Answering questions with data

Mallory Barnes

2025-08-20

This comprehensive resource offers a free, accessible textbook for environmental science students embarking on introductory statistics. The package includes a practical lab manual and a dedicated course website, all provided under a CC BY-SA 4.0 license.

Table of contents

Preface

Second Draft (version 0.1 = August 18th, 2025)

Welcome to the second edition of this Open Educational Resource (OER) textbook, specifically adapted for Environmental Science students enrolled in the SPEA E-538 statistics course at Indiana University (IU).

This textbook is an adaptation of a thorough introductory statistics textbook originally developed for undergraduate Psychology students by Matthew Crump and colleagues (refer to Acknowledgements for more details). As part of IU's Course Materials Fellowship Program (CMFP), I've had the opportunity to mold this material, refining it to serve as a specialized resource for students studying Environmental Science.

Online Textbook:https://malloryb.github.io/statistics_E538/

Citation for original textbook: Crump, M. J. C., Navarro, D. J., & Suzuki, J. (2019, June 5). Answering Questions with Data (Textbook): Introductory Statistics for Psychology Students. https://doi.org/10.17605/OSF.IO/JZE52

All resources are released under a creative commons licence CC BY-SA 4.0. Click the link to read more about the license, or read more below in the license section.

Acknowledgements

I wish to express my deepest appreciation to the contributors of the original textbook, without whom this adaptation would not have been possible. I am deeply grateful for the expertise and vision of Matthew Crump, Alla Chavarga, Anjali Krishnan, Jeffrey Suzuki, and Stephen Volz. Their exceptional groundwork laid the foundation for this project.

My heartfelt thanks also go to the Course Materials Fellowship Program (CMFP) at Indiana University (IU), which has been instrumental in supporting this adaptation effort. The CMFP is an initiative designed to incentivize the discovery, implementation, and creation of cost-effective course materials. Its aim is to foster the use of Open Educational Resources (OERs)—freely accessible and customizable learning materials that make education more equitable and accessible.

I am particularly grateful to Sarah Hare, the Bloomington lead for the CMFP program, and Adam Mazel, a digital publishing librarian at IU. Their guidance, expertise, and steadfast

support have been crucial to this project's success. Their contributions to the advancement of affordable and accessible education are truly commendable.

CC BY-SA 4.0 license

This license means that you are free to:

- Share: copy and redistribute the material in any medium or format
- Adapt: remix, transform, and build upon the material for any purpose, even commercially.

The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:

- Attribution: You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.
- ShareAlike: If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original.
- No additional restrictions: You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

Copying the textbook

This textbook was written in R-Studio, using R Markdown, and compiled into a web-book format using the bookdown package. In general, I thank the larger R community for all of the amazing tools they made, and for making those tools open, so that I could use them to make this thing.

All of the source code for compiling the book is available from the GitHub repository for the original textbook:

https://github.com/CrumpLab/statistics

and my github repository:

https://github.com/malloryb/statistics_E538

In principle, anyone can fork or download the E-538 textbook, or the original textbook, which is what I did. You can load the Rproj file in RStudio, compile the entire book, and then edit the individual .Rmd files for content and style to fit your needs.

If you'd like to contribute to this version, you can submit pull requests on GitHub or use the issues tab to share suggestions.

The vision behind this textbook

The aim of this textbook is twofold: to make core statistical concepts accessible to Environmental Science students, and to promote the use of open-source tools like R as flexible, transparent resources for learning and research.

1 Why Statistics?

Adapted to environmental science by Mallory Barnes. Portions adapted from Chapters 1 and 2 in Navarro, D. J. "Learning Statistics with R." https://compcogscisydney.org/learning-statistics-with-r/

> To call in statisticians after the experiment is done may be no more than asking them to perform a post-mortem examination: They may be able to say what the experiment died of.*
> – Sir Ronald Fisher

1.1 The Role of Statistics in Environmental Science

Many students come into statistics classes with a mix of nerves and low expectations. It is not always the most eagerly anticipated part of an environmental science degree. But statistics is one of the most important tools we have for understanding complex systems. Environmental data are messy: ecosystems shift, climate varies, and our judgments are often biased. Without statistical methods we risk telling convenient stories. With them, we can detect real patterns, test competing explanations, and make stronger arguments for environmental decisions.

1.1.1 Why Numbers Matter

Common sense is useful, but it is prone to bias. When we already believe something to be true, we are more likely to interpret new evidence as supporting it, even if it does not. This tendency can lead us astray in environmental science.

For example, if you are convinced that industrial farming is the main driver of bee declines, you might see every new drop in bee abundance as confirmation of that belief. Maybe you are right, but maybe disease or weather are stronger contributors. Statistics gives us tools to test these ideas systematically. It helps us separate patterns from noise, avoid being misled by our own expectations, and draw conclusions that are more likely to hold up under scrutiny. It's not magic, but it makes our conclusions far more reliable.

Another challenge is that data can tell very different stories depending on how they are aggregated. **Simpson's paradox** occurs when a trend seen in combined data reverses after you split the data into meaningful groups.

Here's a recent example. During the COVID-19 pandemic, early data showed that Italy's overall case fatality rate was higher than China's. But, once researchers *disaggregated by age group,* the trend flipped: within every age group, the fatality rate was actually higher in China. The apparent contradiction was explained by differences in the age distribution of cases (von Kügelgen, Gresele, and Schölkopf 2021).

Using case fatality rate (CFR), here are some illustrative numbers (made up for teaching).

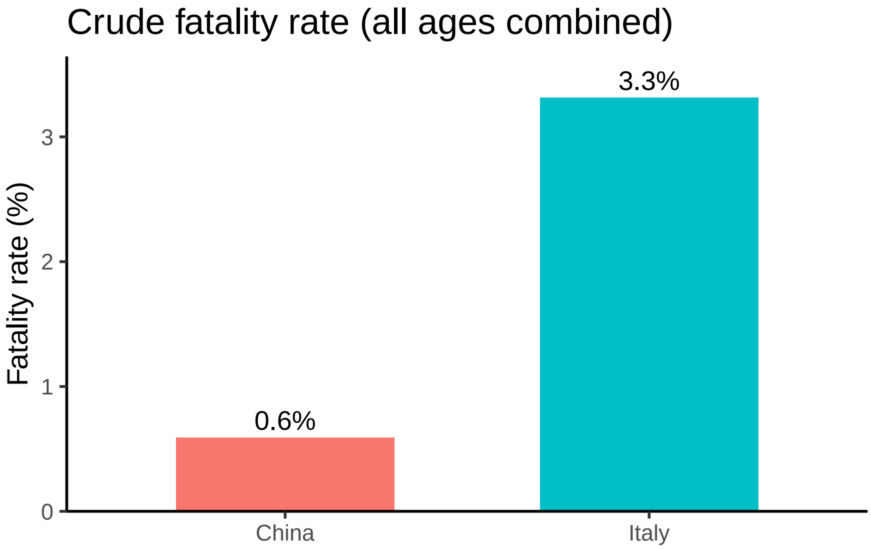


Figure 1.1: Crude case fatality rates. Italy appears to have a higher fatality rate than China when all cases are combined.

At face value, Italy looks worse. But what happens when we stratify by age group? Because age strongly affects COVID fatality, splitting cases into 10-year bands reveals a different picture. When we do, China's fatality rate is higher in every band.

Why the reversal? Italy's cases skew older while China's skew younger, and baseline risk rises steeply with age. The differing age distributions weight the overall averages in opposite ways. Aggregating without age adjustment hid the within-group pattern.

i Takeaway

Statistics will not answer every scientific question, but it gives us a disciplined way to sort signal from bias so our inferences age well.

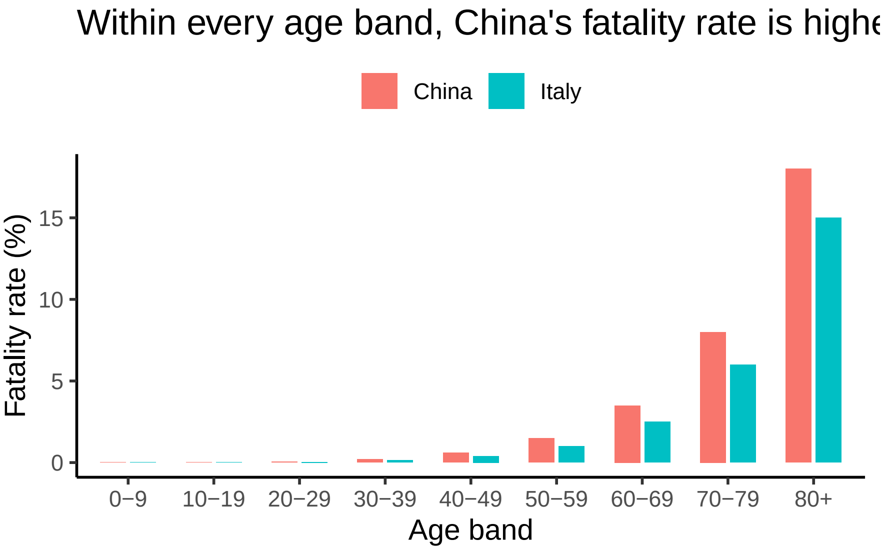


Figure 1.2: Fatality rates by age group (same made-up numbers). Within each age band, China's CFR exceeds Italy's.

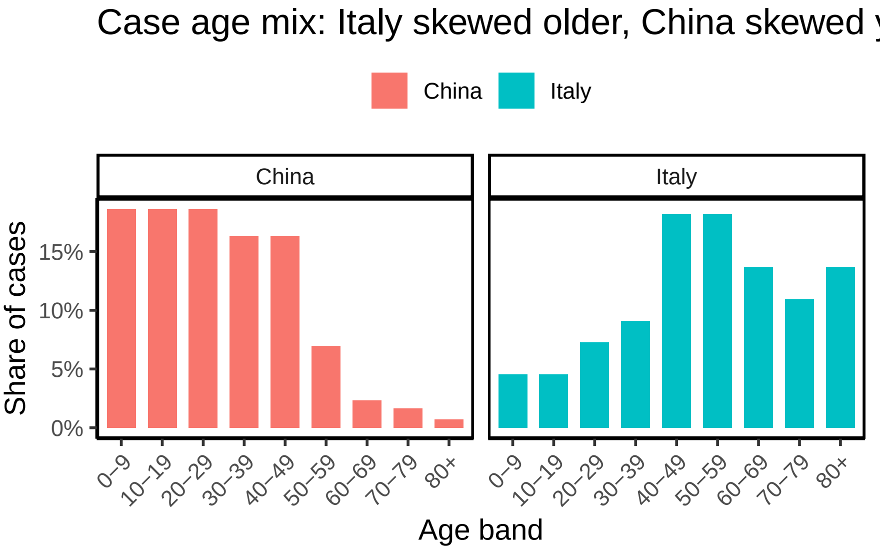


Figure 1.3: Case age distribution by country. Italy's cases are older on average; China's are younger.

1.1.2 What Statistics Add

Environmental science produces a *lot* of data. Every day satellites measure surface temperature, field stations log rainfall, and sensors track air quality. A single study can generate thousands of rows of numbers. Without statistical tools, these numbers would be overwhelming. With them, we can summarize, compare, and test ideas systematically.

Statistics matters here because environmental systems are messy. River flow changes by the hour, forests grow unevenly, and human actions layer on additional variability. Patterns are rarely obvious by eye. Statistics helps us sort signal from noise and ask: is this apparent change real, or just random variation?

You might think, "can't a statistician just handle the math?" But knowing the basics yourself is essential for three reasons:

1. **Design and analysis go together.** A good study starts long before you run a statistical test. If you want to study how fertilizer affects crop yields, your sampling plan and your analysis are inseparable. Poor design, like measuring only in unusually wet fields, can't be rescued by sophisticated analysis.
2. **Understanding the Science:** Scientific papers on climate change, biodiversity, or pollution are built on statistical results. To interpret them, you need to know what the numbers mean.
3. **Practicality:** Hiring a specialist for every question isn't realistic. A working knowledge of statistics makes you more self-sufficient as a scientist.

> *"We are drowning in information, but we are starved for knowledge"*
>
> – Various authors, original probably John Naisbitt

Finally, statistics is not only for researchers. Weather forecasts, air quality alerts, and wildlife population trends are all communicated through numbers. A basic knowledge of statistics helps you tell whether those numbers support the claims being made, or whether someone is stretching the truth. In that sense, statistics is part of being an informed citizen in a data-saturated world.

1.2 Introduction to measurements

Every dataset begins with measurement. Measurement means assigning numbers, labels, or categories to aspects of the world so they can be recorded and analyzed.

Why measurement matters

Environmental science often deals with broad ideas that need to be defined before they can be studied. For example, *soil health* or *forest change* are meaningful, but vague. To analyze them, we have to decide exactly what we mean and how to capture it. That process is called **operationalization**: turning a general concept into something measurable.

Before moving on, let's clarify how operationalization fits with three related terms:

- **Operationalization**: the step where you define exactly how a broad idea will be captured with a measure
- **Measure**: the tool or method used to make observations
- **Variable**: the actual values you record once the measure is applied. Variables are the actual data that end up in our dataset.

Examples:

- Concept of interest: *Soil carbon sequestration*
  - **Operationalization:** concentration of organic carbon in the top 30 cm
  - **Measure:** laboratory analysis of soil samples
  - **Variable:** recorded carbon concentration values for each plot
- Concept of interest: *Deforestation*
  - **Operationalization**: change in forest cover between the current year and five years ago
  - **Measure:** aerial image classification
  - **Variable:** change in forest cover (%) per unit area

Operationalization is rarely straightforward, and there is no single correct way to do it. The best choice depends on the question you are asking and how the data will be used. In many fields, scientists have developed common practices, but every project still requires case-by-case judgment. Even so, some principles of good operationalization apply across studies: be precise about what you mean, how you will measure it, and what values are possible.

1.3 Scales of measurement

As the previous section indicates, the outcome of a measurement is called a **variable**. Not all variables are the same type, and knowing the type matters because it determines which statistical tools make sense. The four classic scales are **nominal, ordinal, interval,** and **ratio**.

1.3.1 Nominal scale

A **nominal scale** variable (also referred to as a **categorical** variable) is one in which the values are just names. They do not have an inherent order, and it makes no sense to average them.

Example: eye color. Eyes can be blue, green, or brown, but none is "greater" than another, and there's no such thing as an "average eye color."

1.3.2 Ordinal scale

Ordinal scale variables have a meaningful order, but the spacing between values is not defined mathematically. You can rank the values, but you can't assume equal steps between them, nor calculate a meaningful average.

Example: finishing position in a race. First place comes before second, and second before third, but you don't know how much faster one runner was than another. Saying the "average place" of the group is 2.3 doesn't mean anything.

1.3.3 Interval scale

Interval variables have equal intervals between values, so differences are meaningful. However, zero is arbitrary, so multiplication and division are not valid.

Example: temperature in degrees Celsius. A difference of 3° means the same regardless of whether it is $7 \to 10$ or $15 \to 18$. But 0° does not mean "no temperature." That zero is defined by the freezing point of water. So while you can say today is 3° warmer than yesterday, you cannot say 20° is "twice as hot" as 10°.

(Alternate example: latitude. A difference of 10° latitude is meaningful, but zero latitude—the equator—doesn't mean "no latitude." Twice 45° is 90°, but that doesn't mean one place has "twice as much latitude" as another.)In contrast to nominal and ordinal scale variables, **interval scale** and ratio scale variables are variables for which the numerical value is genuinely meaningful. In the case of interval scale variables, the *differences* between the numbers are interpretable, but the variable doesn't have a "natural" zero value. You can add and subtract, but ratios don't make sense.

1.3.4 Ratio scale

Ratio scale variables have all the properties of interval variables, plus a true zero. This makes multiplication and division valid, as well as addition and subtraction.

Example: age in years. Zero means no age at all, and someone who is 20 years old really is twice as old as someone who is 10. Differences (20 − 10 = 10 years) and ratios (20 ÷ 10 = 2) are both meaningful.

	can rank	can subtract/add	can multiply/divide	example
nominal				eye color
ordinal	x			race position
interval	x	x		temperature (°C)
ratio	x	x	x	age

1.3.5 Continuous versus discrete variables

Another useful distinction is whether a variable can take on values in between others.

- A **continuous variable** can, in principle, take on any value within a range. For example, consider height. If you are 72 inches tall and your friend Cameron is 71 inches tall, Alan could be 71.4 inches and David 71.49 inches. Because we can always imagine a new value in between two others, height is continuous.
- A **discrete variable** is, in effect, a variable that isn't continuous. Discrete variables have separate, distinct values with nothing in between. Nominal variables are always discrete: there is no eye color that falls "between" green and blue in the same way that 71.4 falls between 71 and 72. Ordinal variables are also discrete: although second place falls between first and third, nothing can logically fall between first and second.

Interval and ratio variables can be either. Height (a ratio variable) is continuous. But the number of people living in a household (a ratio variable) is discrete: you cannot have 4.2 people. Temperature in degrees Celsius (an interval variable) is also continuous. But the year you started college (an interval variable) is discrete: there is no year between 2022 and 2023.

The table below shows how the scales of measurement relate to this distinction. Cells with an "x" mark what is possible.

	continuous	discrete
nominal		x
ordinal		x

	continuous	discrete
interval	x	x
ratio	x	x

1.3.6 A note on real data

These categories are guides, not hard rules. For example, survey responses on a 1–5 “strongly disagree” to “strongly agree” scale are technically **ordinal**, but researchers often treat them as “quasi-interval” because the spacing is assumed to be roughly equal.

1.4 Assessing the reliability of a measurement

When we measure something, two questions matter: *is the measurement consistent, and is it accurate?*

- **Reliability** means consistency. If you measure the same thing again, do you get the same result?
- **Validity** means accuracy. Does the measurement reflect the real thing you care about?

These aren’t the same. A soil moisture probe that is miscalibrated might always read 5% too high. It is highly **reliable**, you get the same number each time, but not **valid**, since it doesn’t reflect true soil moisture. On the other hand, a set of volunteer bird counts might fluctuate from one observer to the next (**low reliability**), but when averaged across many counts, the results may still approximate true abundance (**reasonable validity**).

In practice, a measure that is very unreliable usually ends up being invalid too, because we can’t tell which result is right. This is why reliability is often considered a prerequisite for validity.

- Reliability can show up in different ways:
  - **Over time** (*test–retest*): if you sample soil moisture today and tomorrow under the same conditions, do you get the same value?
  - **Across people** (*inter-rater*): if two field crews count birds at the same site, do their tallies agree?
  - **Across tools** (*parallel forms*): if you use two different rain gauges in the same spot, do they record the same rainfall?
  - **Within a test** (*internal consistency*): if multiple survey questions are meant to capture the same attitude, do they give similar answers?

Not every measurement needs (or can even possibly have) every form of reliability. The point is that **reliability is necessary but not sufficient for validity**: a method can be consistently wrong, but if it isn't consistent at all, it's almost impossible to know whether it's right. We'll discuss how we assess validity a bit later in this chapter.

1.5 The role of variables: predictors and outcomes

One last piece of terminology before we leave variables behind. In most studies we have many variables, but when we analyze them, we usually split them into two roles: the **thing we're trying to explain** and the **thing doing the explaining**. To keep it straight, we use Y for the variable being explained, and a X_1, X_2, etc. for the variables used to explain it.

Traditionally, X is called the **indepdendent variable (IV)** and Y is the **dependent variable (DV).** The logic is that if there's a relationship, Y depends on X. These terms can be clunky and can be confusing because: (a) IVs are rarely actually "independent of everything else" and (b) if there's no relationship, then the DV doesn't actually "depend" on the IV at all.

Alternative terminology is often clearer. In experiments, IVs are **manipulations**, and DVs are **measurements:**

role of the variable	classical name	alternative
"to be explained" (Y)	dependent variable (DV)	measurement
"to do the explaining" (X)	independent variable (IV)	manipulation

Another useful pair is **predictors** and **outcomes**. The idea here is that what we use X to make predict or guess about Y:

role of the variable	classical name	alternative
"to be explained" (Y)	dependent variable (DV)	outcome
"to do the explaining" (X)	independent variable (IV)	predictor

1.6 Experimental and non-experimental research

A central distinction in research is between experimental and non-experimental studies. What matters here is the degree of control the researcher has.

In **experimental research**, the researcher deliberately manipulates something (the predictor/IV) and measures its effect on outcomes (the outcome/DV). The goal is to isolate causal effects. To avoid the problem of “something else” influencing the outcome, researchers try to hold other factors constant. In practice, it’s almost impossible to identify *everything* that might matter, much less keep it constant. The standard solution is **randomization**. Randomization doesn’t eliminate confounds, but it makes them less likely to systematically bias results.

For instance, suppose we wanted to know if smoking causes lung cancer. Observing smokers and non-smokers can only get us so far, because those groups differ in many ways besides smoking, like occupation, income, and diet. A true experiment would require randomly assigning people to smoke or not. You can see how that would be deeply unethical. The same problem comes up in medicine: we know surprisingly little about how certain drugs or exposures affect pregnant people, precisely because we cannot ethically assign them to risky conditions.

In **environmental science**, the limitation isn’t usually ethics but feasibility: there’s no “control planet” we can set aside as a baseline. Nor can researchers manipulate fundamental drivers like precipitation or temperature. As a result, much of the field relies on **non-experimental research** — quasi-experiments, long-term time series, or detailed case studies that allow scientists to tease out effects in complex systems.

Much environmental science fits a **quasi-experimental design**. For example, researchers may wish to study the effects of industrial pollution on a river system but cannot directly control the pollutants emitted. Instead, they observe existing conditions and make careful comparisons, often using statistical tools to account for confounding variables.

Another important quasi-experimental tool is **time series analysis**, since many environmental processes unfold over decades or centuries. Tracking changes in global temperature or sea level requires years of consistent data, and statistical methods help isolate signals from the noise of natural variability.

Environmental science also makes extensive use of **case studies**, which provide in-depth insights into specific events or locations. A case study might explore the aftermath of a natural disaster, the ecology of a threatened habitat, or the consequences of an environmental policy. While a single case is tied to a particular context, well-designed studies can uncover mechanisms that apply more broadly. For example, a drought in one forest might reveal how water potential and transpiration respond to stress, helping researchers anticipate plant responses in other ecosystems.

1.7 Assessing the validity of a study

Earlier we drew a line between **reliability** (consistency) and **validity** (accuracy). Reliability tells us whether our measurement is stable and repeatable. Validity asks the harder question: are we actually measuring what we think we’re measuring, and can we trust the conclusions we draw from it?

1.7.1 Internal validity

Internal validity is about whether the relationship you see is really causal. From an earlier example: if smokers have more lung cancer than non-smokers, does that mean smoking causes cancer? Not necessarily. Smokers may differ from non-smokers in income, occupation, diet, or other ways. These "confounds" muddy causal claims. Experiments with random assignment improve internal validity by balancing out such factors, but in environmental science, we often rely on careful design and statistical controls instead.

1.7.2 External validity

External validity is about generalization. Does what you found in one study hold elsewhere? A soil-warming experiment in one forest plot may not capture how all forests respond. Or water-quality measurements from a single river reach may not represent conditions across the whole watershed. Strong internal validity can still leave you with weak external validity if the setting, population, or conditions are narrow.

1.7.3 Construct validity

Construct validity is about whether your measure really matches the concept. If you care about soil fertility but measure only soil moisture, you have high reliability but weak construct validity: you're consistently measuring the wrong thing. Getting this right requires aligning theory, measurement, and data.

1.8 Confounds, artifacts and other threats to validity

Broadly, the two biggest threats to validity are *confounds* and *artifacts*:

- **Confound**: An additional variable, often unmeasured, that masks or distorts the relationshp between the predictor and the outcome. Confounds threaten *internal validity* because you can't tell whether the predictor really causes the outcome. Two variables are said to be **confounded** if their effects on a response variable cannot be distinguished.
  - **Example:** In a study of tree growth near roads, trees closer to the road might appear to grow faster. But if those same trees also receive more runoff water from pavement, then water availability (not proximity to the road) could explain the growth difference.

- **Artifact**: An aspect of the experimental setup or apparatus that biases the results. An artifact gives the appearance of measuring the phenomenon of interest, but is actually measuring something else introduced by the method itself. Artifacts often undermine *external validity,* and sometimes internal validity too.
  - **Example:** In soft-sediment studies, repeated coring to sample buried organisms disturbs the sediment, exposing animals and disrupting microbial and structural assemblages; later samples then reflect the disturbance caused by earlier sampling rather than true ecological patterns (Skilleter 1996)

Here are some more types of threats to validity:

- **History Effects:** Events outside the study can shift the outcome. Long-term ecosystem monitoring is especially vulnerable: imagine tracking stream chemistry for years, then a wildfire in the watershed alters nutrient fluxes. Your "trend" may just reflect that disturbance rather than the process you meant to study.
- **Repeated Testing Effects:** Even measurements themselves can have an effect. For example, extracting multiple tree-ring cores from the same tree could wound it and reduce growth in later years, influencing the very variable you were trying to measure.
- **Selection bias:** Environmental research often focuses on sites that are easiest to study, like forests near R1 universities, long-established research stations, or accessible watersheds. But those places aren't necessarily representative. If they differ systematically from the broader set of ecosystems, conclusions may not generalize.
- **Differential Attrition:** Long-term field studies depend on cooperation, and dropout can skew results. For example, in a cover-cropping trial, if only the most motivated farmers stick with the practice, the observed benefits won't generalize to all farms.
- **Non-response bias:** Surveys face the same problem: those who answer aren't random. For instance, a survey on attitudes toward conservation easements may mostly attract landowners already supportive of conservation, biasing the results.
- **Regression to the mean:** Extreme cases tend to moderate over time. If you select the fastest-growing trees in a stand for special study, you'll probably find their growth slows later — not necessarily because of your treatment, but simply because growth rates fluctuate and extremes rarely persist.
- **Experimenter bias:** Researchers bring expectations to the field. Choices about where to sample a river, or when to measure soil respiration, can unconsciously favor the expected outcome. The classic cautionary tale is Clever Hans, the horse who "did math" by responding to subtle human cues (Pfungst 1911; Hothersall 2004). The lesson: be aware of how your own expectations can shape results.

- **Reactivity and demand effects:** Sometimes behavior changes because people know they're being observed. A field crew told they are testing the "impact" of a method may record data more carefully than usual, producing results that reflect observer effort rather than the phenomenon itself.

1.8.1 Fraud, deception, and self-deception

> *It is difficult to get a man to understand something, when his salary depends on his not understanding it.*
> – Upton Sinclair

Most scientists are honest, but self-deception is common. Researchers can over-interpret noisy results, design studies that all but guarantee an effect, or (consciously or not) hide inconvenient variables. Publication bias makes it worse: null results often go unreported, so the literature overstates effects. And sometimes, as in Simpson's paradox, aggregated data can mislead if you don't dig deeper. The point isn't to be cynical, it's to recognize that science is done by humans, and humans bring bias, incentives, and blind spots.

1.9 Summary

In this chapter, we have explored essential aspects of research methodology pertinent to environmental statistics:

- **The role of statistics in environmental science:** Statistics does not answer every substantive question, but it provides the discipline to separate signal from noise and make credible inferences from messy data.
- **Operationalization and Measurement**: Defining theoretical constructs and deciding how to measure them is foundational.
- **Scales of measurement and types of variables**: Distinguishing discrete from continuous data, and recognizing nominal, ordinal, interval, and ratio scales.
- **Reliability of a measurement**: If I measure the "same" thing twice, should I expect the same result? In what sense (test–retest, inter-rater, internal consistency)?
- **Terminology: predictors and outcomes**: Can I clearly explain the roles variables play in an analysis (predictor vs. outcome, independent vs. dependent)?
- **Experimental and non-experimental research designs**: Identifying what constitutes an experiment, and how researchers rely on quasi-experiments, time series, and case studies when full control is not possible.

- **Validity and Threats**: Does the study truly measure what it claims to? What pitfalls—such as confounds or artifacts—could bias results?

Study design is a cornerstone of environmental research methodology. While many textbooks provide fuller coverage, e.g., Campbell and Stanley (1963), this chapter highlights the unique challenges of applying these principles in environmental science, where ethical and practical constraints often preclude perfect control. By linking statistics to study design, we see how careful methods make it possible to draw credible inferences from complex, real-world systems.

1.10 Videos

1.10.1 Terms of Statistics

2 Describing Data

> Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise. —John W. Tukey

This chapter is about **descriptive statistics**. These are tools for describing data. Some things to keep in mind as we go along are:

1. There are lots of different ways to describe data
2. There is more than one "correct" way, and you get to choose the most "useful" way for the data that you are describing
3. It is possible to invent new ways of describing data, all of the ways we discuss were previously invented by other people, and they are commonly used because they are useful.
4. Describing data is necessary because there is usually too much of it, so it doesn't make any sense by itself.

2.1 This is what too many numbers looks like

Let's say you wanted to know how happy people are. So, you ask thousands of people on the street how happy they are. You let them pick any number they want from negative infinity to positive infinity. Then you record all the numbers. Now what?

Well, how about you look at the numbers and see if that helps you determine anything about how happy people are. What could the numbers look like. Perhaps something like this:

153	-277	371	247	660	194	753	569	243	90
335	-702	88	483	533	-670	-163	1019	-807	-406
137	18	97	57	-111	-9	223	-309	808	479
149	299	-87	-288	-65	50	313	-571	255	162
215	-482	254	-283	20	-243	8	806	146	53
-222	-509	560	58	-290	-511	-581	-483	-195	540
239	-376	-709	481	447	-496	-504	327	-558	-1445
-259	-785	351	199	-324	311	-623	463	918	1079
112	-451	-243	-448	-99	-709	-211	1156	-327	-124

-23	-5	688	456	415	-50	413	-94	-630	197
361	-423	875	594	77	593	-325	273	-496	-192
274	-192	658	227	92	-182	25	-13	610	576
309	-287	1712	612	77	611	-282	140	-341	425
1287	223	-418	645	-215	931	157	170	338	-910
-1033	256	176	795	462	108	934	-308	1133	944
250	228	479	129	102	-392	872	114	118	-187
0	-437	-59	175	-509	644	26	545	-475	-438
369	90	-446	-46	314	417	289	-130	-608	282
19	109	228	-344	789	-273	-168	73	-504	364
640	1097	947	1180	-588	-356	114	143	252	152
-709	1017	-185	366	-325	41	-147	-285	222	-88
-168	954	-59	196	191	-646	114	733	-335	-83
610	584	676	701	369	-688	-661	852	-108	169
643	191	1126	553	-133	885	-603	421	112	328
-433	-85	-522	301	-359	307	292	147	251	-167
-297	206	92	140	475	1165	708	-89	-270	-233
450	-566	454	421	-1237	284	-191	-276	-91	611
-253	748	-161	-843	634	420	-164	89	79	-259
302	618	-742	24	-765	1021	-293	753	-381	-312
111	215	157	689	541	189	-5	-762	490	117
-537	598	26	714	190	329	636	-215	178	-121
-270	-247	525	725	26	-62	-70	-72	566	693
764	94	588	-366	207	440	844	331	-6	199
-124	428	90	153	392	-882	40	298	-223	8
112	278	160	277	185	-434	465	1376	385	175
-511	-379	-298	927	-680	711	-509	-362	201	45
351	455	-170	429	-682	-338	-82	-187	32	-683
111	639	599	-295	-259	295	316	-25	402	432
392	-478	106	652	-140	667	237	998	-997	668
-312	476	-316	728	350	117	-658	0	812	386
-569	152	211	1064	9	-480	851	-25	863	194
-113	76	642	620	70	439	86	-50	204	-34
-202	-556	270	-175	1016	2	-764	222	-41	116
-210	1139	669	354	357	-201	-495	432	-278	-270
-23	-220	-156	-719	800	125	-243	-424	-153	1074
-330	-272	-617	-294	613	958	206	204	408	414
-710	121	1209	-108	-708	361	27	429	748	-128
-357	402	-142	699	267	-216	704	915	387	488
-375	1217	926	184	-597	561	-118	218	849	-391
-347	737	599	-139	-84	359	-294	221	-13	114

Now, what are you going to with that big pile of numbers? Look at it all day long? When you deal with data, it will deal so many numbers to you that you will be overwhelmed by them. That is why we need ways to describe the data in a more manageable fashion.

The complete description of the data is always the data itself. **Descriptive statistics** and other tools for describing data go one step further to summarize aspects of the data. Summaries are a way to compress the important bits of a thing down to a useful and manageable tidbit. It's like telling your friends why they should watch a movie: you don't replay the entire movie for them, instead you hit the highlights. Summarizing the data is just like a movie preview, only for data.

2.2 Look at the data

We already tried one way of looking at the numbers, and it wasn't useful. Let's look at some other ways of looking at the numbers, using graphs.

2.2.1 Stop, time to plot!

Let's turn all of the numbers into dots, then show them in a graph. Note, when we do this, we have not yet summarized anything about the data. Instead, we just look at all of the data in a visual format, rather than looking at the numbers.

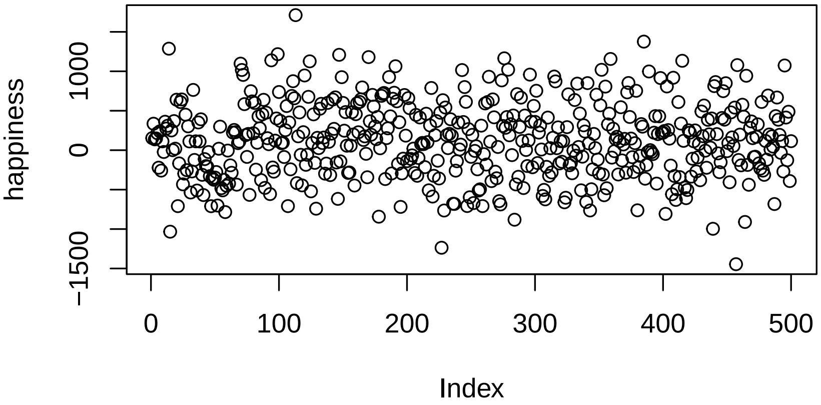


Figure 2.1: Pretend happiness ratings from 500 people

Figure 2.1 shows 500 measurements of happiness. The graph has two axes. The horizontal **x-axis**, going from left to right is labeled "Index". The vertical **y-axis**, going up and down, is labelled "happiness". Each dot represents one measurement of every person's happiness from our pretend study. Before we talk about what we can and cannot see about the data, it is worth mentioning that the way you plot the data will make some things easier to see and some things harder to see. So, what can we now see about the data?

There are lots of dots everywhere. It looks like there are 500 of them because the index goes to 500. It looks like some dots go as high as 1000-1500 and as low as -1500. It looks like there are more dots in the middle-ish area of the plot, sort of spread about 0.

> Take home: we can see all the numbers at once by putting them in a plot, and that is much easier and more helpful than looking at the raw numbers.

OK, so if these dots represent how happy 500 people are, what can we say about those people? First, the dots are kind of all over the place, so different people have different levels of happiness. Are there any trends? Are more people happy than unhappy, or vice-versa? It's hard to see that in the graph, so let's make a different one, called a **histogram.**

2.2.2 Histograms

Making a histogram will be our first act of officially summarizing something about the data. We will no longer look at the individual bits of data, instead we will see how the numbers group together. Let's look at Figure 2.2, a histogram of the happiness data, and then explain it.

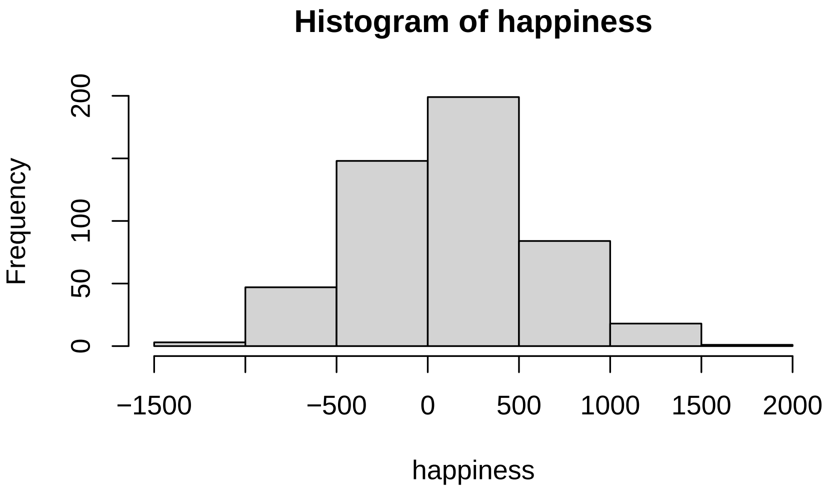


Figure 2.2: A histogram of the happiness ratings

The dots have disappeared, and now we some bars. Each bar is a summary of the dots, representing the number of dots (frequency count) inside a particular range of happiness, also called **bins**. For example, how many people gave a happiness rating between 0 and 500? The fifth bar, the one between 0 and 500 on the x-axis, tells you how many. Look how tall that bar is. How tall is it? The height is shown on the y-axis, which provides a frequency count (the number of dots or data points). It looks like around 150 people said their happiness was between 0-500.

More generally, we see there are many bins on the x-axis. We have divided the data into bins of 500. Bin #1 goes from -2000 to -1500, bin #2 goes from -1500 to -1000, and so on until the last bin. To make the histogram, we just count up the number of data points falling inside each bin, then plot those frequency counts as a function of the bins. Voila, a histogram.

What does the histogram help us see about the data? First, we can see the **shape** of data. The shape of the histogram refers to how it goes up and down. The shape tells us where the data is. For example, when the bars are low we know there isn't much data there. When the bars are high, we know there is more data there. So, where is most of the data? It looks like it's mostly in the middle two bins, between -500 and 500. We can also see the **range** of the data. This tells us the minimums and the maximums of the data. Most of the data is between -1500 and +1500, so no infinite sadness or infinite happiness in our data-set.

When you make a histogram you get to choose how wide each bar will be. For example, below are four different histograms of the very same happiness data. What changes is the width of the bins.

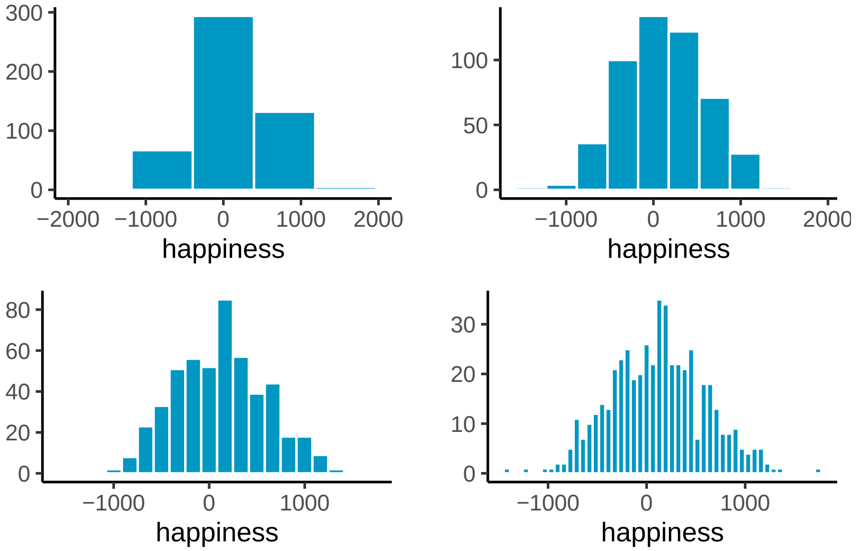


Figure 2.3: Four histograms of the same data using different bin widths

All of the histograms have roughly the same overall shape: From left to right, the bars start off small, then go up, then get small again. In other words, as the numbers get closer to zero, they start to occur more frequently. We see this general trend across all the histograms. But, some aspects of the trend fall apart when the bars get really narrow. For example, although the bars generally get taller when moving from -1000 to 0, there are some exceptions and the bars seem to fluctuate a little bit. When the bars are wider, there are less exceptions to the general trend. How wide or narrow should your histogram be? It's a Goldilocks question. Make it just right for your data.

2.3 Important Ideas: Distribution, Central Tendency, and Variance

Let's introduce three important terms we will use a lot, **distribution**, **central tendency**, and **variance**. These terms are similar to their everyday meanings (although I suspect most people don't say central tendency very often).

Distribution. When you order something from Amazon, where does it come from, and how does it get to your place? That stuff comes from one of Amazon's distribution centers. They distribute all sorts of things by spreading them around to your doorstep. "To Distribute" " is to spread something. Notice, the data in the histogram is distributed, or spread across the bins. We can also talk about a distribution as a noun. The histogram is a distribution of the frequency counts across the bins. Distributions are **very, very, very, very, very** important. They can have many different shapes. They can describe data, like in the histogram above. And as we will learn in later chapters, they can **produce** data. Many times we will be asking questions about where our data came from, and this usually means asking what kind of distribution could have created our data (more on that later.)

Central Tendency is all about sameness: What is common about some numbers? For example, is there anything similar about all of the numbers in the histogram? Yes, we can say that most of them are near 0. There is a tendency for most of the numbers to be centered near 0. Notice we are being cautious about our generalization about the numbers. We are not saying they are all 0. We are saying there is a tendency for many of them to be near zero. There are lots of ways to talk about the central tendency of some numbers. There can even be more than one kind of tendency. For example, if lots of the numbers were around -1000, and a similar large amount of numbers were grouped around 1000, we could say there was two tendencies.

Variance is all about different*ness*: What is different about some numbers?. For example, is there anything different about all of the numbers in the histogram? YES!!! The numbers are not all the same! When the numbers are not all the same, they must vary. So, the variance in the numbers refers to how the numbers are different. There are many ways to summarize the amount of variance in the numbers, and we discuss these very soon.

2.4 Measures of Central Tendency (Sameness)

We've seen that we can get a sense of data by plotting dots in a graph, and by making a histogram. These tools show us what the numbers look like, approximately how big and small they are, and how similar and different they are from another. It is good to get a feeling about the numbers in this way. But, these visual sensitudes are not very precise. In addition to summarizing numbers with graphs, we can summarize numbers using numbers (NO, please not more numbers, we promise numbers can be your friend).

2.4.1 From many numbers to one

Measures of central have one important summary goal: to reduce a pile of numbers to a single number that we can look at. We already know that looking at thousands of numbers is hopeless. Wouldn't it be nice if we could just look at one number instead? We think so. It turns out there are lots of ways to do this. Then, if your friend ever asks the frightening question, "hey, what are all these numbers like?". You can say they are like this one number right here.

But, just like in Indiana Jones and the Last Crusade (highly recommended movie), you must choose your measure of central tendency wisely.

2.4.2 Mode

The **mode** is the most frequently occurring number in your measurement. That is it. How do you find it? You have to count the number of times each number appears in your measure, then whichever one occurs the most, is the mode.

> Example: 1 1 1 2 3 4 5 6

The mode of the above set is 1, which occurs three times. Every other number only occurs once.

OK fine. What happens here:

> Example: 1 1 1 2 2 2 3 4 5 6

Hmm, now 1 and 2 both occur three times each. What do we do? We say there are two modes, and they are 1 and 2.

Why is the mode a measure of central tendency? Well, when we ask, "what are my numbers like", we can say, "most of the number are, like a 1 (or whatever the mode is)".

Is the mode a good measure of central tendency? That depends on your numbers. For example, consider these numbers

1 1 2 3 4 5 6 7 8 9

Here, the mode is 1 again, because there are two 1s, and all of the other numbers occur once. But, are most of the numbers like, a 1. No, they are mostly not 1s.

"Argh, so should I or should I not use the mode? I thought this class was supposed to tell me what to do?". There is no telling you what to do. Every time you use a tool in statistics you have to think about what you are doing and justify why what you are doing makes sense. Sorry.

2.4.3 Median

The **median** is the exact middle of the data. After all, we are asking about central tendency, so why not go to the center of the data and see where we are. What do you mean middle of the data? Let's look at these numbers:

1 5 4 3 6 7 9

Umm, OK. So, three is in the middle? Isn't that kind of arbitrary. Yes. Before we can compute the median, we need to order the numbers from smallest to largest.

1 3 4 **5** 6 7 9

Now, 5 is in the middle. And, by middle we mean in the middle. There are three numbers to the left of 5, and three numbers to the right. So, five is definitely in the middle.

OK fine, but what happens when there aren't an even number of numbers? Then the middle will be missing right? Let's see:

1 2 3 4 5 6

There is no number between 3 and 4 in the data, the middle is empty. In this case, we compute the median by figuring out the number in between 3 and 4. So, the median would be 3.5.

Is the median a good measure of central tendency? Sure, it is often very useful. One property of the median is that it stays in the middle even when some of the other numbers get really weird. For example, consider these numbers:

1 2 3 4 4 4 **5** 6 6 6 7 7 1000

Most of these numbers are smallish, but the 1000 is a big old weird number, very different from the rest. The median is still 5, because it is in the middle of these ordered numbers. We can also see that five is pretty similar to most of the numbers (except for 1000). So, the median does a pretty good job of representing most of the numbers in the set, and it does so even if one or two of the numbers are very different from the others.

Finally, **outlier** is a term will we use to describe numbers that appear in data that are very different from the rest. 1000 is an outlier, because it lies way out there on the number line compared to the other numbers. What to do with outliers is another topic we discuss sometimes throughout this course.

2.4.4 Mean

Have you noticed this is a textbook about statistics that hasn't used a formula yet? That is about to change, but for those of you with formula anxiety, don't worry, we will do our best to explain them.

The **mean** is also called the average. And, we're guessing you might already now what the average of a bunch of numbers is? It's the sum of the numbers, divided by the number of number right? How do we express that idea in a formula? Just like this:

$Mean = \bar{X} = \frac{\sum_{i=1}^{n} x_i}{N}$

"That looks like Greek to me". Yup. The $\sum$ symbol is called **sigma**, and it stands for the operation of summing. The little "i" on the bottom, and the little "n" on the top refers to all of the numbers in the set, from the first number "i" to the last number "n". The letters are just arbitrary labels, called **variables** that we use for descriptive purposes. The x_i refers to individual numbers in the set. We sum up all of the numbers, then divide the sum by N, which is the total number of numbers. Sometimes you will see $\bar{X}$ to refer to the mean of all of the numbers.

In plain English, the formula looks like:

$mean = \frac{\text{Sum of my numbers}}{\text{Count of my numbers}}$

"Well, why didn't you just say that?". We just did.

Let's compute the mean for these five numbers:

> 3 7 9 2 6

Add em up:

> 3+7+9+2+6 = 27

Count em up:

> $i_1 = 3$, $i_2 = 7$, $i_3 = 9$, $i_4 = 2$, $i_5 = 6$; N=5, because i went from 1 to 5

Divide em:

> mean = 27 / 5 = 5.4

Or, to put the numbers in the formula, it looks like this:

$Mean = \bar{X} = \frac{\sum_{i=1}^{n} x_i}{N} = \frac{3+7+9+2+6}{5} = \frac{27}{5} = 5.4$

OK fine, that is how to compute the mean. But, like we imagined, you probably already knew that, and if you didn't that's OK, now you do. What's next?

