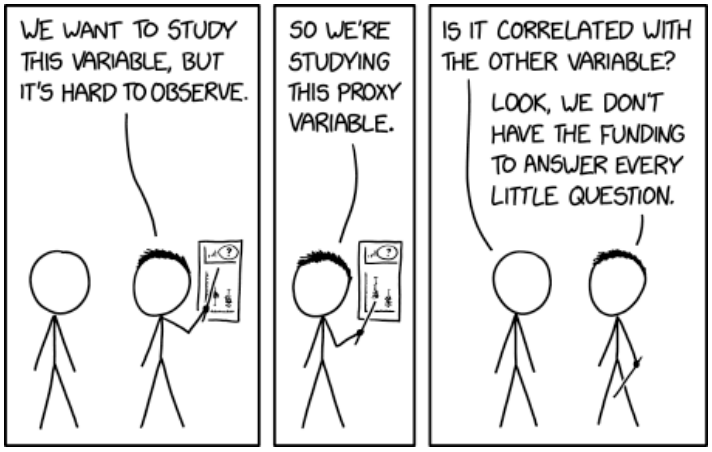


Figure 11.4: xkcd comic 2652 (Proxy Variable) by Randall Munroe. Alt text: Our work has produced great answers. Now someone just needs to figure out which questions they go with. Cartoon reprinted under the CC BY-NC 2.5 license.

Bullock, Green, and Ha (2010) list limitations of the approach and provide examples of instances in which the model does not have a meaningful causal interpretation.

The linear mediation model assumes that the effect of mediation and treatment is additive and that the response measurement is continuous. Consider covariates $\mathbf{Z}$, experimental factor $\mathbf{X}$ corresponding to treatment and postulated mediator variable M, assumed continuous. Given **uncorrelated** unobserved noise variables ε_M and ε_Y, we specify linear regression models,

$$\begin{aligned} M \mid X = x &= c_M + \alpha x + \varepsilon_M, \\ Y \mid X = x, M = m &= c_Y + \beta x + \gamma m + \mathbf{z}^\top \boldsymbol{\omega} + \varepsilon_Y \end{aligned}$$

where we use the contrast-to-reference parametrization so that the reference category for the intercept corresponds to control (group x^*) and x the other category of interest, with α capturing the difference between x and x^*. The model for $Y \mid X, M$ should include additional covariates $\mathbf{z}$ to control for confounding between M and Y if the latter is suspected.

The parameters can be interpreted as the direct (β), indirect ($\alpha\gamma$) and total ($\beta + \alpha\gamma$) effects. To see this, we plug the first equation in the second and obtain the marginal model for Y given treatment X,

$$\begin{aligned} Y \mid X = x &= \underbrace{(c_Y + \gamma c_M)}_{\text{intercept}} + \underbrace{(\beta + \alpha\gamma)}_{\text{total effect}} \cdot x + \underbrace{(\gamma\varepsilon_M + \varepsilon_Y)}_{\text{error}} && (11.1) \\ &= c'_Y + \tau X + \varepsilon'_Y && (11.2) \end{aligned}$$

These parameters can be estimated using structural equation models (SEM), or more typically by running a series of linear regression (ordinary least squares).

The sequential ignorability in the linear mediation models boils down to "no unmeasured confounders" in the relations $X \to Y$, $X \to M$ and $M \to Y$: the first two are satisfied in experimental studies due to randomization, as shown in Figure 11.7. This means $\varepsilon_M \perp\!\!\!\perp \varepsilon_Y$ must be independent and, as a result, error terms should also be uncorrelated.

In the linear mediation model, we can estimate the conditional direct effect corresponding to the product of coefficients $\alpha\gamma$. Absence of mediation implies the product is zero. Baron and Kenny (1986) recommended using Sobel's test statistic, a Wald-test of the form

$$S = \frac{\widehat{\alpha}\widehat{\gamma} - 0}{\mathsf{se}(\widehat{\alpha}\widehat{\gamma})} = \frac{\widehat{\alpha}\widehat{\gamma}}{\sqrt{\widehat{\gamma}^2\mathsf{Va}(\widehat{\alpha}) + \widehat{\alpha}^2\mathsf{Va}(\widehat{\gamma}) + \mathsf{Va}(\widehat{\gamma})\mathsf{Va}(\widehat{\alpha})}} \dot{\sim} \mathsf{No}(0,1)$$

where $\widehat{\alpha}$, $\widehat{\gamma}$ and their variance $\mathsf{Va}(\widehat{\alpha})$ and $\mathsf{Va}(\widehat{\gamma})$ can be obtained from the estimated coefficients and standard errors.[4] The Sobel's statistic S is approximately standard normal in large samples, but the approximation is sometimes crude.

In the linear mediation causal model, we can estimate the total causal effect of X, labelled τ, by running the linear regression of Y on X as there is no confounding affecting treatment X in a completely randomized experimental design. This strategy isn't valid with observational data unless we adjust for confounders. Under no unmeasured confounders and linearity, the product $\alpha\gamma$ is also equal to the difference between the total effect and the **natural direct effect**, $\tau - \beta$.

Baron and Kenny (1986) suggested for X and M continuous breaking down the task in three separate steps:

1) fit a linear regression of M on X to estimate α
2) fit a linear regression of Y on X to estimate τ
3) fit a linear regression of Y on X and M to estimate β and γ.

In the "Baron and Kenny (1986) approach", we test $\mathscr{H}_0 : \alpha = 0$, $\mathscr{H}_0 : \tau = 0$ and $\mathscr{H}_0 : \gamma = 0$ against the two-sided alternative. This approach has caveats since mediation refers to the relation $X \to M \to Y$, so we only need to consider (joint tests of) α and γ (the total effect could be zero because $\beta = -\alpha\gamma$ even if there is mediation). Zhao, Lynch, and Chen (2010) review the typology of mediation.

1. complementary mediation when both direct and indirect effects are of the same sign and non-zero.
2. competitive mediation when direct and indirect effects are of opposite signs.

[4] Sobel derived the asymptotic variance using a first-order Taylor series expansion assuming both α and γ are non-zero (hence the tests!)

3. indirect-only mediation when the direct effect of $X \to Y$ is null, but the effect $X \to M \to Y$ isn't.

Previous definitions popularized by Baron and Kenny (1986) still found in old papers include "full mediation" for instances where $\beta = 0$ and partial mediation if the direct effect is less than the total effect, meaning $\beta < \tau$.

To see this, let's generate data with a binary treatment and normally distributed mediators and response with no confounding (so the data generating process matches exactly the formulation fo Baron–Kenny. We set $\alpha = 2$, $\beta = 1/2$ and $\gamma = 0$. This is an instance where the null is true (X affects both M and Y).