Is the mean a good measure of central tendency? By now, you should know: it depends.

2.4.5 What does the mean mean?

It is not enough to know the formula for the mean, or to be able to use the formula to compute a mean for a set of numbers. We believe in your ability to add and divide numbers. What you really need to know is what the mean really "means". This requires that you know what the mean does, and not just how to do it. Puzzled? Let's explain.

Can you answer this question: What happens when you divide a sum of numbers by the number of numbers? What are the consequences of doing this? What is the formula doing? What kind of properties does the result give us? FYI, the answer is not that we compute the mean.

OK, so what happens when you divide any number by another number? Of course, the key word here is divide. We literally carve the number up top in the numerator into pieces. How many times do we split the top number? That depends on the bottom number in the denominator. Watch:

$\frac{12}{3} = 4$

So, we know the answer is 4. But, what is really going on here is that we are slicing and dicing up 12 aren't we. Yes, and we slicing 12 into three parts. It turns out the size of those three parts is 4. So, now we are thinking of 12 as three different pieces $12 = 4 + 4 + 4$. I know this will be obvious, but what kind of properties do our pieces have? You mean the fours? Yup. Well, obviously they are all fours. Yes. The pieces are all the same size. They are all equal. So, division equalizes the numerator by the denominator…

"Umm, I think I learned this in elementary school, what does this have to do with the mean?". The number on top of the formula for the mean is just another numerator being divided by a denominator isn't it. In this case, the numerator is a sum of all the values in your data. What if it was the sum of all of the 500 happiness ratings? The sum of all of them would just be a single number adding up all the different ratings. If we split the sum up into equal parts representing one part for each person's happiness what would we get? We would get 500 identical and equal numbers for each person. It would be like taking all of the happiness in the world, then dividing it up equally, then to be fair, giving back the same equal amount of happiness to everyone in the world. This would make some people more happy than they were before, and some people less happy right. Of course, that's because it would be equalizing the distribution of happiness for everybody. This process of equalization by dividing something into equal parts is what the **mean** does. See, it's more than just a formula. It's an idea. This is just the beginning of thinking about these kinds of ideas. We will come back to this idea about the mean, and other ideas, in later chapters.

> Pro tip: The mean is the one and only number that can take the place of every number in the data, such that when you add up all the equal parts, you get back the original sum of the data.

2.4.6 All together now

Just to remind ourselves of the mode, median, and mean, take a look at the next histogram in Figure 2.4. We have overlaid the location of the mean (red), median (green), and mode (blue). For this dataset, the three measures of central tendency all give different answers. The mean is the largest because it is influenced by large numbers, even if they occur rarely. The mode and median are insensitive to large numbers that occur infrequently, so they have smaller values.

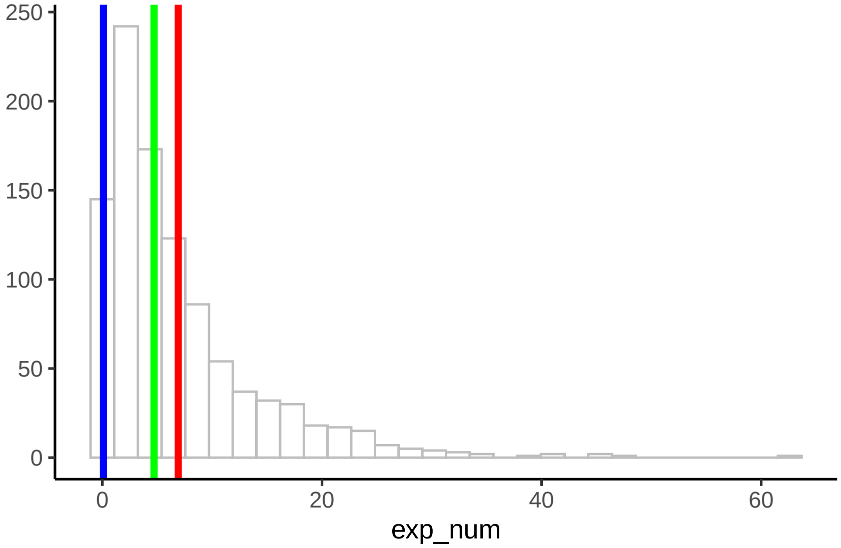


Figure 2.4: A histogram with the mean (red), the median (green), and the mode (blue)

2.5 Measures of Variation (Different*ness*)

What did you do when you wrote essays in high school about a book you read? Probably compare and contrast something right? When you summarize data, you do the same thing. Measures of central tendency give us something like comparing does, they tell us stuff about what is the same. Measures of variation give us something like contrasting does, they tell us stuff about what is different.

First, we note that whenever you see a bunch of numbers that aren't the same, you already know there are some differences. This means the numbers vary, and there is variation in the size of the numbers.

2.5.1 The Range

Consider these 10 numbers, that I already ordered from smallest to largest for you:

1 3 4 5 5 6 7 8 9 24

The numbers have variation, because they are not all the same. We can use the range to describe the width of the variation. The range refers to the **minimum** (smallest value) and **maximum** (largest value) in the set. So, the range would be 1 and 24.

The range is a good way to quickly summarize the boundaries of your data in just two numbers. By computing the range we know that none of the data is larger or smaller than the range. And, it can alert you to outliers. For example, if you are expecting your numbers to be between 1 and 7, but you find the range is 1 - 340,500, then you know you have some big numbers that shouldn't be there, and then you can try to figure out why those numbers occurred (and potentially remove them if something went wrong).

2.5.2 The Difference Scores

It would be nice to summarize the amount of different*ness* in the data. Here's why. If you thought that raw data (lots of numbers) is too big to look at, then you will be frightened to contemplate how many differences there are to look at. For example, these 10 numbers are easy to look at:

1 3 4 5 5 6 7 8 9 24

But, what about the difference between the numbers, what do those look like? We can compute the difference scores between each number, then put them in a matrix like the one below:

	1	3	4	5	5	6	7	8	9	24
1	0	2	3	4	4	5	6	7	8	23
3	-2	0	1	2	2	3	4	5	6	21
4	-3	-1	0	1	1	2	3	4	5	20
5	-4	-2	-1	0	0	1	2	3	4	19
5	-4	-2	-1	0	0	1	2	3	4	19
6	-5	-3	-2	-1	-1	0	1	2	3	18

	1	3	4	5	5	6	7	8	9	24
7	-6	-4	-3	-2	-2	-1	0	1	2	17
8	-7	-5	-4	-3	-3	-2	-1	0	1	16
9	-8	-6	-5	-4	-4	-3	-2	-1	0	15
24	-23	-21	-20	-19	-19	-18	-17	-16	-15	0

We are looking at all of the possible differences between each number and every other number. So, in the top left, the difference between 1 and itself is 0. One column over to the right, the difference between 3 and 1 (3-1) is 2, etc. As you can see, this is a 10x10 matrix, which means there are 100 differences to look at. Not too bad, but if we had 500 numbers, then we would have 500*500 = 250,000 differences to look at (go for it if you like looking at that sort of thing).

Pause for a simple question. What would this matrix look like if all of the 10 numbers in our data were the same number? It should look like a bunch of 0s right? Good. In that case, we could easily see that the numbers have no variation.

But, when the numbers are different, we can see that there is a very large matrix of difference scores. How can we summarize that? How about we apply what we learned from the previous section on measures of central tendency. We have a lot of differences, so we could ask something like, what is the average difference that we have? So, we could just take all of our differences, and compute the mean difference right? What do you think would happen if we did that?

Let's try it out on these three numbers:

    1 2 3

	1	2	3
1	0	1	2
2	-1	0	1
3	-2	-1	0

You might already guess what is going to happen. Let's compute the mean:

mean of difference scores = $\frac{0+1+2-1+0+1-2-1+0}{9} = \frac{0}{9} = 0$

Uh oh, we get zero for the mean of the difference scores. This will always happen whenever you take the mean of the difference scores. We can see that there are some differences between the numbers, so using 0 as the summary value for the variation in the numbers doesn't make much sense.

Furthermore, you might also notice that the matrices of difference scores are redundant. The diagonal is always zero, and numbers on one side of the diagonal are the same as the numbers on the other side, except their signs are reversed. So, that's one reason why the difference scores add up to zero.

These are little problems that can be solved by computing the **variance** and the **standard deviation**. For now, the standard deviation is a just a trick that we use to avoid getting a zero. But, later we will see it has properties that are important for other reasons.

2.5.3 The Variance

Variability, variation, variance, vary, variable, varying, variety. Confused yet? Before we describe **the variance**, we want to you be OK with how this word is used. First, don't forget the big picture. We know that variability and variation refers to the big idea of differences between numbers. We can even use the word variance in the same way. When numbers are different, they have variance.

> ℹ Note
>
> The formulas for variance and standard deviation depend on whether you think your data represents an entire population of numbers, or is sample from the population. We discuss this issue in later on. For now, we divide by N, later we discuss why you will often divide by N-1 instead.

The word **variance** also refers to a specific summary statistic, the sum of the squared deviations from the mean. Hold on what? Plain English please. The variance is the sum of the squared difference scores, where the difference scores are computed between each score and the mean. What are these scores? The scores are the numbers in the data set. Let's see the formula in English first:

$variance = \frac{\text{Sum of squared difference scores}}{\text{Number of Scores}}$

2.5.3.1 Deviations from the mean, Difference scores from the mean

We got a little bit complicated before when we computed the difference scores between all of the numbers in the data. Let's do it again, but in a more manageable way. This time, we calculate the difference between each score and the mean. The idea here is

1. We can figure out how similar our scores are by computing the mean
2. Then we can figure out how different our scores are from the mean

This could tell us, 1) something about whether our scores are really all very close to the mean (which could help us know if the mean is good representative number of the data), and 2) something about how much differences there are in the numbers.

Take a look at this table:

scores	values	mean	Difference_from_Mean
1	1	4.5	-3.5
2	6	4.5	1.5
3	4	4.5	-0.5
4	2	4.5	-2.5
5	6	4.5	1.5
6	8	4.5	3.5
Sums	27	27	0
Means	4.5	4.5	0

The first column shows we have 6 scores in the data set, and the `value` columns shows each score. The sum of the values, and the mean is presented on the last two rows. The sum and the mean were obtained by:

$\frac{1+6+4+2+6+8}{6} = \frac{27}{6} = 4.5$.

The third column `mean`, appears a bit silly. We are just listing the mean once for every score. If you think back to our discussion about the meaning of the mean, then you will remember that it equally distributes the total sum across each data point. We can see that here, if we treat each score as the mean, then every score is a 4.5. We can also see that adding up all of the means for each score gives us back 27, which is the sum of the original values. Also, we see that if we find the mean of the mean scores, we get back the mean (4.5 again).

All of the action is occurring in the fourth column, `Difference_from_Mean`. Here, we are showing the difference scores from the mean, using $X_i - \bar{X}$. In other words, we subtracted the mean from each score. So, the first score, 1, is -3.5 from the mean, the second score, 6, is +1.5 from the mean, and so on.

Now, we can look at our original scores and we can look at their differences from the mean. Notice, we don't have a matrix of raw difference scores, so it is much easier to look at out. But, we still have a problem:

We can see that there are non-zero values in the difference scores, so we know there are a differences in the data. But, when we add them all up, we still get zero, which makes it seem like there are a total of zero differences in the data…Why does this happen…and what to do about it?

2.5.3.2 The mean is the balancing point in the data

One brief pause here to point out another wonderful property of the mean. It is the balancing point in the data. If you take a pen or pencil and try to balance it on your figure so it lays flat what are you doing? You need to find the center of mass in the pen, so that half of it is on one side, and the other half is on the other side. That's how balancing works. One side = the other side.

We can think of data as having mass or weight to it. If we put our data on our bathroom scale, we could figure out how heavy it was by summing it up. If we wanted to split the data down the middle so that half of the weight was equal to the other half, then we could balance the data on top of a pin. The mean of the data tells you where to put the pin. It is the location in the data, where the numbers on the one side add up to the same sum as the numbers on the other side.

If we think this through, it means that the sum of the difference scores from the mean will always add up to zero. This is because the numbers on one side of the mean will always add up to -x (whatever the sum of those numbers is), and the numbers of the other side of the mean will always add up to +x (which will be the same value only positive). And:

$-x + x = 0$, right.

Right.

2.5.3.3 The squared deviations

Some devious someone divined a solution to the fact that differences scores from the mean always add to zero. Can you think of any solutions? For example, what could you do to the difference scores so that you could add them up, and they would weigh something useful, that is they would not be zero?

The devious solution is to square the numbers. Squaring numbers converts all the negative numbers to positive numbers. For example, $2^2 = 4$, and $-2^2 = 4$. Remember how squaring works, we multiply the number twice: $2^2 = 2 * 2 = 4$, and $-2^2 = -2 * -2 = 4$. We use the term **squared deviations** to refer to differences scores that have been squared. Deviations are things that move away from something. The difference scores move away from the mean, so we also call them **deviations**.

Let's look at our table again, but add the squared deviations.

scores	values	mean	Difference_from_Mean	Squared_Deviations
1	1	4.5	-3.5	12.25
2	6	4.5	1.5	2.25
3	4	4.5	-0.5	0.25
4	2	4.5	-2.5	6.25
5	6	4.5	1.5	2.25
6	8	4.5	3.5	12.25
Sums	27	27	0	35.5
Means	4.5	4.5	0	5.91666666666667

OK, now we have a new column called `squared_deviations`. These are just the difference scores squared. So, $-3.5^2 = 12.25$, etc. You can confirm for yourself with your cellphone calculator.

Now that all of the squared deviations are positive, we can add them up. When we do this we create something very special called the sum of squares (SS), also known as the sum of the squared deviations from the mean. We will talk at length about this SS later on in the ANOVA chapter. So, when you get there, remember that you already know what it is, just some sums of some squared deviations, nothing fancy.

2.5.3.4 Finally, the variance

Guess what, we already computed the variance. It already happened, and maybe you didn't notice. "Wait, I missed that, what happened?".

First, see if you can remember what we are trying to do here. Take a pause, and see if you can tell yourself what problem we are trying solve.

> pause

Without further ado, we are trying to get a summary of the differences in our data. There are just as many difference scores from the mean as there are data points, which can be a lot, so it would be nice to have a single number to look at, something like a mean, that would tell us about the average differences in the data.

If you look at the table, you can see we already computed the mean of the squared deviations. First, we found the sum (SS), then below that we calculated the mean = 5.916 repeating. This is **the variance**. The variance is the mean of the sum of the squared deviations:

$variance = \frac{SS}{N}$, where SS is the sum of the squared deviations, and N is the number of observations.

OK, now what. What do I do with the variance? What does this number mean? Good question. The variance is often an unhelpful number to look at. Why? Because it is not in the same scale as the original data. This is because we squared the difference scores before taking the mean. Squaring produces large numbers. For example, we see a 12.25 in there. That's a big difference, bigger than any difference between any two original values. What to do? How can we bring the numbers back down to their original unsquared size?

If you are thinking about taking the square root, that's a ding ding ding, correct answer for you. We can always unsquare anything by taking the square root. So, let's do that to 5.916. $\sqrt{5.916} = 2.4322829$.

2.5.4 The Standard Deviation

Oops, we did it again. We already computed the standard deviation, and we didn't tell you. The standard deviation is the square root of the variance...At least, it is right now, until we complicate matters for you in the next chapter.

Here is the formula for the standard deviation:

standard deviation $= \sqrt{Variance} = \sqrt{\frac{SS}{N}}$.

We could also expand this to say:

standard deviation $= \sqrt{\frac{\sum_i^n (x_i - \bar{x})^2}{N}}$

Don't let those big square root signs put you off. Now, you know what they are doing there. Just bringing our measure of the variance back down to the original size of the data. Let's look at our table again:

scores	values	mean	Difference_from_Mean	Squared_Deviations
1	1	4.5	-3.5	12.25
2	6	4.5	1.5	2.25
3	4	4.5	-0.5	0.25
4	2	4.5	-2.5	6.25
5	6	4.5	1.5	2.25
6	8	4.5	3.5	12.25
Sums	27	27	0	35.5
Means	4.5	4.5	0	5.91666666666667

We measured the standard deviation as 2.4322829. Notice this number fits right in the with differences scores from the mean. All of the scores are kind of in and around + or - 2.4322829. Whereas, if we looked at the variance, 5.916 is just too big, it doesn't summarize the actual differences very well.

What does all this mean? Well, if someone told they had some number with a mean of 4.5 (like the values in our table), and a standard deviation of 2.4322829, you would get a pretty good summary of the numbers. You would know that many of the numbers are around 4.5, and you would know that not all of the numbers are 4.5. You would know that the numbers spread around 4.5. You also know that the spread isn't super huge, it's only + or - 2.4322829 on average. That's a good starting point for describing numbers.

If you had loads of numbers, you could reduce them down to the mean and the standard deviation, and still be pretty well off in terms of getting a sense of those numbers.

2.6 Using Descriptive Statistics with data

Remember, you will be learning how to compute descriptive statistics using software in the labs. Check out the lab manual exercises for descriptives to see some examples of working with real data.

2.7 Rolling your own descriptive statistics

We spent many paragraphs talking about variation in numbers, and how to use calculate the **variance** and **standard deviation** to summarize the average differences between numbers in a data set. The basic process was to 1) calculate some measure of the differences, then 2) average the differences to create a summary. We found that we couldn't average the raw difference scores, because we would always get a zero. So, we squared the differences from the mean, then averaged the squared differences differences. Finally, we square rooted our measure to bring the summary back down to the scale of the original numbers.

Perhaps you haven't heard, but there is more than one way to skin a cat, but we prefer to think of this in terms of petting cats, because some of us love cats. Jokes aside, perhaps you were also thinking that the problem of summing differences scores (so that they don't equal zero), can be solved in more than one way. Can you think of a different way, besides squaring?

2.7.1 Absolute deviations

How about just taking the absolute value of the difference scores. Remember, the absolute value converts any number to a positive value. Check out the following table:

scores	values	mean	Difference_from_Mean	Absolute_Deviations
1	1	4.5	-3.5	3.5
2	6	4.5	1.5	1.5
3	4	4.5	-0.5	0.5
4	2	4.5	-2.5	2.5
5	6	4.5	1.5	1.5
6	8	4.5	3.5	3.5
Sums	27	27	0	13
Means	4.5	4.5	0	2.16666666666667

This works pretty well too. By converting the difference scores from the mean to positive values, we can now add them up and get a non-zero value (if there are differences). Then, we can find the mean of the sum of the absolute deviations. If we were to map the terms sum of squares (SS), variance and standard deviation onto these new measures based off of the absolute deviation, how would the mapping go? For example, what value in the table corresponds to the SS? That would be the sum of absolute deviations in the last column. How about the variance and standard deviation, what do those correspond to? Remember that the variance is mean (SS/N), and the standard deviation is a square-rooted mean ($\sqrt{SS/N}$). In the table above we only have one corresponding mean, the mean of the sum of the absolute deviations. So, we have a **variance** measure that does not need to be square rooted. We might say the mean absolute deviation, is doing double-duty as a variance and a standard-deviation. Neat.

2.7.2 Other sign-inverting operations

In principle, we could create lots of different summary statistics for variance that solve the summing to zero problem. For example, we could raise every difference score to any even numbered power beyond 2 (which is the square). We could use, 4, 6, 8, 10, etc. There is an infinity of even numbers, so there is an infinity of possible variance statistics. We could also use odd numbers as powers, and then take their absolute value. Many things are possible. The important aspect to any of this is to have a reason for what you are doing, and to choose a method that works for the data-analysis problem you are trying to solve. Note also, we bring up this general issue because we want you to understand that statistics is a creative exercise. We invent things when we need them, and we use things that have already been invented when they work for the problem at hand.

2.8 Remember to look at your data

Descriptive statistics are great and we will use them a lot in the course to describe data. You may suspect that descriptive statistics also have some short-comings. This is very true. They are compressed summaries of large piles of numbers. They will almost always be unable to represent all of the numbers fairly. There are also different kinds of descriptive statistics that you could use, and it sometimes not clear which one's you should use.

Perhaps the most important thing you can do when using descriptives is to use them in combination with looking at the data in a graph form. This can help you see whether or not your descriptives are doing a good job of representing the data.

2.8.1 Anscombe's Quartet

To hit this point home, and to get you thinking about the issues we discuss in the next chapter, check this out. It's called Anscombe's Quartet, because these interesting graphs and numbers and numbers were produced by Anscombe (1973). In Figure 2.5 you are looking at pairs of

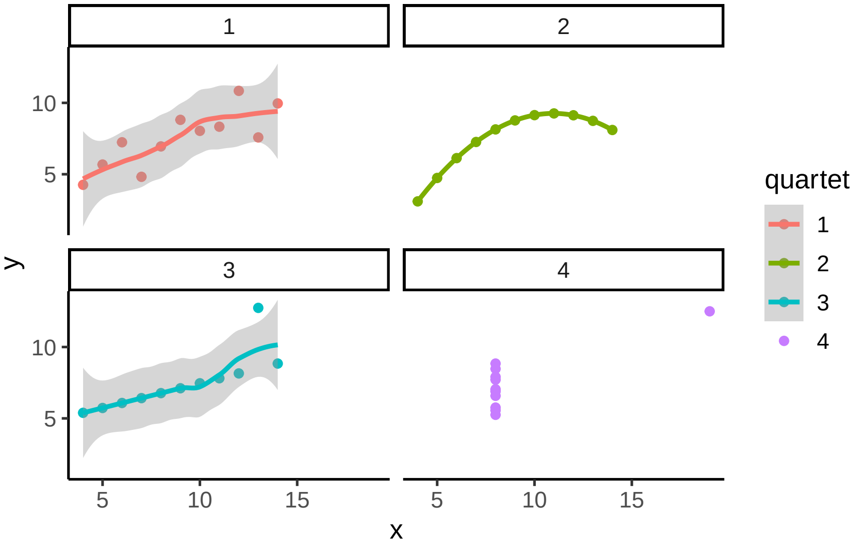


Figure 2.5: Anscombe's Quartet

measurements. Each graph has an X and Y axis, and each point represents two measurements. Each of the graphs looks very different, right?

Well, would you be surprised if I told that the descriptive statistics for the numbers in these graphs are exactly the same? It turns out they do have the same descriptive statistics. In the table below I present the mean and variance for the x-values in each graph, and the mean and the variance for the y-values in each graph.

quartet	mean_x	var_x	mean_y	var_y
1	9	11	7.500909	4.127269
2	9	11	7.500909	4.127629
3	9	11	7.500000	4.122620
4	9	11	7.500909	4.123249

The descriptives are all the same! Anscombe put these special numbers together to illustrate the point of graphing your numbers. If you only look at your descriptives, you don't know what patterns in the data they are hiding. If you look at the graph, then you can get a better understanding.

2.8.2 Datasaurus Dozen

If you thought that Anscombe's quartet was neat, you should take a look at the Datasaurus Dozen (Matejka and Fitzmaurice 2017). Scroll down to see the examples. You will be looking at dot plots. The dot plots show many different patterns, including dinosaurs! What's amazing is that all of the dots have very nearly the same descriptive statistics. Just another reminder to look at your data, it might look like a dinosaur!

2.9 Videos

2.9.1 Measures of center: Mode

2.9.2 Measures of center: Median and Mean

2.9.3 Standard deviation part I

2.9.4 Standard deviation part II

3 Correlation

> Correlation does not equal causation —Every Statistics and Research Methods Instructor Ever

In the last chapter we had some data. It was too much too look at and it didn't make sense. So, we talked about how to look at the data visually using plots and histograms, and we talked about how to summarize lots of numbers so we could determine their central tendencies (sameness) and variability (differentness). And, all was well with the world.

Let's not forget the big reason why we learned about descriptive statistics. The big reason is that we are interested in getting answers to questions using data.

If you are looking for a big theme to think about while you take this course, the theme is: how do we ask and answer questions using data?

For every section in this book, you should be connecting your inner monologue to this question, and asking yourself: How does what I am learning about help me answer questions with data? Advance warning: we know it is easy to forget this stuff when we dive into the details, and we will try to throw you a rope to help you out along the way…remember, we're trying to answer questions with data.

We started Chapter two with some fake data on human happiness, remember? We imagined that we asked a bunch of people to tell us how happy they were, then we looked at the numbers they gave us. Let's continue with this imaginary thought experiment.

What do you get when you ask people to use a number to describe how happy they are? A bunch of numbers. What kind of questions can you ask about those numbers? Well, you can look at the numbers and estimate their general properties as we already did. We would expect those numbers tell us some things we already know. There are different people, and different people are different amounts of happy. You've probably met some of those of really happy people, and really unhappy people, and you yourself probably have some amount of happiness. "Great, thanks Captain Obvious".

Before moving on, you should also be skeptical of what the numbers might mean. For example, if you force people to give a number between 0-100 to rate their happiness, does this number truly reflect how happy that person is? Can a person know how happy they are? Does the question format bias how they give their answer? Is happiness even a real thing? These are all good questions about the **validity** of the construct (happiness itself) and the measure (numbers) you are using to quantify it. For now, though, we will side-step those very important questions, and assume that, happiness is a thing, and our measure of happiness measures something about how happy people are.

OK then, after we have measured some happiness, I bet you can think of some more pressing questions. For example, what causes happiness to go up or down. If you knew the causes of happiness what could you do? How about increase your own happiness; or, help people who are unhappy; or, better appreciate why Eeyore from Winnie the Pooh is unhappy; or, present valid scientific arguments that argue against incorrect claims about what causes happiness. A causal theory and understanding of happiness could be used for all of those things. How can we get there?

Imagine you were an alien observer. You arrived on earth and heard about this thing called happiness that people have. You want to know what causes happiness. You also discover that planet earth has lots of other things. Which of those things, you wonder, cause happiness? How would your alien-self get started on this big question.

As a person who has happiness, you might already have some hunches about what causes changes in happiness. For example things like: weather, friends, music, money, education, drugs, books, movies, beliefs, personality, color of your shoes, eyebrow length, number of cat's you see per day, frequency of subway delay, a lifetime supply of chocolate, et cetera et cetera (as Willy Wonka would say), might all contribute to happiness in someway. There could be many different causes of happiness.

3.1 If something caused something else to change, what would that look like?

Before we go around determining the causes of happiness, we should prepare ourselves with some analytical tools so that we could identify what causation looks like. If we don't prepare ourselves for what we might find, then we won't know how to interpret our own data. Instead, we need to anticipate what the data could look like. Specifically, we need to know what data would look like when one thing does not cause another thing, and what data would look like when one thing does cause another thing. This chapter does some of this preparation. Fair warning: we will find out some tricky things. For example, we can find patterns that look like one thing is causing another, even when that one thing DOES NOT CAUSE the other thing. Hang in there.

3.1.1 Charlie and the Chocolate factory

Let's imagine that a person's supply of chocolate has a causal influence on their level of happiness. Let's further imagine that, like Charlie, the more chocolate you have the more happy you will be, and the less chocolate you have, the less happy you will be. Finally, because we suspect happiness is caused by lots of other things in a person's life, we anticipate that the relationship between chocolate supply and happiness won't be perfect. What do these assumptions mean for how the data should look?

Our first step is to collect some imaginary data from 100 people. We walk around and ask the first 100 people we meet to answer two questions:

1. how much chocolate do you have, and
2. how happy are you.

For convenience, both the scales will go from 0 to 100. For the chocolate scale, 0 means no chocolate, 100 means lifetime supply of chocolate. Any other number is somewhere in between. For the happiness scale, 0 means no happiness, 100 means all of the happiness, and in between means some amount in between.

Here is some sample data from the first 10 imaginary subjects.

subject	chocolate	happiness
1	1	1
2	1	2
3	3	2
4	3	4
5	4	3
6	5	4
7	4	4
8	5	5
9	5	8
10	9	10

We asked each subject two questions so there are two scores for each subject, one for their chocolate supply, and one for their level of happiness. You might already notice some relationships between amount of chocolate and level of happiness in the table. To make those relationships even more clear, let's plot all of the data in a graph.

3.1.2 Scatter plots

When you have two measurements worth of data, you can always turn them into dots and plot them in a scatter plot. A scatter plot has a horizontal x-axis, and a vertical y-axis. You get to choose which measurement goes on which axis. Let's put chocolate supply on the x-axis, and happiness level on the y-axis. Figure 3.1 shows 100 dots for each subject.

You might be wondering, why are there only 100 dots for the data. Didn't we collect 100 measures for chocolate, and 100 measures for happiness, shouldn't there be 200 dots? Nope. Each dot is for one subject, there are 100 subjects, so there are 100 dots.

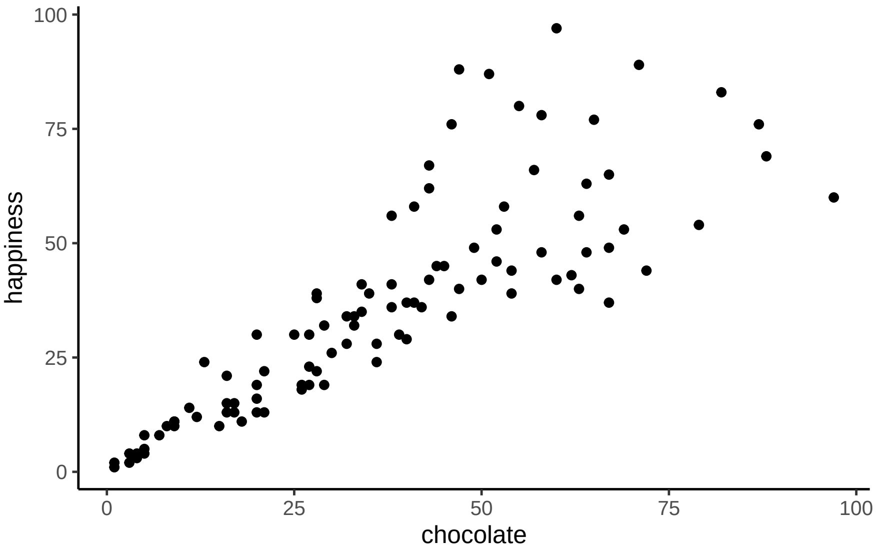


Figure 3.1: Imaginary data showing a positive correlation between amount of chocolate and amount happiness

What do the dots mean? Each dot has two coordinates, an x-coordinate for chocolate, and a y-coordinate for happiness. The first dot, all the way on the bottom left is the first subject in the table, who had close to 0 chocolate and close to zero happiness. You can look at any dot, then draw a straight line down to the x-axis: that will tell you how much chocolate that subject has. You can draw a straight line left to the y-axis: that will tell you how much happiness the subject has.

Now that we are looking at the scatter plot, we can see many things. The dots are scattered around a bit aren't they, hence **scatter plot**. Even when the dot's don't scatter, they're still called scatter plots, perhaps because those pesky dots in real life have so much scatter all the time. More important, the dots show a relationship between chocolate supply and happiness. Happiness is lower for people with smaller supplies of chocolate, and higher for people with larger supplies of chocolate. It looks like the more chocolate you have the happier you will be, and vice-versa. This kind of relationship is called a **positive correlation**.

3.1.3 Positive, Negative, and No-Correlation

Seeing as we are in the business of imagining data, let's imagine some more. We've already imagined what data would look like if larger chocolate supplies increase happiness. We'll show that again in a bit. What do you imagine the scatter plot would look like if the relationship was reversed, and larger chocolate supplies decreased happiness. Or, what do you imagine the scatter plot would look like if there was no relationship, and the amount of chocolate that you have doesn't do anything to your happiness. We invite your imagination to look at Figure 3.2:

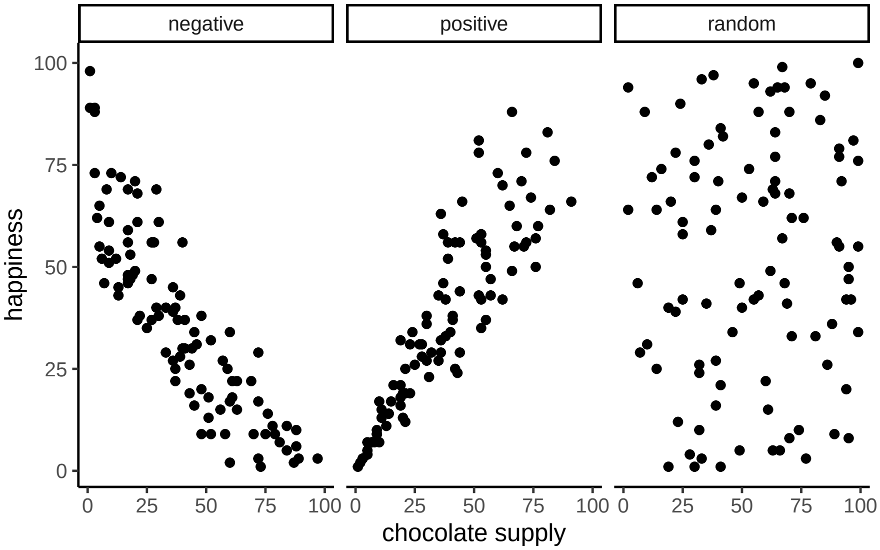


Figure 3.2: Three scatterplots showing negative, positive, and zero correlation

The first panel shows a **negative correlation**. Happiness goes down as chocolate supply increases. Negative correlation occurs when one thing goes up and the other thing goes down; or, when more of X is less of Y, and vice-versa. The second panel shows a **positive correlation**. Happiness goes up as chocolate as chocolate supply increases. Positive correlation occurs when both things go up together, and go down together: more of X is more of Y, and vice-versa. The third panel shows **no correlation**. Here, there doesn't appear to be any obvious relationship between chocolate supply and happiness. The dots are scattered all over the place, the truest of the scatter plots.

> ℹ Note
>
> We are wading into the idea that measures of two things can be related, or correlated with one another. It is possible for the relationships to be more complicated than just going up, or going down. For example, we could have a relationship that where the dots go up for the first half of X, and then go down for the second half.

Zero correlation occurs when one thing is not related in any way to another things: changes in X do not relate to any changes in Y, and vice-versa.

3.2 Pearson's r

"So you've examined your scatter plots and now you might be wondering how to quantify what you see. We've already covered how to generate descriptive statistics for individual variables—think of single measures like happiness levels or chocolate consumption, summarized through means, variances, and so on. But what if you want to capture the relationship between two such variables in a single descriptive statistic? Is that even possible? Karl Pearson to the rescue.

ℹ Note

The stories about the invention of various statistics are very interesting, you can read more about them in the book, "The Lady Tasting Tea" (Salsburg 2001)

There's a statistic for that, and Karl Pearson invented it. Everyone now calls it, "Pearson's r". We will find out later that Karl Pearson was a big-wig editor at Biometrika in the 1930s. He took a hating to another big-wig statistician, Sir Ronald Fisher (who we learn about later), and they had some statistics fights. Even in the stats world, not everyone plays nice in the sandbox.

How does Pearson's r work? Let's look again at the first 10 subjects in our fake experiment:

subject	chocolate	happiness
1	1	1
2	1	2
3	3	2
4	3	4
5	4	3
6	5	4
7	4	4
8	5	5
9	5	8
10	9	10
Sums	40	43
Means	4	4.3

What could we do to these numbers to produce a single summary value that represents the relationship between the chocolate supply and happiness?

3.2.1 The idea of co-variance

"Oh please no, don't use the word variance again". Yes, we're doing it, we're going to use the word variance again, and again, until it starts making sense. Remember what variance means about some numbers. It means the numbers have some change in them, they are not all the same, some of them are big, some are small. We can see that there is variance in chocolate supply across the 10 subjects. We can see that there is variance in happiness across the 10 subjects. We also saw in the scatter plot, that happiness increases as chocolate supply increases; which is a positive relationship, a positive correlation. What does this have to do with variance? Well, it means there is a relationship between the variance in chocolate supply, and the variance in happiness levels. The two measures vary together don't they? When we have two measures that vary together, they are like a happy couple who share their variance. This is what co-variance refers to, the idea that the pattern of varying numbers in one measure is shared by the pattern of varying numbers in another measure.

Co-variance is **very, very, very, very** important. I suspect that the word co-variance is initially confusing, especially if you are not yet fully comfortable with the meaning of variance for a single measure. Nevertheless, we must proceed and use the idea of co-variance over and over again to firmly implant it into your statistical mind (we already said, but redundancy works, it's a thing).

> Pro tip: Three-legged race is a metaphor for co-variance. Two people tie one leg to each other, then try to walk. It works when they co-vary their legs together (positive relationship). They can also co-vary in an unhelpful way, when one person tries to move forward exactly when the other person tries to move backward. This is still co-variance (negative relationship). Funny random walking happens when there is no co-variance. This means one person does whatever they want, and so does the other person. There is a lot of variance, but the variance is shared randomly, so it's just a bunch of legs moving around accomplishing nothing.
>
> Pro tip #2: Successfully playing paddy-cake occurs when two people coordinate their actions so they have positively shared co-variance.

3.3 Turning the numbers into a measure of co-variance

"OK, so if you are saying that co-variance is just another word for correlation or relationship between two measures, I'm good with that. I suppose we would need some way to measure that." Correct, back to our table...notice anything new?

subject	chocolate	happiness	Chocolate_X_Happiness
1	1	1	1
2	1	2	2

subject	chocolate	happiness	Chocolate_X_Happiness
3	3	2	6
4	3	4	12
5	4	3	12
6	5	4	20
7	4	4	16
8	5	5	25
9	5	8	40
10	9	10	90
Sums	40	43	224
Means	4	4.3	22.4

We've added a new column called `Chocolate_X_Happiness`, which translates to Chocolate scores multiplied by Happiness scores. Each row in the new column, is the product, or multiplication of the chocolate and happiness score for that row. Yes, but why would we do this?

Last chapter we took you back to Elementary school and had you think about division. Now it's time to do the same thing with multiplication. We assume you know how that works. One number times another, means taking the first number, and adding it as many times as the second says to do,

$2 * 2 = 2 + 2 = 4$

$2 * 6 = 2 + 2 + 2 + 2 + 2 + 2 = 12$, or $6 + 6 = 12$, same thing.

Yes, you know all that. But, can you bend multiplication to your will, and make it do your bidding when need to solve a problem like summarizing co-variance? Multiplication is the droid you are looking for.

We know how to multiple numbers, and all we have to next is think about the consequences of multiplying sets of numbers together. For example, what happens when you multiply two small numbers together, compared to multiplying two big numbers together? The first product should be smaller than the second product right? How about things like multiplying a small number by a big number? Those products should be in between right?.

Then next step is to think about how the products of two measures sum together, depending on how they line up. Let's look at another table:

scores	X	Y	A	B	XY	AB
1	1	1	1	10	1	10
2	2	2	2	9	4	18
3	3	3	3	8	9	24
4	4	4	4	7	16	28

scores	X	Y	A	B	XY	AB
5	5	5	5	6	25	30
6	6	6	6	5	36	30
7	7	7	7	4	49	28
8	8	8	8	3	64	24
9	9	9	9	2	81	18
10	10	10	10	1	100	10
Sums	55	55	55	55	385	220
Means	5.5	5.5	5.5	5.5	38.5	22

Look at the X and Y column. The scores for X and Y perfectly co-vary. When X is 1, Y is 1; when X is 2, Y is 2, etc. They are perfectly aligned. The scores for A and B also perfectly co-vary, just in the opposite manner. When A is 1, B is 10; when A is 2, B is 9, etc. B is a reversed copy of A.

Now, look at the column XY. These are the products we get when we multiply the values of X across with the values of Y. Also, look at the column AB. These are the products we get when we multiply the values of A across with the values of B. So far so good.

Now, look at the `Sums` for the XY and AB columns. Not the same. The sum of the XY products is 385, and the sum of the AB products is 220. For this specific set of data, the numbers 385 and 220 are very important. They represent the biggest possible sum of products (385), and the smallest possible sum of products (220). There is no way of re-ordering the numbers 1 to 10, say for X, and the numbers 1 to 10 for Y, that would ever produce larger or smaller numbers. Don't believe me? Check this out:

Figure 3.3 shows 1000 computer simulations. I convinced my computer to randomly order the numbers 1 to 10 for X, and randomly order the numbers 1 to 10 for Y. Then, I multiplied X and Y, and added the products together. I did this 1000 times. The dots show the sum of the products for each simulation. The two black lines show the maximum possible sum (385), and the minimum possible sum (220), for this set of numbers. Notice, how all of the dots are in between the maximum and minimum possible values. Told you so.

"OK fine, you told me so…So what, who cares?". We've been looking for a way to summarize the co-variance between two measures right? Well, for these numbers, we have found one, haven't we. It's the sum of the products. We know that when the sum of the products is 385, we have found a perfect, positive correlation. We know, that when the sum of the products is 220, we have found a perfect negative correlation. What about the numbers in between. What could we conclude about the correlation if we found the sum of the products to be 350. Well, it's going to be positive, because it's close to 385, and that's perfectly positive. If the sum of the products was 240, that's going to be negative, because it's close to the perfectly negatively correlating 220. What about no correlation? Well, that's going to be in the middle between 220 and 385 right.

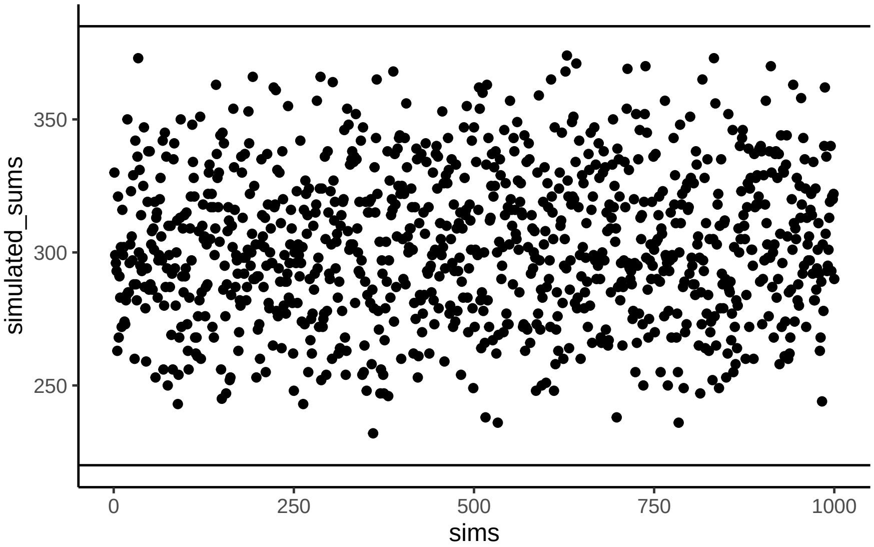


Figure 3.3: Simulated sums of products showing the kinds of values than can be produced by randomly ordering the numbers in X and Y.

We have just come up with a data-specific summary measure for the correlation between the numbers 1 to 10 in X, and the numbers 1 to 10 in Y, it's the sum of the products. We know the maximum (385) and minimum values (220), so we can now interpret any product sum for this kind of data with respect to that scale.

> Pro tip: When the correlation between two measures increases in the positive direction, the sum of their products increases to its maximum possible value. This is because the bigger numbers in X will tend to line up with the bigger numbers in Y, creating the biggest possible sum of products. When the correlation between two measures increases in the negative direction, the sum of their products decreases to its minimum possible value. This is because the bigger numbers in X will tend to line up with the smaller numbers in Y, creating the smallest possible sum of products. When there is no correlation, the big numbers in X will be randomly lined up with the big and small numbers in Y, making the sum of the products, somewhere in the middle.

3.3.1 Co-variance, the measure

We took some time to see what happens when you multiply sets of numbers together. We found that $big * big = bigger$ and $small * small =$ still small, and $big * small =$ in the middle. The purpose of this was to give you some conceptual idea of how the co-variance between two measures is reflected in the sum of their products. We did something very straightforward.

We just multiplied X with Y, and looked at how the product sums get big and small, as X and Y co-vary in different ways.

Now, we can get a little bit more formal. In statistics, **co-variance** is not just the straight multiplication of values in X and Y. Instead, it's the multiplication of the deviations in X from the mean of X, and the deviation in Y from the mean of Y. Remember those difference scores from the mean we talked about last chapter? They're coming back to haunt you know, but in a good way like Casper the friendly ghost.

Let's see what this look like in a table:

subject	chocolate	happiness	C_d	H_d	Cd_x_Hd
1	1	1	-3	-3.3	9.9
2	1	2	-3	-2.3	6.9
3	3	2	-1	-2.3	2.3
4	3	4	-1	-0.3	0.3
5	4	3	0	-1.3	0
6	5	4	1	-0.3	-0.3
7	4	4	0	-0.3	0
8	5	5	1	0.7	0.7
9	5	8	1	3.7	3.7
10	9	10	5	5.7	28.5
Sums	40	43	0	0	52
Means	4	4.3	0	0	5.2

We have computed the deviations from the mean for the chocolate scores (column `C_d`), and the deviations from the mean for the happiness scores (column `H_d`). Then, we multiplied them together (last column). Finally, you can see the mean of the products listed in the bottom right corner of the table, the official **the covariance**.

The formula for the co-variance is:
$cov(X,Y) = \frac{\sum_i^n (x_i - \bar{X})(y_i - \bar{Y})}{N}$

OK, so now we have a formal single number to calculate the relationship between two variables. This is great, it's what we've been looking for. However, there is a problem. Remember when we learned how to compute just the plain old **variance**. We looked at that number, and we didn't know what to make of it. It was squared, it wasn't in the same scale as the original data. So, we square rooted the **variance** to produce the **standard deviation**, which gave us a more interpretable number in the range of our data. The **co-variance** has a similar problem. When you calculate the co-variance as we just did, we don't know immediately know its scale. Is a 3 big? is a 6 big? is a 100 big? How big or small is this thing?

From our prelude discussion on the idea of co-variance, we learned the sum of products between two measures ranges between a maximum and minimum value. The same is true of the co-variance. For a given set of data, there is a maximum possible positive value for the co-variance (which occurs when there is perfect positive correlation). And, there is a minimum possible negative value for the co-variance (which occurs when there is a perfect negative correlation). When there is zero co-variation, guess what happens. Zeroes. So, at the very least, when we look at a co-variation statistic, we can see what direction it points, positive or negative. But, we don't know how big or small it is compared to the maximum or minimum possible value, so we don't know the relative size, which means we can't say how strong the correlation is. What to do?

3.3.2 Pearson's r we there yet

Yes, we are here now. Wouldn't it be nice if we could force our measure of co-variation to be between -1 and +1?

-1 would be the minimum possible value for a perfect negative correlation. +1 would be the maximum possible value for a perfect positive correlation. 0 would mean no correlation. Everything in between 0 and -1 would be increasingly large negative correlations. Everything between 0 and +1 would be increasingly large positive correlations. It would be a fantastic, sensible, easy to interpret system. If only we could force the co-variation number to be between -1 and 1. Fortunately, for us, this episode is brought to you by Pearson's r, which does precisely this wonderful thing.

Let's take a look at a formula for Pearson's r:

$r = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{cov(X,Y)}{SD_X SD_Y}$

We see the symbol σ here, that's more Greek for you. σ is often used as a symbol for the standard deviation (SD). If we read out the formula in English, we see that r is the co-variance of X and Y, divided by the product of the standard deviation of X and the standard deviation of Y. Why are we dividing the co-variance by the product of the standard deviations. This operation has the effect of **normalizing** the co-variance into the range -1 to 1.

> Note
>
> But, we will fill this part in as soon as we can…promissory note to explain the magic. FYI, it's not magic. Brief explanation here is that dividing each measure by its standard deviation ensures that the values in each measure are in the same range as one another.

For now, we will call this mathematical magic. It works, but we don't have space to tell you why it works right now.

It's worth saying that there are loads of different formulas for computing Pearson's r. You can find them by Googling them. We will probably include more of them here, when we get around to it. However, they all give you the same answer. And, they are all not as pretty as each other. Some of them might even look scary. In other statistics textbook you will often find formulas that are easier to use for calculation purposes. For example, if you only had a pen and paper, you might use one or another formula because it helps you compute the answer faster by hand. To be honest, we are not very interested in teaching you how to plug numbers into formulas. We give one lesson on that here: Put the numbers into the letters, then compute the answer. Sorry to be snarky. Nowadays you have a computer that you should use for this kind of stuff. So, we are more interested in teaching you what the calculations mean, rather than how to do them. Of course, every week we are showing you how to do the calculations in lab with computers, because that is important too.

Does Pearson's r really stay between -1 and 1 no matter what? It's true, take a look at the following simulation. Here I randomly ordered the numbers 1 to 10 for an X measure, and did the same for a Y measure. Then, I computed Pearson's r, and repeated this process 1000 times. As you can see from Figure 3.4 all of the dots are between -1 and 1. Neat huh.

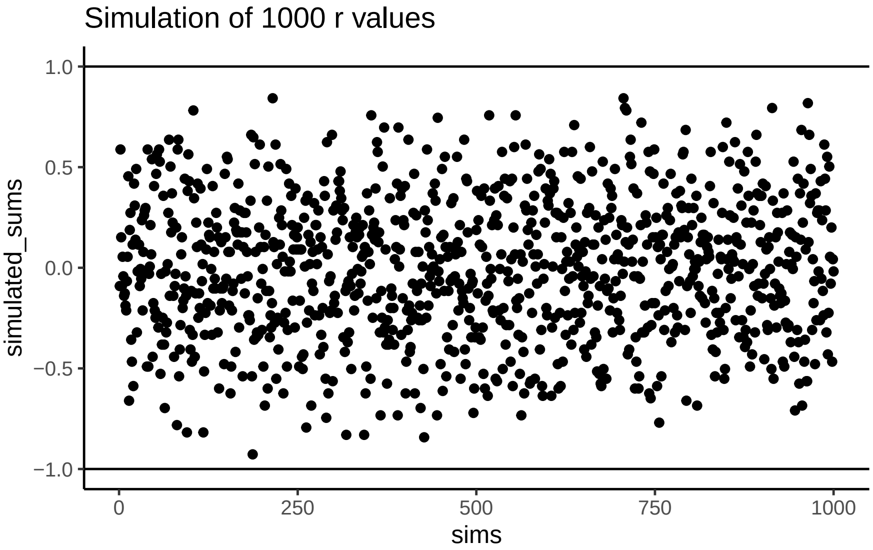


Figure 3.4: A simulation of of correlations. Each dot represents the r-value for the correlation between an X and Y variable that each contain the numbers 1 to 10 in random orders. The figure ilustrates that many r-values can be obtained by this random process

3.4 Examples with Data

In the lab for correlation you will be shown how to compute correlations in real data-sets using software. To give you a brief preview, let's look at some data from the world happiness report

(2018).

This report measured various attitudes across people from different countries. For example, one question asked about how much freedom people thought they had to make life choices. Another question asked how confident people were in their national government. **?@fig-3hrsdata** is a scatterplot showing the relationship between these two measures. Each dot represents means for different countries.

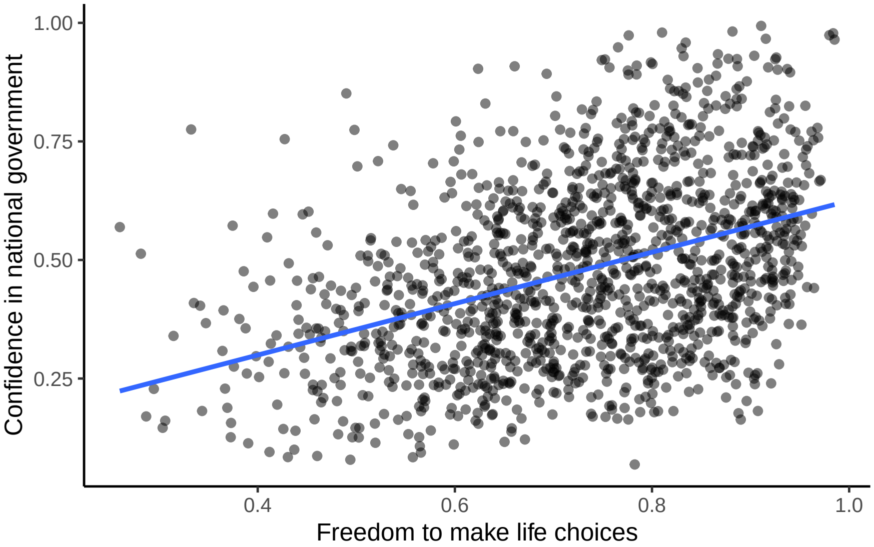


Figure 3.5: Relationship between freedom to make life choices and confidence in national government. Data from the world happiness report for 2018

We put a blue line on the scatterplot to summarize the positive relationship. It appears that as "freedom to make life choices goes up", so to does confidence in national government. It's a positive correlation.

The actual correlation, as measured by Pearson's r is:

```
#> [1] 0.4080963
```

You will do a lot more of this kind of thing in the lab. Looking at the graph you might start to wonder: Does freedom to make life choices cause changes how confident people are in their national government? Our does it work the other way? Does being confident in your national government give you a greater sense of freedom to make life choices? Or, is this just a random relationship that doesn't mean anything? All good questions. These data do not provide the answers, they just suggest a possible relationship.

3.5 Regression: A mini intro

We're going to spend the next little bit adding one more thing to our understanding of correlation. It's called **linear regression**. It sounds scary, and it really is. You'll find out much later in your Statistics education that everything we will be soon be talking about can be thought of as a special case of regression. But, we don't want to scare you off, so right now we just introduce the basic concepts.

First, let's look at a linear regression. This way we can see what we're trying to learn about. Figure 3.6 shows the same scatter plots as before with something new: lines!

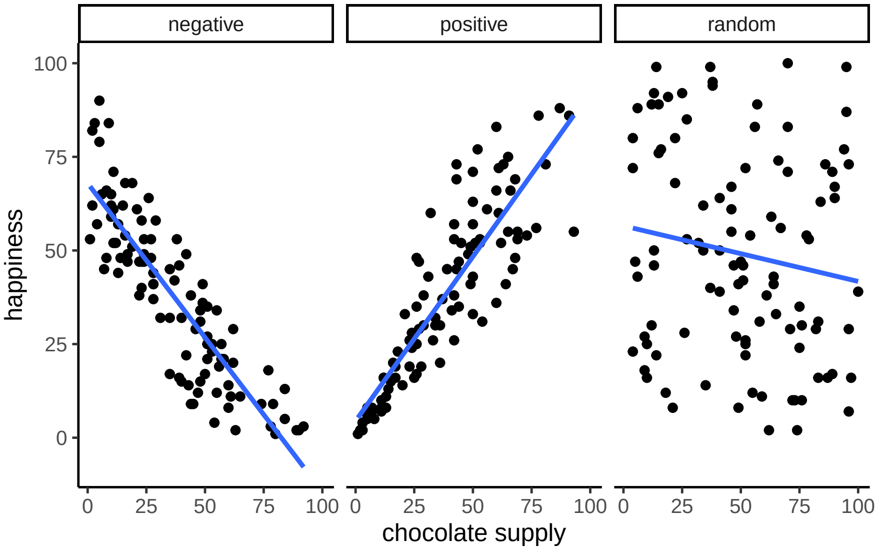


Figure 3.6: Three scatterplots showing negative, positive, and a random correlation (where the r-value is expected to be 0), along with the best fit regression line

3.5.1 The best fit line

Notice anything about these blue lines? Hopefully you can see, at least for the first two panels, that they go straight through the data, just like a kebab skewer. We call these lines **best fit** lines, because according to our definition (soon we promise) there are no other lines that you could draw that would do a better job of going straight throw the data.

One big idea here is that we are using the line as a kind of mean to describe the relationship between the two variables. When we only have one variable, that variable exists on a single dimension, it's 1D. So, it is appropriate that we only have one number, like the mean, to describe it's central tendency. When we have two variables, and plot them together, we now have a two-dimensional space. So, for two dimensions we could use a bigger thing that is 2d, like a line, to summarize the central tendency of the relationship between the two variables.

What do we want out of our line? Well, if you had a pencil, and a printout of the data, you could draw all sorts of straight lines any way you wanted. Your lines wouldn't even have to go through the data, or they could slant through the data with all sorts of angles. Would all of those lines be very good a describing the general pattern of the dots? Most of them would not. The best lines would go through the data following the general shape of the dots. Of the best lines, however, which one is the best? How can we find out, and what do we mean by that? In short, the best fit line is the one that has the least error.

Note

R code for plotting residuals thanks to Simon Jackson's blog post: https://drsimonj.svbtle.com/visualising-residuals

Check out this next plot, it shows a line through some dots. But, it also shows some teeny tiny lines. These lines drop down from each dot, and they land on the line. Each of these little lines is called a **residual**. They show you how far off the line is for different dots. It's measure of error, it shows us just how wrong the line is. After all, it's pretty obvious that not all of the dots are on the line. This means the line does not actually represent all of the dots. The line is wrong. But, the best fit line is the least wrong of all the wrong lines.

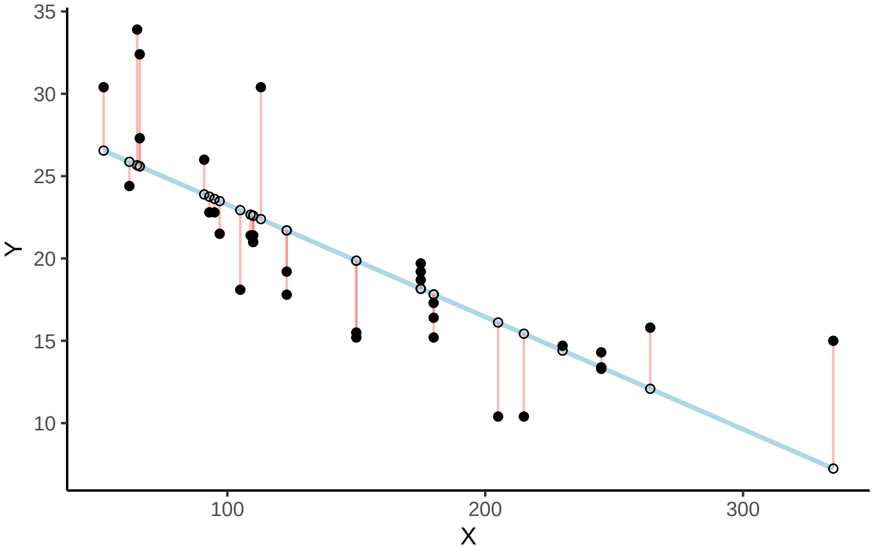


Figure 3.7: Black dots represent data points. The blue line is the best fit regression line. The white dots are repesent the predicted location of each black dot. The red lines show the error between each black dot and the regression line. The blue line is the best fit line because it minimizes the error shown by the red lines.

There's a lot going on in Figure 3.7. First, we are looking at a scatter plot of two variables, an X and Y variable. Each of the black dots are the actual values from these variables. You can see there is a negative correlation here, as X increases, Y tends to decrease. We drew a regression line through the data, that's the blue line. There's these little white dots too. This is where the line thinks the black dots should be. The red lines are the important residuals we've been talking about. Each black dot has a red line that drops straight down, or straight up from the location of the black dot, and lands directly on the line. We can already see that many of the dots are not on the line, so we already know the line is "off" by some amount for each dot. The red line just makes it easier to see exactly how off the line is.

The important thing that is happening here, is that the the blue line is drawn is such a way, that it minimizes the total length of the red lines. For example, if we wanted to know how wrong this line was, we could simply gather up all the red lines, measure how long they are, and then add all the wrongness together. This would give us the total amount of wrongness. We usually call this the error. In fact, we've already talked about this idea before when we discussed standard deviation. What we will actually be doing with the red lines, is computing the sum of the squared deviations from the line. That sum is the total amount of error. Now, this blue line here minimizes the sum of the squared deviations. Any other line would produce a larger total error.

?@fig-3regressionGIF is an animation to see this in action. The animations compares the best fit line in blue, to some other possible lines in black. The black line moves up and down. The red lines show the error between the black line and the data points. As the black line moves toward the best fit line, the total error, depicted visually by the grey area shrinks to it's minimum value. The total error expands as the black line moves away from the best fit line.

Whenever the black line does not overlap with the blue line, it is worse than the best fit line. The blue regression line is like Goldilocks, it's just right, and it's in the middle.

Figure 3.8 shows how the sum of squared deviations (the sum of the squared lengths of the red lines) behaves as we move the line up and down. What's going on here is that we are computing a measure of the total error as the black line moves through the best fit line. This represents the sum of the squared deviations. In other words, we square the length of each red line from the above animation, then we add up all of the squared red lines, and get the total error (the total sum of the squared deviations). The graph below shows what the total error looks like as the black line approaches then moves away from the best fit line. Notice, the dots in this graph start high on the left side, then they swoop down to a minimum at the bottom middle of the graph. When they reach their minimum point, we have found a line that minimizes the total error. This is the best fit regression line.

OK, so we haven't talked about the y-intercept yet. But, what this graph shows us is how the total error behaves as we move the line up and down. The y-intercept here is the thing we change that makes our line move up and down. As you can see the dots go up when we move the line down from 0 to -5, and the dots go up when we move the line up from 0 to +5. The

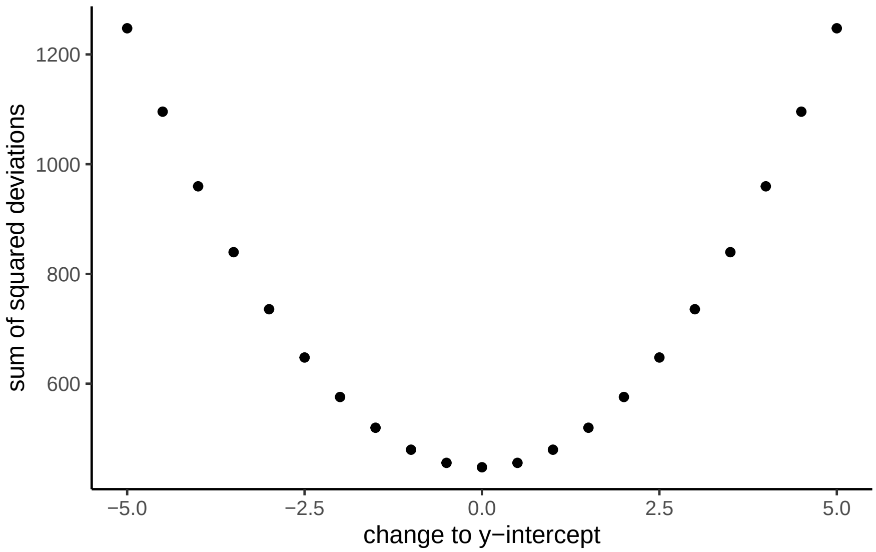


Figure 3.8: A plot of the sum of the squared deviations for different lines moving up and down, through the best fit line. The best fit line occurs at the position that minimizes the sum of the sqaured deviations.

best line, that minimizes the error occurs right in the middle, when we don't move the blue regression line at all.

3.5.2 Lines

OK, fine you say. So, there is one magic line that will go through the middle of the scatter plot and minimize the sum of the squared deviations. How do I find this magic line? We'll show you. But, to be completely honest, you'll almost never do it the way we'll show you here. Instead, it's much easier to use software and make your computer do it for. You'll learn how to that in the labs.

Before we show you how to find the regression line, it's worth refreshing your memory about how lines work, especially in 2 dimensions. Remember this?

$y = ax + b$, or also $y = mx + b$ (sometimes a or m is used for the slope)

This is the formula for a line. Another way of writing it is:

$y = slope * x +$ y-intercept

The slope is the slant of the line, and the y-intercept is where the line crosses the y-axis. Let's look at the lines in Figure 3.9.

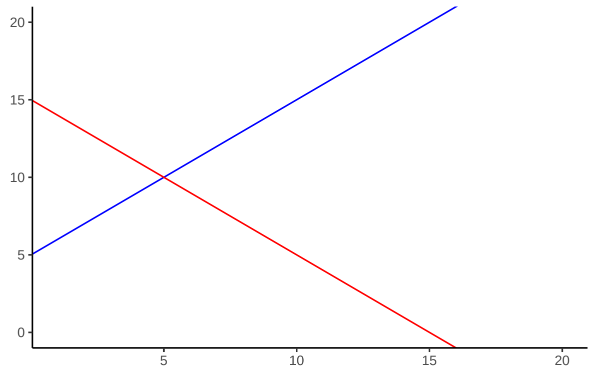


Figure 3.9: Two different lines with different y-intercepts (where the line crosses the y-axis), and different slopes. A positive slope makes the line go up from left to right. A negative slope makes the line go down from left to right.

The formula for the blue line is $y = 1 * x + 5$. Let's talk about that. When x = 0, where is the blue line on the y-axis? It's at five. That happens because 1 times 0 is 0, and then we just have the five left over. How about when x = 5? In that case y =10. You just need the plug in the numbers to the formula, like this:

$y = 1 * x + 5$ $y = 1 * 5 + 5 = 5 + 5 = 10$

The point of the formula is to tell you where y will be, for any number of x. The slope of the line tells you whether the line is going to go up or down, as you move from the left to the right. The blue line has a positive slope of one, so it goes up as x goes up. How much does it go up? It goes up by one for everyone one of x! If we made the slope a 2, it would be much steeper, and go up faster. The red line has a negative slope, so it slants down. This means y goes down, as x goes up. When there is no slant, and we want to make a perfectly flat line, we set the slope to 0. This means that y doesn't go anywhere as x gets bigger and smaller.

That's lines.

3.5.3 Computing the best fit line

If you have a scatter plot showing the locations of scores from two variables, the real question is how can you find the slope and the y-intercept for the best fit line? What are you going to do? Draw millions of lines, add up the residuals, and then see which one was best? That would take forever. Fortunately, there are computers, and when you don't have one around, there's also some handy formulas.

Note

It's worth pointing out just how much computers have changed everything. Before computers everyone had to do these calculations by hand, such a chore! Aside from the deeper mathematical ideas in the formulas, many of them were made for convenience, to speed up hand calculations, because there were no computers. Now that we have computers, the hand calculations are often just an exercise in algebra. Perhaps they build character. You decide.

We'll show you the formulas. And, work through one example by hand. It's the worst, we know. By the way, you should feel sorry for me as I do this entire thing by hand for you.

Here are two formulas we can use to calculate the slope and the intercept, straight from the data. We won't go into why these formulas do what they do. These ones are for "easy" calculation.

$intercept = b = \frac{\sum y \sum x^2 - \sum x \sum xy}{n \sum x^2 - (\sum x)^2}$

$slope = m = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2}$

In these formulas, the x and the y refer to the individual scores. Here's a table showing you how everything fits together.

scores	x	y	x_squared	y_squared	xy
1	1	2	1	4	2
2	4	5	16	25	20
3	3	1	9	1	3
4	6	8	36	64	48
5	5	6	25	36	30
6	7	8	49	64	56
7	8	9	64	81	72
Sums	34	39	200	275	231

We see 7 sets of scores for the x and y variable. We calculated x^2 by squaring each value of x, and putting it in a column. We calculated y^2 by squaring each value of y, and putting it in a column. Then we calculated xy, by multiplying each x score with each y score, and put that in a column. Then we added all the columns up, and put the sums at the bottom. These are all the number we need for the formulas to find the best fit line. Here's what the formulas look like when we put numbers in them:

$intercept = b = \frac{\sum y \sum x^2 - \sum x \sum xy}{n \sum x^2 - (\sum x)^2} = \frac{39*200 - 34*231}{7*200 - 34^2} = -.221$

$slope = m = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2} = \frac{7*231 - 34*39}{7*275 - 34^2} = 1.19$

Great, now we can check our work, let's plot the scores in a scatter plot and draw a line through it with slope = 1.19, and a y-intercept of -.221. As shown in Figure 3.10, the line should go through the middle of the dots.

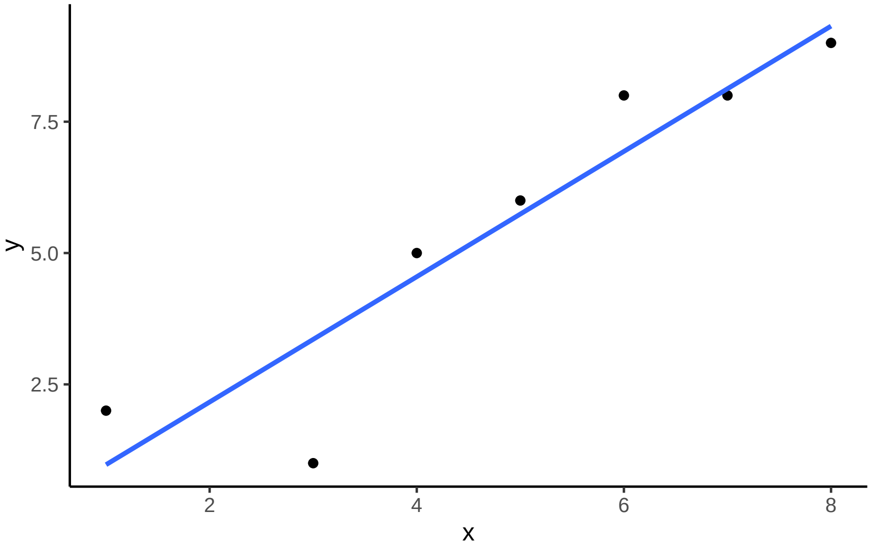


Figure 3.10: An example regression line with confidence bands going through a few data points in a scatterplot

3.6 Interpreting Correlations

What does the presence or the absence of a correlation between two measures mean? How should correlations be interpreted? What kind of inferences can be drawn from correlations? These are all very good questions. A first piece of advice is to use caution when interpreting correlations. Here's why.

3.6.1 Correlation does not equal causation

Perhaps you have heard that correlation does not equal causation. Why not? There are lots of reasons why not. However, before listing some of the reasons let's start with a case where we would expect a causal connection between two measurements. Consider, buying a snake plant for your home. Snake plants are supposed to be easy to take care of because you can mostly ignore them.

Like most plants, snake plants need some water to stay alive. However, they also need just the right amount of water. Imagine an experiment where 1000 snake plants were grown in a house.

Each snake plant is given a different amount of water per day, from zero teaspoons of water per day to 1000 teaspoons of water per day. We will assume that water is part of the causal process that allows snake plants to grow. The amount of water given to each snake plant per day can also be one of our measures. Imagine further that every week the experimenter measures snake plant growth, which will be the second measurement. Now, can you imagine for yourself what a scatter plot of weekly snake plant growth by tablespoons of water would look like?

3.6.1.1 Even when there is causation, there might not be obvious correlation

The first plant given no water at all would have a very hard time and eventually die. It should have the least amount of weekly growth. How about the plants given only a few teaspoons of water per day. This could be just enough water to keep the plants alive, so they will grow a little bit but not a lot. If you are imagining a scatter plot, with each dot being a snake plant, then you should imagine some dots starting in the bottom left hand corner (no water & no plant growth), moving up and to the right (a bit of water, and a bit of growth). As we look at snake plants getting more and more water, we should see more and more plant growth, right? "Sure, but only up to a point". Correct, there should be a trend for a positive correlation with increasing plant growth as amount of water per day increases. But, what happens when you give snake plants too much water? From personal experience, they die. So, at some point, the dots in the scatter plot will start moving back down again. Snake plants that get way too much water will not grow very well.

The imaginary scatter plot you should be envisioning could have an upside U shape. Going from left to right, the dot's go up, they reach a maximum, then they go down again reaching a minimum. Computing Pearson's r for data like this can give you r values close to zero. The scatter plot could look something like Figure 3.11.

Granted this looks more like an inverted V, than an inverted U, but you get the picture right? There is clearly a relationship between watering and snake plant growth. But, the correlation isn't in one direction. As a result, when we compute the correlation in terms of Pearson's r, we get a value suggesting no relationship.

```
#> [1] -0.008825909
```

What this really means is there is no linear relationship that can be described by a single straight line. When we need lines or curves going in more than one direction, we have a nonlinear relationship.

This example illustrates some conundrums in interpreting correlations. We already know that water is needed for plants to grow, so we are rightly expecting there to be a relationship between our measure of amount of water and plant growth. If we look at the first half of the data we see a positive correlation, if we look at the last half of the data we see a negative

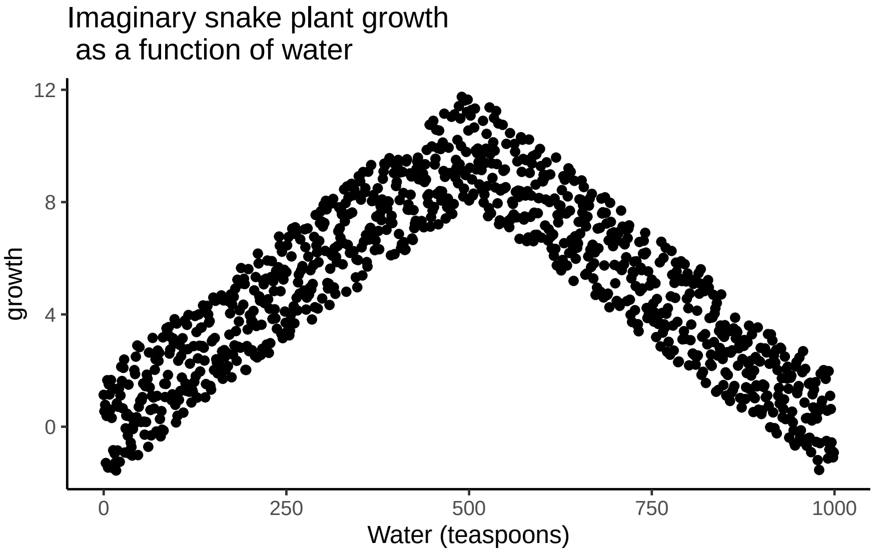


Figure 3.11: Illustration of a possible relationship between amount of water and snake plant growth. Growth goes up with water, but eventually goes back down as too much water makes snake plants die.

correlation, and if we look at all of the data we see no correlation. Yikes. So, even when there is a causal connection between two measures, we won't necessarily obtain clear evidence of the connection just by computing a correlation coefficient.

> Pro Tip: This is one reason why plotting your data is so important. If you see an upside U shape pattern, then a correlation analysis is probably not the best analysis for your data.

3.6.1.2 Confounding variable, or Third variable problem

Anybody can correlate any two things that can be quantified and measured. For example, we could find a hundred people, ask them all sorts of questions like:

1. how happy are you
2. how old are you
3. how tall are you
4. how much money do you make per year
5. how long are your eyelashes
6. how many books have you read in your life
7. how loud is your inner voice

Let's say we found a positive correlation between yearly salary and happiness. Note, we could have just as easily computed the same correlation between happiness and yearly salary. If we found a correlation, would you be willing to infer that yearly salary causes happiness? Perhaps it does play a small part. But, something like happiness probably has a lot of contributing causes. Money could directly cause some people to be happy. But, more likely, money buys people access to all sorts of things, and some of those things might contribute happiness. These "other" things are called **third** variables. For example, perhaps people living in nicer places in more expensive houses are more happy than people in worse places in cheaper houses. In this scenario, money isn't causing happiness, it's the places and houses that money buys. But, even is this were true, people can still be more or less happy in lots of different situations.

The lesson here is that a correlation can occur between two measures because of a third variable that is not directly measured. So, just because we find a correlation, does not mean we can conclude anything about a causal connection between two measurements.

3.6.2 Correlation and Random chance

Another very important aspect of correlations is the fact that they can be produced by random chance. This means that you can find a positive or negative correlation between two measures, even when they have absolutely nothing to do with one another. You might have hoped to find zero correlation when two measures are totally unrelated to each other. Although this certainly happens, unrelated measures can accidentally produce **spurious** correlations, just by chance alone.

Let's demonstrate how correlations can occur by chance when there is no causal connection between two measures. Imagine two participants. One is at the North pole with a lottery machine full of balls with numbers from 1 to 10. The other is at the south pole with a different lottery machine full of balls with numbers from 1 to 10. There are an endless supply of balls in the machine, so every number could be picked for any ball. Each participant randomly chooses 10 balls, then records the number on the ball. In this situation we will assume that there is no possible way that balls chosen by the first participant could causally influence the balls chosen by the second participant. They are on the other side of the world. We should assume that the balls will be chosen by chance alone.

Here is what the numbers on each ball could look like for each participant:

Ball	North_pole	South_pole
1	4	5
2	4	7
3	6	4
4	1	5
5	3	8
6	3	4

Ball	North_pole	South_pole
7	9	5
8	8	7
9	7	9
10	4	3

In this one case, if we computed Pearson's r, we would find that $r = 0.2202107$. But, we already know that this value does not tell us anything about the relationship between the balls chosen in the north and south pole. We know that relationship should be completely random, because that is how we set up the game.

The better question here is to ask what can random chance do? For example, if we ran our game over and over again thousands of times, each time choosing new balls, and each time computing the correlation, what would we find?First, we will find fluctuation. The r value will sometimes be positive, sometimes be negative, sometimes be big and sometimes be small. Second, we will see what the fluctuation looks like. This will give us a window into the kinds of correlations that chance alone can produce. Let's see what happens.

3.6.2.1 Monte-carlo simulation of random correlations

It is possible to use a computer to simulate our game as many times as we want. This process is often termed **monte-carlo simulation**.

Below is a script written for the programming language R. We won't go into the details of the code here. However, let's briefly explain what is going on. Notice, the part that says `for(sim in 1:1000)`. This creates a loop that repeats our game 1000 times. Inside the loop there are variables named `North_pole` and `South_pole`. During each simulation, we sample 10 random numbers (between 1 to 10) into each variable. These random numbers stand for the numbers that would have been on the balls from the lottery machine. Once we have 10 random numbers for each, we then compute the correlation using `cor(North_pole,South_pole)`. Then, we save the correlation value and move on to the next simulation. At the end, we will have 1000 individual Pearson r values.

```r
simulated_correlations <- length(0)
for(sim in 1:1000){
  North_pole <- runif(10,1,10)
  South_pole <- runif(10,1,10)
  simulated_correlations[sim] <- cor(North_pole,South_pole)
}

sim_df <- data.frame(sims=1:1000,simulated_correlations)
```

```r
ggplot(sim_df, aes(x = sims, y = simulated_correlations))+
  geom_point()+
  theme_classic()+
  geom_hline(yintercept = -1)+
  geom_hline(yintercept = 1)+
  ggtitle("Simulation of 1000 r values")
```

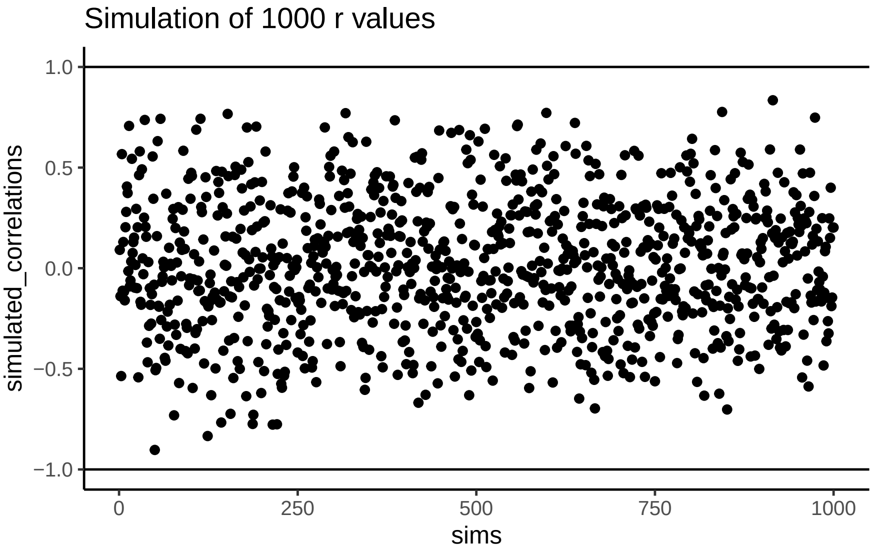


Figure 3.12: Another figure showing a range of r-values that can be obtained by chance.

Figure 3.12 shows the 1000 Pearson r values from the simulation. Does the figure below look familiar to you? We have already conducted a similar kind of simulation before. Each dot in the scatter plot shows the Pearson r for each simulation from 1 to 1000. As you can see the dots are all over of the place, in between the range -1 to 1. The important lesson here is that random chance produced all of these correlations. This means we can find “correlations” in the data that are completely meaningless, and do not reflect any causal relationship between one measure and another.

Let’s illustrate the idea of finding “random” correlations one more time, with a little movie. This time, we will show you a scatter plot of the random values sampled for the balls chosen from the North and South pole. If there is no relationship we should see dots going everywhere. If there happens to be a positive relationship (purely by chance), we should see the dots going from the bottom left to the top right. If there happens to be a negative relationship (purely by chance), we should see the dots going from the top left down to the bottom right.

On more thing to prepare you for the movie. There are three scatter plots below in Figure 3.13, showing negative, positive, and zero correlations between two variables. You’ve already seen this graph before. We are just reminding you that the blue lines are helpful for seeing the correlation.Negative correlations occur when a line goes down from the top left to bottom right. Positive correlations occur when a line goes up from the bottom left to the top right. Zero correlations occur when the line is flat (doesn't go up or down).

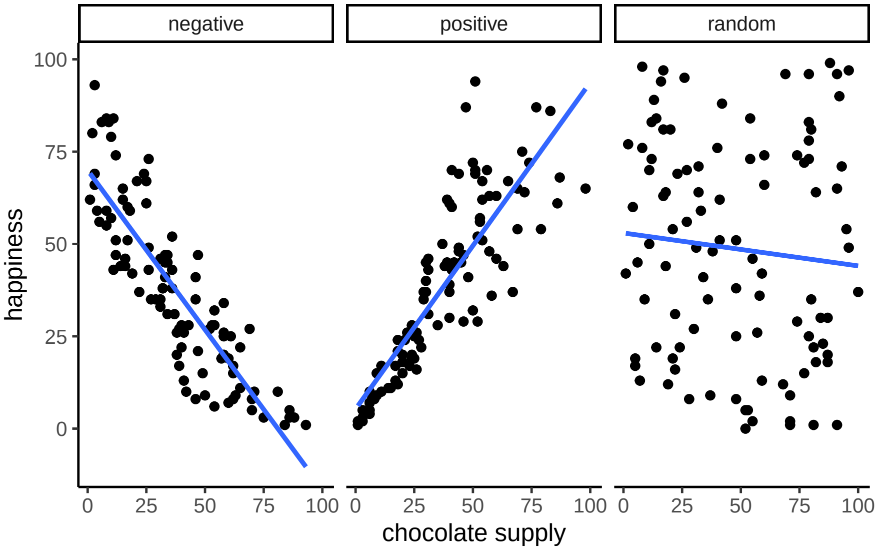


Figure 3.13: A reminder of what positive, negative, and zero correlation looks like

OK, now we are ready for the movie. **?@fig-3randcor10gif** shows the process of sampling two sets of numbers randomly, one for the X variable, and one for the Y variable. Each time we sample 10 numbers for each, plot them, then draw a line through them. Remember, these numbers are all completely random, so we should expect, on average that there should be no correlation between the numbers. However, this is not what happens. You can the line going all over the place. Sometimes we find a negative correlation (line goes down), sometimes we see a positive correlation (line goes up), and sometimes it looks like zero correlation (line is more flat).

You might be thinking this is kind of disturbing. If we know that there should be no correlation between two random variables, how come we are finding correlations? This is a big problem right? I mean, if someone showed me a correlation between two things, and then claimed one thing was related to another, how could know I if it was true. After all, it could be chance! Chance can do that too.

Fortunately, all is not lost. We can look at our simulated data in another way, using a histogram. Remember, just before the movie, we simulated 1000 different correlations using random numbers. By, putting all of those r values into a histogram, we can get a better sense of how chance behaves. We can see what kind of correlations chance is likely or unlikely to produce. Figure 3.14 is a histogram of the simulated r values.

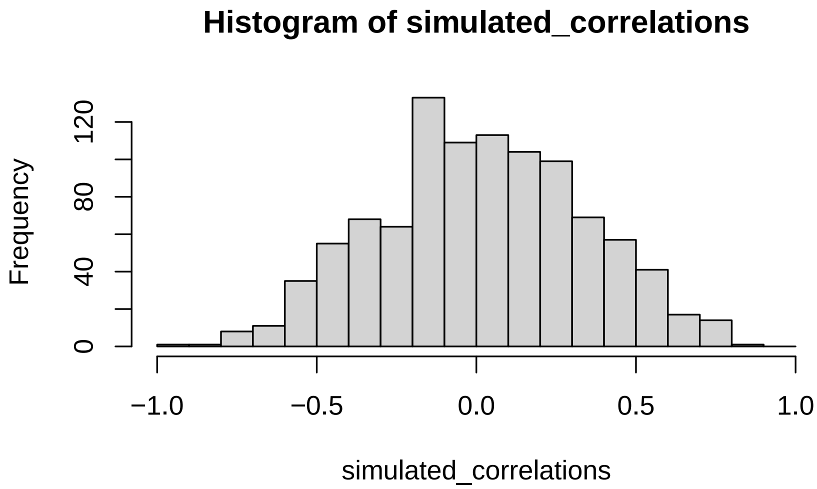


Figure 3.14: A histogram showing the frequency distribution of r-values for completely random values between an X and Y variable (sample-size=10). A rull range of r-values can be obtained by chance alone. Larger r-values are less common than smaller r-values

Notice that this histogram is not flat. Most of the simulated r values are close to zero. Notice, also that the bars get smaller as you move away from zero in the positive or negative direction. The general take home here is that chance can produce a wide range of correlations. However, not all correlations happen very often. For example, the bars for -1 and 1 are very small. Chance does not produce nearly perfect correlations very often. The bars around -.5 and .5 are smaller than the bars around zero, as medium correlations do not occur as often as small correlations by chance alone.

You can think of this histogram as the window of chance. It shows what chance often does, and what it often does not do. If you found a correlation under these very same circumstances (e.g., measured the correlation between two sets of 10 random numbers), then you could consult this window. What should you ask the window? How about, could my observed correlation (the one that you found in your data) have come from this window. Let's say you found a correlation of $r = .1$. Could a .1 have come from the histogram? Well, look at the histogram around where the .1 mark on the x-axis is. Is there a big bar there? If so, this means that chance produces this value fairly often. You might be comfortable with the inference: Yes, this .1 could have been produced by chance, because it is well inside the window of chance. How about $r = .5$? The bar is much smaller here, you might think, "well, I can see that chance does produce .5 some times, so chance could have produced my .5. Did it? Maybe, maybe not, not sure". Here, your confidence in a strong inference about the role of chance might start getting a bit shakier.

How about an $r = .95$?. You might see that the bar for .95 is very very small, perhaps too small to see. What does this tell you? It tells you that chance does not produce .95 very often, hardly if at all, pretty much never. So, if you found a .95 in your data, what would you infer? Perhaps you would be comfortable inferring that chance did not produce your .95, after .95 is mostly outside the window of chance.

3.6.2.2 Increasing sample-size decreases opportunity for spurious correlation

Before moving on, let's do one more thing with correlations. In our pretend lottery game, each participant only sampled 10 balls each. We found that this could lead to a range of correlations between the numbers randomly drawn from either sides of the pole. Indeed, we even found some correlations that were medium to large in size. If you were a researcher who found such correlations, you might be tempted to believe there was a relationship between your measurements. However, we know in our little game, that those correlations would be spurious, just a product of random sampling.

The good news is that, as a researcher, you get to make the rules of the game. You get to determine how chance can play. This is all a little bit metaphorical, so let's make it concrete.

We will see what happens in four different scenarios. First, we will repeat what we already did. Each participant will draw 10 balls, then we compute the correlation, and do this over 1000 times and look at a histogram. Second, we will change the game so each participant draws 50 balls each, and then repeat our simulation. Third, and fourth, we will change the game so each participant draws 100 balls each, and then 1000 balls each, and repeat etc.

Figure 3.15 shows four different histograms of the Pearson r values in each of the different scenarios. Each scenario involves a different sample-size, from, 10, 50, 100 to 1000.

By inspecting the four histograms you should notice a clear pattern. The width or range of each histogram shrinks as the sample-size increases. What is going on here? Well, we already know that we can think of these histograms as windows of chance. They tell us which r values occur fairly often, which do not. When our sample-size is 10, lots of different r values happen. That histogram is very flat and spread out. However, as the sample-size increases, we see that the window of chance gets pulled in. For example, by the time we get to 1000 balls each, almost all of the Pearson r values are very close to 0.

One take home here, is that increasing sample-size narrows the window of chance. So, for example, if you ran a study involving 1000 samples of two measures, and you found a correlation of .5, then you can clearly see in the bottom right histogram that .5 does not occur very often by chance alone. In fact, there is no bar, because it didn't happen even once in the simulation. As a result, when you have a large sample size like n = 1000, you might be more confident that your observed correlation (say of .5) was not a spurious correlation. If chance is not producing your result, then something else is.

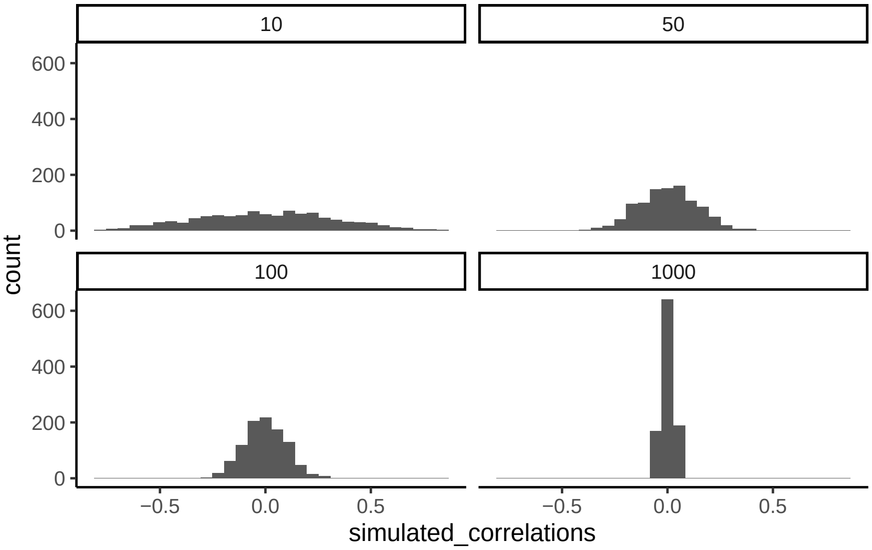


Figure 3.15: Four histograms showing the frequency distributions of r-values between completely random X and Y variables as a function of sample-size. The width of the distributions shrink as sample-size increases. Smaller sample-sizes are more likely to produce a wider range of r-values by chance. Larger sample-sizes always produce a narrow range of small r-values

Finally, notice how your confidence about whether or not chance is mucking about with your results depends on your sample size. If you only obtained 10 samples per measurement, and found $r = .5$, you should not be as confident that your correlation reflects a real relationship. Instead, you can see that r's of .5 happen fairly often by chance alone.

> Pro tip: when you run an experiment you get to decide how many samples you will collect, which means you can choose to narrow the window of chance. Then, if you find a relationship in the data you can be more confident that your finding is real, and not just something that happened by chance.

3.6.3 Some more movies

Let's ingrain these idea with some more movies. When our sample-size is small (N is small), sampling error can cause all sort "patterns" in the data. This makes it possible, and indeed common, for "correlations" to occur between two sets of numbers. When we increase the sample-size, sampling error is reduced, making it less possible for "correlations" to occur just by chance alone. When N is large, chance has less of an opportunity to operate.

3.6.3.1 Watching how correlation behaves when there is no correlation

Below we randomly sample numbers for two variables, plot them, and show the correlation using a line. There are four panels, each showing the number of observations in the samples, from 10, 50, 100, to 1000 in each sample.

Remember, because we are randomly sampling numbers, there should be no relationship between the X and Y variables. But, as we have been discussing, because of chance, we can sometimes observe a correlation (due to chance). The important thing to watch is how the line behaves across the four panels in **?@fig-3corRandfour**. The line twirls around in all directions when the sample size is 10. It is also moves around quite a bit when the sample size is 50 or 100. It still moves a bit when the sample size is 1000, but much less. In all cases we expect that the line should be flat, but every time we take new samples, sometimes the line shows us pseudo patterns.

Which line should you trust? Well, hopefully you can see that the line for 1000 samples is the most stable. It tends to be very flat every time, and it does not depend so much on the particular sample. The line with 10 observations per sample goes all over the place. The take home here, is that if someone told you that they found a correlation, you should want to know how many observations they hand in their sample. If they only had 10 observations, how could you trust the claim that there was a correlation? You can't!!! Not now that you know samples that are that small can do all sorts of things by chance alone. If instead, you found out the sample was very large, then you might trust that finding a little bit more. For example, in the above movie you can see that when there are 1000 samples, we never see a strong or weak correlation; the line is always flat. This is because chance almost never produces strong correlations when the sample size is very large.

In the above example, we sampled numbers random numbers from a uniform distribution. Many examples of real-world data will come from a normal or approximately normal distribution. We can repeat the above, but sample random numbers from the same normal distribution. There will still be zero actual correlation between the X and Y variables, because everything is sampled randomly. **?@fig-3normCorfour** shows the same behavior. The computed correlation for small sample-sizes fluctuate wildly, and large sample sizes do not.

OK, so what do things look like when there actually is a correlation between variables?