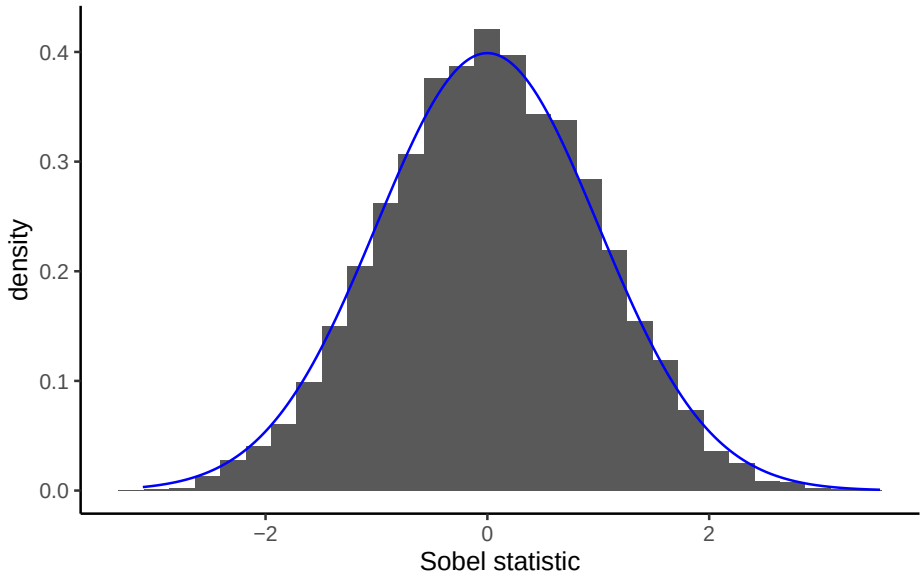


Figure 11.5: Null distribution of Sobel's statistic against approximate asymptotic normal distribution with 20 observations $\alpha = 0$, $\gamma = 0.1$ and normal random errors.

If we knew exactly the model that generated X, M, Y and the relations between them, we could simulate multiple datasets like in Figure 11.5 with $n = 20$ observations and compare the test statistic we obtained with the simulation-based null distribution with $\alpha\gamma = 0$. In practice we do not know the model that generated the data and furthermore we have a single dataset at hand. An alternative is the bootstrap, a form of simulation-based inference. The latter is conceptually easy to understand: we generate new datasets by resampling from the ones observed (as if it was the population). Since we want the sample size to be identical and our objective is to get heterogeneity, we sample with replacement: from one bootstrap dataset to the next, we will have multiple copies, or none, of each observation. See Efron and Tibshirani (1993) and Davison and Hinkley (1997) for a more thorough treatment of

the bootstrap and alternative sampling schemes for regression models. The nonparametric bootstrap procedure advocated by Preacher and Hayes (2004) consists in repeating the following B times:

1) sample n observations **with replacement**, i.e., a tuple (X_i, M_i, Y_i), from the original data .
2) compute the natural indirect effect $\widehat{\alpha} \cdot \widehat{\gamma}$ on each simulated sample

For a two-sided test at level α, compute the $\alpha/2$ and $1-\alpha/2$ quantiles of the bootstrap statistics $\{\widehat{\alpha}_b\widehat{\gamma}_b\}_{b=1}^B$. For example, if the level is $\alpha = 5\%$ and we generate $B = 1000$ bootstrap samples, the percentile confidence intervals bounds are the 25th and 975th ordered observations.

Nowadays, the asymptotic approximation (sometimes misnamed delta method[5]) has fallen out of fashion among practitioners, who prefer the nonparametric bootstrap coupled with the percentile method.

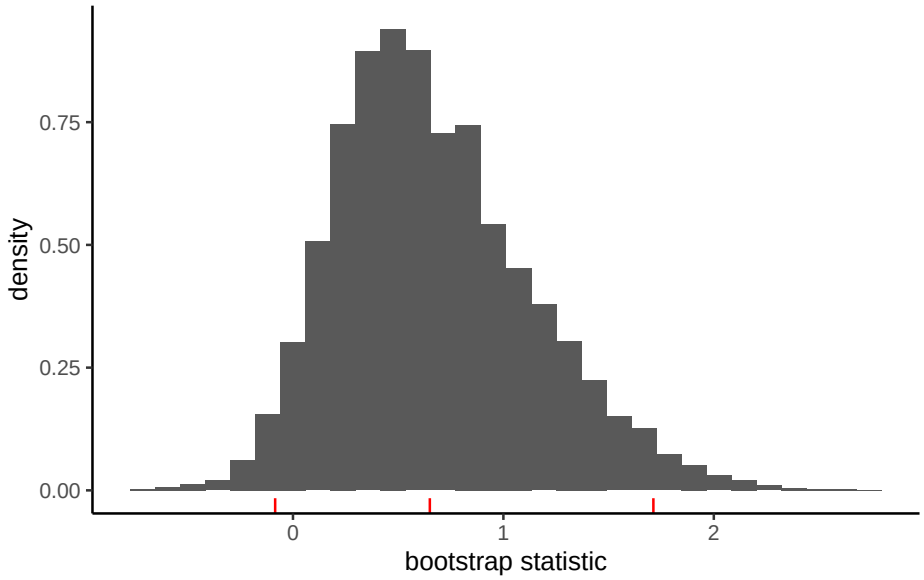


Figure 11.6: Bootstrap distribution of indirect effect with estimate and percentile 95% confidence intervals (vertical red lines).

The nonparametric percentile bootstrap confidence interval for $\alpha\gamma$ is [-0.08, 1.71] and thus we fail to reject the null hypothesis $\mathscr{H}_0 : \alpha\gamma = 0$.

[5]The latter is the name of the method used to derive the denominator of Sobel's statistic.

The inference scheme is popular, but we could also rely on different models and use the definitions of the causal effects to perform simulation-based inference; see Appendix D of Imai, Keele, and Tingley (2010). The latter fit two models for the mediator and the outcome, then estimate parameters. In large samples, the parameter estimators are approximately multivariate Gaussian and we can simulate parameters, use the parametric model to generate new data and potential outcomes. The average causal mediation effect can be estimated empirically based on the simulated potential outcomes.

11.3.1 Model assumptions

We can unpack the model assumptions for the linear mediation model.

1. The *no unmeasured confounders* assumption. Plainly stated, there are no unobserved confounders and thus the error terms ε_M and ε_Y are independent. Additionally, when we consider observational data, we must make sure there is hidden confounders affecting either the $M \to X$ and the $X \to Y$ relation, as shown in Figure 11.7. We can include covariates in the regression models to palliate to this, but we must only include the minimal set of confounders (and no additional mediator or collider chain).