3.6.3.2 Watching correlations behave when there really is a correlation

Sometimes there really are correlations between two variables that are not caused by chance. **?@fig-3realcorFour** shows a movie of four scatter plots. Each shows the correlation between two variables. Again, we change the sample-size in steps of 10, 50 100, and 1000. The data have been programmed to contain a real positive correlation. So, we should expect that the line will be going up from the bottom left to the top right. However, there is still variability in the data. So this time, sampling error due to chance will fuzz the correlation. We know it is there, but sometimes chance will cause the correlation to be eliminated.

Notice that in the top left panel (sample-size = 10), the line is twirling around much more than the other panels. Every new set of samples produces different correlations. Sometimes, the line even goes flat or downward. However, as we increase sample-size, we can see that the line doesn't change very much, it is always going up showing a positive correlation.

The main takeaway here is that even when there is a positive correlation between two things, you might not be able to see it if your sample size is small. For example, you might get unlucky with the one sample that you measured. Your sample could show a negative correlation, even when the actual correlation is positive! Unfortunately, in the real world we usually only have the sample that we collected, so we always have to wonder if we got lucky or unlucky. Fortunately, if you want to remove luck, all you need to do is collect larger samples. Then you will be much more likely to observe the real pattern, rather the pattern that can be introduced by chance.

3.7 Summary

In this section we have talked about correlation, and started to build some intuitions about **inferential statistics**, which is the major topic of the remaining chapters. For now, the main ideas are:

1. We can measure relationships in data using things like correlation

2. The correlations we measure can be produced by numerous things, so they are hard to to interpret
3. Correlations can be produced by chance, so have the potential to be completely meaningless.
4. However, we can create a model of exactly what chance can do. The model tells us whether chance is more or less likely to produce correlations of different sizes
5. We can use the chance model to help us make decisions about our own data. We can compare the correlation we found in our data to the model, then ask whether or not chance could have or was likely to have produced our results.

4 Probability, Sampling, and Estimation

Sections 4.1 & 4.9 - Adapted text by Danielle Navarro Section 4.10 - 4.11 & 4.13 - Mix of Matthew Crump & Danielle Navarro Section 4.12 - 4.13 - Adapted text by Danielle Navarro, all sections modified by Mallory Barnes.

> I have studied many languages-French, Spanish and a little Italian, but no one told me that Statistics was a foreign language. —Charmaine J. Forde

Up to this point in the book, we've discussed some of the key ideas in experimental design, and we've talked a little about how you can summarize a data set. To a lot of people, this is all there is to statistics: it's about calculating averages, collecting all the numbers, drawing pictures, and putting them all in a report somewhere. Kind of like stamp collecting, but with numbers. However, statistics covers much more than that. In fact, descriptive statistics is one of the smallest parts of statistics, and one of the least powerful. The bigger and more useful part of statistics is that it provides tools **that let you make inferences about data**.

Once you start thinking about statistics in these terms – that statistics is there to help us draw inferences from data – you start seeing examples of it everywhere. For instance, here's a tiny extract from a newspaper article in the Sydney Morning Herald (30 Oct 2010):

> "I have a tough job," the Premier said in response to a poll which found her government is now the most unpopular Labor administration in polling history, with a primary vote of just 23 per cent.

This kind of remark is entirely unremarkable in the papers or in everyday life, but let's have a think about what it entails. A polling company has conducted a survey, usually a pretty big one because they can afford it. I'm too lazy to track down the original survey, so let's just imagine that they called 1000 voters at random, and 230 (23%) of those claimed that they intended to vote for the party. For the 2010 Federal election, the Australian Electoral Commission reported 4,610,795 enrolled voters in New South Whales; so the opinions of the remaining 4,609,795 voters (about 99.98% of voters) remain unknown to us. Even assuming that no-one lied to the polling company the only thing we can say with 100% confidence is that the true primary vote is somewhere between 230/4610795 (about 0.005%) and 4610025/4610795 (about 99.83%). So, on what basis is it legitimate for the polling company, the newspaper, and the readership to conclude that the ALP primary vote is only about 23%?

The answer to the question is pretty obvious: if I call 1000 people at random, and 230 of them say they intend to vote for the ALP, then it seems very unlikely that these are the **only** 230 people out of the entire voting public who actually intend to do so. In other words, we assume that the data collected by the polling company is pretty representative of the population at large. But how representative? Would we be surprised to discover that the true ALP primary vote is actually 24%? 29%? 37%? At this point everyday intuition starts to break down a bit. No-one would be surprised by 24%, and everybody would be surprised by 37%, but it's a bit hard to say whether 29% is plausible. We need some more powerful tools than just looking at the numbers and guessing.

Inferential statistics provides the tools that we need to answer these sorts of questions, and since these kinds of questions lie at the heart of the scientific enterprise, they take up the lions share of every introductory course on statistics and research methods. However, our tools for making statistical inferences are 1) built on top of **probability theory**, and 2) require an understanding of how samples behave when you take them from distributions (defined by probability theory…). So, this chapter has two main parts. A brief introduction to probability theory, and an introduction to sampling from distributions.

4.1 How are probability and statistics different?

Before we start talking about probability theory, it's helpful to spend a moment thinking about the relationship between probability and statistics. The two disciplines are closely related but they're not identical. Probability theory is "the doctrine of chances". It's a branch of mathematics that tells you how often different kinds of events will happen. For example, all of these questions are things you can answer using probability theory:

- What are the chances of a fair coin coming up heads 10 times in a row?
- If I roll two six sided dice, how likely is it that I'll roll two sixes?
- How likely is it that five cards drawn from a perfectly shuffled deck will all be hearts?
- What are the chances that I'll win the lottery?

Notice that all of these questions have something in common. In each case the "truth of the world" is known, and my question relates to the "what kind of events" will happen. In the first question I **know** that the coin is fair, so there's a 50% chance that any individual coin flip will come up heads. In the second question, I **know** that the chance of rolling a 6 on a single die is 1 in 6. In the third question I **know** that the deck is shuffled properly. And in the fourth question, I **know** that the lottery follows specific rules. You get the idea. The critical point is that probabilistic questions start with a known *model* of the world, and we use that model to do some calculations.

The underlying model can be quite simple. For instance, in the coin flipping example, we can write down the model like this: $P(\text{heads}) = 0.5$ which you can read as “the probability of heads is 0.5”.

As we’ll see later, in the same way that percentages are numbers that range from 0% to 100%, probabilities are just numbers that range from 0 to 1. When using this probability model to answer the first question, I don’t actually know exactly what’s going to happen. Maybe I’ll get 10 heads, like the question says. But maybe I’ll get three heads. That’s the key thing: in probability theory, the **model** is known, but the **data** are not.

So that’s probability. What about statistics? Statistical questions work the other way around. In statistics, we know the truth about the world. All we have is the data, and it is from the data that we want to **learn** the truth about the world. Statistical questions tend to look more like these:

- If my friend flips a coin 10 times and gets 10 heads, are they playing a trick on me?
- If five cards off the top of the deck are all hearts, how likely is it that the deck was shuffled?
- If the lottery commissioner’s spouse wins the lottery, how likely is it that the lottery was rigged?

This time around, the only thing we have are data. What I **know** is that I saw my friend flip the coin 10 times and it came up heads every time. And what I want to *infer* is whether or not I should conclude that what I just saw was actually a fair coin being flipped 10 times in a row, or whether I should suspect that my friend is playing a trick on me. The data I have look like this:

```
H H H H H H H H H H
```

and what I’m trying to do is work out which “model of the world” I should put my trust in. If the coin is fair, then the model I should adopt is one that says that the probability of heads is 0.5; that is, $P(\text{heads}) = 0.5$. If the coin is not fair, then I should conclude that the probability of heads is **not** 0.5, which we would write as $P(\text{heads}) \neq 0.5$. In other words, the statistical inference problem is to figure out which of these probability models is right. Clearly, the statistical question isn’t the same as the probability question, but they’re deeply connected to one another. Because of this, a good introduction to statistical theory will start with a discussion of what probability is and how it works.

4.2 What does probability mean?

Let's start with the first of these questions. What is "probability"? It might seem surprising to you, but while statisticians and mathematicians (mostly) agree on what the **rules** of probability are, there's much less of a consensus on what the word really **means**. It seems weird because we're all very comfortable using words like "chance", "likely", "possible" and "probable", and it doesn't seem like it should be a very difficult question to answer. If you had to explain "probability" to a five year old, you could do a pretty good job. But if you've ever had that experience in real life, you might walk away from the conversation feeling like you didn't quite get it right, and that (like many everyday concepts) it turns out that you don't **really** know what it's all about.

So I'll have a go at it. Let's suppose I want to bet on a soccer game between two teams of robots, **Arduino Arsenal** and **C Milan**. After thinking about it, I decide that there is an 80% probability that **Arduino Arsenal** winning. What do I mean by that? Here are three possibilities…

- They're robot teams, so I can make them play over and over again, and if I did that, **Arduino Arsenal** would win 8 out of every 10 games on average.
- For any given game, I would only agree that betting on this game is only "fair" if a \$1 bet on **C Milan** gives a \$5 payoff (i.e. I get my \$1 back plus a \$4 reward for being correct), as would a \$4 bet on **Arduino Arsenal** (i.e., my \$4 bet plus a \$1 reward).
- My subjective "belief" or "confidence" in an **Arduino Arsenal** victory is four times as strong as my belief in a **C Milan** victory.

Each of these seems sensible. However they're not identical, and not every statistician would endorse all of them. The reason is that there are different statistical ideologies (yes, really!) and depending on which one you subscribe to, you might say that some of those statements are meaningless or irrelevant. In this section, I give a brief introduction the two main approaches that exist in the literature. These are by no means the only approaches, but they're the two big ones.

4.2.1 The frequentist view

The first of the two major approaches to probability, and the more dominant one in statistics, is referred to as the *frequentist view*, and it defines probability as a *long-run frequency.* Suppose we were to try flipping a fair coin, over and over again. By definition, this is a coin that has $P(H) = 0.5$. What might we observe? One possibility is that the first 20 flips might look like this:

```
T,H,H,H,H,T,T,H,H,H,H,T,H,H,T,T,T,T,T,H
```

In this case 11 of these 20 coin flips (55%) came up heads. Now suppose that I'd been keeping a running tally of the number of heads (which I'll call N_H) that I've seen, across the first N flips, and calculate the proportion of heads N_H/N every time. Here's what I'd get (I did literally flip coins to produce this!):

number of flips	1	2	3	4	5	6	7	8	9	10
number of heads	0	1	2	3	4	4	4	5	6	7
proportion	.00	.50	.67	.75	.80	.67	.57	.63	.67	.70

number of flips	11	12	13	14	15	16	17	18	19	20
number of heads	8	8	9	10	10	10	10	10	10	11
proportion	.73	.67	.69	.71	.67	.63	.59	.56	.53	.55

Notice that at the start of the sequence, the **proportion** of heads fluctuates wildly, starting at .00 and rising as high as .80. Later on, one gets the impression that it dampens out a bit, with more and more of the values actually being pretty close to the "right" answer of .50. This is the frequentist definition of probability in a nutshell: flip a fair coin over and over again, and as N grows large (approaches infinity, denoted $N \rightarrow \infty$), the proportion of heads will converge to 50%. There are some subtle technicalities that the mathematicians care about, but qualitatively speaking, that's how the frequentists define probability. Unfortunately, I don't have an infinite number of coins, or the infinite patience required to flip a coin an infinite number of times. However, I do have a computer, and computers excel at mindless repetitive tasks. So I asked my computer to simulate flipping a coin 1000 times, and then drew a picture of what happens to the proportion N_H/N as N increases. Actually, I did it four times, just to make sure it wasn't a fluke. The results are shown in Figure 4.1. As you can see, the **proportion of observed heads** eventually stops fluctuating, and settles down; when it does, the number at which it finally settles is the true probability of heads.

The frequentist definition of probability has some desirable characteristics. First, it is objective: the probability of an event is **necessarily** grounded in the world. The only way that probability statements can make sense is if they refer to (a sequence of) events that occur in the physical universe. Second, it is unambiguous: any two people watching the same sequence of events unfold, trying to calculate the probability of an event, must inevitably come up with the same answer.

However, it also has undesirable characteristics. Infinite sequences don't exist in the physical world. Suppose you picked up a coin from your pocket and started to flip it. Every time it lands, it impacts on the ground. Each impact wears the coin down a bit; eventually, the coin will be destroyed. So, one might ask whether it really makes sense to pretend that an "infinite" sequence of coin flips is even a meaningful concept, or an objective one. We can't say that an "infinite sequence" of events is a real thing in the physical universe, because the physical universe doesn't allow infinite anything.

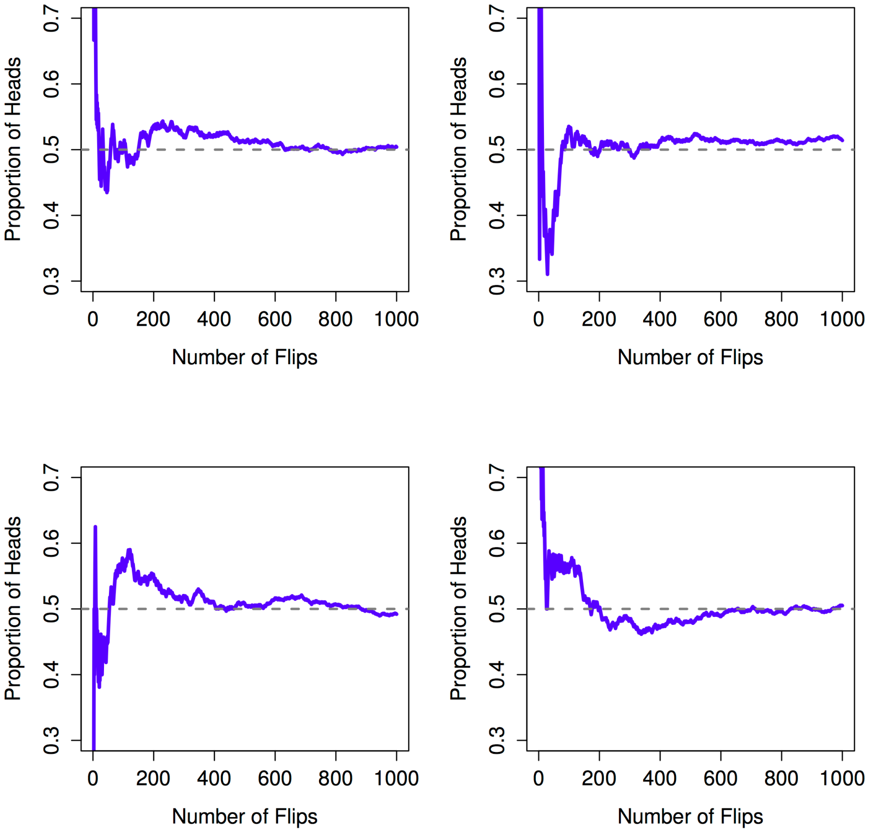


Figure 4.1: An illustration of how frequentist probability works. If you flip a fair coin over and over again, the proportion of heads that you've seen eventually settles down, and converges to the true probability of 0.5. Each panel shows four different simulated experiments: in each case, we pretend we flipped a coin 1000 times, and kept track of the proportion of flips that were heads as we went along. Although none of these sequences actually ended up with an exact value of .5, if we'd extended the experiment for an infinite number of coin flips they would have.

More seriously, the frequentist definition has a narrow scope. There are lots of things out there that human beings are happy to assign probability to in everyday language, but cannot (even in theory) be mapped onto a hypothetical sequence of events. For instance, if a meteorologist comes on TV and says, "the probability of rain in Adelaide on 2 November 2048 is 60%" we humans are happy to accept this. But it's not clear how to define this in frequentist terms. There's only one city of Adelaide, and only 2 November 2048. There's no infinite sequence of events here, just a once-off thing. Frequentist probability genuinely **forbids** us from making probability statements about a single event. From the frequentist perspective, it will either rain tomorrow or it will not; there is no "probability" that attaches to a single non-repeatable event. Now, it should be said that there are some very clever tricks that frequentists can use to get around this. One possibility is that what the meteorologist means is something like this: "There is a category of days for which I predict a 60% chance of rain; if we look only across those days for which I make this prediction, then on 60% of those days it will actually rain". It's very weird and counter intuitive to think of it this way, but you do see frequentists do this sometimes.

4.2.2 The Bayesian view

The **Bayesian view** of probability is often called the subjectivist view, and it is a minority view among statisticians, but one that has been steadily gaining traction for the last several decades. There are many flavors of Bayesianism, making hard to say exactly what "the" Bayesian view is. The most common way of thinking about subjective probability is to define the probability of an event as the **degree of belief** that an intelligent and rational agent assigns to that truth of that event. From that perspective, probabilities don't exist in the world, but rather in the thoughts and assumptions of people and other intelligent beings. However, in order for this approach to work, we need some way of operationalising "degree of belief". One way that you can do this is to formalize it in terms of "rational gambling", though there are many other ways. Suppose that I believe that there's a 60% probability of rain tomorrow. If someone offers me a bet: if it rains tomorrow, then I win \$5, but if it doesn't rain then I lose \$5. Clearly, from my perspective, this is a pretty good bet. On the other hand, if I think that the probability of rain is only 40%, then it's a bad bet to take. Thus, we can operationalize the notion of a "subjective probability" in terms of what bets I'm willing to accept.

What are the advantages and disadvantages to the Bayesian approach? The main advantage is that it allows you to assign probabilities to any event you want to. You don't need to be limited to those events that are repeatable. The main disadvantage (to many people) is that we can't be purely objective – specifying a probability requires us to specify an entity that has the relevant degree of belief. This entity might be a human, an alien, a robot, or even a statistician, but there has to be an intelligent agent out there that believes in things. To many people this is uncomfortable: it seems to make probability arbitrary. While the Bayesian approach does require that the agent in question be rational (i.e., obey the rules of probability), it does allow everyone to have their own beliefs; I can believe the coin is fair and you don't have to, even though we're both rational. The frequentist view doesn't allow any two observers to attribute different probabilities to the same event: when that happens, then at least one of them must be wrong. The Bayesian view does not prevent this from occurring. Two observers with different background knowledge can legitimately hold different beliefs about the same event. In short, where the frequentist view is sometimes considered to be too narrow (forbids lots of things that that we want to assign probabilities to), the Bayesian view is sometimes thought to be too broad (allows too many differences between observers).

4.2.3 What's the difference? And who is right?

Now that you've seen each of these two views independently, it's useful to make sure you can compare the two. Go back to the hypothetical robot soccer game at the start of the section. What do you think a frequentist and a Bayesian would say about these three statements? Which statement would a frequentist say is the correct definition of probability? Which one would a Bayesian do? Would some of these statements be meaningless to a frequentist or a Bayesian? If you've understood the two perspectives, you should have some sense of how to answer those questions.

Okay, assuming you understand the different, you might be wondering which of them is **right**? Honestly, I don't know that there is a right answer. As far as I can tell there's nothing mathematically incorrect about the way frequentists think about sequences of events, and there's nothing mathematically incorrect about the way that Bayesians define the beliefs of a rational agent. In fact, when you dig down into the details, Bayesians and frequentists actually agree about a lot of things. Many frequentist methods lead to decisions that Bayesians agree a rational agent would make. Many Bayesian methods have very good frequentist properties.

For the most part, I'm a pragmatist so I'll use any statistical method that I trust. As it turns out, that makes me prefer Bayesian methods, for reasons I'll explain towards the end of the book, but I'm not fundamentally opposed to frequentist methods. Not everyone is quite so relaxed. For instance, consider Sir Ronald Fisher, one of the towering figures of 20th century statistics and a vehement opponent to all things Bayesian, whose paper on the mathematical foundations of statistics referred to Bayesian probability as "an impenetrable jungle [that] arrests progress towards precision of statistical concepts" Fisher (1922, 311). Or the psychologist Paul Meehl, who suggests that relying on frequentist methods could turn you into "a potent but sterile intellectual rake who leaves in his merry path a long train of ravished maidens but no viable scientific offspring" Meehl (1967, 114). The history of statistics, as you might gather, is not devoid of entertainment.

4.3 Basic probability theory

Ideological arguments between Bayesians and frequentists notwithstanding, it turns out that people mostly agree on the rules that probabilities should obey. There are lots of different ways of arriving at these rules. The most commonly used approach is based on the work of Andrey Kolmogorov, one of the great Soviet mathematicians of the 20th century. I won't go into a lot of detail, but I'll try to give you a bit of a sense of how it works. And in order to do so, I'm going to have to talk about my pants.

4.3.1 Introducing probability distributions

One of the disturbing truths about my life is that I only own 5 pairs of pants: three pairs of jeans, the bottom half of a suit, and a pair of tracksuit pants. Even sadder, I've given them names: I call them X_1, X_2, X_3, X_4 and X_5. I really do: that's why they call me Mister Imaginative. Now, on any given day, I pick out exactly one of pair of pants to wear. Not even I'm so stupid as to try to wear two pairs of pants, and thanks to years of training I never go outside without wearing pants anymore. If I were to describe this situation using the language of probability theory, I would refer to each pair of pants (i.e., each X) as an *elementary event*. The key characteristic of elementary events is that every time we make an observation (e.g., every time I put on a pair of pants), then the outcome will be one and only one of these events. Like I said, these days I always wear exactly one pair of pants, so my pants satisfy this constraint. Similarly, the set of all possible events is called a *sample space*. Granted, some people would call it a "wardrobe", but that's because they're refusing to think about my pants in probabilistic terms. Sad.

Okay, now that we have a sample space (a wardrobe), which is built from lots of possible elementary events (pants), what we want to do is assign a *probability* of one of these elementary events. For an event X, the probability of that event $P(X)$ is a number that lies between 0 and 1. The bigger the value of $P(X)$, the more likely the event is to occur. So, for example, if $P(X) = 0$, it means the event X is impossible (i.e., I never wear those pants). On the other hand, if $P(X) = 1$ it means that event X is certain to occur (i.e., I always wear those pants). For probability values in the middle, it means that I sometimes wear those pants. For instance, if $P(X) = 0.5$ it means that I wear those pants half of the time.

At this point, we're almost done. The last thing we need to recognize is that "something always happens". Every time I put on pants, I really do end up wearing pants (crazy, right?). What this somewhat trite statement means, in probabilistic terms, is that the probabilities of the elementary events need to add up to 1. This is known as the *law of total probability*, not that any of us really care. More importantly, if these requirements are satisfied, then what we have is a *probability distribution*. For example, this is an example of a probability distribution

Which pants?	Label	Probability
Blue jeans	X_1	$P(X_1) = .5$
Grey jeans	X_2	$P(X_2) = .3$
Black jeans	X_3	$P(X_3) = .1$
Black suit	X_4	$P(X_4) = 0$
Blue tracksuit	X_5	$P(X_5) = .1$

Each of the events has a probability that lies between 0 and 1, and if we add up the probability of all events, they sum to 1. Awesome. We can even draw a nice bar graph to visualize this distribution, as shown in Figure 4.2. And at this point, we've all achieved something. You've learned what a probability distribution is, and I've finally managed to find a way to create a graph that focuses entirely on my pants. Everyone wins!

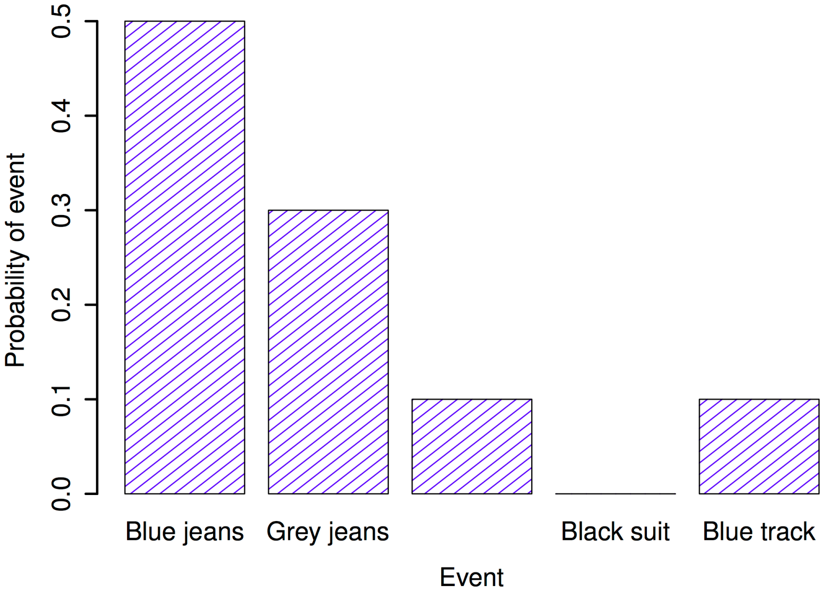


Figure 4.2: A visual depiction of the pants probability distribution. There are five elementary events, corresponding to the five pairs of pants that I own. Each event has some probability of occurring: this probability is a number between 0 to 1. The sum of these probabilities is 1.

The only other thing that I need to point out is that probability theory allows you to talk about *non elementary events* as well as elementary ones. The easiest way to illustrate the concept is with an example. In the pants example, it's perfectly legitimate to refer to the probability that I wear jeans. In this scenario, the "Dan wears jeans" event said to have happened as long as the elementary event that actually did occur is one of the appropriate ones; in this case "blue jeans", "black jeans" or "grey jeans". In mathematical terms, we defined the "jeans" event E to correspond to the set of elementary events (X_1, X_2, X_3). If any of these elementary events occurs, then E is also said to have occurred. Having decided to write down the definition of the E this way, it's pretty straightforward to state what the probability $P(E)$ is: we just add everything up. In this particular case

$$P(E) = P(X_1) + P(X_2) + P(X_3)$$

and, since the probabilities of blue, grey and black jeans respectively are .5, .3 and .1, the probability that I wear jeans is equal to .9.

At this point you might be thinking that this is all terribly obvious and simple and you'd be right. All we've really done is wrap some basic mathematics around a few common sense intuitions. However, from these simple beginnings it's possible to construct some extremely powerful mathematical tools. I'm definitely not going to go into the details in this book, but what I will do is list some of the other rules that probabilities satisfy. These rules can be derived from the simple assumptions that I've outlined above, but since we don't actually use these rules for anything in this book, I won't do so here.

Table 4.4: Some basic rules that probabilities must satisfy. You don't really need to know these rules in order to understand the analyses that we'll talk about later in the book, but they are important if you want to understand probability theory a bit more deeply.

English	Notation		Formula
not A	$P(\neg A)$	$=$	$1 - P(A)$
A or B	$P(A \cup B)$	$=$	$P(A) + P(B) - P(A \cap B)$
A and B	$P(A \cap B)$	$=$	$P(A\|B) P(B)$

Now that we have the ability to "define" non-elementary events in terms of elementary ones, we can actually use this to construct (or, if you want to be all mathematicallish, "derive") some of the other rules of probability. These rules are listed above, and while I'm pretty confident that very few of my readers actually care about how these rules are constructed, I'm going to show you anyway: even though it's boring and you'll probably never have a lot of use for these derivations, if you read through it once or twice and try to see how it works, you'll find that probability starts to feel a bit less mysterious, and with any luck a lot less daunting. So here goes. Firstly, in order to construct the rules I'm going to need a sample space X that consists of a bunch of elementary events x, and two non-elementary events, which I'll call A and B. Let's say:

$$\begin{array}{rcl} X &=& (x_1, x_2, x_3, x_4, x_5) \\ A &=& (x_1, x_2, x_3) \\ B &=& (x_3, x_4) \end{array}$$

To make this a bit more concrete, let's suppose that we're still talking about the pants distribution. If so, A corresponds to the event "jeans", and B corresponds to the event "black":

$$\begin{array}{rcl} \text{“jeans”} & = & (\text{“blue jeans”}, \text{“grey jeans”}, \text{“black jeans”}) \\ \text{“black”} & = & (\text{“black jeans”}, \text{“black suit”}) \end{array}$$

So now let's start checking the rules that I've listed in the table.

In the first line, the table says that

$$P(\neg A) = 1 - P(A)$$

and what it **means** is that the probability of "not A" is equal to 1 minus the probability of A. A moment's thought (and a tedious example) make it obvious why this must be true. If A corresponds to the even that I wear jeans (i.e., one of x_1 or x_2 or x_3 happens), then the only meaningful definition of "not A" (which is mathematically denoted as $\neg A$) is to say that $\neg A$ consists of **all** elementary events that don't belong to A. In the case of the pants distribution it means that $\neg A = (x_4, x_5)$, or, to say it in English: "not jeans" consists of all pairs of pants that aren't jeans (i.e., the black suit and the blue tracksuit). Consequently, every single elementary event belongs to either A or $\neg A$, but not both. Okay, so now let's rearrange our statement above:

$$P(\neg A) + P(A) = 1$$

which is a trite way of saying either I do wear jeans or I don't wear jeans: the probability of "not jeans" plus the probability of "jeans" is 1. Mathematically:

$$\begin{array}{rcl} P(\neg A) & = & P(x_4) + P(x_5) \\ P(A) & = & P(x_1) + P(x_2) + P(x_3) \end{array}$$

so therefore

$$\begin{array}{rcl} P(\neg A) + P(A) & = & P(x_1) + P(x_2) + P(x_3) + P(x_4) + P(x_5) \\ & = & \sum_{x \in X} P(x) \\ & = & 1 \end{array}$$

Excellent. It all seems to work.

Wow, I can hear you saying. That's a lot of xs to tell me the freaking obvious. And you're right: this **is** freaking obvious. The whole **point** of probability theory to to formalize and mathematize a few very basic common sense intuitions. So let's carry this line of thought forward a bit further. In the last section I defined an event corresponding to **not** A, which I denoted $\neg A$. Let's now define two new events that correspond to important everyday concepts: A **and** B, and A **or** B. To be precise:

English statement:	Mathematical notation:
"A and B" both happen	$A \cap B$
at least one of "A or B" happens	$A \cup B$

Since A and B are both defined in terms of our elementary events (the xs) we're going to need to try to describe $A \cap B$ and $A \cup B$ in terms of our elementary events too. Can we do this? Yes we can The only way that both A and B can occur is if the elementary event that we observe turns out to belong to both A and B. Thus "$A \cap B$" includes only those elementary events that belong to both A and B...

$$\begin{array}{rcl} A & = & (x_1, x_2, x_3) \\ B & = & (x_3, x_4) \\ A \cap B & = & (x_3) \end{array}$$

So, um, the only way that I can wear "jeans" (x_1, x_2, x_3) and "black pants" (x_3, x_4) is if I wear "black jeans" (x_3). Another victory for the bloody obvious.

At this point, you're not going to be at all shocked by the definition of $A \cup B$, though you're probably going to be extremely bored by it. The only way that I can wear "jeans" or "black pants" is if the elementary pants that I actually do wear belongs to A or to B, or to both. So...

$$\begin{array}{rcl} A & = & (x_1, x_2, x_3) \\ B & = & (x_3, x_4) \\ A \cup B & = & (x_1, x_2, x_3, x_4) \end{array}$$

Oh yeah baby. Mathematics at its finest.

So, we've defined what we mean by $A \cap B$ and $A \cup B$. Now let's assign probabilities to these events. More specifically, let's start by verifying the rule that claims that:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Using our definitions earlier, we know that $A \cup B = (x_1, x_2, x_3, x_4)$, so

$$P(A \cup B) = P(x_1) + P(x_2) + P(x_3) + P(x_4)$$

and making similar use of the fact that we know what elementary events belong to A, B and $A \cap B$....

$$\begin{array}{rcl} P(A) & = & P(x_1) + P(x_2) + P(x_3) \\ P(B) & = & P(x_3) + P(x_4) \\ P(A \cap B) & = & P(x_3) \end{array}$$

and therefore

$$\begin{array}{rcl} P(A) + P(B) - P(A \cap B) & = & P(x_1) + P(x_2) + P(x_3) + P(x_3) + P(x_4) - P(x_3) \\ & = & P(x_1) + P(x_2) + P(x_3) + P(x_4) \\ & = & P(A \cup B) \end{array}$$

Done.

The next concept we need to define is the notion of "B given A", which is typically written $B|A$. Here's what I mean: suppose that I get up one morning, and put on a pair of pants. An elementary event x has occurred. Suppose further I yell out to my wife (who is in the other room, and so cannot see my pants) "I'm wearing jeans today!". Assuming that she believes that I'm telling the truth, she knows that A is true. **Given** that she knows that A has happened, what is the **conditional probability** that B is also true? Well, let's think about what she knows. Here are the facts:

- **The non-jeans events are impossible**. If A is true, then we know that the only possible elementary events that could have occurred are x_1, x_2 and x_3 (i.e.,the jeans). The non-jeans events x_4 and x_5 are now impossible, and must be assigned probability zero. In other words, our **sample space** has been restricted to the jeans events. But it's still the case that the probabilities of these these events **must** sum to 1: we know for sure that I'm wearing jeans.
- **She's learned nothing about which jeans I'm wearing**. Before I made my announcement that I was wearing jeans, she already knew that I was five times as likely to be wearing blue jeans ($P(x_1) = 0.5$) than to be wearing black jeans ($P(x_3) = 0.1$). My announcement doesn't change this… I said **nothing** about what color my jeans were, so it must remain the case that $P(x_1)/P(x_3)$ stays the same, at a value of 5.

There's only one way to satisfy these constraints: set the impossible events to have zero probability (i.e., $P(x|A) = 0$ if x is not in A), and then divide the probabilities of all the others by $P(A)$. In this case, since $P(A) = 0.9$, we divide by 0.9. This gives:

which pants?	elementary event	old prob, $P(x)$	new prob, $P(x\|A)$
blue jeans	x_1	0.5	0.556
grey jeans	x_2	0.3	0.333
black jeans	x_3	0.1	0.111
black suit	x_4	0	0
blue tracksuit	x_5	0.1	0

In mathematical terms, we say that

$$P(x|A) = \frac{P(x)}{P(A)}$$

if $x \in A$, and $P(x|A) = 0$ otherwise. And therefore…

$$\begin{aligned} P(B|A) &= P(x_3|A) + P(x_4|A) \\ &= \frac{P(x_3)}{P(A)} + 0 \\ &= \frac{P(x_3)}{P(A)} \end{aligned}$$

Now, recalling that $A \cap B = (x_3)$, we can write this as

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

and if we multiply both sides by $P(A)$ we obtain:

$$P(A \cap B) = P(B|A)P(A)$$

which is the third rule that we had listed in the table.

4.4 The binomial distribution

As you might imagine, probability distributions vary enormously, and there's an enormous range of distributions out there. However, they aren't all equally important. In fact, the vast majority of the content in this book relies on one of five distributions: the binomial distribution, the normal distribution, the t distribution, the χ^2 ("chi-square") distribution and the F distribution. Given this, what I'll do over the next few sections is provide a brief introduction to all five of these, paying special attention to the binomial and the normal. I'll start with the binomial distribution, since it's the simplest of the five.

4.4.1 Introducing the binomial

The theory of probability originated in the attempt to describe how games of chance work, so it seems fitting that our discussion of the *binomial distribution* should involve a discussion of rolling dice and flipping coins. Let's imagine a simple "experiment": in my hand I'm holding 20 identical six-sided dice. On one face of each die there's a picture of a skull; the other five faces are all blank. If I proceed to roll all 20 dice, what's the probability that I'll get exactly 4 skulls? Assuming that the dice are fair, we know that the chance of any one die coming up skulls is 1 in 6; to say this another way, the skull probability for a single die is approximately .167. This is enough information to answer our question, so let's have a look at how it's done.

As usual, we'll want to introduce some names and some notation. We'll let N denote the number of dice rolls in our experiment; which is often referred to as the *size parameter* of our binomial distribution. We'll also use θ to refer to the the probability that a single die comes up skulls, a quantity that is usually called the *success probability* of the binomial. Finally, we'll use X to refer to the results of our experiment, namely the number of skulls I get when I roll the dice. Since the actual value of X is due to chance, we refer to it as a *random variable.* In any case, now that we have all this terminology and notation, we can use it to state the problem a little more precisely. The quantity that we want to calculate is the probability that $X = 4$ given that we know that $\theta = .167$ and $N = 20$. The general "form" of the thing I'm interested in calculating could be written as

$$P(X \mid \theta, N)$$

and we're interested in the special case where $X = 4$, $\theta = .167$ and $N = 20$. There's only one more piece of notation I want to refer to before moving on to discuss the solution to the problem. If I want to say that X is generated randomly from a binomial distribution with parameters θ and N, the notation I would use is as follows:

$$X \sim \text{Binomial}(\theta, N)$$

Yeah, yeah. I know what you're thinking: notation, notation, notation. Really, who cares? Very few readers of this book are here for the notation, so I should probably move on and talk about how to use the binomial distribution. To that end, Figure Figure 4.3 plots the binomial probabilities for all possible values of X for our dice rolling experiment, from $X = 0$ (no skulls) all the way up to $X = 20$ (all skulls). Note that this is basically a bar chart, and is no different to the "pants probability" plot I drew in Figure 4.2. On the horizontal axis we have all the possible events, and on the vertical axis we can read off the probability of each of those events. So, the probability of rolling 4 skulls out of 20 times is about 0.20 (the actual answer is 0.2022036, as we'll see in a moment). In other words, you'd expect that to happen about 20% of the times you repeated this experiment.

4.4.2 Working with the binomial distribution in R

R has a function called `dbinom` that calculates binomial probabilities for us. The main arguments to the function are

- `x` This is a number, or vector of numbers, specifying the outcomes whose probability you're trying to calculate.
- `size` This is a number telling R the size of the experiment.
- `prob` This is the success probability for any one trial in the experiment.

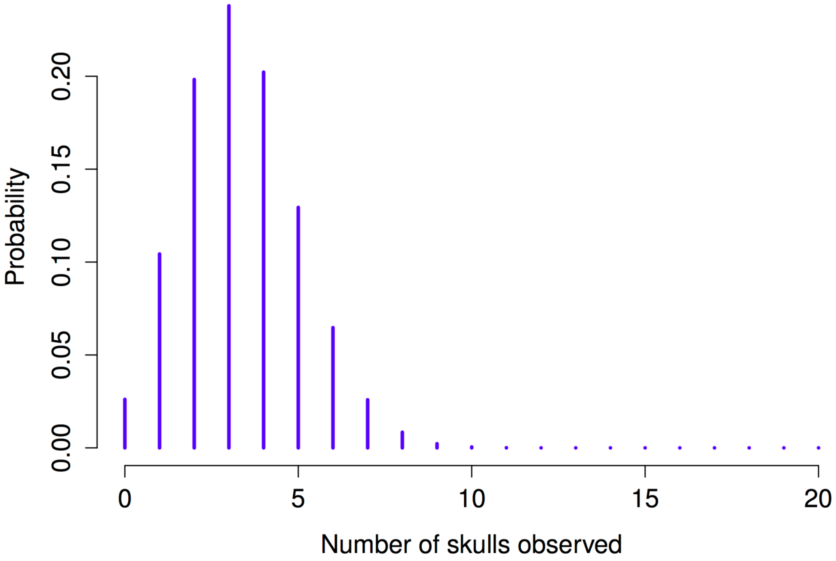


Figure 4.3: The binomial distribution with size parameter of N =20 and an underlying success probability of 1/6. Each vertical bar depicts the probability of one specific outcome (i.e., one possible value of X). Because this is a probability distribution, each of the probabilities must be a number between 0 and 1, and the heights of the bars must sum to 1 as well.

So, in order to calculate the probability of getting skulls, from an experiment of trials, in which the probability of getting a skull on any one trial is … well, the command I would use is simply this:

```
dbinom( x = 4, size = 20, prob = 1/6 )
#> [1] 0.2022036
```

To give you a feel for how the binomial distribution changes when we alter the values of θ and N, let's suppose that instead of rolling dice, I'm actually flipping coins. This time around, my experiment involves flipping a fair coin repeatedly, and the outcome that I'm interested in is the number of heads that I observe. In this scenario, the success probability is now $\theta = 1/2$. Suppose I were to flip the coin $N = 20$ times. In this example, I've changed the success probability, but kept the size of the experiment the same. What does this do to our binomial distribution?

Well, as Figure 4.4 *a* shows, the main effect of this is to shift the whole distribution, as you'd expect. Okay, what if we flipped a coin $N = 100$ times? Well, in that case, we get Figure 4.4 *b*. The distribution stays roughly in the middle, but there's a bit more variability in the possible outcomes.

At this point, I should probably explain the name of the `dbinom` function. Obviously, the "binom" part comes from the fact that we're working with the binomial distribution, but the "d" prefix is probably a bit of a mystery. In this section I'll give a partial explanation: specifically, I'll explain why there is a prefix. As for why it's a "d" specifically, you'll have to wait until the next section. What's going on here is that R actually provides **four** functions in relation to the binomial distribution. These four functions are `dbinom`, `pbinom`, `rbinom` and `qbinom`, and each one calculates a different quantity of interest. Not only that, R does the same thing for **every** probability distribution that it implements. No matter what distribution you're talking about, there's a `d` function, a `p` function, `r` a function and a `q` function.

Let's have a look at what all four functions do. Firstly, all four versions of the function require you to specify the `size` and `prob` arguments: no matter what you're trying to get R to calculate, it needs to know what the parameters are. However, they differ in terms of what the other argument is, and what the output is. So let's look at them one at a time.

- The `d` form we've already seen: you specify a particular outcome `x`, and the output is the probability of obtaining exactly that outcome. (the "d" is short for *density*, but ignore that for now).
- The `p` form calculates the *cumulative probability*. You specify a particular quantile `q` , and it tells you the probability of obtaining an outcome **smaller than or equal to** `q`.
- The `q` form calculates the *quantiles* of the distribution. You specify a probability value `p`, and it gives you the corresponding percentile. That is, the value of the variable for which there's a probability `p` of obtaining an outcome lower than that value.

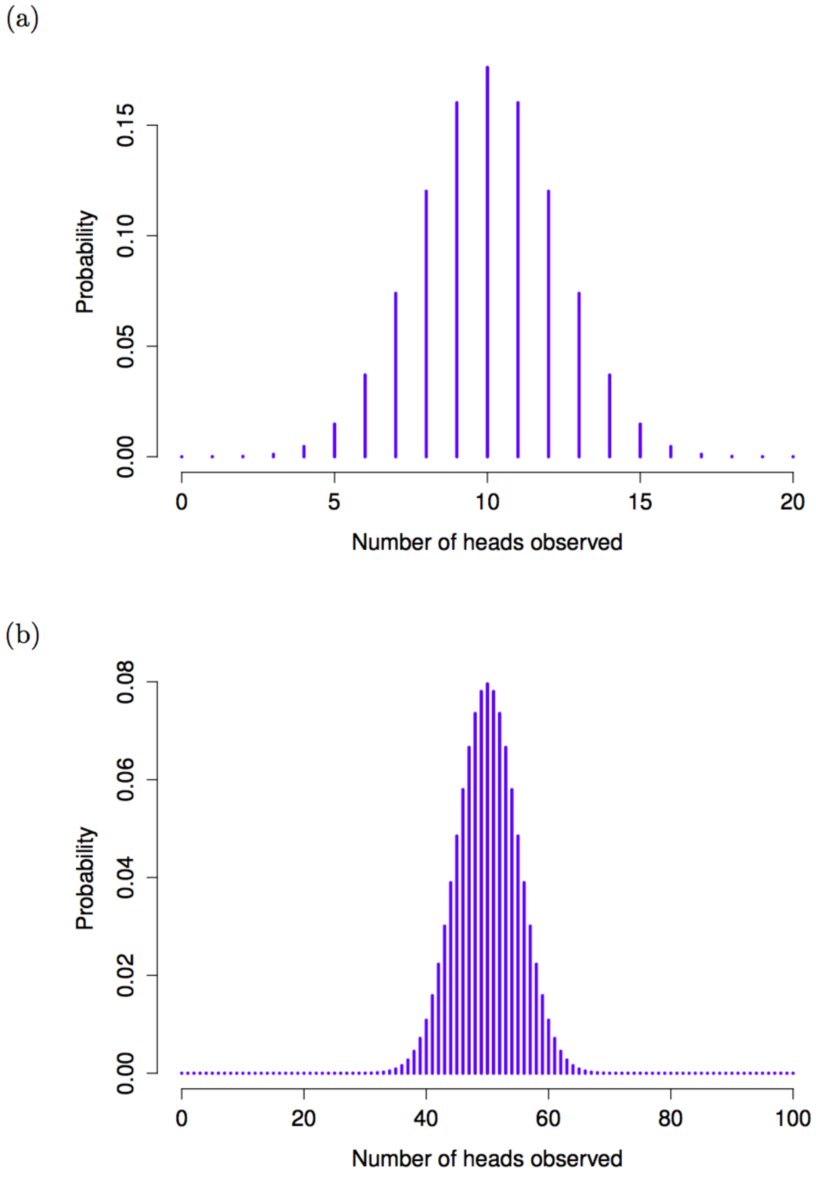


Figure 4.4: Two binomial distributions, involving a scenario in which I'm flipping a fair coin, so the underlying success probability is 1/2. In panel (a), we assume I'm flipping the coin N = 20 times. In panel (b) we assume that the coin is flipped N = 100 times.

- The `r` form is a *random number generator*: specifically, it generates `n` random outcomes from the distribution.

This is a little abstract, so let's look at some concrete examples. Again, we've already covered `dbinom` so let's focus on the other three versions. We'll start with `pbinom`, and we'll go back to the skull-dice example. Again, I'm rolling 20 dice, and each die has a 1 in 6 chance of coming up skulls. Suppose, however, that I want to know the probability of rolling 4 **or fewer** skulls. If I wanted to, I could use the `dbinom` function to calculate the exact probability of rolling 0 skulls, 1 skull, 2 skulls, 3 skulls and 4 skulls and then add these up, but there's a faster way. Instead, I can calculate this using the `pbinom` function. Here's the command:

```
pbinom( q= 4, size = 20, prob = 1/6)
#> [1] 0.7687492
```

In other words, there is a 76.9% chance that I will roll 4 or fewer skulls. Or, to put it another way, R is telling us that a value of 4 is actually the 76.9th percentile of this binomial distribution.

Next, let's consider the `qbinom` function. Let's say I want to calculate the 75th percentile of the binomial distribution. If we're sticking with our skulls example, I would use the following command to do this:

```
qbinom( p = 0.75, size = 20, prob = 1/6 )
#> [1] 4
```

Hm. There's something odd going on here. Let's think this through. What the `qbinom` function appears to be telling us is that the 75th percentile of the binomial distribution is 4, even though we saw from the function that 4 is **actually** the 76.9th percentile. And it's definitely the `pbinom` function that is correct. I promise. The weirdness here comes from the fact that our binomial distribution doesn't really **have** a 75th percentile. Not really. Why not? Well, there's a 56.7% chance of rolling 3 or fewer skulls (you can type `pbinom(3, 20, 1/6)` to confirm this if you want), and a 76.9% chance of rolling 4 or fewer skulls. So there's a sense in which the 75th percentile should lie "in between" 3 and 4 skulls. But that makes no sense at all! You can't roll 20 dice and get 3.9 of them come up skulls. This issue can be handled in different ways: you could report an in between value (or **interpolated** value, to use the technical name) like 3.9, you could round down (to 3) or you could round up (to 4).

The `qbinom` function rounds upwards: if you ask for a percentile that doesn't actually exist (like the 75th in this example), R finds the smallest value for which the the percentile rank is **at least** what you asked for. In this case, since the "true" 75th percentile (whatever that would mean) lies somewhere between 3 and 4 skulls, R Rounds up and gives you an answer of 4. This subtlety is tedious, I admit, but thankfully it's only an issue for discrete distributions like the binomial. The other distributions that I'll talk about (normal, t, χ^2 and F) are all continuous, and so R can always return an exact quantile whenever you ask for it.

Finally, we have the random number generator. To use the `rbinom` function, you specify how many times R should "simulate" the experiment using the `n` argument, and it will generate random outcomes from the binomial distribution. So, for instance, suppose I were to repeat my die rolling experiment 100 times. I could get R to simulate the results of these experiments by using the following command:

```
rbinom( n = 100, size = 20, prob = 1/6 )
#>   [1] 5 2 3 3 5 5 3 2 2 2 3 2 2 3 3 2 3 1 3 2 2 2 3 5 3 5 5 6 3 5 1 1 2 2 3 5 3
#>  [38] 3 6 2 4 6 1 1 4 3 3 4 4 3 4 3 2 2 6 0 3 4 1 1 5 2 3 4 3 3 4 1 3 6 2 4 4 2
#>  [75] 5 3 7 3 3 5 2 4 3 4 2 4 1 4 4 1 1 3 4 1 6 3 6 4 1 1
```

As you can see, these numbers are pretty much what you'd expect given the distribution shown in Figure 4.3 . Most of the time I roll somewhere between 1 to 5 skulls. There are a lot of subtleties associated with random number generation using a computer, but for the purposes of this book we don't need to worry too much about them.

4.5 The normal distribution

While the binomial distribution is conceptually the simplest distribution to understand, it's not the most important one. That particular honor goes to the *normal distribution*, which is also referred to as "the bell curve" or a "Gaussian distribution".

A normal distribution is described using two parameters, the mean of the distribution μ and the standard deviation of the distribution σ. The notation that we sometimes use to say that a variable X is normally distributed is as follows:

$$X \sim \text{Normal}(\mu, \sigma)$$

Of course, that's just notation. It doesn't tell us anything interesting about the normal distribution itself. The mathematical formula for the normal distribution is:

$$\underline{\text{Normal}}$$

$$p(X \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(X-\mu)^2}{2\sigma^2} \right)$$

Figure 4.6: Formula for the normal distribution

The formula is important enough that everyone who learns statistics should at least look at it, but since this is an introductory text I don't want to focus on it to much. Instead, we look

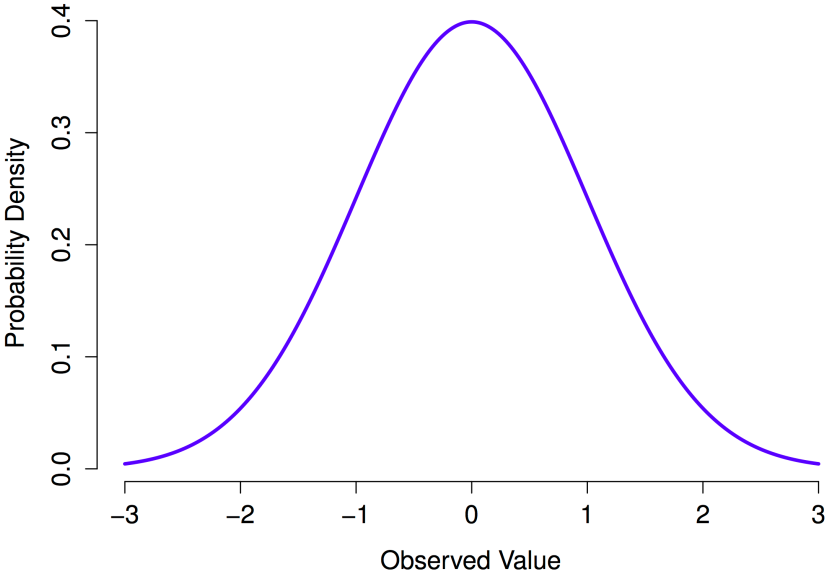


Figure 4.5: The normal distribution with mean = 0 and standard deviation = 1. The x-axis corresponds to the value of some variable, and the y-axis tells us something about how likely we are to observe that value. However, notice that the y-axis is labelled Probability Density and not Probability. There is a subtle and somewhat frustrating characteristic of continuous distributions that makes the y axis behave a bit oddly: the height of the curve here isn't actually the probability of observing a particular x value. On the other hand, it is true that the heights of the curve tells you which x values are more likely (the higher ones!).

at how R can be used to work with normal distributions. The R functions for the normal distribution are *dnorm()*, *pnorm()*, *qnorm()* and *rnorm()*. However, they behave in pretty much exactly the same way as the corresponding functions for the binomial distribution, so there's not a lot that you need to know. The only thing that I should point out is that the argument names for the parameters are *mean* and *sd*. In pretty much every other respect, there's nothing else to add.

Instead of focusing on the maths, let's try to get a sense for what it means for a variable to be normally distributed. To that end, have a look at Figure 4.5, which plots a normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$. You can see where the name "bell curve" comes from: it looks a bit like a bell. Notice that, unlike the plots that I drew to illustrate the binomial distribution, the picture of the normal distribution in Figure Figure 4.5 shows a smooth curve instead of "histogram-like" bars. This isn't an arbitrary choice: the normal distribution is continuous, whereas the binomial is discrete. For instance, in the die rolling example from the last section, it was possible to get 3 skulls or 4 skulls, but impossible to get 3.9 skulls.

With this in mind, let's see if we can get an intuition for how the normal distribution works. First, let's have a look at what happens when we play around with the parameters of the distribution. One parameter we can change is the mean. This will shift the distribution to the right or left. The animation in **?@fig-4normalMeanShift** shows a normal distribution with mean = 0, moving up and down from mean = 0 to mean = 5. Note, when you change the mean the whole shape of the distribution does not change, it just shifts from left to right. In the animation the normal distribution bounces up and down a little, but that's just a quirk of the animation (plus it looks fun that way).

In contrast, if we increase the standard deviation while keeping the mean constant, the peak of the distribution stays in the same place, but the distribution gets wider. The animation in **?@fig-4normalSDShift** shows what happens when you start with a small standard deviation (sd = 0.5), and move to larger and larger standard deviation (up to sd = 5). As you can see, the distribution spreads out and becomes wider as the standard deviation increases.

Notice that when we widen the distribution the height of the peak shrinks. This has to happen: in the same way that the heights of the bars that we used to draw a discrete binomial distribution have to *sum* to 1, the total *area under the curve* for the normal distribution must equal 1. Before moving on, I want to point out one important characteristic of the normal distribution. Irrespective of what the actual mean and standard deviation are, 68.3% of the area falls within 1 standard deviation of the mean. Similarly, 95.4% of the distribution falls within 2 standard deviations of the mean, and 99.7% of the distribution is within 3 standard deviations.

4.5.1 Probability density

There's something I've been trying to hide throughout my discussion of the normal distribution, something that some introductory textbooks omit completely. They might be right to do so: this "thing" that I'm hiding is weird and counter intuitive even by the admittedly distorted standards that apply in statistics. Fortunately, it's not something that you need to understand at a deep level in order to do basic statistics: rather, it's something that starts to become important later on when you move beyond the basics. So, if it doesn't make complete sense, don't worry: try to make sure that you follow the gist of it.

Throughout my discussion of the normal distribution, there's been one or two things that don't quite make sense. Perhaps you noticed that the y-axis in these figures is labelled "Probability Density" rather than density. Maybe you noticed that I used $p(X)$ instead of $P(X)$ when giving the formula for the normal distribution. Maybe you're wondering why R uses the "d" prefix for functions like *dnorm()*. And maybe, just maybe, you've been playing around with the *dnorm()* function, and you accidentally typed in a command like this:

```
dnorm( x = 1, mean = 1, sd = 0.1 )
#> [1] 3.989423
```

And if you've done the last part, you're probably very confused. I've asked R to calculate the probability that *x = 1*, for a normally distributed variable with *mean = 1* and standard deviation *sd = 0.1*; and it tells me that the probability is 3.99. But, as we discussed earlier, probabilities *can't* be larger than 1. So either I've made a mistake, or that's not a probability.

As it turns out, the second answer is correct. What we've calculated here isn't actually a probability: it's something else. To understand what that something is, you have to spend a little time thinking about what it really *means* to say that X is a continuous variable. Let's say we're talking about the temperature outside. The thermometer tells me it's 23 degrees, but I know that's not really true. It's not *exactly* 23 degrees. Maybe it's 23.1 degrees, I think to myself. But I know that that's not really true either, because it might actually be 23.09 degrees. But, I know that… well, you get the idea. The tricky thing with genuinely continuous quantities is that you never really know exactly what they are.

Now think about what this implies when we talk about probabilities. Suppose that tomorrow's maximum temperature is sampled from a normal distribution with mean 23 and standard deviation 1. What's the probability that the temperature will be *exactly* 23 degrees? The answer is "zero", or possibly, "a number so close to zero that it might as well be zero". Why is this?

It's like trying to throw a dart at an infinitely small dart board: no matter how good your aim, you'll never hit it. In real life you'll never get a value of exactly 23. It'll always be something like 23.1 or 22.99998 or something. In other words, it's completely meaningless to talk about the probability that the temperature is exactly 23 degrees. However, in everyday language, if

I told you that it was 23 degrees outside and it turned out to be 22.9998 degrees, you probably wouldn't call me a liar. Because in everyday language, "23 degrees" usually means something like "somewhere between 22.5 and 23.5 degrees". And while it doesn't feel very meaningful to ask about the probability that the temperature is exactly 23 degrees, it does seem sensible to ask about the probability that the temperature lies between 22.5 and 23.5, or between 20 and 30, or any other range of temperatures.

The point of this discussion is to make clear that, when we're talking about continuous distributions, it's not meaningful to talk about the probability of a specific value. However, what we *can* talk about is the **probability that the value lies within a particular range of values**. To find out the probability associated with a particular range, what you need to do is calculate the "area under the curve".

Okay, so that explains part of the story. I've explained a little bit about how continuous probability distributions should be interpreted (i.e., area under the curve is the key thing), but I haven't actually explained what the *dnorm()* function actually calculates. Equivalently, what does the formula for $p(x)$ that I described earlier actually mean? Obviously, $p(x)$ doesn't describe a probability, but what is it? The name for this quantity $p(x)$ is a *probability density*, and in terms of the plots we've been drawing, it corresponds to the *height* of the curve. The densities themselves aren't meaningful in and of themselves: but they're "rigged" to ensure that the *area* under the curve is always interpretable as genuine probabilities. To be honest, that's about as much as you really need to know for now.

4.6 Other useful distributions

There are many other useful distributions, these include the `t` distribution, the `F` distribution, and the chi squared distribution. We will soon discover more about the `t` and `F` distributions when we discuss t-tests and ANOVAs in later chapters.

4.7 Summary of Probability

We've talked what probability means, and why statisticians can't agree on what it means. We talked about the rules that probabilities have to obey. And we introduced the idea of a probability distribution, and spent a good chunk talking about some of the more important probability distributions that statisticians work with. We talked about things like this:

- Probability theory versus statistics
- Frequentist versus Bayesian views of probability
- Basics of probability theory
- Binomial distribution, normal distribution

As you'd expect, this coverage is by no means exhaustive. Probability theory is a large branch of mathematics in its own right, entirely separate from its application to statistics and data analysis. As such, there are thousands of books written on the subject and universities generally offer multiple classes devoted entirely to probability theory. Even the "simpler" task of documenting standard probability distributions is a big topic.Fortunately for you, very little of this is necessary. You're unlikely to need to know dozens of statistical distributions when you go out and do real world data analysis, and you definitely won't need them for this book, but it never hurts to know that there's other possibilities out there.

Picking up on that last point, there's a sense in which this whole chapter is something of a digression. Many statistics classes skim over this content very quickly (I know mine did), and even the more advanced classes will often "forget" to revisit the basic foundations of the field. Many academics would not know the difference between probability and density, and until recently very few would have been aware of the difference between Bayesian and frequentist probability. However, I think it's important to understand these things before moving onto the applications. For example, there are a lot of rules about what you're "allowed" to say when doing statistical inference, and many of these can seem arbitrary and weird. However, they start to make sense if you understand that there is this Bayesian/frequentist distinction.

4.8 Samples, populations and sampling

Remember, the role of descriptive statistics is to concisely summarize what we **do** know. In contrast, the purpose of inferential statistics is to "learn what we do not know from what we do". What kinds of things would we like to learn about? And how do we learn them? These are the questions that lie at the heart of inferential statistics, and they are traditionally divided into two "big ideas": estimation and hypothesis testing. The goal in this chapter is to introduce the first of these big ideas, estimation theory, but we'll talk about sampling theory first because estimation theory doesn't make sense until you understand sampling. So, this chapter divides into sampling theory, and how to make use of sampling theory to discuss how statisticians think about estimation. We have already done lots of sampling, so you are already familiar with some of the big ideas.

Sampling theory plays a huge role in specifying the assumptions upon which your statistical inferences rely. And in order to talk about "making inferences" the way statisticians think about it, we need to be a bit more explicit about what it is that we're drawing inferences **from** (the sample) and what it is that we're drawing inferences **about** (the population).

In almost every situation of interest, what we have available to us as researchers is a **sample** of data. We might have run experiment with some number of participants; a polling company might have phoned some number of people to ask questions about voting intentions; etc. Regardless: the data set available to us is finite, and incomplete. We can't possibly get every person in the world to do our experiment; a polling company doesn't have the time or the money to ring up every voter in the country etc. In our earlier discussion of descriptive statistics, this sample was the only thing we were interested in. Our only goal was to find ways of describing, summarizing and graphing that sample. This is about to change.

4.8.1 Defining a population

A sample is a concrete thing. You can open up a data file, and there's the data from your sample. A **population**, on the other hand, is a more abstract idea. It refers to the set of all possible people, or all possible observations, that you want to draw conclusions about, and is generally **much** bigger than the sample. In an ideal world, the researcher would begin the study with a clear idea of what the population of interest is, since the process of designing a study and testing hypotheses about the data that it produces does depend on the population about which you want to make statements. However, that doesn't always happen in practice: usually the researcher has a fairly vague idea of what the population is and designs the study as best he/she can on that basis.

Sometimes it's easy to state the population of interest. For instance, in the "polling company" example, the population consisted of all voters enrolled at the a time of the study – millions of people. The sample was a set of 1000 people who all belong to that population. In most situations the situation is much less simple. In a typical a psychological experiment, determining the population of interest is a bit more complicated. Suppose I run an experiment using 100 undergraduate students as my participants. My goal, as a cognitive scientist, is to try to learn something about how the mind works. So, which of the following would count as "the population":

- All of the undergraduate psychology students at the University of Adelaide?
- Undergraduate psychology students in general, anywhere in the world?
- Australians currently living?
- Australians of similar ages to my sample?
- Anyone currently alive?
- Any human being, past, present or future?
- Any biological organism with a sufficient degree of intelligence operating in a terrestrial environment?
- Any intelligent being?

Each of these defines a real group of mind-possessing entities, all of which might be of interest to me as a cognitive scientist, and it's not at all clear which one ought to be the true population of interest.

4.8.2 Simple random samples

Irrespective of how we define the population, the critical point is that the sample is a subset of the population, and our goal is to use our knowledge of the sample to draw inferences about the properties of the population. The relationship between the two depends on the **procedure** by which the sample was selected. This procedure is referred to as a **sampling method**, and it is important to understand why it matters.

To keep things simple, imagine we have a bag containing 10 chips. Each chip has a unique letter printed on it, so we can distinguish between the 10 chips. The chips come in two colors, black and white.

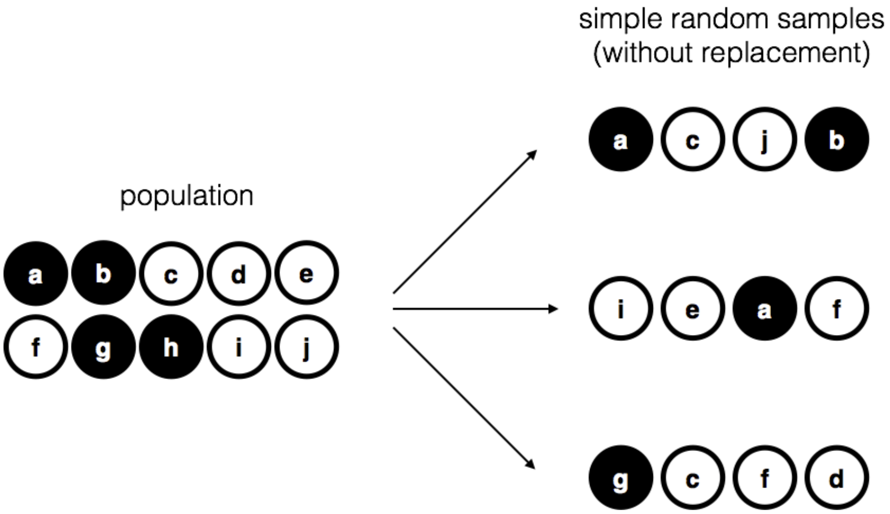


Figure 4.7: Simple random sampling without replacement from a finite population

This set of chips is the population of interest, and it is depicted graphically on the left of Figure 4.7.

As you can see from looking at the picture, there are 4 black chips and 6 white chips, but of course in real life we wouldn't know that unless we looked in the bag. Now imagine you run the following "experiment": you shake up the bag, close your eyes, and pull out 4 chips without putting any of them back into the bag. First out comes the *a* chip (black), then the *c* chip (white), then *j* (white) and then finally *b* (black). If you wanted, you could then put all the chips back in the bag and repeat the experiment, as depicted on the right hand side of Figure 4.7. Each time you get different results, but the procedure is identical in each case. The fact that the same procedure can lead to different results each time, we refer to it as a **random** process. However, because we shook the bag before pulling any chips out, it seems reasonable to think that every chip has the same chance of being selected. A procedure in which every member of the population has the same chance of being selected is called a **simple random sample**. The fact that we did **not** put the chips back in the bag after pulling them out means that you can't observe the same thing twice, and in such cases the observations are said to have been sampled **without replacement**.

To help make sure you understand the importance of the sampling procedure, consider an alternative way in which the experiment could have been run. Suppose that my 5-year old son had opened the bag, and decided to pull out four black chips without putting any of them back in the bag. This **biased** sampling scheme is depicted in Figure 4.8.

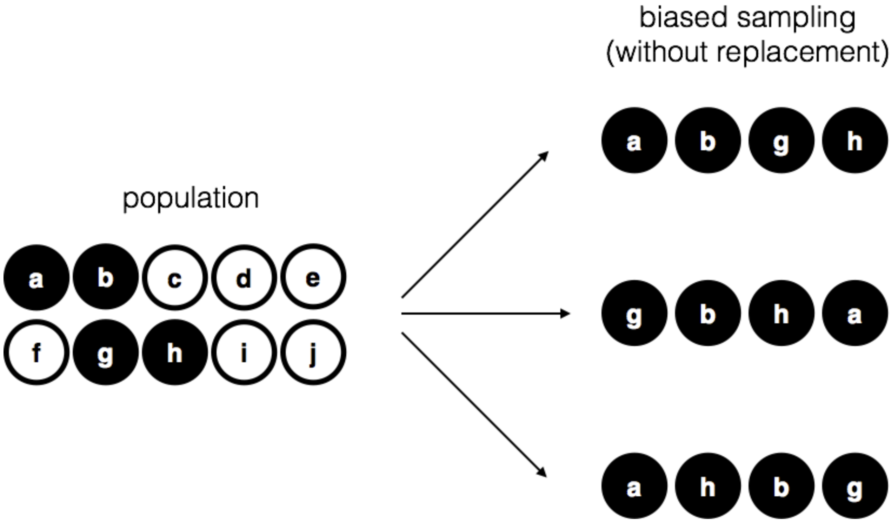


Figure 4.8: Biased sampling without replacement from a finite populations.

Now consider the evidentiary value of seeing 4 black chips and 0 white chips. Clearly, it depends on the sampling scheme, does it not? If you know that the sampling scheme is biased to select only black chips, then a sample that consists of only black chips doesn't tell you very much about the population! For this reason, statisticians really like it when a data set can be considered a simple random sample, because it makes the data analysis **much** easier.

A third procedure is worth mentioning. This time around we close our eyes, shake the bag, and pull out a chip. This time, however, we record the observation and then put the chip back in the bag. Again we close our eyes, shake the bag, and pull out a chip. We then repeat this procedure until we have 4 chips. Data sets generated in this way are still simple random samples, but because we put the chips back in the bag immediately after drawing them it is referred to as a sample **with replacement**. The difference between this situation and the first one is that it is possible to observe the same population member multiple times, as illustrated in Figure 4.9.

Most psychology experiments tend to be sampling without replacement, because the same person is not allowed to participate in the experiment twice. However, most statistical theory is based on the assumption that the data arise from a simple random sample **with** replacement. In real life, this very rarely matters. If the population of interest is large (e.g., has more than

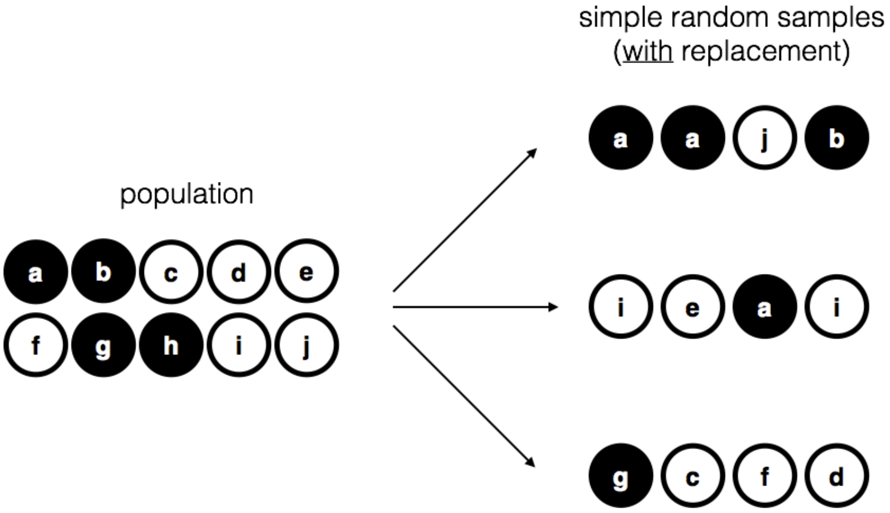


Figure 4.9: Simple random sampling with replacement from a finite population.

10 entities!) the difference between sampling with- and without- replacement is too small to be concerned with. The difference between simple random samples and biased samples, on the other hand, is not such an easy thing to dismiss.

4.8.3 Most samples are not simple random samples

As you can see from looking at the list of possible populations that I showed above, it is almost impossible to obtain a simple random sample from most populations of interest. When I run experiments, I'd consider it a minor miracle if my participants turned out to be a random sampling of the undergraduate psychology students at Adelaide university, even though this is by far the narrowest population that I might want to generalize to. A thorough discussion of other types of sampling schemes is beyond the scope of this book, but to give you a sense of what's out there I'll list a few of the more important ones:

- **Stratified sampling**. Suppose your population is (or can be) divided into several different sub-populations, or **strata**. Perhaps you're running a study at several different sites, for example. Instead of trying to sample randomly from the population as a whole, you instead try to collect a separate random sample from each of the strata. Stratified sampling is sometimes easier to do than simple random sampling, especially when the population is already divided into the distinct strata. It can also be more efficient that simple random sampling, especially when some of the sub-populations are rare. For instance, when studying schizophrenia it would be much better to divide the population into two strata (schizophrenic and not-schizophrenic), and then sample an equal number of people from each group. If you selected people randomly, you would get so few schizophrenic people in the sample that your study would be useless. This specific kind of of stratified sampling is referred to as **oversampling** because it makes a deliberate attempt to over-represent rare groups.

- **Snowball sampling** is a technique that is especially useful when sampling from a "hidden" or hard to access population, and is especially common in social sciences. For instance, suppose the researchers want to conduct an opinion poll among transgender people. The research team might only have contact details for a few trans folks, so the survey starts by asking them to participate (stage 1). At the end of the survey, the participants are asked to provide contact details for other people who might want to participate. In stage 2, those new contacts are surveyed. The process continues until the researchers have sufficient data. The big advantage to snowball sampling is that it gets you data in situations that might otherwise be impossible to get any. On the statistical side, the main disadvantage is that the sample is highly non-random, and non-random in ways that are difficult to address. On the real life side, the disadvantage is that the procedure can be unethical if not handled well, because hidden populations are often hidden for a reason. I chose transgender people as an example here to highlight this: if you weren't careful you might end up outing people who don't want to be outed (very, very bad form), and even if you don't make that mistake it can still be intrusive to use people's social networks to study them. It's certainly very hard to get people's informed consent **before** contacting them, yet in many cases the simple act of contacting them and saying "hey we want to study you" can be hurtful. Social networks are complex things, and just because you can use them to get data doesn't always mean you should.
- **Convenience sampling** is more or less what it sounds like. The samples are chosen in a way that is convenient to the researcher, and not selected at random from the population of interest. Snowball sampling is one type of convenience sampling, but there are many others. A common example in psychology are studies that rely on undergraduate psychology students. These samples are generally non-random in two respects: firstly, reliance on undergraduate psychology students automatically means that your data are restricted to a single sub-population. Secondly, the students usually get to pick which studies they participate in, so the sample is a self selected subset of psychology students not a randomly selected subset. In real life, most studies are convenience samples of one form or another. This is sometimes a severe limitation, but not always.

4.8.4 How much does it matter if you don't have a simple random sample?

Okay, so real world data collection tends not to involve nice simple random samples. Does that matter? A little thought should make it clear to you that it **can** matter if your data are not a simple random sample: just think about the difference between Figure 4.7 and Figure 4.8. However, it's not quite as bad as it sounds. Some types of biased samples are entirely unproblematic. For instance, when using a stratified sampling technique you actually **know** what the bias is because you created it deliberately, often to **increase** the effectiveness of your study, and there are statistical techniques that you can use to adjust for the biases you've introduced (not covered in this book!). So in those situations it's not a problem.

More generally though, it's important to remember that random sampling is a means to an end, not the end in itself. Let's assume you've relied on a convenience sample, and as such you can assume it's biased. A bias in your sampling method is only a problem if it causes you to draw the wrong conclusions. When viewed from that perspective, I'd argue that we don't need the sample to be randomly generated in **every** respect: we only need it to be random with respect to the psychologically-relevant phenomenon of interest. Suppose I'm doing a study looking at working memory capacity. In study 1, I actually have the ability to sample randomly from all human beings currently alive, with one exception: I can only sample people born on a Monday. In study 2, I am able to sample randomly from the Australian population. I want to generalize my results to the population of all living humans. Which study is better? The answer, obviously, is study 1. Why? Because we have no reason to think that being "born on a Monday" has any interesting relationship to working memory capacity. In contrast, I can think of several reasons why "being Australian" might matter. Australia is a wealthy, industrialized country with a very well-developed education system. People growing up in that system will have had life experiences much more similar to the experiences of the people who designed the tests for working memory capacity. This shared experience might easily translate into similar beliefs about how to "take a test", a shared assumption about how psychological experimentation works, and so on. These things might actually matter. For instance, "test taking" style might have taught the Australian participants how to direct their attention exclusively on fairly abstract test materials relative to people that haven't grown up in a similar environment; leading to a misleading picture of what working memory capacity is.

There are two points hidden in this discussion. Firstly, when designing your own studies, it's important to think about what population you care about, and try hard to sample in a way that is appropriate to that population. In practice, you're usually forced to put up with a "sample of convenience" (e.g., psychology lecturers sample psychology students because that's the least expensive way to collect data, and our coffers aren't exactly overflowing with gold), but if so you should at least spend some time thinking about what the dangers of this practice might be.

Secondly, if you're going to criticize someone else's study because they've used a sample of convenience rather than laboriously sampling randomly from the entire human population, at least have the courtesy to offer a specific theory as to **how** this might have distorted the results. Remember, everyone in science is aware of this issue, and does what they can to alleviate it. Merely pointing out that "the study only included people from group BLAH" is entirely unhelpful, and borders on being insulting to the researchers, who are aware of the issue. They just don't happen to be in possession of the infinite supply of time and money required to construct the perfect sample. In short, if you want to offer a responsible critique of the sampling process, then be **helpful**. Rehashing the blindingly obvious truisms that I've been rambling on about in this section isn't helpful.

4.8.5 Population parameters and sample statistics

Okay. Setting aside the thorny methodological issues associated with obtaining a random sample, let's consider a slightly different issue. Up to this point we have been talking about populations the way a scientist might. To a psychologist, a population might be a group of people. To an ecologist, a population might be a group of bears. In most cases the populations that scientists care about are concrete things that actually exist in the real world.

Statisticians, however, are a funny lot. On the one hand, they **are** interested in real world data and real science in the same way that scientists are. On the other hand, they also operate in the realm of pure abstraction in the way that mathematicians do. As a consequence, statistical theory tends to be a bit abstract in how a population is defined. In much the same way that psychological researchers operationalize our abstract theoretical ideas in terms of concrete measurements, statisticians operationalize the concept of a "population" in terms of mathematical objects that they know how to work with. You've already come across these objects they're called probability distributions (remember, the place where data comes from).

The idea is quite simple. Let's say we're talking about IQ scores. To a psychologist, the population of interest is a group of actual humans who have IQ scores. A statistician "simplifies" this by operationally defining the population as the probability distribution depicted in Figure 4.10 *a*.

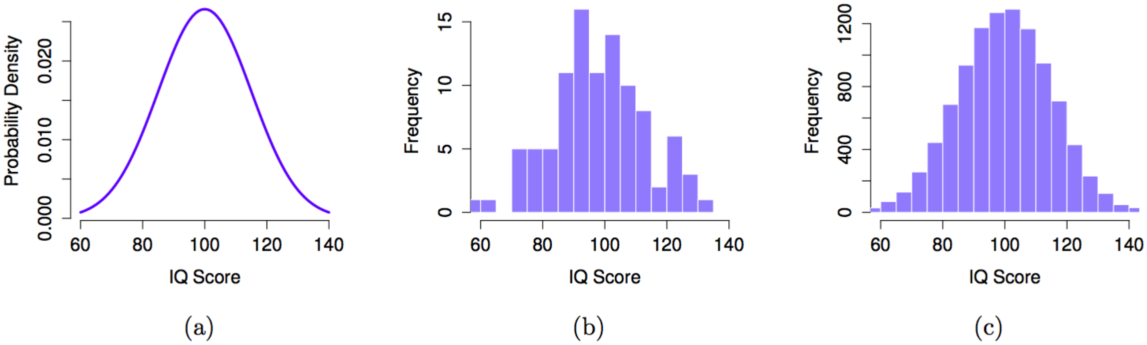


Figure 4.10: The population distribution of IQ scores (panel a) and two samples drawn randomly from it. In panel b we have a sample of 100 observations, and panel c we have a sample of 10,000 observations.

IQ tests are designed so that the average IQ is 100, the standard deviation of IQ scores is 15, and the distribution of IQ scores is normal. These values are referred to as the **population parameters** because they are characteristics of the entire population. That is, we say that the population mean μ is 100, and the population standard deviation σ is 15.

Now suppose we collect some data. We select 100 people at random and administer an IQ test, giving a simple random sample from the population. The sample would consist of a collection of numbers like this:

```
106 101 98 80 74 ... 107 72 100
```

Each of these IQ scores is sampled from a normal distribution with mean 100 and standard deviation 15. So if I plot a histogram of the sample, I get something like the one shown in Figure 4.10 *b*. As you can see, the histogram is **roughly** the right shape, but it's a very crude approximation to the true population distribution shown in Figure 4.10 *a*. The mean of the sample is fairly close to the population mean 100 but not identical. In this case, it turns out that the people in the sample have a mean IQ of 98.5, and the standard deviation of their IQ scores is 15.9. These **sample statistics** are properties of the data set, and although they are fairly similar to the true population values, they are not the same. **In general, sample statistics are the things you can calculate from your data set, and the population parameters are the things you want to learn about.** Later on in this chapter we'll talk about how you can estimate population parameters using your sample statistics and how to work out how confident you are in your estimates but before we get to that there's a few more ideas in sampling theory that you need to know about.

4.9 The law of large numbers

We just looked at the results of one fictitious IQ experiment with a sample size of $N = 100$. The results were somewhat encouraging: the true population mean is 100, and the sample mean of 98.5 is a pretty reasonable approximation to it. In many scientific studies that level of precision is perfectly acceptable, but in other situations you need to be a lot more precise. If we want our sample statistics to be much closer to the population parameters, what can we do about it?

The obvious answer is to collect more data. Suppose that we ran a much larger experiment, this time measuring the IQ's of 10,000 people. We can simulate the results of this experiment using R, using the **rnorm()** function, which generates random numbers sampled from a normal distribution. For an experiment with a sample size of **n = 10000**, and a population with **mean = 100** and **sd = 15**, R produces our fake IQ data using these commands:

```r
IQ <- rnorm(n=10000, mean=100, sd=15) #generate IQ scores
IQ <- round(IQ) # make round numbers
```

Cool, we just generated 10,000 fake IQ scores. Where did they go? Well, they went into the variable IQ on my computer. You can do the same on your computer too by copying the above code. 10,000 numbers is too many numbers to look at. We can look at the first 100 like this:

```
print(IQ[1:100])
#>   [1]  88 106 111  92 106  88  96 106  96 109 115 109  83 118  98  90 101 118
#>  [19] 112 101 109  97 103  89  63  85  95 106  98 105 104  87  97 107  98  80
#>  [37]  96  93  98 112 102 109 107  91 103  97 110  85 102  78 110 122  76  96
#>  [55]  92  95 114 101  92  82  86  89  89  77  98 115 100 117 118 104 117  93
#>  [73] 107 109  84  95  81 114  87  71 119  95 113 120 112  87 105 100 102 128
#>  [91]  92  73  86  97 114  84 123  89  94  97
```

We can compute the mean IQ using the command **mean(IQ)** and the standard deviation using the command **sd(IQ)**, and draw a histogram using **hist()**. The histogram of this much larger sample is shown in Figure @ref(fig:IQdist)c. Even a moment's inspections makes clear that the larger sample is a much better approximation to the true population distribution than the smaller one. This is reflected in the sample statistics: the mean IQ for the larger sample turns out to be 99.9, and the standard deviation is 15.1. These values are now very close to the true population.

I feel a bit silly saying this, but the thing I want you to take away from this is that large samples generally give you better information. I feel silly saying it because it's so bloody obvious that it shouldn't need to be said. In fact, it's such an obvious point that when Jacob Bernoulli – one of the founders of probability theory – formalized this idea back in 1713, he was kind of a jerk about it. Here's how he described the fact that we all share this intuition:

> **For even the most stupid of men, by some instinct of nature, by himself and without any instruction (which is a remarkable thing), is convinced that the more observations have been made, the less danger there is of wandering from one's goal** (see Stigler, 1986, p65).

Okay, so the passage comes across as a bit condescending (not to mention sexist), but his main point is correct: it really does feel obvious that more data will give you better answers. The question is, why is this so? Not surprisingly, this intuition that we all share turns out to be correct, and statisticians refer to it as the **law of large numbers**. The law of large numbers is a mathematical law that applies to many different sample statistics, but the simplest way to think about it is as a law about averages. The sample mean is the most obvious example of a statistic that relies on averaging (because that's what the mean is… an average), so let's look at that. **When applied to the sample mean, what the law of large numbers states is that as the sample gets larger, the sample mean tends to get closer to the true population mean.** Or, to say it a little bit more precisely, as the sample size "approaches" infinity (written as $N \rightarrow \infty$) the sample mean approaches the population mean ($\bar{X} \rightarrow \mu$).

I don't intend to subject you to a proof that the law of large numbers is true, but it's one of the most important tools for statistical theory. The law of large numbers is the thing we can use to justify our belief that collecting more and more data will eventually lead us to the truth. For any particular data set, the sample statistics that we calculate from it will be wrong, but the law of large numbers tells us that if we keep collecting more data those sample statistics will tend to get closer and closer to the true population parameters.

4.10 Sampling distributions and the central limit theorem

The law of large numbers is a very powerful tool, but it's not going to be good enough to answer all our questions. Among other things, all it gives us is a "long run guarantee". In the long run, if we were somehow able to collect an infinite amount of data, then the law of large numbers guarantees that our sample statistics will be correct. But as John Maynard Keynes famously argued in economics, a long run guarantee is of little use in real life:

> [**The] long run is a misleading guide to current affairs. In the long run we are all dead. Economists set themselves too easy, too useless a task, if in tempestuous seasons they can only tell us, that when the storm is long past, the ocean is flat again.** Keynes (1923, 80)

As in economics, so too in psychology and statistics. It is not enough to know that we will **eventually** arrive at the right answer when calculating the sample mean. Knowing that an infinitely large data set will tell me the exact value of the population mean is cold comfort when my **actual** data set has a sample size of $N = 100$. In real life, then, we must know something about the behavior of the sample mean when it is calculated from a more modest data set!

4.10.1 Sampling distribution of the sample means

"Oh no, what is the sample distribution of the sample means? Is that even allowed in English?". Yes, unfortunately, this is allowed. The **sampling distribution of the sample means** is the next most important thing you will need to understand. IT IS SO IMPORTANT THAT IT IS NECESSARY TO USE ALL CAPS. It is only confusing at first because it's long and uses sampling and sample in the same phrase.

Don't worry, we've been prepping you for this. You know what a distribution is right? It's where numbers comes from. It makes some numbers occur more or less frequently, or the same as other numbers. You know what a sample is right? It's the numbers we take from a distribution. So, what could the sampling distribution of the sample means refer to?

First, what do you think the sample means refers to? Well, if you took a sample of numbers, you would have a bunch of numbers…then, you could compute the mean of those numbers. The sample mean is the mean of the numbers in the sample. That is all. So, what is this distribution you speak of? Well, what if you took a bunch of samples, put one here, put one there, put some other ones other places. You have a lot of different samples of numbers. You could compute the mean for each them. Then you would have a bunch of means. What do those means look like? Well, if you put them in a histogram, you could find out. If you did that, you would be looking at (roughly) a distribution, AKA **the sampling distribution of the sample means**.

“I’m following along sort of, why would I want to do this instead of watching Netflix…”. Because, the sampling distribution of the sample means gives you another window into chance. A very useful one that you can control, just like your remote control, by pressing the right design buttons.

4.10.2 Seeing the pieces

To make a sampling distribution of the sample means, we just need the following:

1. A distribution to take numbers from
2. A bunch of different samples from the distribution
3. The means of each of the samples
4. Get all of the sample means, and plot them in a histogram

Question for yourself: What do you think the sampling distribution of the sample means will look like? Will it tend to look the shape of the distribution that the samples came from? Or not? Good question, think about it.

Let’s do those four things. We will sample numbers from the uniform distribution. Figure 4.11 shows the uniform distribution for sampling the set of integers from 1 to 10:

?@fig-4sample20unif animates the process of taking a bunch of samples from the uniform distribution. We will set our sample-size to 20. It’s easier to see how the sample mean behaves in a movie. Each histogram shows a new sample. The red line shows where the mean of the sample is. The samples are all very different from each other, but the red line doesn’t move around very much, it always stays near the middle. However, the red line does move around a little bit, and this variance is what we call the sampling distribution of the sample mean.

OK, what have we got here? We have an animation of 10 different samples. Each sample has 20 observations and these are summarized in each of histograms that show up in the animation. Each histogram has a red line. The red line shows you where the mean of each sample is located. So, we have found the sample means for the 10 different samples from a uniform distribution.

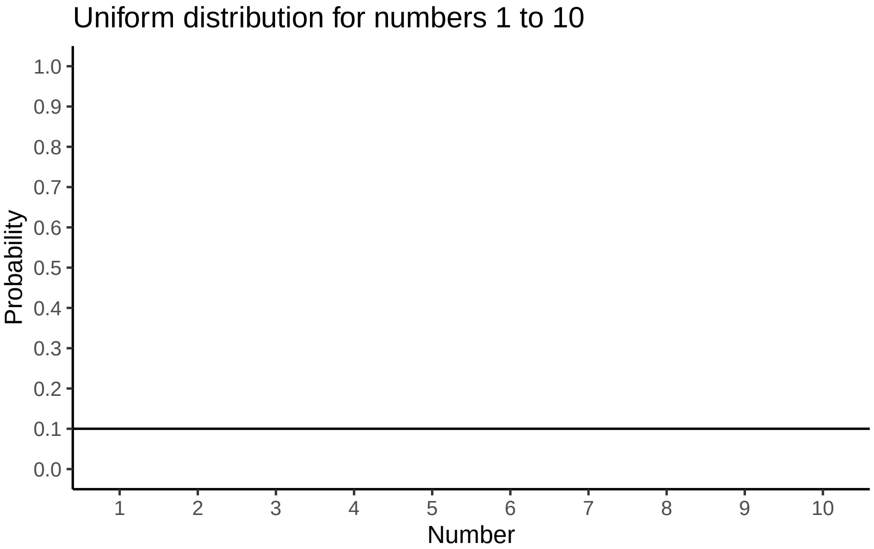


Figure 4.11: A uniform distribution illustrating the probabilites of sampling the numbers 1 to 10. In a uniform distribution, all numbers have an equal probability of being sampled, so the line is flat indicating all numbers have the same probability

First question. Are the sample means all the same? The answer is no. They are all kind of similar to each other though, they are all around five plus or minus a few numbers. This is interesting. Although all of our samples look pretty different from one another, the means of our samples look more similar than different.

Second question. What should we do with the means of our samples? Well, how about we collect them them all, and then plot a histogram of them. This would allow us to see what the distribution of the sample means looks like. The next histogram is just this. Except, rather than taking 10 samples, we will take 10,000 samples. For each of them we will compute the means. So, we will have 10,000 means. Figure 4.12 shows the histogram of the sample means:

"Wait what? This doesn't look right. I thought we were taking samples from a uniform distribution. Uniform distributions are flat. THIS DOES NOT LOOK LIKE A FLAT DISTRIBTUION, WHAT IS GOING ON, AAAAAGGGHH.". We feel your pain.

Remember, we are looking at the distribution of sample means. It is indeed true that the distribution of sample means does not look the same as the distribution we took the samples from. Our distribution of sample means goes up and down. In fact, this will almost always be the case for distributions of sample means. This fact is called the **central limit theorem**, which we talk about later.

For now, let's talk about about what's happening. Remember, we have been sampling numbers between the range 1 to 10. We are supposed to get each number with roughly equal frequency,

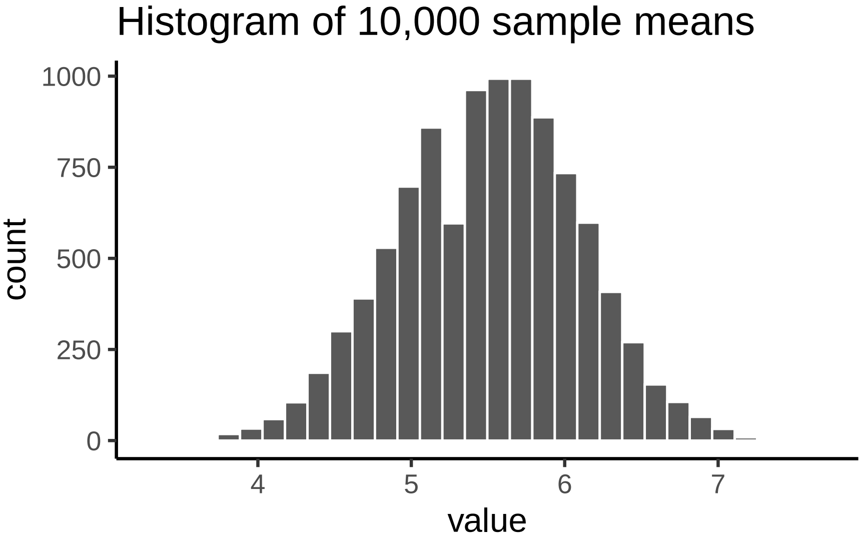


Figure 4.12: A histogram showing the sample means for 10,000 samples, each size 20, from the uniform distribution of numbers from 1 to 10. The expected mean is 5.5, and the histogram is centered on 5.5. The mean of each sample is not always 5.5 because of sampling error or chance

because we are sampling from a uniform distribution. So, let's say we took a sample of 10 numbers, and happened to get one of each from 1 to 10.

```
1 2 3 4 5 6 7 8 9 10
```

What is the mean of those numbers? Well, its 1+2+3+4+5+6+7+8+9+10 = 55 / 10 = 5.5. Imagine if we took a bigger sample, say of 20 numbers, and again we got exactly 2 of each number. What would the mean be? It would be (1+2+3+4+5+6+7+8+9+10)*2 = 110 / 20 = 5.5. Still 5.5. You can see here, that the mean value of our uniform distribution is 5.5. Now that we know this, we might expect that most of our samples will have a mean near this number. We already know that every sample won't be perfect, and it won't have exactly an equal amount of every number. So, we will expect the mean of our samples to vary a little bit. The histogram that we made shows the variation. Not surprisingly, the numbers vary around the value 5.5.

4.10.3 Sampling distributions exist for any sample statistic!

One thing to keep in mind when thinking about sampling distributions is that **any** sample statistic you might care to calculate has a sampling distribution. For example, suppose that each time you sampled some numbers from an experiment you wrote down the largest number in the experiment. Doing this over and over again would give you a very different sampling distribution, namely the **sampling distribution of the maximum**. You could calculate the smallest number, or the mode, or the median, of the variance, or the standard deviation, or anything else from your sample. Then, you could repeat many times, and produce the sampling distribution of those statistics. Neat!