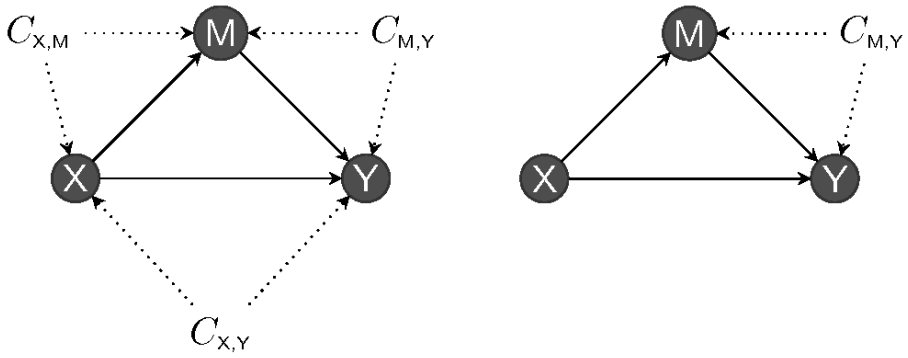


Figure 11.7: Directed acyclic graph representing observational settings (left) and experimental settings in which assignment is random given covariates measured pre-treatments (right). The 'no unmeasured confounder' assumption postulates such confounders are included (with the correct parametric form) in the regression models.

Another problem would be to claim that variable M is a mediator when in truth part of the effect on the outcome is due to change in another mediator. Figure 11.8 shows an instance with no confounding, but multiple mediators, say M_1 and M_2: the latter mediates the relation $M_1 \to Y$. The linear mediation model would capture the total effect of M_1, but it would be incorrect to claim that the mediation effect on X is due to M_1.

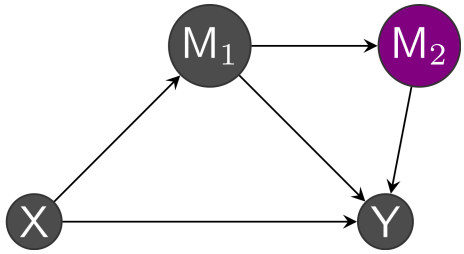


Figure 11.8: Directed acyclic graph showing multiple mediators.

2. The **linearity assumption** implies that the linear models are correctly specified and that the effect of the covariates are linear. This means that, *ceteris paribus*, the effect of an increase of one unit of M on Y is the same regardless of the value of M. It also doesn't depend on the level of other variables (no interaction). If the model isn't properly specified, the linear model coefficients will capture the best linear predictor given the design and the coefficients may not be meaningful.

The linearity assumption also implies that the effect is the same for every individual, so there is no treatment heterogeneity or measurement error which could lead to attenuation bias.

Following Bullock, Green, and Ha (2010), we index the regression equations by individual i

$$M_i \mid X_i = x = c_M + \alpha_i x + \varepsilon_{Mi},$$
$$Y_i \mid X_i = x, M_i = m = c_Y + \beta_i x + \gamma_i m + \mathbf{z}_i^\top \boldsymbol{\omega} + \varepsilon_{Yi}$$

If α_i differ from one observation to the next, the average estimated by the regression could be positive, negative or null. Even in the latter case, we could have γ_i and α_i positive for some observation, or both negatives so that they cancel out even if there is complementary mediation.

We could easily expand the model to include nonlinear effects (not for the treatment or mediator) and potential interactions, but the linear model approach will be limited. Imai, Keele, and Tingley (2010) details a more general approach for parametric models based on simulations, as well as a nonparametric approach. Both are less restrictive than the Baron and Kenny (1986) model.

The main benefit of experimental designs is that it deconfounds the relation between treatment and other variables, but adding spurious variables could create feedback and lead to inconsistent conclusions. It's not clear that the mediator can be manipulated experimentally, and even if it could be to estimate the γ, one must make sure the relation is the same absent of X. For example, we could estimate the indirect effect term by manipulating jointly

Table 11.1: Coefficients of the mediation (left) and outcome (right) models.

term	estimate	std.error
(Intercept)	5.23	0.27
consistency	0.48	0.24
familiarity	0.08	0.06

term	estimate	std.error
(Intercept)	3.04	0.53
fluency	0.60	0.09
consistency	0.29	0.22
familiarity	0.09	0.05

(if possible) (X, M) but even then the linearity assumption must hold for the estimates to correspond to our linear causal mediation model.

11.3.2 Example

Study 2 of Lee and Choi (2019) focus on inconsistency of photos and text descriptions for online descriptions and how this impact the product evaluation.

The experimental variable is the consistency of the product description and depiction, with `fluency` leading to "processing disfluency" that is expected to impact negatively judgment ratings. Familiarity with the product brand and product is included as covariate in both mediator and outcome model (see Table 1 of Lee and Choi (2019)).[6]

```
data(LC19_S2, package = "hecedsm")
YsMX <- lm(prodeval ~ fluency + consistency + familiarity,
           data = LC19_S2)
MsX <- lm(fluency ~ consistency + familiarity,
           data = LC19_S2)
```

```
Registered S3 methods overwritten by 'broom':
  method            from
  tidy.glht         jtools
  tidy.summary.glht jtools
```

We can extract the effects directly from the outcome: the natural indirect effect estimate is $\widehat{\alpha}\widehat{\gamma} = 0.48 \times 0.60$ and the direct effect is $\widehat{\beta} = 0.29$.

To get confidence intervals, we can use the `mediate` package (Tingley et al. 2014). The function requires the parametric model for the mediation and outcome, as well as a series

[6] Incidentally, reporting the coefficients allows in a Table allows one to retro-engineer the model and ensure reproducibility.

Table 11.2: Linear causal mediation analysis: parameter estimates, nonparametric bootstrap 95% confidence intervals and p-values with the percentile method based on 5000 bootstrap samples.

	estimate	lower 95% CI	upper 95% CI	p-value
ACME	0.286	0.025	0.615	0.037
ADE	0.287	-0.166	0.732	0.218
total effect	0.573	0.037	1.080	0.030
prop. mediated	0.499	0.211	5.956	0.057

of specification (the number of bootstrap samples, the type of confidence interval, the names of the levels for categorical treatment, etc.)

```
set.seed(80667)
library(mediation, quietly = TRUE)
linmed <- mediate(
  model.m = MsX,
  model.y = YsMX,
  treat = "consistency",
  mediator = "fluency",
  sims = 5000L,
  boot = TRUE,
  boot.ci.type = "bca", # bias-corrected and accelerated (bca)
  control.value = "inconsistent",
  treat.value = "consistent")
```

Using the `summary` method, we can print the table of estimates and confidence intervals. We can see that the results are consistent with those reported in the article.