Just for fun here are some different sampling distributions for different statistics. We will take a normal distribution with mean = 100, and standard deviation =20. Then, we'll take lots of samples with n = 50 (50 observations per sample). We'll save all of the sample statistics, then plot their histograms in Figure 4.13. Let's do it:

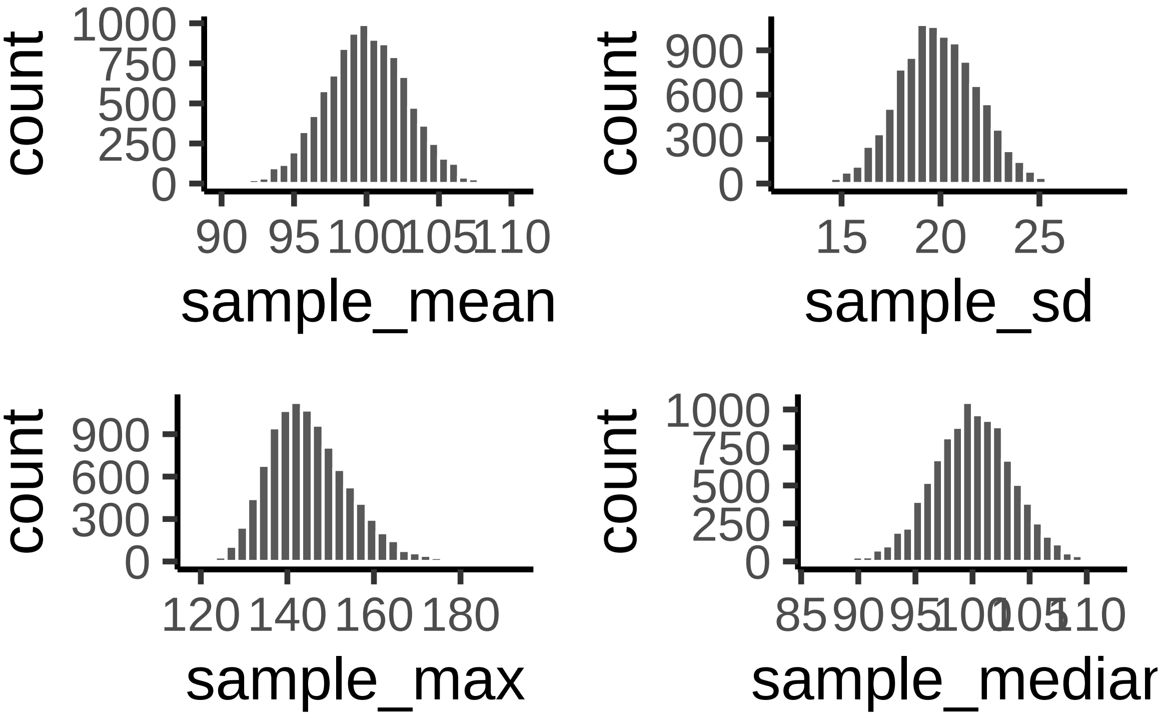


Figure 4.13: Each panel shows a histogram of a different sampling statistic

We just computed 4 different sampling distributions, for the mean, standard deviation, maximum value, and the median. If you just look quickly at these histograms you might think they all basically look the same. Hold up now. It's very important to look at the x-axes. They are different. For example, the sample mean goes from about 90 to 110, whereas the standard deviation goes from 15 to 25.

These sampling distributions are super important, and worth thinking about. What should you think about? Well, here's a clue. These distributions are telling you what to expect from your sample. Critically, they are telling you what you should expect from a sample, when you take one from the specific distribution that we used (normal distribution with mean =100 and

SD = 20). What have we learned. We've learned a tonne. We've learned that we can expect our sample to have a mean somewhere between 90 and 108ish. Notice, the sample means are never more extreme. We've learned that our sample will usually have some variance, and that the the standard deviation will be somewhere between 15 and 25 (never much more extreme than that). We can see that sometime we get some big numbers, say between 120 and 180, but not much bigger than that. And, we can see that the median is pretty similar to the mean. If you ever took a sample of 50 numbers, and your descriptive statistics were inside these windows, then perhaps they came from this kind of normal distribution. If your sample statistics are very different, then your sample probably did not come this distribution. By using simulation, we can find out what samples look like when they come from distributions, and we can use this information to make inferences about whether our sample came from particular distributions.

4.11 The central limit theorem

OK, so now you've seen lots of sampling distributions, and you know what the sampling distribution of the mean is. Here, we'll focus on **how the sampling distribution of the mean changes as a function of sample size.**

Intuitively, you already know part of the answer: if you only have a few observations, the sample mean is likely to be quite inaccurate (you've already seen it bounce around): if you replicate a small experiment and recalculate the mean you'll get a very different answer. In other words, the sampling distribution is quite wide. If you replicate a large experiment and recalculate the sample mean you'll probably get the same answer you got last time, so the sampling distribution will be very narrow.

Let's give ourselves a nice movie to see everything in action. We're going to sample numbers from a normal distribution. **?@fig-4samplingmean** has four panels, each panel represents a different sample size (n), including sample-sizes of 10, 50, 100, and 1000. The red line shows the shape of the normal distribution. The grey bars show a histogram of each of the samples that we take. The red line shows the mean of an individual sample (the middle of the grey bars). As you can see, the red line moves around a lot, especially when the sample size is small (10).

The new bits are the blue bars and the blue lines. The blue bars represent the sampling distribution of the sample mean. For example, in the panel for sample-size 10, we see a bunch of blue bars. This is a histogram of 10 sample means, taken from 10 samples of size 10. In the 50 panel, we see a histogram of 50 sample means, taken from 50 samples of size 50, and so on. The blue line in each panel is the mean of the sample means ("aaagh, it's a mean of means", yes it is).

What should you notice? Notice that the range of the blue bars shrinks as sample size increases. The sampling distribution of the mean is quite wide when the sample-size is 10, it narrows as sample-size increases to 50 and 100, and it's just one bar, right in the middle when sample-size goes to 1000. What we are seeing is that the mean of the sampling distribution approaches the mean of the population as sample-size increases.

So, the sampling distribution of the mean is another distribution, and it has some variance. It varies more when sample-size is small, and varies less when sample-size is large. We can quantify this effect by calculating the standard deviation of the sampling distribution, which is referred to as the **standard error**. The standard error of a statistic is often denoted SE, and since we're usually interested in the standard error of the sample **mean**, we often use the acronym SEM. As you can see just by looking at the movie, as the sample size N increases, the SEM decreases.

Okay, so that's one part of the story. However, there's something we've been glossing over a little bit. We've seen it already, but it's worth looking at it one more time. Here's the thing: **no matter what shape your population distribution is**, as N increases the sampling distribution of the mean starts to look more like a normal distribution. This is the central limit theorem.

To see the central limit theorem in action, we are going to look at some histograms of sample means from different kinds of distributions. It is very important to recognize that you are looking at distributions of sample means, not distributions of individual samples.

Here we go, Figure 4.14 shows sampling from a normal distribution. The red line is the normal distribution where each sample is drawn from. The mean for each sample of numbers is computed, and the distribution of sample means is shown by the blue bars. Note that the shape of red line and the blue bars are similar, they both look like a normal distribution.

Let's do it again. This time we will sample from a flat uniform distribution shown by the red line. However, Figure 4.15 shows the distribution of sample means represented by the blue bars is not flat, it looks like a normal distribution.

One more time with an exponential distribution (shown in red) where smaller numbers are more likely to be sampled than larger numbers. Even though way more of the numbers in a given sample will be smaller than larger, according to Figure 4.16 the sampling distribution of the mean does not look the red line. Instead, the sampling distribution of the mean looks like a bell-shaped normal curve. This is the central limit theorem in action.

On the basis of these figures, it seems like we have evidence for all of the following claims about the sampling distribution of the mean:

- The mean of the sampling distribution is the same as the mean of the population
- The standard deviation of the sampling distribution (i.e., the standard error) gets smaller as the sample size increases
- The shape of the sampling distribution becomes normal as the sample size increases

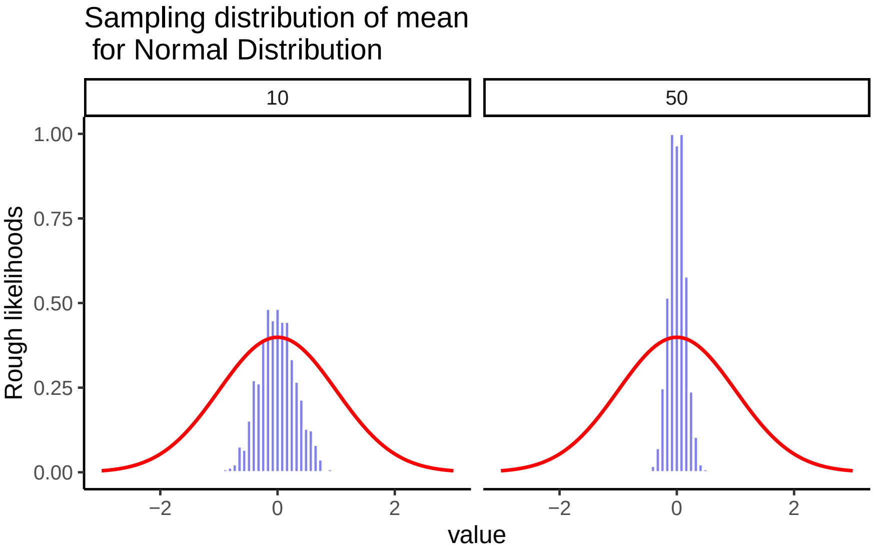


Figure 4.14: Comparison of two normal distributions, and histograms for the sampling distribution of the mean for different samples-sizes. The range of sampling distribution of the mean shrinks as sample-size increases

As it happens, not only are all of these statements true, there is a very famous theorem in statistics that proves all three of them, known as the **central limit theorem**. Among other things, the central limit theorem tells us that if the population distribution has mean μ and standard deviation σ, then the sampling distribution of the mean also has mean μ, and the standard error of the mean is

$$\text{SEM} = \frac{\sigma}{\sqrt{N}}$$

Because we divide the population standard deviation σ by the square root of the sample size N, the SEM gets smaller as the sample size increases. It also tells us that the shape of the sampling distribution becomes normal.

This result is useful for all sorts of things. It tells us why large experiments are more reliable than small ones, and because it gives us an explicit formula for the standard error it tells us **how much** more reliable a large experiment is. It tells us why the normal distribution is, well, **normal**. In real experiments, many of the things that we want to measure are actually averages of lots of different quantities (e.g., arguably, "general" intelligence as measured by IQ is an average of a large number of "specific" skills and abilities), and when that happens, the averaged quantity should follow a normal distribution. Because of this mathematical law, the normal distribution pops up over and over again in real data.

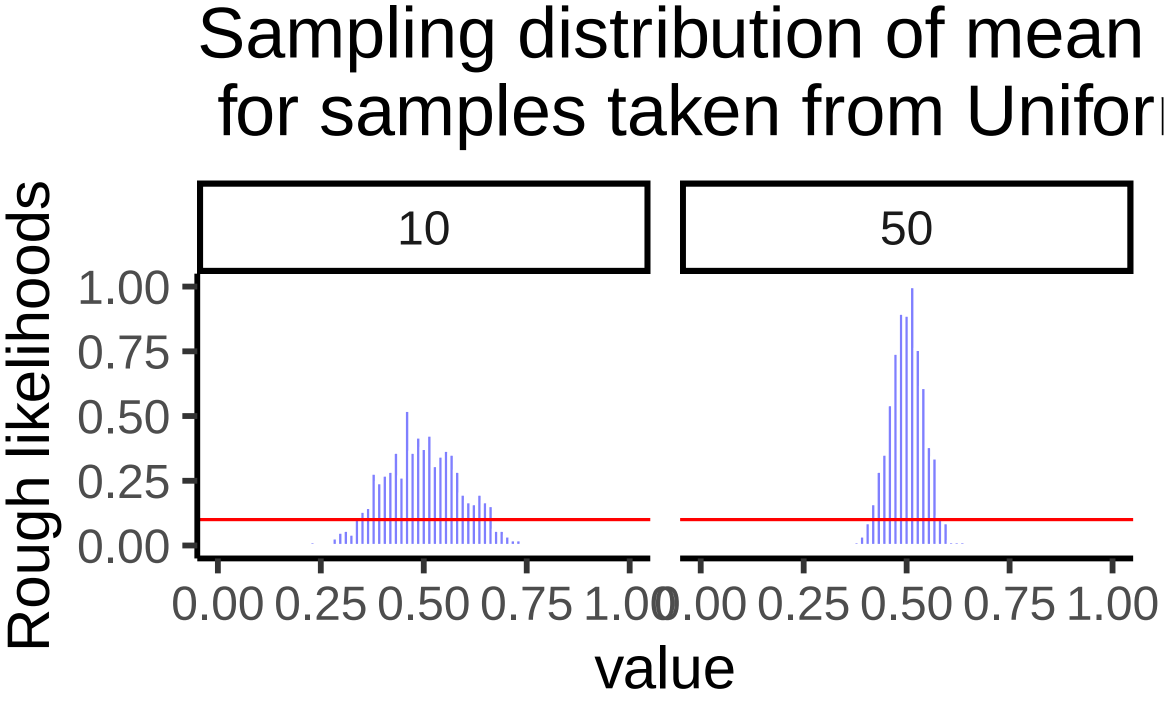


Figure 4.15: Illustration that the shape of the sampling distribution of the mean is normal, even when the samples come from a non-normal (uniform in this case) distribution

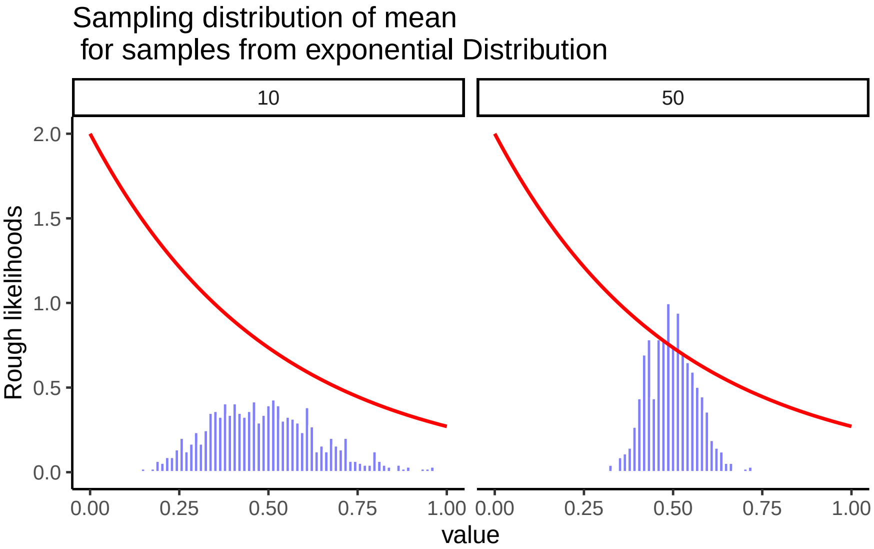


Figure 4.16: Illustration that the shape of the sampling distribution of the mean is normal, even when the samples come from an exponential distribution

4.12 z-scores

We are now in a position to combine some of things we've been talking about in this chapter, and introduce you to a new tool, **z-scores**. It turns out we won't use **z-scores** very much in this textbook. However, you can't take a class on statistics and not learn about **z-scores**.

We are going to look at a normal distribution in Figure 4.17, and draw lines through the distribution at 0, +/- 1, +/-2, and +/- 3 standard deviations from the mean:

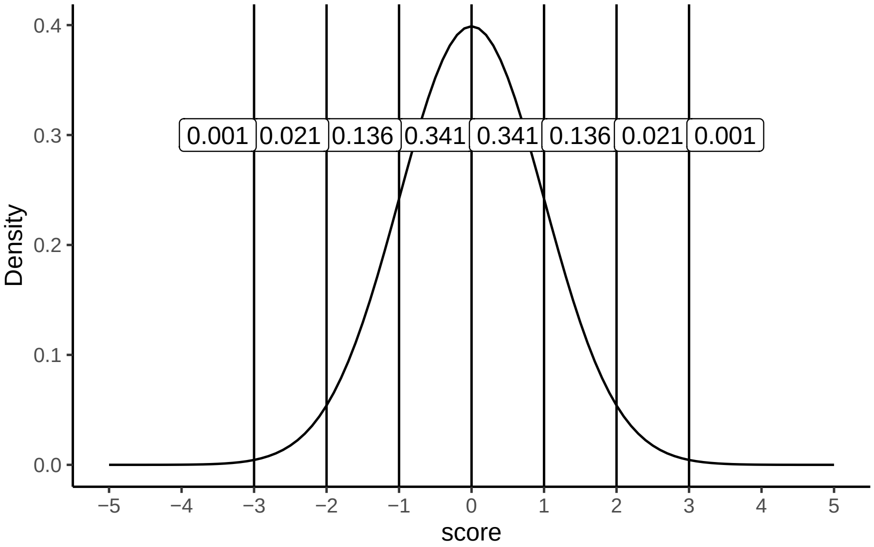


Figure 4.17: A normal distribution. Each line represents a standard deviation from the mean. The labels show the proportions of scores that fall between each bar.

The figure shows a normal distribution with mean = 0, and standard deviation = 1. We've drawn lines at each of the standard deviations: -3, -2, -1, 0, 1, 2, and 3. We also show some numbers in the labels, in between each line. These numbers are proportions. For example, we see the proportion is .341 for scores that fall between the range 0 and 1. Scores between 0 and 1 occur 34.1% of the time. Scores in between -1 and 1, occur 68.2% of the time, that's more than half of the scores. Scores between 1 and occur about 13.6% of the time, and scores between 2 and 3 occur even less, only 2.1% of the time.

Normal distributions always have these properties, even when they have different means and standard deviations. For example, take a look at the normal distribution in Figure 4.18 that has a mean = 100, and standard deviation = 25.

Now we are looking at a normal distribution with mean = 100 and standard deviation = 25. Notice that the region between 100 and 125 contains 34.1% of the scores. This region is 1 standard deviation away from the mean (the standard deviation is 25, the mean is 100, so 25 is one whole standard deviation away from 100). As you can see, the very same proportions

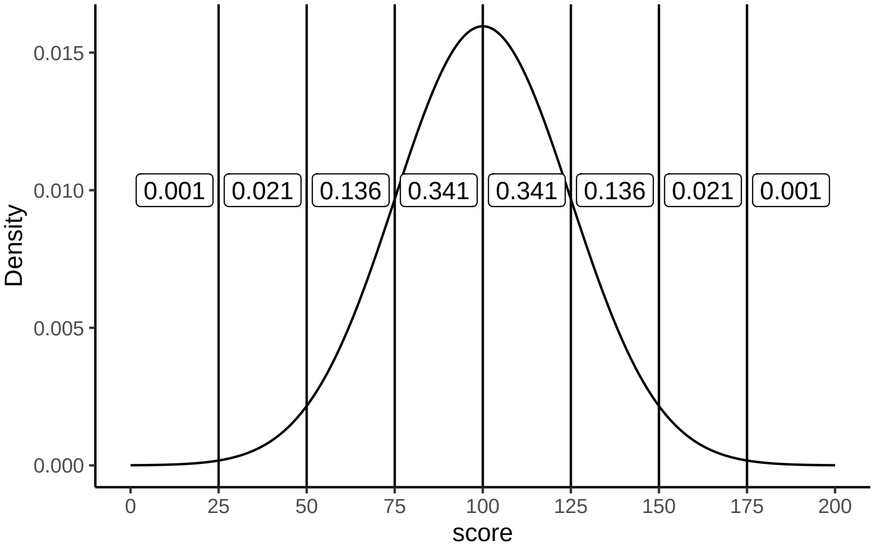


Figure 4.18: A normal distribution. Each line represents a standard deviation from the mean. The labels show the proportions of scores that fall between each bar.

occur between each of the standard deviations, as they did when our standard deviation was set to 1 (with a mean of 0).

4.12.1 Idea behind z-scores

Sometimes it can be convenient to transform your original scores into different scores that are easier to work with. For example, if you have a bunch of proportions, like .3, .5, .6, .7, you might want to turn them into percentages like 30%, 50%, 60%, and 70%. To do that you multiply the proportions by a constant of 100. If you want to turn percentages back into proportions, you divide by a constant of 100. This kind of transformation just changes the scale of the numbers from between 0-1, and between 0-100. Otherwise, the pattern in the numbers stays the same.

The idea behind z-scores is a similar kind of transformation. The idea is to express each raw score in terms of it's standard deviation. For example, if I told you I got a 75% on test, you wouldn't know how well I did compared to the rest of the class. But, if I told you that I scored 2 standard deviations above the mean, you'd know I did quite well compared to the rest of the class, because you know that most scores (if they are distributed normally) fall below 2 standard deviations of the mean.

We also know, now thanks to the central limit theorem, that many of our measures, such as sample means, will be distributed normally. So, it can often be desirable to express the raw scores in terms of their standard deviations.

Let's see how this looks in a table without showing you any formulas. We will look at some scores that come from a normal distribution with mean = 100, and standard deviation = 25. We will list some raw scores, along with the z-scores

raw	z
25	-3
50	-2
75	-1
100	0
125	1
150	2
175	3

Remember, the mean is 100, and the standard deviation is 25. How many standard deviations away from the mean is a score of 100? The answer is 0, it's right on the mean. You can see the z-score for 100, is 0. How many standard deviations is 125 away from the mean? Well the standard deviation is 25, 125 is one whole 25 away from 100, that's a total of 1 standard deviation, so the z-score for 125 is 1. The z-score for 150 is 2, because 150 is two 25s away from 100. The z-score for 50 is -2, because 50 is two 25s away from 100 in the opposite direction. All we are doing here is re-expressing the raw scores in terms of how many standard deviations they are from the mean. Remember, the mean is always right on target, so the center of the z-score distribution is always 0.

4.12.2 Calculating z-scores

To calculate z-scores all you have to do is figure out how many standard deviations from the mean each number is. Let's say the mean is 100, and the standard deviation is 25. You have a score of 97. How many standard deviations from the mean is 97?

First compute the difference between the score and the mean:

$97 - 100 = -3$

Alright, we have a total difference of -3. How many standard deviations does -3 represent if 1 standard deviation is 25? Clearly -3 is much smaller than 25, so it's going to be much less than 1. To figure it out, just divide -3 by the standard deviation.

$\frac{-3}{25} = -.12$

Our z-score for 97 is -.12.

Here's the general formula:

$z = \frac{\text{raw score} - \text{mean}}{\text{standard deviation}}$

So, for example if we had these 10 scores from a normal distribution with mean = 100, and standard deviation =25

```
#>  [1] 133.41  58.21  98.65  81.62  79.00 124.15 100.80 133.23 131.76 103.53
```

The z-scores would be:

```
#>  [1]  1.3364 -1.6716 -0.0540 -0.7352 -0.8400  0.9660  0.0320  1.3292  1.2704
#> [10]  0.1412
```

Once you have the z-scores, you could use them as another way to describe your data. For example, now just by looking at a score you know if it is likely or unlikely to occur, because you know how the area under the normal curve works. z-scores between -1 and 1 happen pretty often, scores greater than 1 or -1 still happen fairly often, but not as often. And, scores bigger than 2 or -2 don't happen very often. This is a convenient thing to do if you want to look at your numbers and get a general sense of how often they happen.

Usually you do not know the mean or the standard deviation of the population that you are drawing your sample scores from. So, you could use the mean and standard deviation of your sample as an estimate, and then use those to calculate z-scores.

Finally, z-scores are also called **standardized scores**, because each raw score is described in terms of it's standard deviation. This may well be the last time we talk about z-scores in this book. You might wonder why we even bothered telling you about them. First, it's worth knowing they are a thing. Second, they become important as your statistical prowess becomes more advanced. Third, some statistical concepts, like correlation, can be re-written in terms of z-scores, and this illuminates aspects of those statistics. Finally, they are super useful when you are dealing with a normal distribution that has a known mean and standard deviation.

4.13 Estimating population parameters

Let's pause for a moment to get our bearings. We're about to go into the topic of **estimation**. What is that, and why should you care? First, population parameters are things about a distribution. For example, distributions have means. The mean is a parameter of the distribution. The standard deviation of a distribution is a parameter. Anything that can describe a distribution is a potential parameter.

OK fine, who cares? This I think, is a really good question. There are some good concrete reasons to care. And there are some great abstract reasons to care. Unfortunately, most of the time in research, it's the abstract reasons that matter most, and these can be the most difficult to get your head around.

4.13.1 Concrete population parameters

First some concrete reasons. There are real populations out there, and sometimes you want to know the parameters of them. For example, if you are a shoe company, you would want to know about the population parameters of feet size. As a first pass, you would want to know the mean and standard deviation of the population. If your company knew this, and other companies did not, your company would do better (assuming all shoes are made equal). Why would your company do better, and how could it use the parameters? Here's one good reason. As a shoe company you want to meet demand with the right amount of supply. If you make too many big or small shoes, and there aren't enough people to buy them, then you're making extra shoes that don't sell. If you don't make enough of the most popular sizes, you'll be leaving money on the table. Right? Yes. So, what would be an optimal thing to do? Perhaps, you would make different amounts of shoes in each size, corresponding to how the demand for each shoe size. You would know something about the demand by figuring out the frequency of each size in the population. You would need to know the population parameters to do this.

Fortunately, it's pretty easy to get the population parameters without measuring the entire population. Who has time to measure every-bodies feet? Nobody, that's who. Instead, you would just need to randomly pick a bunch of people, measure their feet, and then measure the parameters of the sample. If you take a big enough sample, we have learned that the sample mean gives a very good estimate of the population mean. We will learn shortly that a version of the standard deviation of the sample also gives a good estimate of the standard deviation of the population. Perhaps shoe-sizes have a slightly different shape than a normal distribution. Here too, if you collect a big enough sample, the shape of the distribution of the sample will be a good estimate of the shape of the populations. All of these are good reasons to care about estimating population parameters. But, do you run a shoe company? Probably not.

4.13.2 Abstract population parameters

Even when we think we are talking about something concrete in Psychology, it often gets abstract right away. Instead of measuring the population of feet-sizes, how about the population of human happiness. We all think we know what happiness is, everyone has more or less of it, there are a bunch of people, so there must be a population of happiness right? Perhaps, but it's not very concrete. The first problem is figuring out how to measure happiness. Let's use a questionnaire. Consider these questions:

> How happy are you right now on a scale from 1 to 7? How happy are you in general on a scale from 1 to 7? How happy are you in the mornings on a scale from 1 to 7? How happy are you in the afternoons on a scale from 1 to 7?

1. = very unhappy
2. = unhappy
3. = sort of unhappy

4. = in the middle
5. = sort of happy
6. = happy
7. = very happy

Forget about asking these questions to everybody in the world. Let's just ask them to lots of people (our sample). What do you think would happen? Well, obviously people would give all sorts of answers right. We could tally up the answers and plot them in a histogram. This would show us a distribution of happiness scores from our sample. "Great, fantastic!", you say. Yes, fine and dandy.

So, on the one hand we could say lots of things about the people in our sample. We could say exactly who says they are happy and who says they aren't, after all they just told us!

But, what can we say about the larger population? Can we use the parameters of our sample (e.g., mean, standard deviation, shape etc.) to estimate something about a larger population. Can we infer how happy everybody else is, just from our sample? HOLD THE PHONE.

4.13.2.1 Complications with inference

Before listing a bunch of complications, let me tell you what I think we can do with our sample. Provided it is big enough, our sample parameters will be a pretty good estimate of what another sample would look like. Because of the following discussion, this is often all we can say. But, that's OK, as you see throughout this book, we can work with that!

Problem 1: Multiple populations: If you looked at a large sample of questionnaire data you will find evidence of multiple distributions inside your sample. People answer questions differently. Some people are very cautious and not very extreme. Their answers will tend to be distributed about the middle of the scale, mostly 3s, 4s, and 5s. Some people are very bi-modal, they are very happy and very unhappy, depending on time of day. These people's answers will be mostly 1s and 2s, and 6s and 7s, and those numbers look like they come from a completely different distribution. Some people are entirely happy or entirely unhappy. Again, these two "populations" of people's numbers look like two different distributions, one with mostly 6s and 7s, and one with mostly 1s and 2s. Other people will be more random, and their scores will look like a uniform distribution. So, is there a single population with parameters that we can estimate from our sample? Probably not. Could be a mixture of lots of populations with different distributions.

Problem 2: What do these questions measure?: If the whole point of doing the questionnaire is to estimate the population's happiness, we really need wonder if the sample measurements actually tell us anything about happiness in the first place. Some questions: Are people accurate in saying how happy they are? Does the measure of happiness depend on the scale, for example, would the results be different if we used 0-100, or -100 to +100, or no numbers? Does the measure of happiness depend on the wording in the question? Does a measure like this one tell us everything we want to know about happiness (probably not), what is it missing (who knows? probably lots). In short, nobody knows if these kinds of questions measure what we want them to measure. We just hope that they do. Instead, we have a very good idea of the kinds of things that they actually measure. It's really quite obvious, and staring you in the face. Questionnaire measurements measure how people answer questionnaires. In other words, how people behave and answer questions when they are given a questionnaire. This might also measure something about happiness, when the question has to do about happiness. But, it turns out people are remarkably consistent in how they answer questions, even when the questions are total nonsense, or have no questions at all (just numbers to choose!) Maul (2017).

The take home complications here are that we can collect samples, but in Psychology, we often don't have a good idea of the populations that might be linked to these samples. There might be lots of populations, or the populations could be different depending on who you ask. Finally, the "population" might not be the one you want it to be.

4.13.3 Experiments and Population parameters

OK, so we don't own a shoe company, and we can't really identify the population of interest in Psychology, can't we just skip this section on estimation? After all, the "population" is just too weird and abstract and useless and contentious. HOLD THE PHONE AGAIN!

It turns out we can apply the things we have been learning to solve lots of important problems in research. These allow us to answer questions with the data that we collect. Parameter estimation is one of these tools. We just need to be a little bit more creative, and a little bit more abstract to use the tools.

Here is what we know already. The numbers that we measure come from somewhere, we have called this place "distributions". Distributions control how the numbers arrive. Some numbers happen more than others depending on the distribution. We assume, even if we don't know what the distribution is, or what it means, that the numbers came from one. Second, when get some numbers, we call it a sample. This entire chapter so far has taught you one thing. When your sample is big, it resembles the distribution it came from. And, when your sample is big, it will resemble very closely what another big sample of the same thing will look like. We can use this knowledge!

Very often as Psychologists what we want to know is what causes what. We want to know if X causes something to change in Y. Does eating chocolate make you happier? Does studying improve your grades? There a bazillions of these kinds of questions. And, we want answers to them.

I've been trying to be mostly concrete so far in this textbook, that's why we talk about silly things like chocolate and happiness, at least they are concrete. Let's give a go at being abstract. We can do it.

So, we want to know if X causes Y to change. What is X? What is Y? X is something you change, something you manipulate, the independent variable. Y is something you measure. So, we will be taking samples from Y. "Oh I get it, we'll take samples from Y, then we can use the sample parameters to estimate the population parameters of Y!" NO, not really, but yes sort of. We will take sample from Y, that is something we absolutely do. In fact, that is really all we ever do, which is why talking about the population of Y is kind of meaningless. We're more interested in our samples of Y, and how they behave.

So, what would happen if we removed X from the universe altogether, and then took a big sample of Y. We'll pretend Y measures something in a Psychology experiment. So, we know right away that Y is variable. When we take a big sample, it will have a distribution (because Y is variable). So, we can do things like measure the mean of Y, and measure the standard deviation of Y, and anything else we want to know about Y. Fine. What would happen if we replicated this measurement. That is, we just take another random sample of Y, just as big as the first. What should happen is that our first sample should look a lot like our second example. After all, we didn't do anything to Y, we just took two big samples twice. Both of our samples will be a little bit different (due to sampling error), but they'll be mostly the same. The bigger our samples, the more they will look the same, especially when we don't do anything to cause them to be different. In other words, we can use the parameters of one sample to estimate the parameters of a second sample, because they will tend to be the same, especially when they are large.

We are now ready for step two. You want to know if X changes Y. What do you do? You make X go up and take a big sample of Y then look at it. You make X go down, then take a second big sample of Y and look at it. Next, you compare the two samples of Y. If X does nothing then what should you find? We already discussed that in the previous paragraph. If X does nothing, then both of your big samples of Y should be pretty similar. However, if X does something to Y, then one of your big samples of Y will be different from the other. You will have changed something about Y. Maybe X makes the mean of Y change. Or maybe X makes the variation in Y change. Or, maybe X makes the whole shape of the distribution change. If we find any big changes that can't be explained by sampling error, then we can conclude that something about X caused a change in Y! We could use this approach to learn about what causes what!

The very important idea is still about estimation, just not population parameter estimation exactly. We know that when we take samples they naturally vary. So, when we estimate a parameter of a sample, like the mean, we know we are off by some amount. When we find that two samples are different, we need to find out if the size of the difference is consistent with what sampling error can produce, or if the difference is bigger than that. If the difference is bigger, then we can be confident that sampling error didn't produce the difference. So, we can confidently infer that something else (like an X) did cause the difference. This bit of abstract thinking is what most of the rest of the textbook is about. Determining whether there is a difference caused by your manipulation. There's more to the story, there always is. We can get more specific than just, is there a difference, but for introductory purposes, we will focus on the finding of differences as a foundational concept.

4.13.4 Interim summary

We've talked about estimation without doing any estimation, so in the next section we will do some estimating of the mean and of the standard deviation. Formally, we talk about this as using a sample to estimate a parameter of the population. Feel free to think of the "population" in different ways. It could be concrete population, like the distribution of feet-sizes. Or, it could be something more abstract, like the parameter estimate of what samples usually look like when they come from a distribution.

4.13.5 Estimating the population mean

Suppose we go to Brooklyn and 100 of the locals are kind enough to sit through an IQ test. The average IQ score among these people turns out to be $\bar{X} = 98.5$. So what is the true mean IQ for the entire population of Brooklyn? Obviously, we don't know the answer to that question. It could be 97.2, but if could also be 103.5. Our sampling isn't exhaustive so we cannot give a definitive answer. Nevertheless if forced to give a "best guess" I'd have to say 98.5. That's the essence of statistical estimation: giving a best guess. We're using the sample mean as the best guess of the population mean.

In this example, estimating the unknown population parameter is straightforward. I calculate the sample mean, and I use that as my **estimate of the population mean**. It's pretty simple, and in the next section we'll explain the statistical justification for this intuitive answer. However, for the moment let's make sure you recognize that the sample statistic and the estimate of the population parameter are conceptually different things. A sample statistic is a description of your data, whereas the estimate is a guess about the population. With that in mind, statisticians often use different notation to refer to them. For instance, if true population mean is denoted μ, then we would use $\hat{\mu}$ to refer to our estimate of the population mean. In contrast, the sample mean is denoted $\bar{X}$ or sometimes m. However, in simple random samples, the estimate of the population mean is identical to the sample mean: if I observe a sample mean of $\bar{X} = 98.5$, then my estimate of the population mean is also $\hat{\mu} = 98.5$. To help keep the notation clear, here's a handy table:

Symbol	What is it?	Do we know what it is?
$\bar{X}$	Sample mean	Yes, calculated from the raw data
μ	True population mean	Almost never known for sure
$\hat{\mu}$	Estimate of the population mean	Yes, identical to the sample mean

4.13.6 Estimating the population standard deviation

So far, estimation seems pretty simple, and you might be wondering why I forced you to read through all that stuff about sampling theory. In the case of the mean, our estimate of the population parameter (i.e. $\hat{\mu}$) turned out to identical to the corresponding sample statistic (i.e. $\bar{X}$). However, that's not always true. To see this, let's have a think about how to construct an **estimate of the population standard deviation**, which we'll denote $\hat{\sigma}$. What shall we use as our estimate in this case? Your first thought might be that we could do the same thing we did when estimating the mean, and just use the sample statistic as our estimate. That's almost the right thing to do, but not quite.

Here's why. Suppose I have a sample that contains a single observation. For this example, it helps to consider a sample where you have no intuitions at all about what the true population values might be, so let's use something completely fictitious. Suppose the observation in question measures the **cromulence** of my shoes. It turns out that my shoes have a cromulence of 20. So here's my sample:

```
20
```

This is a perfectly legitimate sample, even if it does have a sample size of $N = 1$. It has a sample mean of 20, and because every observation in this sample is equal to the sample mean (obviously!) it has a sample standard deviation of 0. As a description of the **sample** this seems quite right: the sample contains a single observation and therefore there is no variation observed within the sample. A sample standard deviation of $s = 0$ is the right answer here. But as an estimate of the **population** standard deviation, it feels completely insane, right? Admittedly, you and I don't know anything at all about what "cromulence" is, but we know something about data: the only reason that we don't see any variability in the **sample** is that the sample is too small to display any variation! So, if you have a sample size of $N = 1$, it **feels** like the right answer is just to say "no idea at all".

Notice that you **don't** have the same intuition when it comes to the sample mean and the population mean. If forced to make a best guess about the population mean, it doesn't feel completely insane to guess that the population mean is 20. Sure, you probably wouldn't feel very confident in that guess, because you have only the one observation to work with, but it's still the best guess you can make.

Let's extend this example a little. Suppose I now make a second observation. My data set now has $N = 2$ observations of the cromulence of shoes, and the complete sample now looks like this:

```
20, 22
```

This time around, our sample is **just** large enough for us to be able to observe some variability: two observations is the bare minimum number needed for any variability to be observed! For our new data set, the sample mean is $\bar{X} = 21$, and the sample standard deviation is $s = 1$. What intuitions do we have about the population? Again, as far as the population mean goes, the best guess we can possibly make is the sample mean: if forced to guess, we'd probably guess that the population mean cromulence is 21. What about the standard deviation? This is a little more complicated. The sample standard deviation is only based on two observations, and if you're at all like me you probably have the intuition that, with only two observations, we haven't given the population "enough of a chance" to reveal its true variability to us. It's not just that we suspect that the estimate is **wrong**: after all, with only two observations we expect it to be wrong to some degree. The worry is that the error is **systematic**.

If the error is systematic, that means it is **biased**. For example, imagine if the sample mean was always smaller than the population mean. If this was true (it's not), then we couldn't use the sample mean as an estimator. It would be biased, we'd be using the wrong number.

It turns out the sample standard deviation is a **biased estimator** of the population standard deviation. We can sort of anticipate this by what we've been discussing. When the sample size is 1, the standard deviation is 0, which is obviously to small. When the sample size is 2, the standard deviation becomes a number bigger than 0, but because we only have two sample, we suspect it might still be too small. Turns out this intuition is correct.

It would be nice to demonstrate this somehow. There are in fact mathematical proofs that confirm this intuition, but unless you have the right mathematical background they don't help very much. Instead, what I'll do is use R to simulate the results of some experiments. With that in mind, let's return to our IQ studies. Suppose the true population mean IQ is 100 and the standard deviation is 15. I can use the **rnorm()** function to generate the the results of an experiment in which I measure $N = 2$ IQ scores, and calculate the sample standard deviation. If I do this over and over again, and plot a histogram of these sample standard deviations, what I have is the **sampling distribution of the standard deviation**. I've plotted this distribution in Figure 4.19.

Even though the true population standard deviation is 15, the average of the **sample** standard deviations is only 8.5. Notice that this is a very different from when we were plotting sampling distributions of the sample mean, those were always centered around the mean of the population.

Now let's extend the simulation. Instead of restricting ourselves to the situation where we have a sample size of $N = 2$, let's repeat the exercise for sample sizes from 1 to 10. If we plot the average sample mean and average sample standard deviation as a function of sample size, you get the following results.

@fig-estimatorbiasA shows the sample mean as a function of sample size. Notice it's a flat line. The sample mean doesn't underestimate or overestimate the population mean. It is an unbiased estimate!

Figure 4.21 shows the sample standard deviation as a function of sample size. Notice it is not a flat line. The sample standard deviation systematically underestimates the population standard deviation!

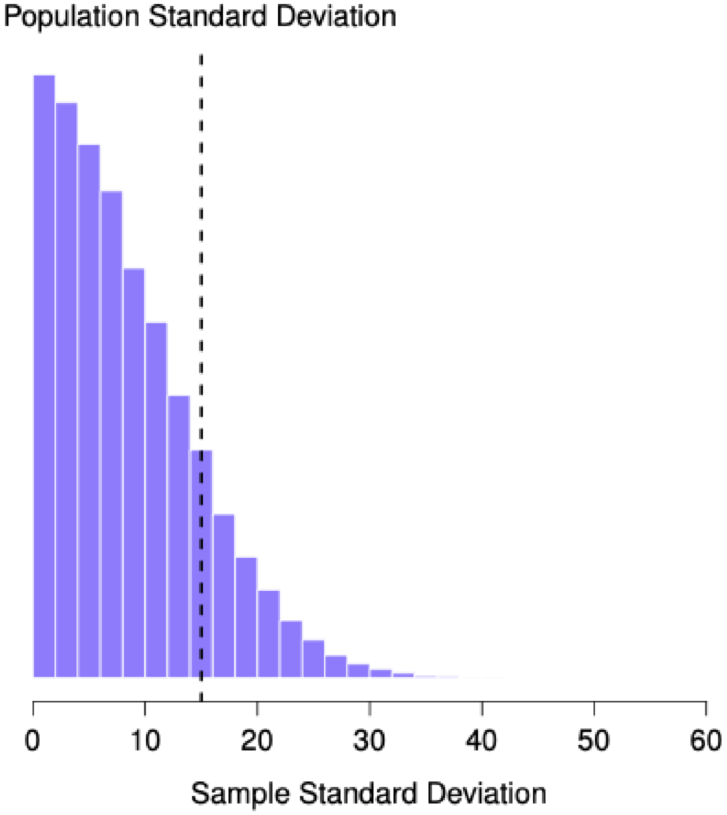


Figure 4.19: The sampling distribution of the sample standard deviation for a two IQ scores experiment. The true population standard deviation is 15 (dashed line), but as you can see from the histogram, the vast majority of experiments will produce a much smaller sample standard deviation than this. On average, this experiment would produce a sample standard deviation of only 8.5, well below the true value! In other words, the sample standard deviation is a biased estimate of the population standard deviation.

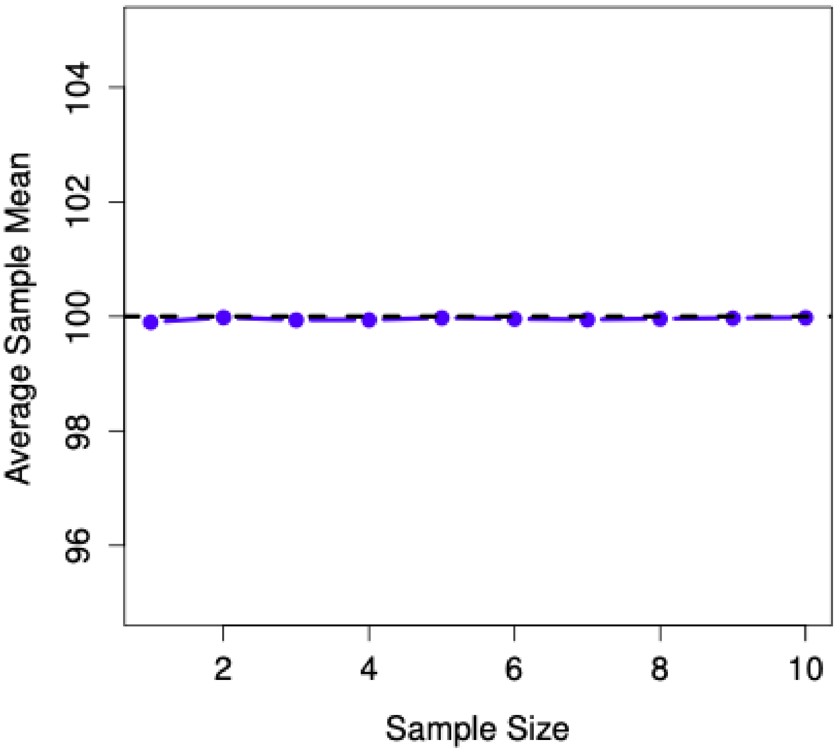


Figure 4.20: An illustration of the fact that the sample mean is an unbiased estimator of the population mean.

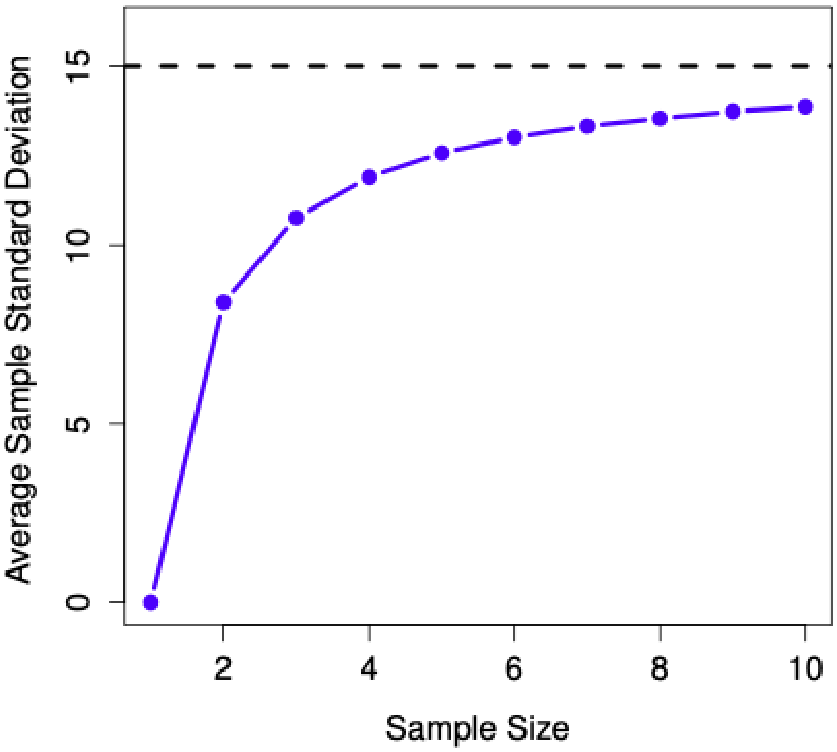


Figure 4.21: An illustration of the fact that the the sample standard deviation is a biased estimator of the population standard deviation.

In other words, if we want to make a “best guess” ($\hat{\sigma}$, our estimate of the population standard deviation) about the value of the population standard deviation σ, we should make sure our guess is a little bit larger than the sample standard deviation s.

The fix to this systematic bias turns out to be very simple. Here’s how it works. Before tackling the standard deviation, let’s look at the variance. If you recall from the second chapter, the sample variance is defined to be the average of the squared deviations from the sample mean. That is:

$$s^2 = \frac{1}{N} \sum_{i=1}^{N} (X_i - \bar{X})^2$$

The sample variance s^2 is a biased estimator of the population variance σ^2. But as it turns out, we only need to make a tiny tweak to transform this into an unbiased estimator. All we have to do is divide by $N - 1$ rather than by N. If we do that, we obtain the following formula:

$$\hat{\sigma}^2 = \frac{1}{N-1} \sum_{i=1}^{N} (X_i - \bar{X})^2$$

This is an unbiased estimator of the population variance σ.

A similar story applies for the standard deviation. If we divide by $N - 1$ rather than N, our estimate of the population standard deviation becomes:

$$\hat{\sigma} = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (X_i - \bar{X})^2}$$

.

It is worth pointing out that software programs make assumptions **for you**, about which variance and standard deviation **you** are computing. Some programs automatically divide by $N - 1$, some do not. You need to check to figure out what they are doing. Don’t let the software tell you what to do. Software is for you telling it what to do.

One final point: in practice, a lot of people tend to refer to $\hat{\sigma}$ (i.e., the formula where we divide by $N - 1$) as the **sample** standard deviation. Technically, this is incorrect: the **sample** standard deviation should be equal to s (i.e., the formula where we divide by N). These aren’t the same thing, either conceptually or numerically. One is a property of the sample, the other is an estimated characteristic of the population. However, in almost every real life application, what we actually care about is the estimate of the population parameter, and so people always report $\hat{\sigma}$ rather than s.

ℹ Note

Note, whether you should divide by N or N-1 also depends on your philosophy about what you are doing. For example, if you don’t think that what you are doing is estimating a population parameter, then why would you divide by N-1? Also, when N is large, it

> doesn't matter too much. The difference between a big N, and a big N-1, is just -1.

This is the right number to report, of course, it's that people tend to get a little bit imprecise about terminology when they write it up, because "sample standard deviation" is shorter than "estimated population standard deviation". It's no big deal, and in practice I do the same thing everyone else does. Nevertheless, I think it's important to keep the two **concepts** separate: it's never a good idea to confuse "known properties of your sample" with "guesses about the population from which it came". The moment you start thinking that s and $\hat{\sigma}$ are the same thing, you start doing exactly that.

To finish this section off, here's another couple of tables to help keep things clear:

Symbol	What is it?	Do we know what it is?
s^2	Sample variance	Yes, calculated from the raw data
σ^2	Population variance	Almost never known for sure
$\hat{\sigma}^2$	Estimate of the population variance	Yes, but not the same as the sample variance

4.14 Estimating a confidence interval

> Statistics means never having to say you're certain – Unknown origin

Up to this point in this chapter, we've outlined the basics of sampling theory which statisticians rely on to make guesses about population parameters on the basis of a sample of data. As this discussion illustrates, one of the reasons we need all this sampling theory is that every data set leaves us with some of uncertainty, so our estimates are never going to be perfectly accurate. The thing that has been missing from this discussion is an attempt to **quantify** the amount of uncertainty in our estimate. It's not enough to be able guess that the mean IQ of undergraduate psychology students is 115 (yes, I just made that number up). We also want to be able to say something that expresses the degree of certainty that we have in our guess. For example, it would be nice to be able to say that there is a 95% chance that the true mean lies between 109 and 121. The name for this is a **confidence interval** for the mean.

Armed with an understanding of sampling distributions, constructing a confidence interval for the mean is actually pretty easy. Here's how it works. Suppose the true population mean is μ and the standard deviation is σ. I've just finished running my study that has N participants, and the mean IQ among those participants is $\bar{X}$. We know from our discussion of the central limit theorem that the sampling distribution of the mean is approximately normal. We also know from our discussion of the normal distribution that there is a 95% chance that a normally-distributed quantity will fall within two standard deviations of the true mean. To be more precise, we can use the **qnorm()** function to compute the 2.5th and 97.5th percentiles of the normal distribution

qnorm(p = c(.025, .975)) [1] -1.959964 1.959964

Okay, so I lied earlier on. The more correct answer is that a 95% chance that a normally-distributed quantity will fall within 1.96 standard deviations of the true mean.

Next, recall that the standard deviation of the sampling distribution is referred to as the standard error, and the standard error of the mean is written as SEM. When we put all these pieces together, we learn that there is a 95% probability that the sample mean $\bar{X}$ that we have actually observed lies within 1.96 standard errors of the population mean. Oof, that is a lot of mathy talk there. We'll clear it up, don't worry.

Mathematically, we write this as:

$$\mu - (1.96 \times \text{SEM}) \ \leq \ \bar{X} \ \leq \ \mu + (1.96 \times \text{SEM})$$

where the SEM is equal to $\sigma / \sqrt{N}$, and we can be 95% confident that this is true.

However, that's not answering the question that we're actually interested in. The equation above tells us what we should expect about the sample mean, given that we know what the population parameters are. What we **want** is to have this work the other way around: we want to know what we should believe about the population parameters, given that we have observed a particular sample. However, it's not too difficult to do this. Using a little high school algebra, a sneaky way to rewrite our equation is like this:

$$\bar{X} - (1.96 \times \text{SEM}) \ \leq \ \mu \ \leq \ \bar{X} + (1.96 \times \text{SEM})$$

What this is telling is is that the range of values has a 95% probability of containing the population mean μ. We refer to this range as a **95% confidence interval**, denoted CI_{95}. In short, as long as N is sufficiently large – large enough for us to believe that the sampling distribution of the mean is normal – then we can write this as our formula for the 95% confidence interval:

$$\text{CI}_{95} = \bar{X} \pm \left(1.96 \times \frac{\sigma}{\sqrt{N}} \right)$$

Of course, there's nothing special about the number 1.96: it just happens to be the multiplier you need to use if you want a 95% confidence interval. If I'd wanted a 70% confidence interval, I could have used the **qnorm()** function to calculate the 15th and 85th quantiles:

qnorm(p = c(.15, .85)) [1] -1.036433 1.036433

and so the formula for CI_{70} would be the same as the formula for CI_{95} except that we'd use 1.04 as our magic number rather than 1.96.

4.14.1 A slight mistake in the formula

As usual, I lied. The formula that I've given above for the 95% confidence interval is approximately correct, but I glossed over an important detail in the discussion. Notice my formula requires you to use the standard error of the mean, SEM, which in turn requires you to use the true population standard deviation σ.

Yet, before we stressed the fact that we don't actually **know** the true population parameters. Because we don't know the true value of σ, we have to use an estimate of the population standard deviation $\hat{\sigma}$ instead. This is pretty straightforward to do, but this has the consequence that we need to use the quantiles of the t-distribution rather than the normal distribution to calculate our magic number; and the answer depends on the sample size. Plus, we haven't really talked about the t distribution yet.

When we use the t distribution instead of the normal distribution, we get bigger numbers, indicating that we have more uncertainty. And why do we have that extra uncertainty? Well, because our estimate of the population standard deviation $\hat{\sigma}$ might be wrong! If it's wrong, it implies that we're a bit less sure about what our sampling distribution of the mean actually looks like… and this uncertainty ends up getting reflected in a wider confidence interval.

4.15 Summary

In this chapter I've covered two main topics. The first half of the chapter talks about sampling theory, and the second half talks about how we can use sampling theory to construct estimates of the population parameters. The section breakdown looks like this:

- Basic ideas about samples, sampling and populations
- Statistical theory of sampling: the law of large numbers, sampling distributions and the central limit theorem.
- Estimating means and standard deviations
- confidence intervals

As always, there's a lot of topics related to sampling and estimation that aren't covered in this chapter, but for an introductory psychology class this is fairly comprehensive I think. For most applied researchers you won't need much more theory than this. One big question that I haven't touched on in this chapter is what you do when you don't have a simple random sample. There is a lot of statistical theory you can draw on to handle this situation, but it's well beyond the scope of this book.

4.16 Videos

4.16.1 Introduction to Probability

Jeff has several more videos on probability that you can view on his statistics playlist.

4.16.2 Chebychev's Theorem

4.16.3 Z-scores

4.16.4 Normal Distribution I

4.16.5 Normal Distribution II

5 Foundations for inference

> Data and data sets are not objective; they are creations of human design. We give numbers their voice, draw inferences from them, and define their meaning through our interpretations. —Katie Crawford

So far we have been talking about describing data and looking for possible relationships between things we measure. We began with the problem of having too many numbers and discussed how they could be summarized with descriptive statistics, and communicated in graphs. We also looked at the idea of relationships between things. If one thing causes change in another thing, then if we measure how one thing goes up and down we should find that other thing goes up and down, or does something systematically following the first thing. At the end of the chapter on correlation, we showed how correlations, which imply a relationship between two things, are very difficult to interpret. Why? Because an observed correlation can be caused by a hidden third variable, or could be a spurious finding "caused" by random chance. In the last chapter, we talked about sampling from distributions, and we saw how samples can be different because of random error introduced by the sampling process.

Now we begin our journey into **inferential statistics**. These are tools used to make inferences about where our data came from, and to make inferences about what causes what.

In this chapter we provide some foundational ideas. We will stay mostly at a conceptual level, and use lots of simulations like we did in the last chapters. In the remaining chapters we formalize the intuitions built here to explain how some common inferential statistics work.

5.1 Brief review of Experiments

In chapter one we talked a about research methods and experiments. Experiments are a structured way of collecting data that can permit inferences about causality. If we wanted to know whether something like watching cats on YouTube increases happiness we would need an experiment. We already found out that just finding a bunch of people and measuring number of hours watching cats, and level of happiness, and correlating the two will not permit inferences about causation. For one, the causal flow could be reversed. Maybe being happy causes people to watch more cat videos. We need an experiment.

An experiment has two parts. A manipulation and a measurement. The manipulation is under the control of the experimenter. Manipulations are also called **independent variables**. For example, we could manipulate time spent watching cat videos: 1 hour versus 2 hours of cat videos. The measurement is the data that is collected. We could measure how happy people are after watching cat videos on a scale from 1 to 100. Measurements are also called **dependent variables**. So, in a basic experiment like the one above, we take measurements of happiness from people in one of two experimental conditions defined by the independent variable. Let's say we ran 50 subjects. 25 subjects would be randomly assigned to watch 1 hour of cat videos, and the other 25 subjects would be randomly assigned to watch 2 hours of cat videos. We would measure happiness for each subject at the end of the videos. Then we could look at the data.

What would we want to look at? If watching cat videos caused a change in happiness, then we would expect the measures of happiness for people watching 1 hour of cat videos to be different from the measures of happiness for people watching 2 hours of cat videos. If watching cat videos does not change happiness, then we would expect no differences in measures of happiness between conditions. Causal forces cause change, and the experiment is set up to detect the change.

Now we can state one overarching question, how do we know if the data changed between conditions? If we can be confident that there was a change between conditions, we can infer that our manipulation caused a changed in the measurement. If we cannot be confident there was a change, then we cannot infer that our manipulation caused a change in the measurement. We need to build some change detection tools so we can know a change when we find one.

"Hold on, if we are just looking for a change, wouldn't that be easy to see by looking at the numbers and seeing if they are different, what's so hard about that?". Good question. Now we must take a detour. The short answer is that there will always be change in the data (remember variance).

5.2 The data came from a distribution

In the last chapter we discussed samples and distributions, and the idea that you can take samples from distributions. So, from now on when you see a bunch of numbers, you should wonder, "where did these numbers come from?". What caused some kinds of numbers to happen more than other kinds of numbers. The answer to this question requires us to again veer off into the abstract world of distributions. A **distribution** a place where numbers can come from. The distribution sets the constraints. It determines what numbers are likely to occur, and what numbers are not likely to occur. Distributions are abstract ideas. But, they can be made concrete, and we can draw them with pictures that you have seen already, called histograms.

The next bit might seem slightly repetitive from the previous chapter. We again look at sampling numbers from a uniform distribution. We show that individual samples can look quite different from each other. Much of the beginning part of this chapter will already be familiar to you, but we take the concepts in a slightly different direction. The direction is how to make inferences about the role of chance in your experiment.

5.2.1 Uniform distribution

As a reminder from last chapter, Figure 5.1 shows that the shape of a uniform distribution is completely flat.

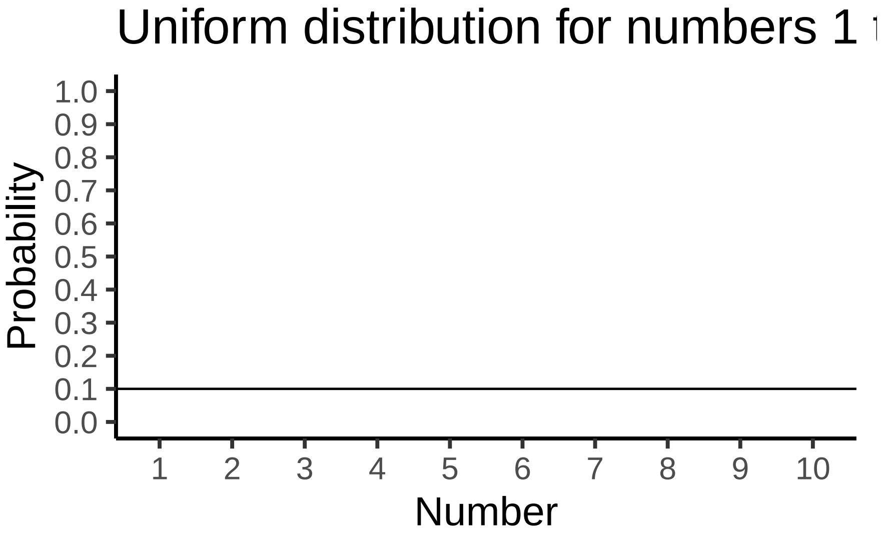


Figure 5.1: Uniform distribution showing that the numbers from 1 to 10 have an equal probability of being sampled

OK, so that doesn't look like much. What is going on here? The y-axis is labelled `probability`, and it goes from 0 to 1. The x-axis is labelled `Number`, and it goes from one to 10. There is a horizontal line drawn straight through. This line tells you the probability of each number from 1 to 10. Notice the line is flat. This means all of the numbers have the same probability of occurring. More specifically, there are 10 numbers from 1 to 10 (1,2,3,4,5,6,7,8,9,10), and they all have an equal chance of occurring. 1/10 = .1, which is the probability indicated by the horizontal line.

"So what?". Imagine that this uniform distribution is a number generating machine. It spits out numbers, but it spits out each number with the probability indicated by the line. If this distribution was going to start spitting out numbers, it would spit out 10% 1s, 10% 2s, 10% 3s, and so on, up to 10% 10s. Wanna see what that would look like? Let's make it spit out 100 numbers and put them in Table 5.1.

Table 5.1: 100 numbers randomly sampled from a uniform distribution.

2	6	9	4	7	7	7	4	6	3
5	5	7	2	3	3	6	9	7	6
7	5	2	2	8	4	4	9	10	4
5	2	5	2	8	8	6	10	1	6
4	9	9	8	5	7	5	6	3	7
10	2	3	3	7	2	5	5	4	7
8	2	1	10	5	9	5	9	6	3
9	9	4	2	10	4	6	7	3	4
7	2	5	5	7	3	7	9	6	6
3	8	9	9	6	2	2	1	2	1

We used the uniform distribution to generate these numbers. Officially, we call this **sampling** from a **distribution**. Sampling is what you do at a grocery store when there is free food. You can keep taking more. However, if you take all of the samples, then what you have is called the **population**. We'll talk more about samples and populations as we go along.

Because we used the uniform distribution to create numbers, we already know where our numbers came from. However, we can still pretend for the moment that someone showed up at your door, showed you these numbers, and then you wondered where they came from. Can you tell just by looking at these numbers that they came from a uniform distribution? What would need to look at? Perhaps you would want to know if all of the numbers occur with roughly equal frequency, after all they should have right? That is, if each number had the same chance of occurring, we should see that each number occurs roughly the same number of times.

We already know what a histogram is, so we can put our sample of 100 numbers into a histogram and see what the counts look like. If all of the numbers from 1 to 10 occur with equal frequency, then each individual number should occur about 10 times. Figure 5.2 shows the histogram:

Uh oh, as you can see, not all of the number occurred 10 times each. All of the bars are not the same height. This shows that randomly sampling numbers from this distribution does not guarantee that our numbers will be exactly like the distribution they came from. We can call this sampling error, or sampling variability.

5.2.2 Not all samples are the same, they are usually quite different

Let's look at sampling error more closely. We will sample 20 numbers from the uniform distribution. We should expect that each number between 1 and 10 occurs about two times each. As before, this expectation can be visualized in a histogram. To get a better sense of

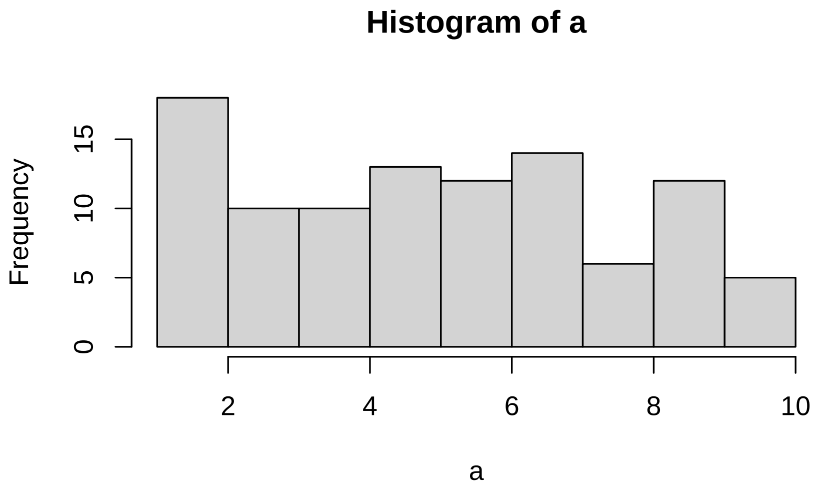


Figure 5.2: Histogram of 100 numbers randomly sampled from the uniform distribution containing the integers from 1 to 10

sampling error, let's repeat the above process ten times. Figure 5.3 has 10 histograms, each showing what 10 different samples of twenty numbers looks like:

You might notice right away that none of the histograms are the same. Even though we are randomly taking 20 numbers from the very same uniform distribution, each sample of 20 numbers comes out different. This is sampling variability, or sampling error.

?@fig-5expectedUnif shows an animated version of the process of repeatedly choosing 20 new random numbers and plotting a histogram. The horizontal line shows the flat-line shape of the uniform distribution. The line crosses the y-axis at 2; and, we expect that each number (from 1 to 10) should occur about 2 times each in a sample of 20. However, each sample bounces around quite a bit, due to random chance.

Looking at the above histograms shows us that figuring out where our numbers came from can be difficult. In the real world, our measurements are samples. We usually only have the luxury of getting one sample of measurements, rather than repeating our own measurements 10 times or more. If you look at the histograms, you will see that some of them look like they could have come from the uniform distribution: most of the bars are near two, and they all fall kind of on a flat line. But, if you happen to look at a different sample, you might see something that is very bumpy, with some numbers happening way more than others. This could suggest to you that those numbers did not come from a uniform distribution (they're just too bumpy). But let me remind you, all of these samples came from a uniform distribution, this is what samples from that distribution look like. This is what chance does to samples, it makes the individual data points noisy.

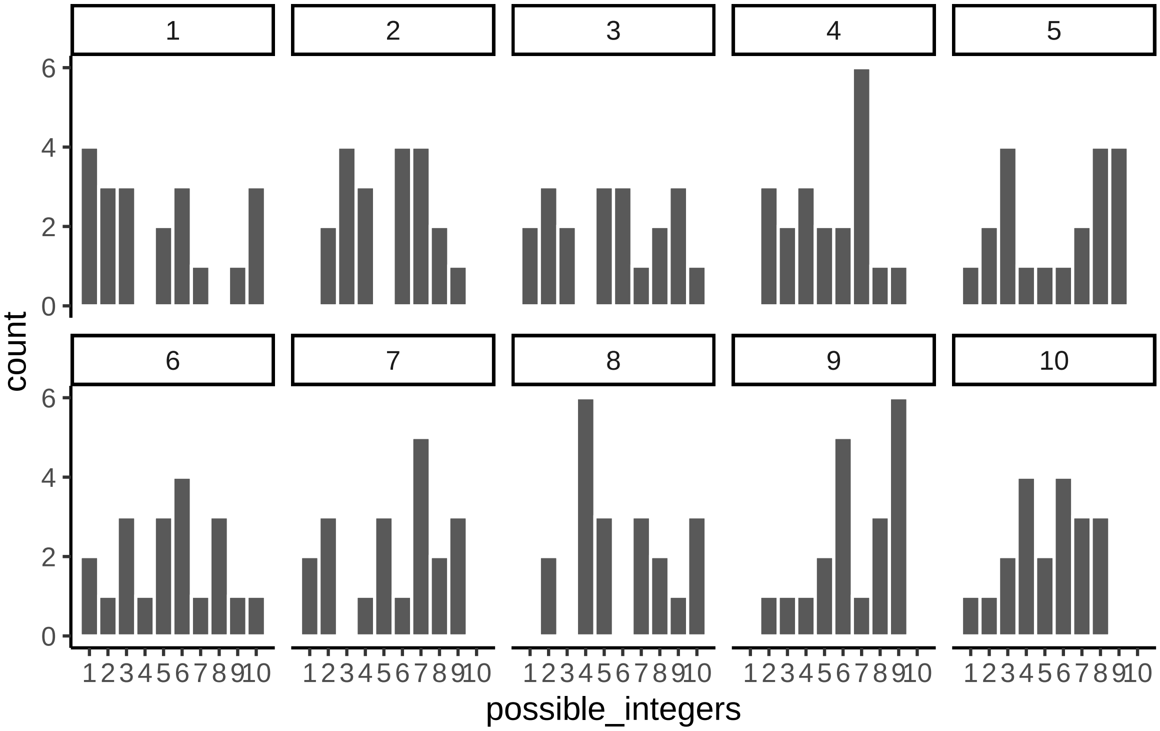


Figure 5.3: Histograms for 10 different samples from the uniform distribution. Each sample contains 20 numbers. The histograms all look quite different. The differences between the samples are due to sampling error or random chance.

5.2.3 Large samples are more like the distribution they came from

Let's refresh the question. Which of the two samples in Figure 5.4 do you think came from a uniform distribution?

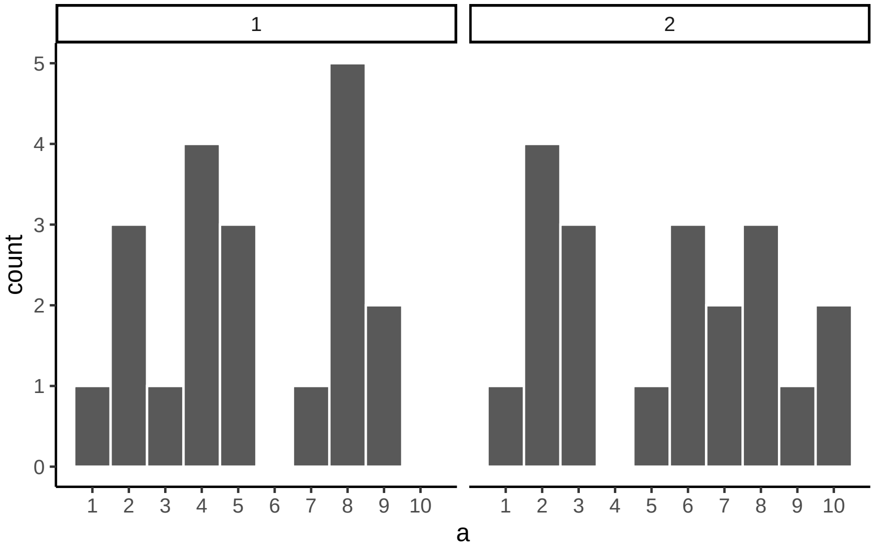


Figure 5.4: Which of these two samples came from a uniform distribution?

The answer is that they both did. But, neither of them look like they did.

Can we improve things, and make it easier to see if a sample came from a uniform distribution? Yes, we can. All we need to do is increase the **sample-size**. We will often use the letter `n` to refer to sample-size. N is the number of observations in the sample.

So let's increase the number of observations in each sample from 20 to 100. We will again create 10 samples (each with 100 observations), and make histograms for each of them. All of these samples will be drawn from the very same uniform distribution. This, means we should expect each number from 1 to 10 to occur about 10 times in each sample. The histograms are shown in Figure 5.5.

Again, most of these histograms don't look very flat, and all of the bars seem to be going up or down, and they are not exactly at 10 each. So, we are still dealing with sampling error. It's a pain. It's always there.

Let's bump up the N from 100 to 1000 observations per sample. Now we should expect every number to appear about 100 times each. What happens?

Figure 5.6 shows the histograms are starting to flatten out. The bars are still not perfectly at 100, because there is still sampling error (there always will be). But, if you found a histogram

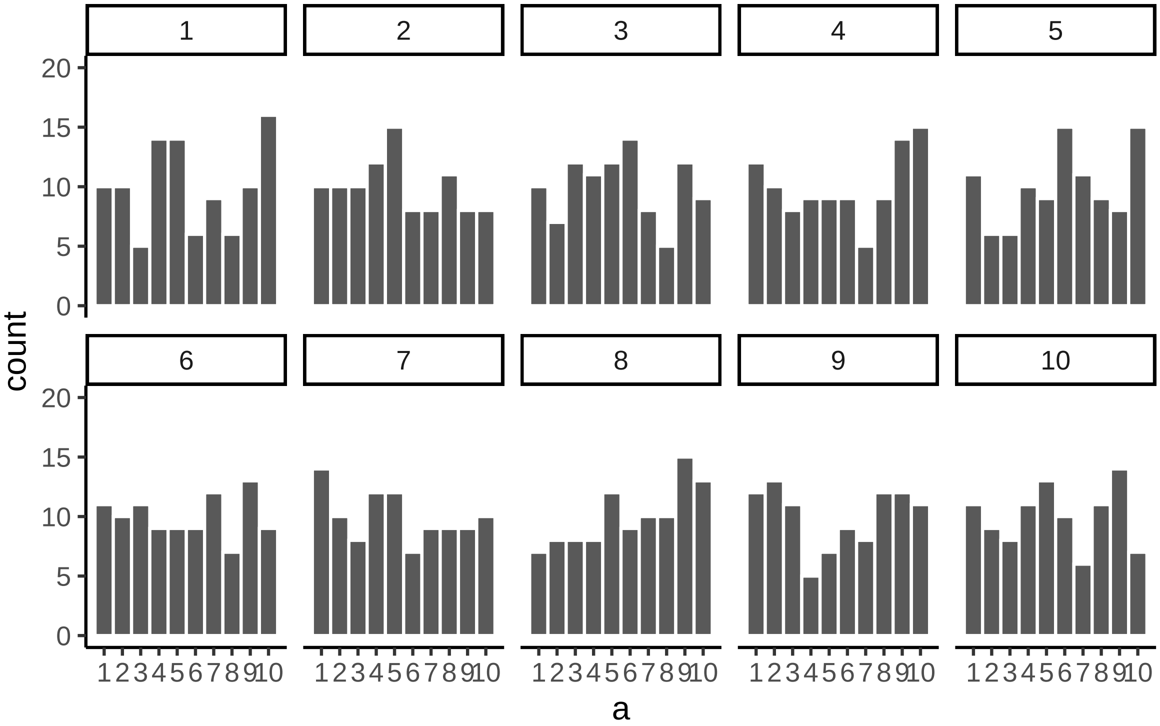


Figure 5.5: Histograms for different samples from a uniform distribution. N = 100 for each sample.

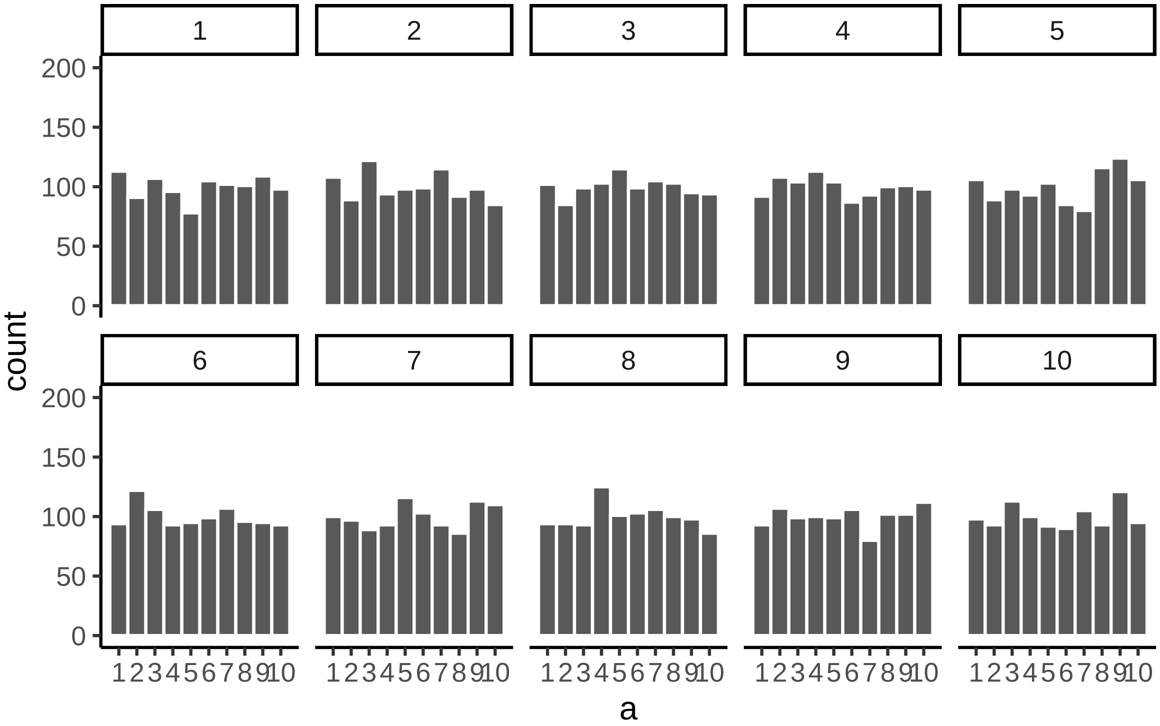


Figure 5.6: Histograms for different samples from a uniform distribution. N = 1000 for each sample.

that looked flat and knew that the sample contained many observations, you might be more confident that those numbers came from a uniform distribution.

Just for fun let's make the samples really big. Say 100,000 observations per sample. Here, we should expect that each number occurs about 10,000 times each. What happens?

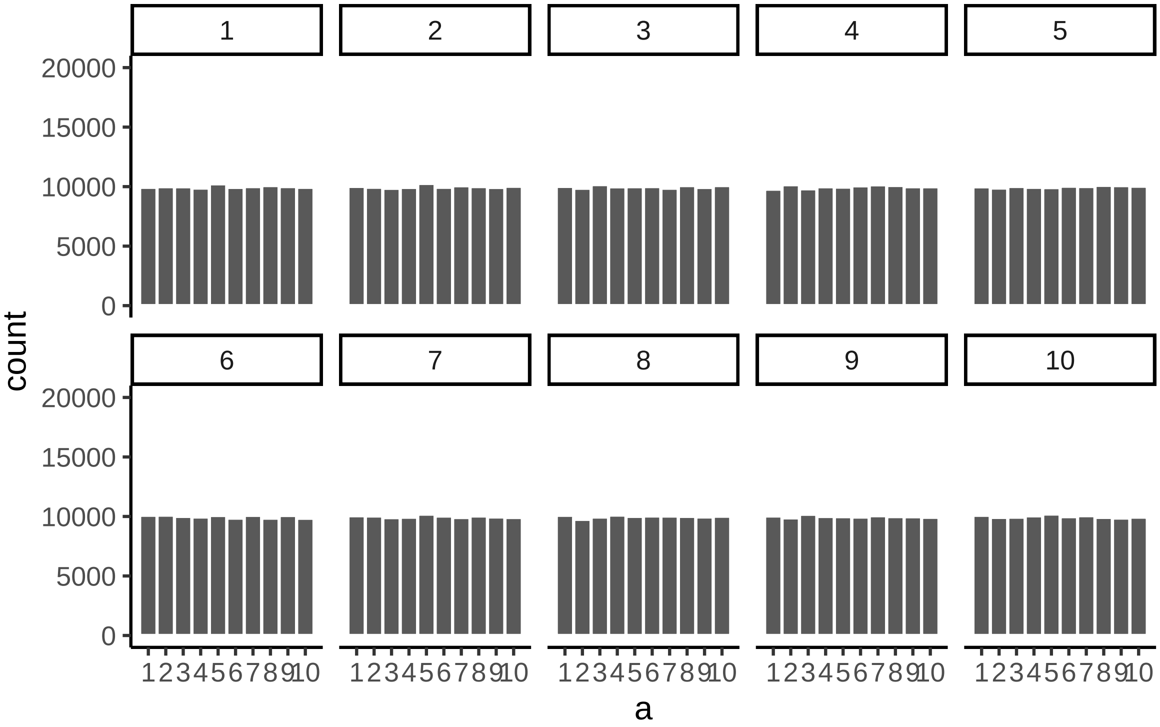


Figure 5.7: Histograms for different samples from a uniform distribution. N = 100,000 for each sample.

Figure 5.7 shows that the histograms for each sample are starting to look the same. They all have 100,000 observations, and this gives chance enough opportunity to equally distribute the numbers, roughly making sure that they all occur very close to the same amount of times. As you can see, the bars are all very close to 10,000, which is where they should be if the sample came from a uniform distribution.

Pro tip

The pattern behind a sample will tend to stabilize as sample-size increases. Small samples will have all sorts of patterns because of sampling error (chance).

Before getting back to the topic of experiments that we started with, let's ask two more questions. First, which of the two samples in Figure 5.8 do you think came from a uniform distribution? FYI, each of these samples had 20 observations each.

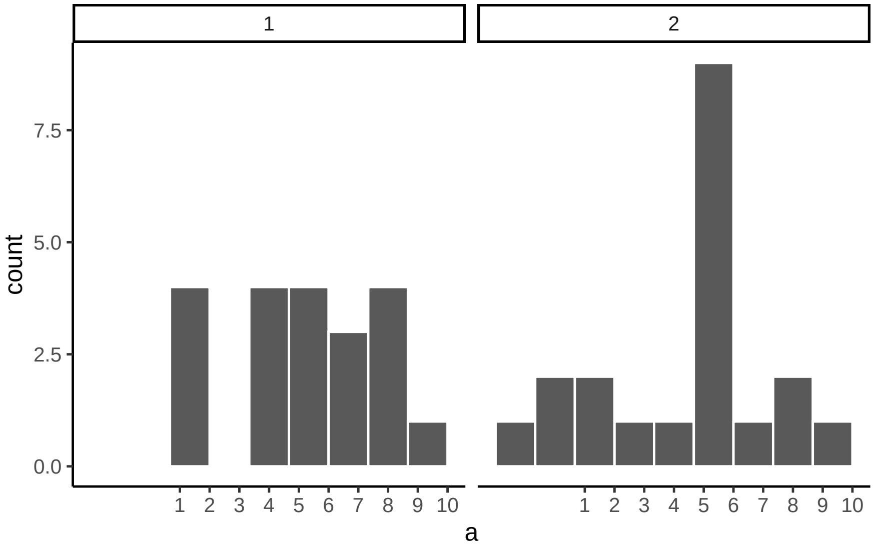


Figure 5.8: Which of these samples came from a uniform distribution?

If you are not confident in the answer, this is because **sampling error** (randomness) is fuzzing with the histograms.

Here is the very same question, only this time we will take 1,000 observations for each sample. Which histogram in Figure 5.9 do you think came from a uniform distribution, which one did not?

Now that we have increased N, we can see the pattern in each sample becomes more obvious. The histogram for sample 1 has bars near 100, not perfectly flat, but it resembles a uniform distribution. The histogram for sample 2 is not flat looking at all.

Congratulations to Us! We have just made some statistical inferences without using formulas!

"We did?" Yes, by looking at our two samples we have inferred that sample 2 did not come from a uniform distribution. We have also inferred that sample 1 could have come form a uniform distribution. Fantastic. These are the same kinds of inferences we will be making for the rest of the course. We will be looking at some numbers, wondering where they came from, then we will arrange the numbers in such a way so that we can make inferences about the kind of distribution they came from. That's it.

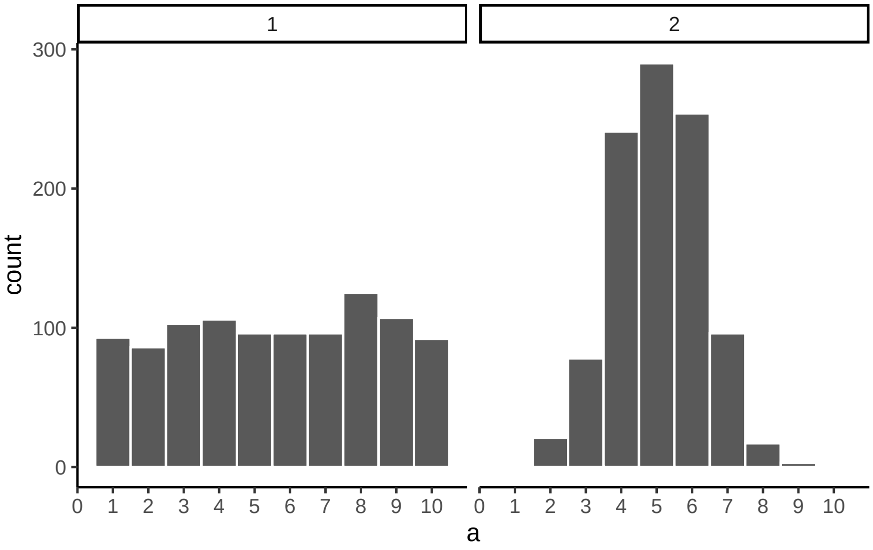


Figure 5.9: Which of these samples came from a uniform distribution?

5.3 Is there a difference?

Let’s get back to experiments. In an experiment we want to know if an independent variable (our manipulation) causes a change in a dependent variable (measurement). If this occurs, then we will expect to see some differences in our measurement as a function of the manipulation.

Consider the light switch example:

Light Switch Experiment: You manipulate the switch up (condition 1 of independent variable), light goes on (measurement). You manipulate the switch down (condition 2 of independent variable), light goes off (another measurement). The measurement (light) changes (goes off and on) as a function of the manipulation (moving switch up or down).

You can see the change in measurement between the conditions, it is as obvious as night and day. So, when you conduct a manipulation, and can see the difference (change) in your measure, you can be pretty confident that your manipulation is causing the change.

> note: to be cautious we can say “something” about your manipulation is causing the change, it might not be what you think it is if your manipulation is very complicated and involves lots of moving parts.

5.3.1 Chance can produce differences

Do you think random chance can produce the appearance of differences, even when there really aren't any? I hope so. We have already shown that the process of sampling numbers from a distribution is a chancy process that produces different samples. Different samples are different, so yes, chance can produce differences. This can muck up our interpretation of experiments.

Let's conduct a fictitious experiment where we expect to find no differences, because we will manipulate something that shouldn't do anything. Here's the set-up:

You are the experimenter standing in front of a gumball machine. It is very big, has thousands of gumballs. 50% of the gumballs are green, and 50% are red. You want to find out if picking gumballs with your right hand vs. your left hand will cause you to pick more green gumballs. Plus, you will be blindfolded the entire time. The independent variable is Hand: right hand vs. left hand. The dependent variable is the measurement of the color of each gumball.

You run the experiment as follows. 1) put on blind fold. 2) pick 10 gumballs randomly with left hand, set them aside. 3) pick 10 gumballs randomly with right hand, set them aside. 4) count the number of green and red gumballs chosen by your left hand, and count the number of green and red gumballs chosen by your right hand. Hopefully you will agree that your hands will not be able to tell the difference between the gumballs. If you don't agree, we will further stipulate the gumballs are completely identical in every way except their color, so it would be impossible to tell them apart using your hands. So, what should happen in this experiment?

"Umm, maybe you get 5 red gum balls and 5 green balls from your left hand, and also from your right hand?". Sort of yes, this is what you would usually get. But, it is not all that you can get. Here is some data showing what happened from one pretend experiment:

hand	gumball
left	0
left	0
left	0
left	1
left	1
left	0
left	0
left	0
left	1
left	0
right	1
right	0
right	1
right	0

hand	gumball
right	0
right	0
right	1
right	0
right	1
right	0

“What am I looking at here”. This is a long-format table. Each row is one gumball. The first column tells you what hand was used. The second column tells you what kind of gumball. We will say 1s stand for green gum balls, and 0s stand for red gumballs. So, did your left hand cause you to pick more green gumballs than your right hand?

It would be easier to look at the data using a bar graph (Figure 5.10). To keep things simple, we only count the green gumballs (the other gumballs must be red). So, all we need to do is sum up the 1s. The 0s won’t add anything.

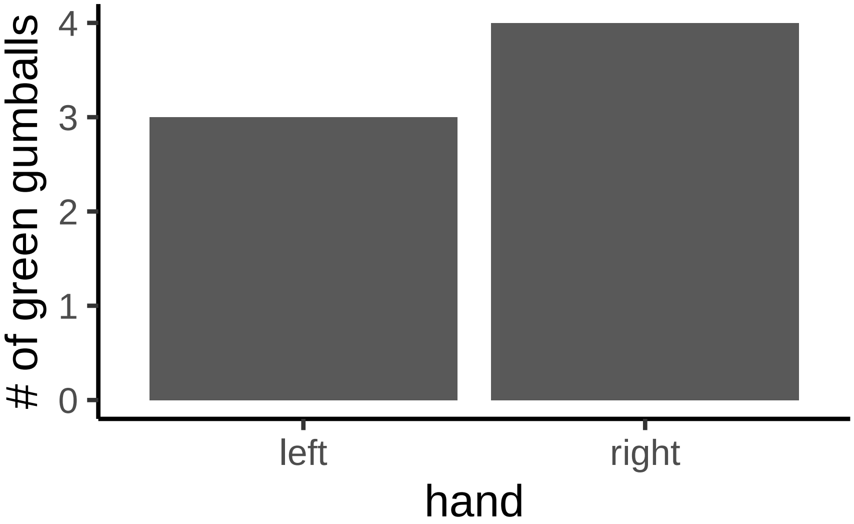


Figure 5.10: Counts of green gumballs picked randomly by each hand.

Oh look, the bars are not the same. One hand picked more green gum balls than the other. Does this mean that one of your hands secretly knows how to find green gumballs? No, it’s just another case of sampling error, that thing we call luck or chance. The difference here is caused by chance, not by the manipulation (which hand you use). **Major problem for inference alert**. We run experiments to look for differences so we can make inferences about whether our manipulations cause change in our measures. However, this example demonstrates that we can find differences by chance. How can we know if a difference is real, or just caused by chance?

5.3.2 Differences due to chance can be simulated

Remember when we showed that chance can produce correlations. We also showed that chance is restricted in its ability to produce correlations. For example, chance more often produces weak correlations than strong correlations. Remember the window of chance? We found out before that correlations falling outside the window of chance were very unlikely. We can do the same thing for differences. Let's find out just what chance can do in our experiment. Once we know what chance is capable of we will be in a better position to judge whether our manipulation caused a difference, or whether it could have been chance.

The first thing to do is pretend you conduct the gumball experiment 10 times in a row. This will produce 10 different sets of results. Figure 5.11 shows bar graphs for each replication of the experiment. Now we can look at whether the left hand chose more green gumballs than red gumballs.

These 10 experiments give us a better look at what chance can do. It should also mesh well with your expectations. If everything is determined by chance (as we have made it so), then sometimes your left hand will choose more green balls, sometimes your right hand will choose more green gumballs, and sometimes they will choose the same amount of gumballs. Right? Right.

5.4 Chance makes some differences more likely than others

OK, we have seen that chance can produce differences here. But, we still don't have a good idea about what chance usually does and doesn't do. For example, if we could find the window of opportunity here, we would be able find out that chance usually does not produce differences of a certain large size. If we knew what the size was, then if we ran experiment and our difference was bigger than what chance can do, we could be confident that chance did not produce our difference.

Let's think about our measure of green balls in terms of a difference. For example, in each experiment we counted the green balls for the left and right hand. What we really want to know is if there is a difference between them. So, we can calculate the **difference score**. Let's decide that the difference score = # of green gumballs in left hand - # of green gumballs in right hand. Figure 5.12 redraws the 10 bar graphs from above; however, now there is only one bar for each experiment. This bar represents the difference in number of green gumballs drawn by the left and right hand.

Missing bars mean that there were an equal number of green gumballs chosen by the left and right hands (difference score is 0). A positive value means that more green gumballs were

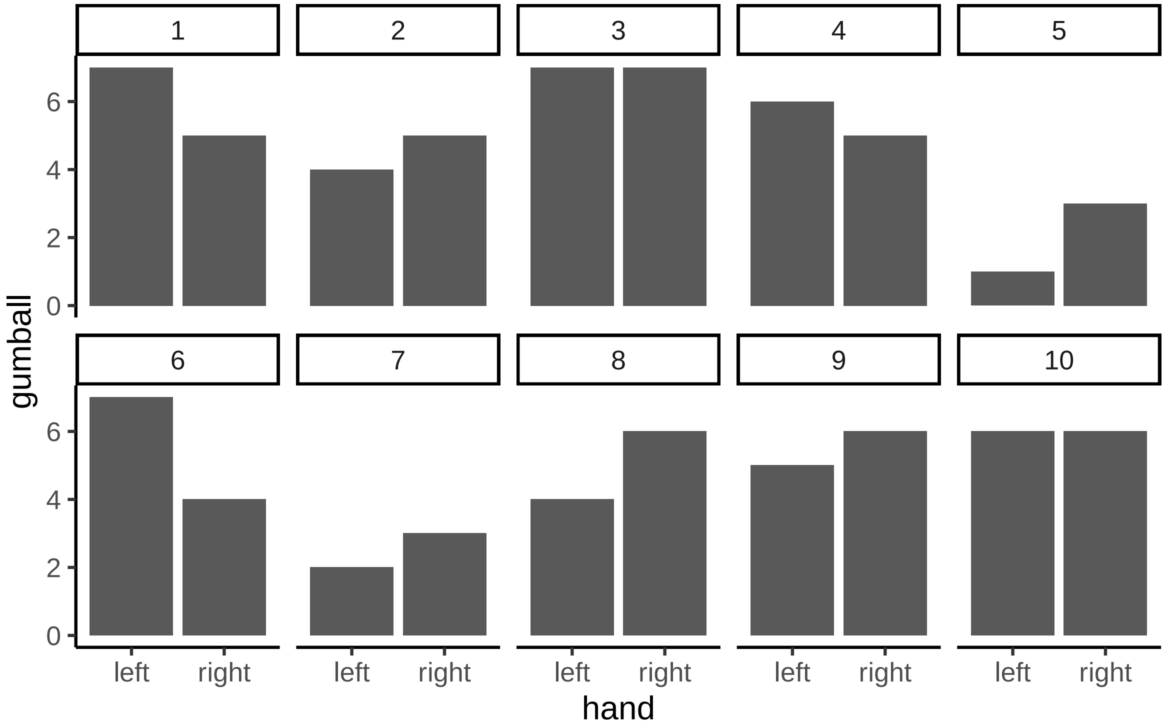


Figure 5.11: 10 simulated replications of picking gumballs. Each replication gives a slightly different answer. Any difference between the bars is due to chance, or sampling error. This shows that chance alone can produce differences, just by the act of sampling.