11.4 Moderation and interactions

The causal effect $Y \mid \mathrm{do}(X)$ may be a misleading summary if another variable modifies the relation: for example, the perception of gender discrimination or racism may depend on the person background and experience and this may impact the effect of the manipulation. Such variables, say W, thus have an interactive effect with the experimental factor X, termed moderator in psychology.

In a blocking design, covariates are included that have an impact on the outcome to filter

out variability, but with the assumption that they do not influence the effect of treatment. With moderators, we include the interaction.

If we have an experimental factor X which is binary or categorical, the resulting model is a simple analysis of variance model and we can test the significance of the interaction term to assess the moderating effect of W.

If W is a mean-centered continuous variable and X a categorical variable with $k = 1, \ldots, K$ levels using the sum-to-zero parametrization, the linear model

$$\mathsf{E}\{Y \mid \mathsf{do}(X) = k, W\} = \mu + \alpha_k + (\beta + \gamma_k)W + \varepsilon,$$

includes different slopes for W in each experimental group, as well as different intercepts $(\mu + \alpha_k)$ for group k.

As an example, we consider Garcia et al. (2010). These data are from a study on gender discrimination. Participants were put with a file where a women was turned down promotion in favour of male colleague despite her being clearly more experimented and qualified. The authors manipulated the decision of the participant to this decision, either choosing not to challenge the decision (no protest), a request to reconsider based on individual qualities of the applicants (individual) and a request to reconsider based on abilities of women (collective). All items were measured using scales constructed using items measured using Likert scales ranging from strongly disagree (1) to strongly agree (7).

The postulated mediator variable is `sexism`, the average of 6 Likert scales for the Modern Sexism Scale assessing pervasiveness of gender discrimination. We consider participants' evaluation of the appropriateness of the response of the fictional character.

We fit the linear model with the interaction and display the observed slopes

```
data(GSBE10, package = "hecedsm")
lin_moder <- lm(respeval ~ protest*sexism,
                data = GSBE10)
summary(lin_moder) # coefficients
car::Anova(lin_moder, type = 3) # tests
```

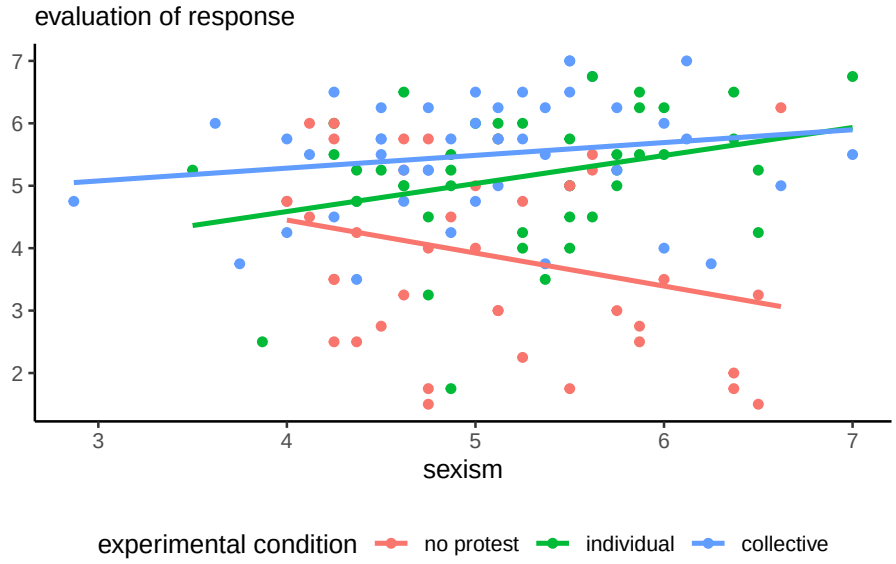
Because of the interaction, comparing the levels of the experimental factor only makes sense if we fix the value of sexism (since the slopes are not parallel) and won't necessarily be reliable outside of the range of observed values of sexism. We could look at quantiles and differences at the mean sexism,[7] or one standard deviation away.

We may be interested in the range of values of the predictor W for which the difference between treatments is not statistically significant if we only have a binary treatment. The

[7] This is the default with `emmeans`

Table 11.3: Analysis of variance table for the linear moderation model.

term	sumsq	df	statistic	p.value
protest	6.34	2	2.45	.091
sexism	6.59	1	5.09	.026
protest:sexism	12.49	2	4.82	.010
Residuals	159.22	123		



Johnson–Neyman method (Johnson and Neyman 1936) considers this range, but this leads to multiple testing problems since we probe the model repeatedly. Esarey and Sumner (2018) offer a method that provides control for the false discovery rate.

To illustrate the method, we dichotomize the manipulation pooling individual and collective protests, since these are the most similar.

```
library(interactions)
db <- GSBE10 |>
  dplyr::mutate(
    protest = as.integer(protest != "no protest"))
lin_moder2 <- lm(respeval ~ protest*sexism, data = db)
jn <- johnson_neyman(
  model = lin_moder2, # linear model
```

```
  pred = protest, # binary experimental factor
  modx = sexism, # moderator
  control.fdr = TRUE,
  mod.range = range(db$sexism))
jn$plot
```

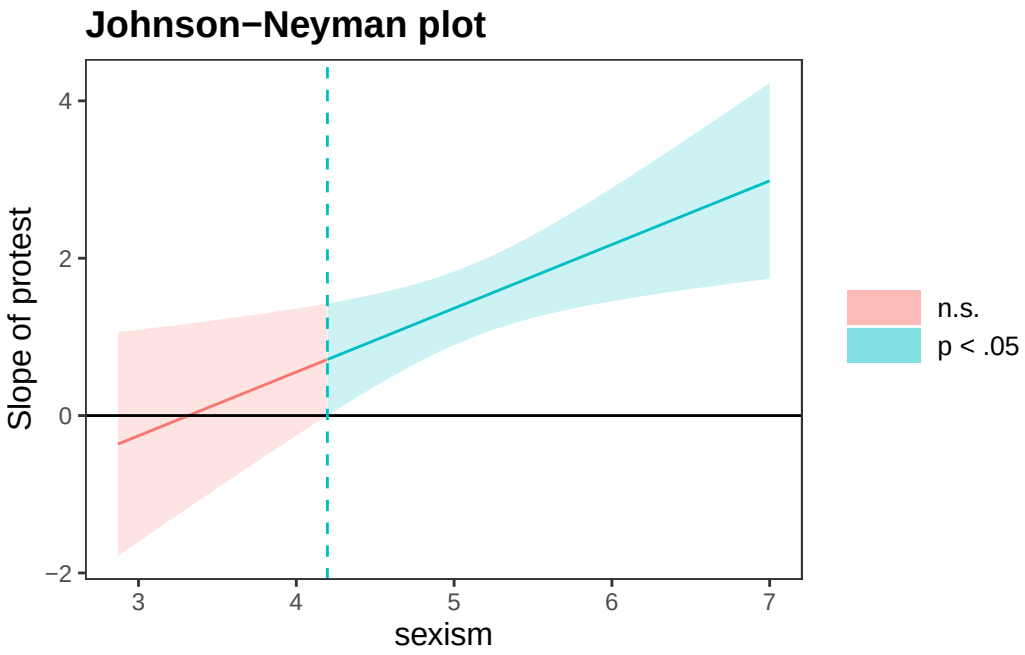


Figure 11.9: Johnson–Neyman plot for difference between protest and no protest as a function of sexism.

The cutoff value is 4.20 with control for the false discovery rate and 4.15 without. The interval is not extended beyond the range of value for sexism, as these are not possible given the Likert scale (which starts at 1). In this example, the moderator is not experimentally manipulated, but it could be. More complicated mediation models could include interactions between treatment effects or moderators and covariates, with external variables, leading to moderated mediation. Interactions can be considered for pretty much any statistical model, but the usual assumptions need to hold for inference to be approximately valid.

12 Nonparametric tests

In small samples or in the presence of very skewed outcome responses, often combined with extreme observations, the conclusions drawn from the large-sample approximations for t-tests or analysis of variance models need not hold. This chapter presents **nonparametric tests**.

If our responses are numeric (or at least ordinal, such as those measured by Likert scales), we could subtitute them by their ranks. Ranks give the relative ordering in a sample of size n, where rank 1 denotes the smallest observation and rank n the largest. Ranks are not affected by outliers and are more robust (contrary to averages), but discard information. For example, ranking the set of four observations (8,2,1,2) gives ranks (4, 2.5., 1, 2.5) if we assign the average rank to ties.

When are nonparametric tests used? The answer is that they are robust (meaning their conclusions are less affected) by departure from distributional assumptions (e.g., data are normally distributed) and by outliers. In large samples, the central limit theorem kicks in and the behaviour of most group average is normal. However, in small samples, the quality of the p-value approximate depends more critically on whether the model assumptions hold or not.

All of what has been covered so far is part of parametric statistics: we assume summary statistics behave in a particular way and utilize the probabilistic model from which these originate to describe the range of likely outcomes under the null hypothesis. As ranks are simply numbers between 1 to n (if there are no ties), no matter how data are generated, we can typically assess the repartition of those integers under the null hypothesis. There is no free lunch: while rank-based tests require fewer modelling assumptions, they have lower power than their parametric counterparts *if* the assumption underlying these tests are validated.

In short: the more assumptions you are willing to assume, the more information you can squeeze out of your problem. However, the inference can be fragile so you have to decide on a trade-off between efficiency (keeping all numerical records) and robustness (e.g., keeping only the signs or the ranking of the data).

The following list nonparametric tests and their popular parametric equivalent.

- sign test: an alternative to a one-sample t-test (also valid for paired measurements, where we subtract the two to get a single numeric number and we rank differences). Only uses the sign of the difference, minimal assumptions but not powerful
- Wilcoxon's signed rank test: idem, but using the ranks of the observations
- Mann–Whitney U or Wilcoxon's rank-sum test: the nonparametric analog of two-sample t-test, which ranks all observations in the sample and compares them between groups. For between-subject designs

These can be extended with repeated measurements to more than two groups:

- Friedman's rank sum test for completely randomized block design: ranks are computed within each block (one block, one experimental factor) and we consider the sum of the ranks for each treatment level. Equivalent of sign test with more samples; also Quade's test
- Kruskal–Wallis test: one-way analysis of variance model with ranks, obtained by pooling all observations, computing the ranks, and splitting them by experimental condition.

For more than 15 observations, the normal, student or Fisher approximation obtained by running the tests from the linear model or ANOVA function yield more or less the same benchmarks for all useful purposes: see J. K. Lindeløv cheatsheet and examples for indications.

12.1 Wilcoxon signed rank test

The most common use of the signed rank test is for paired samples for which the response is measured on a numeric or ordinal scale. Let Y_{ij} denote measurement j of person i and the matching observation Y_{kj}. For each pair $i = 1, \ldots, n$, we can compute the difference $D_i = Y_{ij} - Y_{ik}$.[1] If we assume there is no difference between the distributions of the values taken, then the distribution of the difference D_j is symmetric around zero under the null hypothesis.[2] The statistic tests thus tests whether the median is zero.[3]

Once we have the new differences $D_1, \ldots, D_n$, we take absolute values and compute their ranks, $R_i = \mathsf{rank}|D_i|$. The test statistic is formed by computing the sum of the ranks R_i associated with positive differences $D_i > 0$. How does this test statistic work as a summary of evidence? If there was no difference we expect roughly half of the centered observations or paired difference to be positive and half to be negative. The sum of positive ranks should

[1] With one sample, we postulate a median μ_0 and set $D_i = Y_i - \mu_0$.
[2] We could subtract likewise μ_0 from the paired difference if we assume the distributions are μ_0 units apart.
[3] When using ranks, we cannot talk about the mean of the distribution, but rather about quantiles.

be close to the average rank: for a two-sided test, large or small sums are evidence against the null hypothesis that there is no difference between groups.

In management sciences, Likert scales are frequently used as response. The drawback of this approach, unless the response is the average of multiple questions, is that there will be ties and potentially zero differences $D_j = 0$. There are subtleties associated with testing, since the signed rank assumes that all differences are either positive or negative. The `coin` package in **R** deals correctly with such instances, but it is important to specify the treatment of such values.[4]

We consider a within-subject design from Brodeur et al. (2021), who conducted an experiment at Tech3Lab to check distraction while driving from different devices including smartwatches using a virtual reality environment. The authors wanted to investigate whether smartwatches were more distracting than cellphones while driving. Using a simulator, they ran a within-subject design where each participant was assigned to a distraction (phone, using a speaker, texting while driving or smartwatch) while using a driving simulator. The response is the number of road safety violations conducted on the segment. Each task was assigned in a random order. The data can be found in the `BRLS21_T3` dataset in package `hecedsm`.

A quick inspection reveals that the data are balanced with four tasks and 31 individuals. We can view the within-subject design with a single replication as a complete block design (with `id` as block) and `task` as experimental manipulation. The data here are clearly far from normally distributed and there are notable outliers in the upper right tail. While conclusions probably wouldn't be affected by using an analysis of variance to compare the average time per task, but it may be easier to convince reviewers that the findings are solid by ressorting to nonparametric procedures.

Both the Friedman and the Quade tests are obtained by computing ranks within each block (participant) and then performing a two-way analysis of variance. The Friedman test is less powerful than Quade's with a small number of groups. Both are applicable for block designs with a single factor.

```
data(BRLS21_T3, package = "hecedsm")
friedman <- coin::friedman_test(
  nviolation ~ task | id,
```

[4]For example, are zero difference discarded prior to ranking, as suggested by Wilcoxon, or kept for the ranking and discarded after, as proposed by Pratt (1959)? We also need to deal with ties, as the distribution of numbers changes with ties. If this seems complicated to you, well it is… so much that the default implementation in **R** is unreliable. Charles Geyer illustrate the problems with the *zero fudge*, but the point is quite technical. His notes make a clear case that you can't trust default software, even if it's been sitting around for a long time.

```
  data = BRLS21_T3)
quade <- coin::quade_test(
  nviolation ~ task | id,
  data = BRLS21_T3)
eff_size <- effectsize::kendalls_w(
  x = "nviolation",
  groups = "task",
  blocks = "id",
  data = BRLS21_T3)
```

The Friedman test is obtained by replacing observations by the rank within each block (so rather than the number of violations per task, we compute the rank among the four tasks). The Friedman's test statistic is 18.97 and is compared to a benchmark χ^2_3 distribution, yielding a p-value of 3×10^{-4}.

We can also obtain effect sizes for the rank test, termed Kendall's W. A value of 1 indicates complete agreement in the ranking: here, this would occur if the ranking of the number of violations was the same for each participant. The estimated agreement (effect size) is 0.2.

The test reveals significant differences in the number of road safety violations across tasks. We could therefore perform all pairwise differences using the signed-rank test and adjust p-values to correct for the fact we have performed six hypothesis tests.

To do this, we modify the data and map them to wide-format (each line corresponds to an individual). We can then feed the data to compute differences, here for `phone` vs `watch`. We could proceed likewise for the five other pairwise comparisons and then adjust p-values.

```
smartwatch <- tidyr::pivot_wider(
  data = BRLS21_T3,
  names_from = task,
  values_from = nviolation)
coin::wilcoxsign_test(phone ~ watch,
                      data = smartwatch)
```

```
    Asymptotic Wilcoxon-Pratt Signed-Rank Test

data:  y by x (pos, neg)
     stratified by block
Z = 0.35399, p-value = 0.7233
alternative hypothesis: true mu is not equal to 0
```

You can think of the test as performing a paired t-test for the 31 signed ranks $R_i = \mathsf{sign}(D_i)\mathsf{rank}(|D_i|)$ and testing whether the mean is zero. The p-value obtained by doing this after discarding zeros is 0.73, which is pretty much the same as the more complicated approximation.

12.2 Wilcoxon rank sum test and Kruskal–Wallis test

These testing procedures are the nonparametric analog of the one-way analysis in a between-subject design. One could be interested in computing the differences between experimental conditions (pairwise) or overall if there are $K \geq 2$ experimental conditions. To this effect, we simply pool all observations, rank them and compare the average rank in each group. We can track what should be the repartition of data if there was no difference between groups (all ranks should be somehow uniformly distributed among the K groups). If there are groups with larger averages than others, than this is evidence.

In the two-sample case, we may also be interested in providing an estimator of the difference between condition. To this effect, we can compute the average of pairwise differences between observations of each pair of groups: those are called Walsh's averages. The Hodges–Lehmann estimate of location is simply the median of Walsh's averages and we can use the Walsh's averages themselves to obtain a confidence interval.

Brucks and Levav (2022) measure the attention of participants based on condition using an eyetracker. We compare the time spend looking at the partner by experimental condition (face-to-face or videoconferencing). The authors used a Kruskal–Wallis test, but this is equivalent to Wilcoxon's rank-sum test.

```
data(BL22_E, package = "hecedsm")
mww <- coin::wilcox_test(
  partner_time ~ cond,
  data = BL22_E,
  conf.int = TRUE)
welch <- t.test(partner_time ~ cond,
  data = BL22_E,
  conf.int = TRUE)
mww
```

```
  Asymptotic Wilcoxon-Mann-Whitney Test
```

```
data:  partner_time by cond (f2f, video)
Z = -6.4637, p-value = 1.022e-10
alternative hypothesis: true mu is not equal to 0
95 percent confidence interval:
 -50.694 -25.908
sample estimates:
difference in location
               -37.808
```

The output of the test includes, in addition to the p-value for the null hypothesis that both median time are the same, a confidence interval for the time difference (in seconds). The Hodges–Lehmann estimate of location is -37.81 seconds, with a 95% confidence interval for the difference of $[-50.69, -25.91]$ seconds.

These can be compared with the usual Welch's two-sample t-test with unequal variance. The estimated mean difference is -39.69 seconds for face-to-face vs group video, with a 95% confidence interval of $[-52.93, -26.45]$.

In either case, it's clear that the videoconferencing translates into longer time spent gazing at the partner than in-person meetings.

13 References

Achilles, C. M., Helen Pate Bain, Fred Bellott, Jayne Boyd-Zaharias, Jeremy Finn, John Folger, John Johnston, and Elizabeth Word. 2008. "Tennessee's Student Teacher Achievement Ratio (STAR) project." Harvard Dataverse. https://doi.org/10.7910/DVN/SIWH9F.

Aguinis, Herman, Isabel Villamor, and Ravi S. Ramani. 2021. "MTurk Research: Review and Recommendations." *Journal of Management* 47 (4): 823–37. https://doi.org/10.1177/0149206320969787.

Alexander, Rohan. 2022. *Telling Stories with Data*. CRC Press. https://www.tellingstorieswithdata.com/.

Barnett, Adrian Gerard, and Jonathan D Wren. 2019. "Examination of CIs in Health and Medical Journals from 1976 to 2019: An Observational Study." *BMJ Open* 9 (11). https://doi.org/10.1136/bmjopen-2019-032506.

Baron, R. M., and D. A. Kenny. 1986. "The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations." *Journal of Personality and Social Psychology* 51 (6): 1173–82. https://doi.org/10.1037/0022-3514.51.6.1173.

Baumann, James F., Nancy Seifert-Kessell, and Leah A. Jones. 1992. "Effect of Think-Aloud Instruction on Elementary Students' Comprehension Monitoring Abilities." *Journal of Reading Behavior* 24 (2): 143–72. https://doi.org/10.1080/10862969209547770.

Berger, Paul, Robert Maurer, and Giovana B Celli. 2018. *Experimental Design with Application in Management, Engineering, and the Sciences*. 2nd ed. Springer. https://doi.org/10.1007/978-3-319-64583-4.

Box, G. E. P., W. G. A. Hunter, and J. S. Hunter. 1978. *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building*. Wiley.

Brodeur, Mathieu, Perrine Ruer, Pierre-Majorique Léger, and Sylvain Sénécal. 2021. "Smartwatches Are More Distracting Than Mobile Phones While Driving: Results from an Experimental Study." *Accident Analysis & Prevention* 149: 105846. https://doi.org/10.1016/j.aap.2020.105846.

Brook, Robert H., Emmett B. Keeler, Kathleen N. Lohr, Joseph P. Newhouse, John E. Ware, William H. Rogers, Allyson Ross Davies, et al. 2006. *The Health Insurance Experiment: A Classic RAND Study Speaks to the Current Health Care Reform Debate*. Santa Monica, CA: RAND Corporation.

Brucks, Melanie S., and Jonathan Levav. 2022. "Virtual Communication Curbs Creative Idea Generation." *Nature* 605 (7908): 108–12. https://doi.org/10.1038/s41586-022-04643-

y.
Bullock, J. G., Green D. P., and S. E. Ha. 2010. "Journal of Personality and Social Psychology" 98 (4): 550–58. https://doi.org/10.1037/a0018933.
Card, David, and Alan B. Krueger. 1994. "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania." *The American Economic Review* 84 (4): 772–93. http://www.jstor.org/stable/2118030.
Claerbout, Jon F., and Martin Karrenbach. 1992. "Electronic documents give reproducible research a new meaning." *SEG Technical Program Expanded Abstracts*. https://doi.org/https://doi.org/10.1190/1.1822162.
Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. New York: Routledge. https://doi.org/10.4324/9780203771587.
Cox, David R. 1958. *Planning of Experiments*. New York, NY: Wiley.
Crump, M. J. C., D. J. Navarro, and J. Suzuki. 2019. *Answering Questions with Data: Introductory Statistics for Psychology Students*. https://doi.org/10.17605/OSF.IO/JZE52.
Davison, A. C., and D. V. Hinkley. 1997. *Bootstrap Methods and Their Application*. New York, NY: Cambridge University Press.
Efron, B., and R. J. Tibshirani. 1993. *An Introduction to the Bootstrap*. Boca Raton, FL: CRC Press.
Elliott, Emily M., Candice C. Morey, Angela M. AuBuchon, Nelson Cowan, Chris Jarrold, Eryn J. Adams, Meg Attwood, et al. 2021. "Multilab Direct Replication of Flavell, Beach, and Chinsky (1966): Spontaneous Verbal Rehearsal in a Memory Task as a Function of Age." *Advances in Methods and Practices in Psychological Science* 4 (2): 1–20. https://doi.org/10.1177/25152459211018187.
Esarey, Justin, and Jane Lawrence Sumner. 2018. "Marginal Effects in Interaction Models: Determining and Controlling the False Positive Rate." *Comparative Political Studies* 51 (9): 1144–76. https://doi.org/10.1177/0010414017730080.
Faul, Franz, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. "G*Power 3: A Flexible Statistical Power Analysis Program for the Social, Behavioral, and Biomedical Sciences." *Behavior Research Methods* 39 (2): 175–91. https://doi.org/10.3758/BF03193146.
Fisher, Ronald A. 1926. "The Arrangement of Field Experiments." *Journal of the Ministry of Agriculture* 33: 503–15. https://doi.org/10.23637/rothamsted.8v61q.
Flavell, John H., David R. Beach, and Jack M. Chinsky. 1966. "Spontaneous Verbal Rehearsal in a Memory Task as a Function of Age." *Child Development* 37 (2): 283–99.
Garcia, Donna M., Michael T. Schmitt, Nyla R. Branscombe, and Naomi Ellemers. 2010. "Women's Reactions to Ingroup Members Who Protest Discriminatory Treatment: The Importance of Beliefs about Inequality and Response Appropriateness." *European Journal of Social Psychology* 40 (5): 733–45. https://doi.org/10.1002/ejsp.644.
Gelman, Andrew. 2005. "Analysis of Variance — Why It Is More Important Than Ever." *The Annals of Statistics* 33 (1): 1–53. https://doi.org/10.1214/009053604000001048.
Gelman, Andrew, and Eric Loken. 2014. "The Statistical Crisis in Science." *American*

Scientist 102: 460–65.

Hariton, Eduardo, and Joseph J Locascio. 2018. "Randomised Controlled Trials – the Gold Standard for Effectiveness Research." *BJOG: An International Journal of Obstetrics & Gynaecology* 125 (13): 1716–16. https://doi.org/https://doi.org/10.1111/1471-0528.15199.

Hedges, Larry V. 1981. "Distribution Theory for Glass's Estimator of Effect Size and Related Estimators." *Journal of Educational Statistics* 6 (2): 107–28. https://doi.org/10.3102/10769986006002107.

Imai, Kosuke, Luke Keele, and Dustin Tingley. 2010. "A General Approach to Causal Mediation Analysis." *Psychological Methods* 15 (4): 309–34. https://doi.org/10.1037/a0020761.

Ioannidis, John P. A. 2005. "Why Most Published Research Findings Are False." *PLOS Medicine* 2 (8). https://doi.org/10.1371/journal.pmed.0020124.

Johnson, P. O., and J. Neyman. 1936. "Tests of Certain Linear Hypotheses and Their Application to Some Educational Problems." *Statistical Research Memoirs* 1: 57–93.

Jones, Damon, David Molitor, and Julian Reif. 2019. "What Do Workplace Wellness Programs Do? Evidence from the Illinois Workplace Wellness Study." *The Quarterly Journal of Economics* 134 (4): 1747–91. https://doi.org/10.1093/qje/qjz023.

Keppel, G., and T. D. Wickens. 2004. *Design and Analysis: A Researcher's Handbook*. Pearson Prentice Hall.

Kohavi, Ron, and Stefan Thomke. 2017. "The Surprising Power of Online Experiments." *Harvard Business Review* September–October: 74–82. https://hbr.org/2017/09/the-surprising-power-of-online-experiments.

Kowal, Agata AND Kochan-Wójcik, Marta AND Groyecka-Bernard. 2021. "When and How Does the Number of Children Affect Marital Satisfaction? An International Survey." *PLOS ONE* 16 (4): 1–14. https://doi.org/10.1371/journal.pone.0249516.

Lakens, Daniel. 2013. "Calculating and Reporting Effect Sizes to Facilitate Cumulative Science: A Practical Primer for t-Tests and ANOVAs." *Frontiers in Psychology* 4: 863. https://doi.org/10.3389/fpsyg.2013.00863.

Läuter, J. 1978. "Sample Size Requirements for the T^2 Test of MANOVA (Tables for One-Way Classification)." *Biometrical Journal* 20 (4): 389–406. https://doi.org/10.1002/bimj.4710200410.

Lee, Kiljae, and Jungsil Choi. 2019. "Image-Text Inconsistency Effect on Product Evaluation in Online Retailing." *Journal of Retailing and Consumer Services* 49: 279–88. https://doi.org/10.1016/j.jretconser.2019.03.015.

Liu, Peggy J., SoYon Rim, Lauren Min, and Kate E. Min. 2022+. "The Surprise of Reaching Out: Appreciated More Than We Think." *Journal of Personality and Social Psychology*, 2022+. https://doi.org/10.1037/pspi0000402.

Maglio, Sam J., and Evan Polman. 2014. "Spatial Orientation Shrinks and Expands Psychological Distance." *Psychological Science* 25 (7): 1345–52. https://doi.org/10.1177/0956797614530571.

Magnusson, Kristoffer. 2021. “Interpreting Cohen’s d Effect Size: An Interactive Visualization.” https://rpsychologist.com/cohend/.

McElreath, R. 2020. *Statistical Rethinking: A Bayesian Course with Examples in r and Stan*. 2nd ed. CRC Press. https://www.routledge.com/Statistical-Rethinking-A-Bayesian-Course-with-Examples-in-R-and-STAN/McElreath/p/book/9780367139919.

McQuire, Cheryl, R. Daniel, L. Hurt, A. Kemp, and S. Paranjothy. 2020. “The Causal Web of Foetal Alcohol Spectrum Disorders: A Review and Causal Diagram.” *European Child & Adolescent Psychiatry* 29 (5): 575–94. https://doi.org/10.1007/s00787-018-1264-3.

Nosek, Brian, Johanna Cohoon, Mallory Kidwell, and Jeffrey Spies. 2015. “Estimating the Reproducibility of Psychological Science.” *Science* 349 (6251). https://doi.org/10.1126/science.aac4716.

Oehlert, Gary. 2000. *A First Course in Design and Analysis of Experiments*. W. H. Freeman. http://users.stat.umn.edu/~gary/Book.html.

Pearl, Judea. 2014. “Interpretation and Identification of Causal Mediation.” *Psychological Methods* 19 (4): 459–81. https://doi.org/10.1037/a0036434.

Pearl, Judea, Maria Glymour, and Nicholas Jewell. 2016. *Causal Inference in Statistics: A Primer*. Chichester, UK: Wiley.

Pratt, John W. 1959. “Remarks on Zeros and Ties in the Wilcoxon Signed Rank Procedures.” *Journal of the American Statistical Association* 54 (287): 655–67. https://doi.org/10.1080/01621459.1959.10501526.

Preacher, Kristopher J., and Andrew F. Hayes. 2004. “SPSS and SAS Procedures for Estimating Indirect Effects in Simple Mediation Models.” *Behavior Research Methods, Instruments & Computers* 36: 717–31. https://doi.org/10.3758/BF03206553.

Rosen, B., and T. H. Jerdee. 1974. “Influence of Sex Role Stereotypes on Personnel Decisions.” *Journal of Applied Psychology* 59: 9–14.

Sharma, Eesha, Stephanie Tully, and Cynthia Cryder. 2021. “Psychological Ownership of (Borrowed) Money.” *Journal of Marketing Research* 58 (3): 497–514. https://doi.org/10.1177/0022243721993816.

Sokolova, Tatiana, Aradhna Krishna, and Tim Döring. 2023. “Paper Meets Plastic: The Perceived Environmental Friendliness of Product Packaging.” *Journal of Consumer Research* 50 (3): 468–91. https://doi.org/10.1093/jcr/ucad008.

Song, Zirui, and Katherine Baicker. 2019. “Effect of a Workplace Wellness Program on Employee Health and Economic Outcomes: A Randomized Clinical Trial.” *JAMA* 321 (15): 1491–501. https://doi.org/10.1001/jama.2019.3307.

Steiger, James H. 2004. “Beyond the F Test: Effect Size Confidence Intervals and Tests of Close Fit in the Analysis of Variance and Contrast Analysis.” *Psychological Methods* 9: 164–82. https://doi.org/10.1037/1082-989X.9.2.164.

Stekelenburg, Aart van, Gabi Schaap, Harm Veling, and Moniek Buijzen. 2021. “Boosting Understanding and Identification of Scientific Consensus Can Help to Correct False Beliefs.” *Psychological Science* 32 (10): 1549–65. https://doi.org/10.1177/09567976211007788.

Tingley, Dustin, Teppei Yamamoto, Kentaro Hirose, Luke Keele, and Kosuke Imai. 2014.

"mediation: R Package for Causal Mediation Analysis." *Journal of Statistical Software* 59 (5): 1–38. https://doi.org/10.18637/jss.v059.i05.

VanderWeele, Tyler. 2015. *Explanation in Causal Inference: Methods for Mediation and Interaction*. New York, NY: Oxford University Press.

Weissgerber, Tracey L., Natasa M. Milic, Stacey J. Winham, and Vesna D. Garovic. 2015. "Beyond Bar and Line Graphs: Time for a New Data Presentation Paradigm." *PLOS Biology* 13 (4): 1–10. https://doi.org/10.1371/journal.pbio.1002128.

Weissgerber, Tracey L., Stacey J. Winham, Ethan P. Heinzen, Jelena S. Milin-Lazovic, Oscar Garcia-Valencia, Zoran Bukumiric, Marko D. Savic, Vesna D. Garovic, and Natasa M. Milic. 2019. "Reveal, Don't Conceal." *Circulation* 140 (18): 1506–18. https://doi.org/10.1161/CIRCULATIONAHA.118.037777.

Yates, F. 1964. "Sir Ronald Fisher and the Design of Experiments." *Biometrics* 20 (2): 307–21. http://www.jstor.org/stable/2528399.

Zhao, Xinshu, Jr. Lynch John G., and Qimei Chen. 2010. "Reconsidering Baron and Kenny: Myths and Truths about Mediation Analysis." *Journal of Consumer Research* 37 (2): 197–206. https://doi.org/10.1086/651257.

Zwet, Erik W. van, and Eric A. Cator. 2021. "The Significance Filter, the Winner's Curse and the Need to Shrink." *Statistica Neerlandica*. https://doi.org/https://doi.org/10.1111/stan.12241.