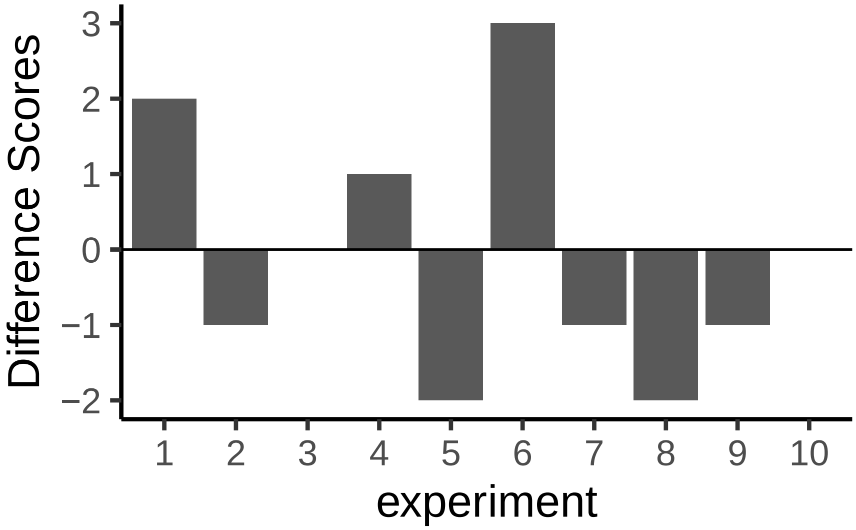


Figure 5.12: A look at the differences between number of each kind of gumball for the different replications. The difference should be zero, but sampling error produces non-zero differences.

chosen by the left than right hand. A negative value means that more green gumballs were chosen by the right than left hand. Note that if we decided (and we get to decide) to calculate the difference in reverse (right hand - left hand), the signs of the differences scores would flip around.

We are starting to see more of the differences that chance can produce. The difference scores are mostly between -2 to +2. We could get an even better impression by running this pretend experiment 100 times instead of only 10 times. The results are shown in Figure 5.13.

Ooph, we just ran so many simulated experiments that the x-axis is unreadable, but it goes from 1 to 100. Each bar represents the difference of number of green balls chosen randomly by the left or right hand. Beginning to notice anything? Look at the y-axis, this shows the size of the difference. Yes, there are lots of bars of different sizes, this shows us that many kinds of differences do occur by chance. However, the y-axis is also restricted. It does not go from -10 to +10. Big differences greater than 5 or -5 don't happen very often.

Now that we have a method for simulating differences due to chance, let's run 10,000 simulated experiments. But, instead of plotting the differences in a bar graph for each experiment, how about we look at the histogram of difference scores. The histogram in Figure 5.14 provides a clearer picture about which differences happen most often, and which ones do not. This will be another window into observing what kinds of differences chance is capable of producing.

Our computer simulation allows us to force chance to operate hundreds of times, each time it produces a difference. We record the difference, then at the end of the simulation we plot the

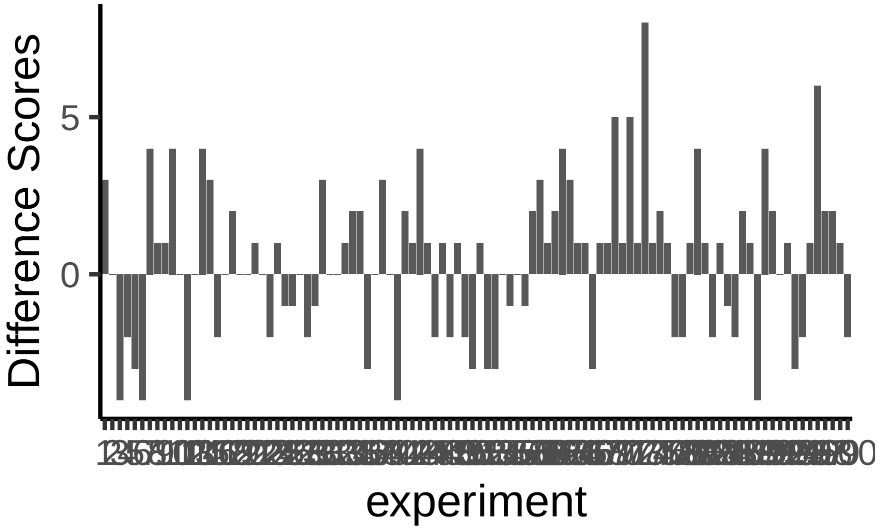


Figure 5.13: Replicating the experiment 100 times, and looking at the differences each time. There are mnay kinds of differences that chance alone can produce.

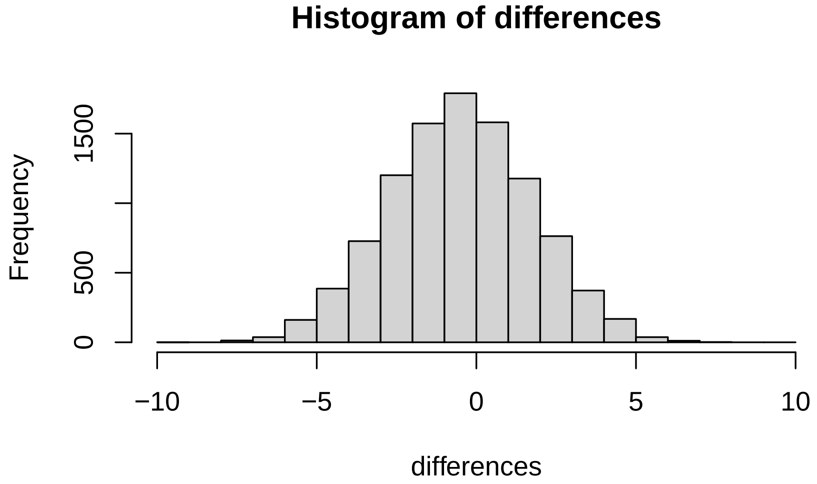


Figure 5.14: A histogram of the differences obtained by chance over 10,000 replications. The most frequency difference is 0, which is what we expect by chance. But the differences can be as large as -10 or +10. Larger differences occur less often by chance. Chance can't do everything.

histogram of the differences. The histogram begins to show us the where the differences came from. Remember the idea that numbers come from a distribution, and the distribution says how often each number occurs. We are looking at one of these distributions. It is showing us that chance produces some differences more often than others. First, chance usually produces 0 differences, that's the biggest bar in the middle. Chance also produces larger differences, but as the differences get larger (positive or negative), they occur less frequently. The shape of this histogram is your chance window, it tells you what chance can do, it tells you what chance usually does, and what it usually does not do.

You can use this chance window to help you make inferences. If you ran yourself in the gumball experiment and found that your left hand chose 2 more green gumballs than red gumballs, would you conclude that you left hand was special, and caused you to choose more green gumballs? Hopefully not. You could look at the chance window and see that differences of size +2 do happen fairly often by chance alone. You should not be surprised if you got a +2 difference. However, what if your left chose 5 more green gumballs than red gumballs. Well, chance doesn't do this very often, you might think something is up with your left hand. If you got a whopping 9 more green gumballs than red gumballs, you might really start to wonder. This is the kind of thing that could happen (it's possible), but virtually never happens by chance. When you get things that almost never happen by chance, you can be more confident that the difference reflects a causal force that is not chance.

5.5 The Crump Test

We are going to be doing a lot of inference throughout the rest of this course. Pretty much all of it will come down to one question. Did chance produce the differences in my data? We will be talking about experiments mostly, and in experiments we want to know if our manipulation caused a difference in our measurement. But, we measure things that have natural variability, so every time we measure things we will always find a difference. We want to know if the difference we found (between our experimental conditions) could have been produced by chance. If chance is a very unlikely explanation of our observed difference, we will make the inference that chance did not produce the difference, and that something about our experimental manipulation did produce the difference. This is it (for this textbook).

> ℹ Note
>
> Statistics is not only about determining whether chance could have produced a pattern in the observed data. The same tools we are talking about here can be generalized to ask whether any kind of distribution could have produced the differences. This allows comparisons between different models of the data, to see which one was the most likely, rather than just rejecting the unlikely ones (e.g., chance). But, we'll leave those advanced topics for another textbook.

This chapter is about building intuitions for making these kinds of inferences about the role of chance in your data. It's not clear to me what are the best things to say, to build up your intuitions for how to do statistical inference. So, this chapter tries different things, some of them standard, and some of them made up. What you are about to read, is a made up way of doing statistical inference, without using the jargon that we normally use to talk about it. The goal is to do things without formulas, and without probabilities, and just work with some ideas using simulations to see what happens. We will look at what chance can do, then we will talk about what needs to happen in your data in order for you to be confident that chance didn't do it.

5.5.1 Intuitive methods

Warning, this is an unofficial statistical test made up by Matt Crump. It makes sense to him (me), and if it turns out someone else already made this up, then Crump didn't do his homework, and we will change the name of this test to it's original author later on. The point of this test is to show how simple arithmetic operations that you already understand can be used to create a statistic tool for inference. This test uses:

1. Sampling numbers randomly from a distribution
2. Adding and subtracting
3. Division, to find the mean
4. Counting
5. Graphing and drawing lines
6. NO FORMULAS

5.5.2 Part 1: Frequency based intuition about occurrence

Question: How many times does something need to happen for it to happen a lot? Or, how many times does something need to happen for it to happen not very much, or even really not at all? Small enough for you to not worry about it at all happening to you?

Would you go outside everyday if you thought that you would get hit by lightning 1 out of 10 times? I wouldn't. You'd probably be hit by lightning more than once per month, you'd be dead pretty quickly. 1 out of 10 is a lot (to me, maybe not to you, there's no right answer here).

Would you go outside everyday if you thought that you would get hit by lightning 1 out of every 100 days? Jeez, that's a tough one. What would I even do? If I went out everyday I'd probably be dead in a year! Maybe I would go out 2 or 3 times per year, I'm risky like that, but I'd probably live longer if I stayed at home forever. It would massively suck.

Would you go outside everyday if you thought you would get hit by lightning 1 out of every 1000 days? Well, you'd probably be dead in 3-6 years if you did that. Are you a gambler? Maybe go out once per month, still sucks.

Would you go outside everyday if you thought lightning would get you 1 out every 10,000 days? 10,000 is a bigger number, harder to think about. It translates to getting hit about once every 27 years. Ya, I'd probably go out 150 days per year, and keep my fingers crossed.

Would you go outside everyday if you thought lightning would get you 1 out every 100,000 days? How many years is that? It's about 273 years. With those odds, I'd probably go out all the time and forget about being hit by lightning. It doesn't happen very often.

The point of considering these questions is to get a sense for yourself of what happens a lot, and what doesn't happen a lot, and how you would make important decisions based on what happens a lot and what doesn't.

5.5.3 Part 2: Simulating chance

This next part could happen a bunch of ways, I'll make loads of assumptions that I won't defend, and I won't claim the Crump test has problems. I will claim it helps us make an inference about whether chance could have produced some differences in data. We've already been introduced to simulating things, so we'll do that again. Here is what we will do. I am a cognitive psychologist who happens to be measuring X. Because of prior research in the field, I know that when I measure X, my samples will tend to have a particular mean and standard deviation. Let's say the mean is usually 100, and the standard deviation is usually 15. In this case, I don't care about using these numbers as estimates of the population parameters, I'm just thinking about what my samples usually look like. What I want to know is how they behave when I sample them. I want to see what kind of samples happen a lot, and what kind of samples don't happen a lot. Now, I also live in the real world, and in the real world when I run experiments to see what changes X, I usually only have access to some number of participants, who I am very grateful too, because they participate in my experiments. Let's say I usually can run 20 subjects in each condition in my experiments. Let's keep the experiment simple, with two conditions, so I will need 40 total subjects.

I would like to learn something to help me with inference. One thing I would like to learn is what the sampling distribution of the sample mean looks like. This distribution tells me what kinds of mean values happen a lot, and what kinds don't happen very often. But, I'm actually going to skip that bit. Because what I'm really interested in is what the **sampling distribution of the difference between my sample means** looks like. After all, I am going to run an experiment with 20 people in one condition, and 20 people in the other. Then I am going to calculate the mean for group A, and the mean for group B, and I'm going to look a the difference. I will probably find a difference, but my question is, did my manipulation cause this difference, or is this the kind of thing that happens a lot by chance. If I knew what chance can do, and how often it produces differences of particular sizes, I could look at the difference I observed, then look at what chance can do, and then I can make a decision! If my difference doesn't happen a lot (we'll get to how much not a lot is in a bit), then I might be willing to believe that my manipulation caused a difference. If my difference happens all the time by chance alone, then I wouldn't be inclined to think my manipulation caused the difference, because it could have been chance.

So, here's what we'll do, even before running the experiment. We'll do a simulation. We will sample numbers for group A and Group B, then compute the means for group A and group B, then we will find the difference in the means between group A and group B. But, we will do one very important thing. We will pretend that we haven't actually done a manipulation. If we do this (do nothing, no manipulation that could cause a difference), then we know that **only sampling error** could cause any differences between the mean of group A and group B. We've eliminated all other causes, only chance is left. By doing this, we will be able to see exactly what chance can do. More importantly, we will see the kinds of differences that occur a lot, and the kinds that don't occur a lot.

Before we do the simulation, we need to answer one question. How much is a lot? We could pick any number for a lot. I'm going to pick 10,000. That is a lot. If something happens only 1 times out 10,000, I am willing to say that is not a lot.

OK, now we have our number, we are going to simulate the possible mean differences between group A and group B that could arise by chance. We do this 10,000 times. This gives chance a lot of opportunity to show us what it does do, and what it does not do.

This is what I did: I sampled 20 numbers into group A, and 20 into group B. The numbers both came from the same normal distribution, with mean = 100, and standard deviation = 15. Because the samples are coming from the same distribution, we expect that on average they will be similar (but we already know that samples differ from one another). Then, I compute the mean for each sample, and compute the difference between the means. I save the **mean difference score**, and end up with 10,000 of them. Then, I draw the histogram in Figure 5.15.

ℹ Note

Of course, we might recognize that chance could do a difference greater than 15. We just didn't give it the opportunity. We only ran the simulation 10,000 times. If we ran it a million times, maybe a difference greater than 15 or even 20 would happen a couple times. If we ran it a bazillion gazillion times, maybe a difference greater than 30 would happen a couple times. If we go out to infinity, then chance might produce all sorts of bigger differences once in a while. But, we've already decided that 1/10,000 is not a lot. So things that happen 0 out of 10,000 times, like differences greater than 15, are considered to be extremely unlikely.

Now we can see what chance can do to the size of our mean difference. The x-axis shows the size of the mean difference. We took our samples from the sample distribution, so the

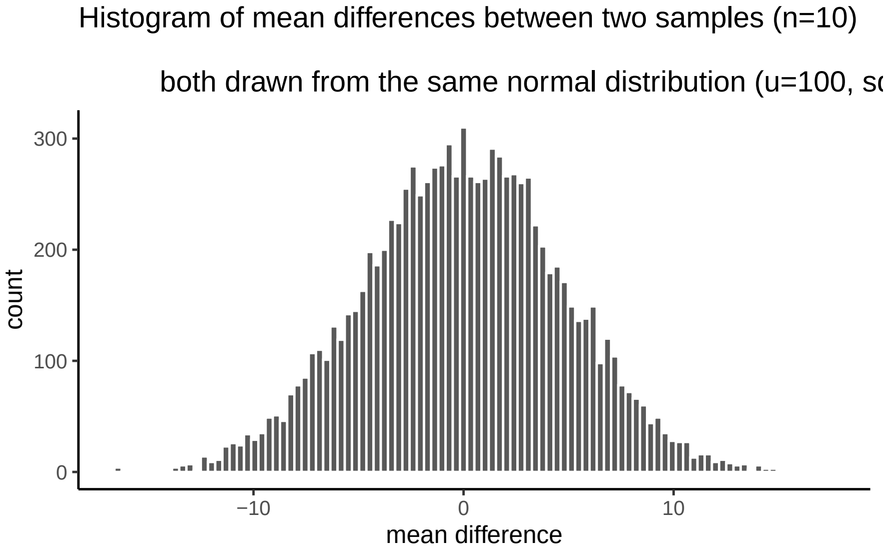


Figure 5.15: Histogram of mean differences arising by chance.

difference between them should usually be 0, and that's what we see in the histogram.

Pause for a second. Why should the mean differences usually be zero, wasn't the population mean = 100, shouldn't they be around 100? No. The mean of group A will tend to be around 100, and the mean of group B will tend be around 100. So, the difference score will tend to be 100-100 = 0. That is why we expect a mean difference of zero when the samples are drawn from the same population.

So, differences near zero happen the most, that's good, that's what we expect. Bigger or smaller differences happen increasingly less often. Differences greater than 15 or -15 never happen at all. For our purposes, it looks like chance only produces differences between -15 to 15.

OK, let's ask a couple simple questions. What was the biggest negative number that occurred in the simulation? We'll use R for this. All of the 10,000 difference scores are stored in a variable I made called `difference`. If we want to find the minimum value, we use the `min` function. Here's the result.

```
min(difference)
#> [1] -16.49206
```

OK, so what was the biggest positive number that occurred? Let's use the `max` function to find out. It finds the biggest (maximum) value in the variable. FYI, we've just computed the range, the minimum and maximum numbers in the data. Remember we learned that before. Anyway, here's the max.

```
max(difference)
#> [1] 17.43542
```

Both of these extreme values only occurred once. Those values were so rare we couldn't even see them on the histogram, the bar was so small. Also, these biggest negative and positive numbers are pretty much the same size if you ignore their sign, which makes sense because the distribution looks roughly symmetrical.

So, what can we say about these two numbers for the min and max? We can say the min happens 1 times out of 10,000. We can say the max happens 1 times out of 10,000. Is that a lot of times? Not to me. It's not a lot.

So, how often does a difference of 30 (much larger larger than the max) occur out of 10,000. We really can't say, 30s didn't occur in the simulation. Going with what we got, we say 0 out of 10,000. That's never.

We're about to move into part three, which involves drawing decision lines and talking about them. The really important part about part 3 is this. What would you say if you ran this experiment once, and found a mean difference of 30? I would say it happens 0 times of out 10,000 by chance. I would say chance did not produce my difference of 30. That's what I would say. We're going to expand upon this right now.

5.5.4 Part 3: Judgment and Decision-making

Remember, we haven't even conducted an experiment. We're just simulating what could happen if we did conduct an experiment. We made a histogram. We can see that chance produces some differences more than others, and that chance never produced really big differences. What should we do with this information?

What we are going to do is talk about judgment and decision making. What kind of judgment and decision making? Well, when you finally do run an experiment, you will get two means for group A and B, and then you will need to make some judgments, and perhaps even a decision, if you are so inclined. You will need to judge whether chance (sampling error) could have produced the difference you observed. If you judge that it did it not, you might make the decision to tell people that your experimental manipulation actually works. If you judge that it could have been chance, you might make a different decision. These are important decisions for researchers. Their careers can depend on them. Also, their decisions matter for the public. Nobody wants to hear fake news from the media about scientific findings.

So, what we are doing is preparing to make those judgments. We are going to draw up a plan, before we even see the data, for how we will make judgments and decisions about what we find. This kind of planning is extremely important, because we discuss in part 4, that your planning can help you design an even better experiment than the one you might have been intending to run. This kind of planning can also be used to interpret other people's results, as a way of double-checking checking whether you believe those results are plausible.

The thing about judgement and decision making is that reasonable people disagree about how to do it, unreasonable people really disagree about it, and statisticians and researchers disagree about how to do it. I will propose some things that people will disagree with. That's OK, these things still make sense. And, the disagreeable things point to important problems that are very real for any "real" statistical inference test.

Let's talk about some objective facts from our simulation of 10,000 things that we definitely know to be true. For example, we can draw some lines on the graph, and label some different regions. We'll talk about two kinds of regions.

1. Region of chance. Chance did it. Chance could have done it
2. Region of not chance. Chance didn't do it. Chance couldn't have done it.

The regions are defined by the minimum value and the maximum value. Chance never produced a smaller or bigger number. The region inside the range is what chance did do, and the the region outside the range on both sides is what chance never did. It looks like Figure 5.16:

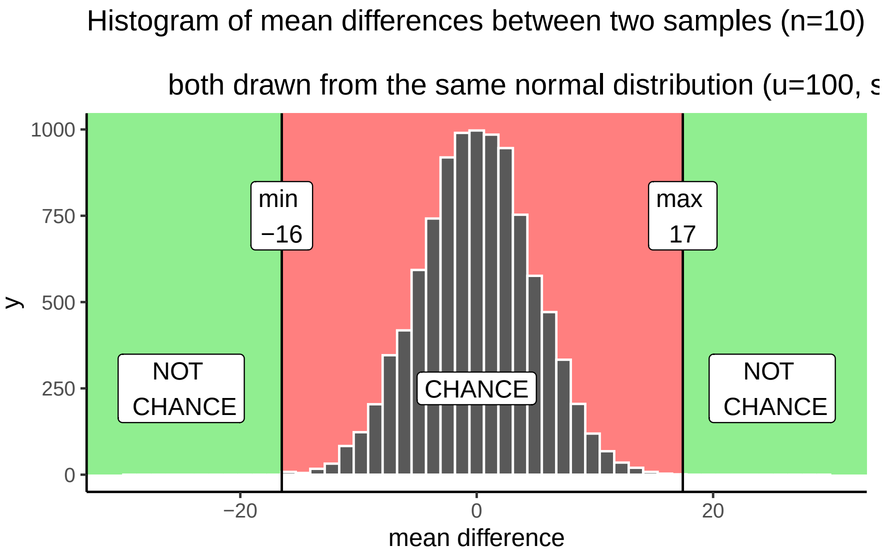


Figure 5.16: Applying decision boundaries to the histogrm of mean differences. The boundaries identify what differences chance did or did not produce in the simulation.

We have just drawn some lines, and shaded some regions, and made one plan we could use to make decisions. How would the decisions work. Let's say you ran the experiment and found a mean difference between groups A and B of 25. Where is 25 in the figure? It's in the green part. What does the green part say? NOT CHANCE. What does this mean. It means chance never made a difference of 25. It did that 0 out of 10,000 times. If we found a difference of 25, perhaps we could confidently conclude that chance did not cause the difference. If I found a difference of 25 with this kind of data, I'd be pretty confident chance did not cause the difference; and, I would give myself license to consider that my experimental manipulation may be causing the difference.

What about a difference of +10? That's in the red part, where chance lives. Chance could have done a difference of +10 because we can see that it did do that sometimes. The red part is the window of what chance did in our simulation. Anything inside the window could have been a difference caused by chance. If I found a difference of +10, I'd say, "ya, it coulda been chance." I would also be less confident that the difference was only caused by my experimental manipulation.

Statistical inference could be this easy. The number you get from your experiment could be in the chance window (then you can't rule out chance as a cause), or it could be outside the chance window (then you can rule out chance). Case closed. Let's all go home.

5.5.4.1 Grey areas

So what's the problem? Depending on who you are, and what kinds of risks you're willing to take, there might not be a problem. But, if you are just even a little bit risky then there is a problem that makes clear judgments about the role of chance difficult. We would like to say chance did or did not cause our difference. But, we're really always in the position of admitting that it could have sometimes, or wouldn't have most times. These are wishy washy statements, they are in between yes or no. That's OK. Grey is a color too, let's give grey some respect.

"What grey areas are you talking about?, I only see red or green, am I grey blind?". Let's look at where some grey areas might be. I say might be, because people disagree about where the grey is. People have different comfort levels with grey. Figure 5.17 shows my opinion on grey areas.

I made two grey areas, and they are reddish grey, because we are still in the chance window. There are question marks (?) in the grey areas. Why? The question marks reflect some uncertainty that we have about those particular differences. For example, if you found a difference that was in a grey area, say a 15. 15 is less than the maximum, which means chance did create differences of around 15. But, differences of 15 don't happen very often.

What can you conclude or say about this 15 you found? Can you say without a doubt that chance did not produce the difference? Of course not, you know that chance could have. Still, it's one of those things that doesn't happen a lot. That makes chance an unlikely explanation. Instead of thinking that chance did it, you might be willing to take a risk and say that your experimental manipulation caused the difference. You'd be making a bet that it wasn't chance…but, could be a safe bet, since you know the odds are in your favor.

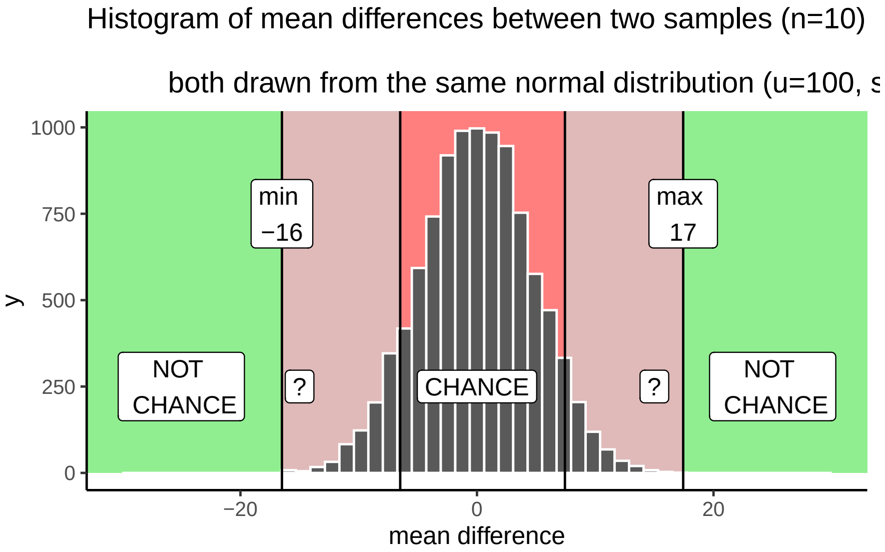


Figure 5.17: The question marks refer to an area where you have some uncertainty. Differences inside the question mark region do not happen very often by chance. When you find differences of these sizes, should you reject the idea that chance caused your difference? You will always have some uncertainty associated with this decision because it is clear that chance could have caused the difference. But, chance usually does not produce differences of these sizes.

You might be thinking that your grey areas aren't the same as the ones I've drawn. Maybe you want to be more conservative, and make them smaller. Or, maybe you're more risky, and would make them bigger. Or, maybe you'd add some grey area going in a little bit to the green area (after all, chance could probably produce some bigger differences sometimes, and to avoid those you would have to make the grey area go a bit into the green area).

Another thing to think about is your decision policy. What will you do, when your observed difference is in your grey area? Will you always make the same decision about the role of chance? Or, will you sometimes flip-flop depending on how you feel. Perhaps, you think that there shouldn't be a strict policy, and that you should accept some level of uncertainty. The difference you found could be a real one, or it might not. There's uncertainty, hard to avoid that.

So let's illustrate one more kind of strategy for making decisions. We just talked about one that had some lines, and some regions. This makes it seem like a binary choice: we can either rule out, or not rule out the role of chance. Another perspective is that everything is a different shade of grey, like in Figure 5.18.

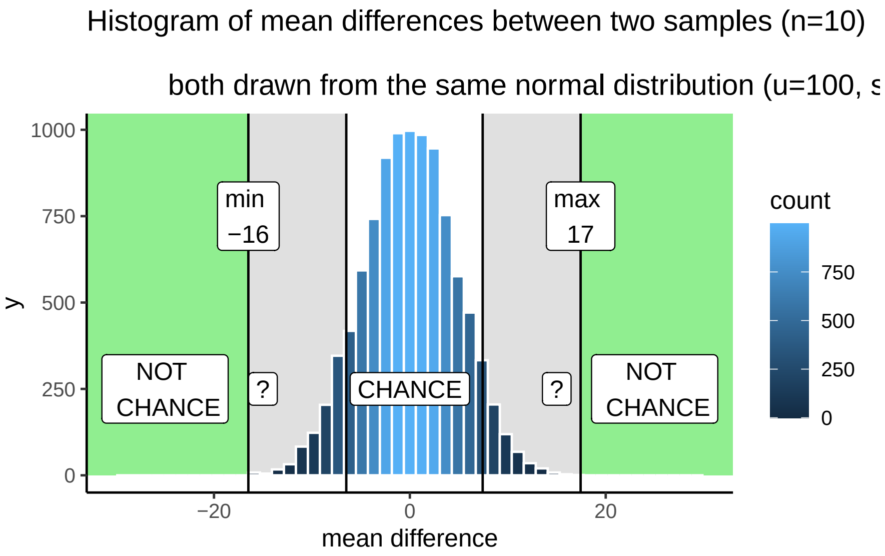


Figure 5.18: The shading of the blue bars indicates levels of confidence in whether a difference could have been produced by chance. Darker bars represent increased confidence that the difference was not produced by chance. Bars get darker as the mean difference increases in absolute value.

OK, so I made it shades of blue (because it was easier in R). Now we can see two decision plans at the same time. Notice that as the bars get shorter, they also get become a darker stronger blue. The color can be used as a guide for your confidence. That is, your confidence in the belief that your manipulation caused the difference rather than chance. If you found a difference near a really dark bar, those don't happen often by chance, so you might be really confident that chance didn't do it. If you find a difference near a slightly lighter blue bar, you might be slightly less confident. That is all. You run your experiment, you get your data, then you have some amount of confidence that it wasn't produced by chance. This way of thinking is elaborated to very interesting degrees in the Bayesian world of statistics. We don't wade too much into that, but mention it a little bit here and there. It's worth knowing it's out there.

5.5.4.2 Making decisions and being wrong

No matter how you plan to make decisions about your data, you will always be prone to making some mistakes. You might call one finding real, when in fact it was caused by chance. This is called a **type I** error, or a false positive. You might ignore one finding, calling it chance, when in fact it wasn't chance (even though it was in the window). This is called a **type II** error, or a false negative.

How you make decisions can influence how often you make errors over time. If you are a researcher, you will run lots of experiments, and you will make some amount of mistakes over time. If you do something like the very strict method of only accepting results as real when they are in the "no chance" zone, then you won't make many type I errors. Pretty much all of your result will be real. But, you'll also make type II errors, because you will miss things real things that your decision criteria says are due to chance. The opposite also holds. If you are willing to be more liberal, and accept results in the grey as real, then you will make more type I errors, but you won't make as many type II errors. Under the decision strategy of using these cutoff regions for decision-making there is a necessary trade-off. The Bayesian view get's around this a little bit. Bayesians talk about updating their beliefs and confidence over time. In that view, all you ever have is some level of confidence about whether something is real, and by running more experiments you can increase or decrease your level of confidence. This, in some fashion, avoids some trade-off between type I and type II errors.

Regardless, there is another way to reduce type I and type II errors, and to increase your confidence in your results, even before you do the experiment. It's called "knowing how to design a good experiment".

5.5.5 Part 4: Experiment Design

We've seen what chance can do. Now, let's venture into an experiment. We make a change between ecosystems A and B, gather the data, assess the average outcomes, and then observe the variance. Then we keep our fingers crossed, hoping that the variance is significant enough to be beyond natural fluctuations. Yes, nature keeps us guessing.

Here's the catch, we aren't always certain about the magnitude of our environmental interventions. So, even if an intervention induces a change, pinning down its exact magnitude can be challenging. And that's the essence of our experiment. Many interventions in Environmental Science might not cause large-scale shifts. This poses a challenge in identifying these subtle, yet potentially crucial, environmental effects. In a hypothetical scenario, introducing a certain pollinator species might influence plant growth, but to what extent? If the difference is marginal, differentiating between natural variation and the effect of our intervention becomes tricky. Let's say our intervention involves introducing shade in one ecosystem versus none in the other. While shade can influence plant growth, if the effect is only marginal, it becomes hard to ascertain if it wasn't just a natural occurrence. And, it's not straightforward to intensify the shading to amplify its impact, without risking other unintended consequences.

EXPERIMENT DESIGN TO THE RESCUE! Newsflash, it is often possible to change how you run your experiment so that it is **more sensitive** to smaller effects. How do you think we can do this? Here is a hint. It's the stuff you learned about the sampling distribution of the sample mean, and the role of sample-size. What happens to the sampling distribution of the sample mean when N (sample size)? The distribution gets narrower and narrower, and starts to look the a single number (the hypothetical mean of the hypothetical population). That's great. If you switch to thinking about mean difference scores, like the distribution we created in this test, what do you think will happen to that distribution as we increase N? It will will also shrink. As we increase N to infinity, it will shrink to 0. Which means that, when N is infinity, chance never produces any differences at all. We can use this.

For example, we could run our experiment with 20 subjects in each group. Or, we could decide to invest more time and run 40 subjects in each group, or 80, or 150. When you are the experimenter, you get to decide the design. These decisions matter big time. Basically, the more subjects you have, the more sensitive your experiment. With bigger N, you will be able to reliably detect smaller mean differences, and be able to confidently conclude that chance did not produce those small effects.

Check out the histograms in Figure 5.19. This is the same simulation as before, but with four different sample-sizes: 20, 40, 80, 160. We are doubling our sample-size across each simulation just to see what happens to the width of the chance window.

There you have it. The **sampling distribution of the mean differences** shrinks toward 0 as sample-size increases. This means if you run an experiment with a larger sample-size, you will be able to detect smaller mean differences, and be confident they aren't due to chance. Table 5.3 contains the minimum and maximum values that chance produced across the four sample-sizes:

Table 5.3: The smallest and largest mean differences produced by chance as a function of sample-size.

sample_size	smallest	largest
20	-23.43563	22.95795
40	-15.62940	15.37448
80	-11.80469	11.73815
160	-10.62788	11.61204

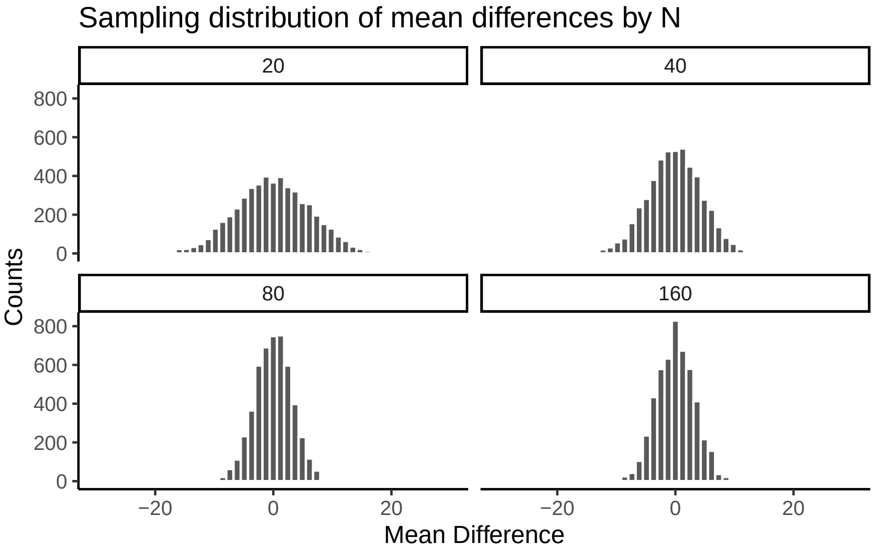


Figure 5.19: The range or width of the differences produced by chance shrinks as sample-size increases.

The table shows the range of chance behavior is very wider for smaller N and narrower for larger N. Consider what this narrowing means for your experiment design. For example, one aspect of the design is the choice of sample size, N, or in a psychology experiment the number of participants.

If it turns out your manipulation will cause a difference of +11, then what should you do? Run an experiment with N = 20 people? I hope not. If you did that, you could get a mean difference of +11 fairly often by chance. However, if you ran the experiment with 160 people, then you would definitely be able to say that +11 was not due to chance, it would be outside the range of what chance can do. You could even consider running the experiment with 80 subjects. A +11 there wouldn't happen often by chance, and you'd be cost-effective, spending less time on the experiment.

The point is: **the design of the experiment determines the sizes of the effects it can detect**. If you want to detect a small effect. Make your sample size bigger. It's really important to say this is not the only thing you can do. You can also make your cell-sizes bigger. For example, often times we take several measurements from a single subject. The more measurements you take (cell-size), the more stable your estimate of the subject's mean. We discuss these issues more later. You can also make a stronger manipulation, when possible.

5.5.6 Part 5: I have the power

> By the power of greyskull, I HAVE THE POWER - He-man

The last topic in this section is called **power**. Later we will define power in terms of some particular ideas about statistical inference. Here, we will just talk about the big idea. And, we'll show how to make sure your design has 100% power. Because, why not. Why run a design that doesn't have the power?

The big idea behind power is the concept of sensitivity. The concept of sensitivity assumes that there is something to be sensitive to. That is, there is some real difference that can be measured. So, the question is, how sensitive is your experiment? We've already seen that the number of subjects (sample-size), changes the sensitivity of the design. More subjects = more sensitivity to smaller effects.

Let's take a look at one more plot. What we will do is simulate a measure of sensitivity across a whole bunch of sample sizes, from 10 to 300. We'll do this in steps of 10. For each simulation, we'll compute the mean differences as we have done. But, rather than showing the histogram, we'll just compute the smallest value and the largest value. This is a pretty good measure of the outer reach of chance. Then we'll plot those values as a function of sample size and see what we've got.

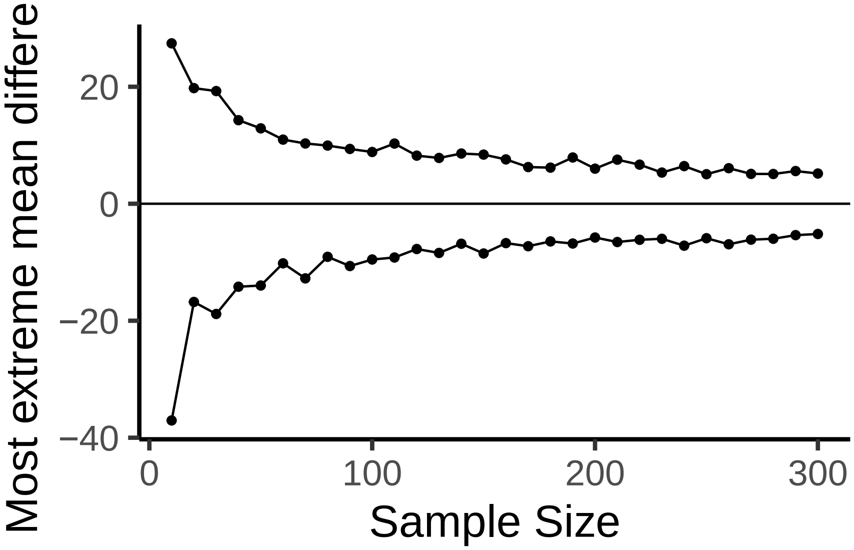


Figure 5.20: A graph of the maximum and minimum mean differences produced by chance as a function of sample-size. The range narrows as sample-size increases showing that chance alone produces a smaller range of mean differences as sample-size increases.

Figure 5.20 shows a reasonably precise window of sensitivity as a function of sample size. For each sample size, we can see the maximum difference that chance produced and the minimum difference. In those simulations, chance never produced bigger or smaller differences. So, each design is sensitive to any difference that is underneath the bottom line, or above the top line.

Here's another way of putting it. Which of the sample sizes will be sensitive to a difference of +10 or -10. That is, if a difference of +10 or -10 was observed, then we could very confidently say that the difference was not due to chance, because according to these simulations, chance never produced differences that big. To help us see which ones are sensitive, Figure 5.21 draws horizontal lines at -10 and +10.

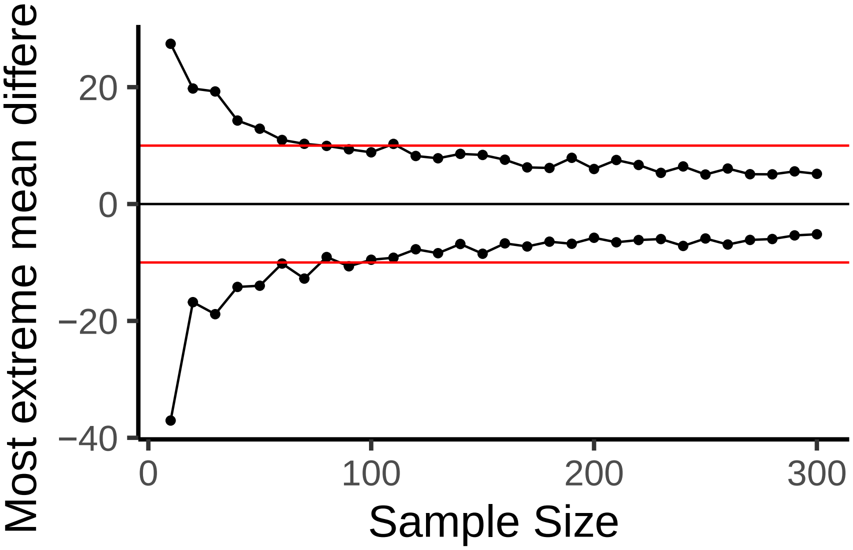


Figure 5.21: The red line represents the size of a mean difference that a researcher may be interested in detecting. All of the dots outside (above or below) the red line represent designs with small sample-sizes. When a difference of 10 occurs for these designs, we can rule out chance with confidence. The dots between the red lines represent designs with larger sample-sizes. These designs never produce differences as large as 10, so when those differences occur, we can be confident chance did not produce them.

Based on visual guesstimation, the designs with sample-size >= 100 are all sensitive to real differences of 10. Designs with sample-size > 100 all failed to produce extreme differences outside of the red lines by chance alone. If these designs were used, and if an effect of 10 or larger was observed, then we could be confident that chance alone did not produce the effect. Designing your experiment so that you know it is sensitive to the thing you are looking for is the big idea behind power.

5.5.7 Summary of Crump Test

What did we learn from this so-called fake Crump test that nobody uses? Well, we learned the basics of what we'll be doing moving forward. And, we did it all without any hard math or formulas. We sampled numbers, we computed means, we subtracted means between experimental conditions, then we repeated that process many times and counted up the mean differences and put them in a histogram. This showed us what chance do in an experiment. Then, we discussed how to make decisions around these facts. And, we showed how we can control the role of chance just by changing things like sample size.

5.6 The randomization test (permutation test)

Welcome to the first official inferential statistic in this textbook. Up till now we have been building some intuitions for you. Next, we will get slightly more formal and show you how we can use random chance to tell us whether our experimental finding was likely due to chance or not. We do this with something called a randomization test. The ideas behind the randomization test are the very same ideas behind the rest of the inferential statistics that we will talk about in later chapters. And, surprise, we have already talked about all of the major ideas already. Now, we will just put the ideas together, and give them the name **randomization test**.

Here's the big idea. When you run an experiment and collect some data you get to find out what happened that one time. But, because you ran the experiment only once, you don't get to find out what **could have happened**. The randomization test is a way of finding out what **could have happened**. And, once you know that, you can compare **what did happen** in your experiment, with **what could have happened**.

5.6.1 Pretend example does chewing gum improve your grades?

Let's say you run an experiment to find out if chewing gum causes students to get better grades on statistics exams. You randomly assign 20 students to the chewing gum condition, and 20 different students to the no-chewing gum condition. Then, you give everybody statistics tests and measure their grades. If chewing gum causes better grades, then the chewing gum group should have higher grades on average than the group who did not chew gum.

Let's say the data looked like this:

student	gum	no_gum
1	88	51
2	84	59
3	86	60
4	75	58
5	72	51
6	96	74
7	99	57
8	100	53

student	gum	no_gum
9	96	59
10	79	89
11	85	52
12	74	44
13	99	55
14	72	78
15	97	64
16	82	47
17	83	42
18	75	68
19	82	51
20	86	52
Sums	1710	1164
Means	85.5	58.2

So, did the students chewing gum do better than the students who didn't chew gum? Look at the mean test performance at the bottom of the table. The mean for students chewing gum was 85.5, and the mean for students who did not chew gum was 58.2. Just looking at the means, it looks like chewing gum worked!

"STOP THE PRESSES, this is silly". We already know this is silly because we are making pretend data. But, even if this was real data, you might think, "Chewing gum won't do anything, this difference could have been caused by chance, I mean, maybe the better students just happened to be put into the chewing group, so because of that their grades were higher, chewing gum didn't do anything…". We agree. But, let's take a closer look. We already know how the data come out. What we want to know is how they could have come out, what are all the possibilities?

For example, the data would have come out a bit different if we happened to have put some of the students from the gum group into the no gum group, and vice versa. Think of all the ways you could have assigned the 40 students into two groups, there are lots of ways. And, the means for each group would turn out differently depending on how the students are assigned to each group.

Practically speaking, it's not possible to run the experiment every possible way, that would take too long. But, we can nevertheless estimate how all of those experiments might have turned out using simulation.

Here's the idea. We will take the 40 measurements (exam scores) that we found for all the students. Then we will randomly take 20 of them and pretend they were in the gum group, and we'll take the remaining 20 and pretend they were in the no gum group. Then we can compute the means again to find out what would have happened. We can keep doing this over and over again. Every time computing what happened in that version of the experiment.

5.6.1.1 Doing the randomization

Before we do that, let's show how the randomization part works. We'll use fewer numbers to make the process easier to look at. Here are the first 5 exam scores for students in both groups.

student	gum	no_gum
1	88	51
2	84	59
3	86	60
4	75	58
5	72	51
Sums	405	279
Means	81	55.8

Things could have turned out differently if some of the subjects in the gum group were switched with the subjects in the no gum group. Here's how we can do some random switching. We will do this using R.

```
all_scores        <- c(gum[1:5],no_gum[1:5])
randomize_scores <- sample(all_scores)
new_gum           <- randomize_scores[1:5]
new_no_gum        <- randomize_scores[6:10]
print(new_gum)
#> [1] 51 86 88 75 51
print(new_no_gum)
#> [1] 59 58 60 84 72
```

We have taken the first 5 numbers from the original data, and put them all into a variable called `all_scores`. Then we use the `sample` function in R to shuffle the scores. Finally, we take the first 5 scores from the shuffled numbers and put them into a new variable called `new_gum`. Then, we put the last five scores into the variable `new_no_gum`. Then we printed them, so we can see them.

If we do this a couple of times and put them in a table, we can indeed see that the means for gum and no gum would be different if the subjects were shuffled around. Check it out:

student	gum	no_gum	gum2	no_gum2	gum3	no_gum3
1	88	51	75	60	51	72
2	84	59	58	72	86	84
3	86	60	59	88	60	88
4	75	58	86	51	51	59
5	72	51	51	84	75	58
Sums	405	279	329	355	323	361
Means	81	55.8	65.8	71	64.6	72.2

5.6.1.2 Simulating the mean differences across the different randomizations

In our pretend experiment we found that the mean for students chewing gum was 85.5, and the mean for students who did not chew gum was 58.2. The mean difference (gum - no gum) was 27.3. This is a pretty big difference. This is **what did happen**. But, **what could have happened**? If we tried out all of the experiments where different subjects were switched around, what does the distribution of the possible mean differences look like? Let's find out. This is what the randomization test is all about.

When we do our randomization test we will measure the mean difference in exam scores between the gum group and the no gum group. Every time we randomize we will save the mean difference.

Let's look at a short animation of what is happening in the randomization test. **?@fig-5randtest** shows data from a different fake experiment, but the principles are the same. We'll return to the gum no gum experiment after the animation. The animation is showing three important things. First, the purple dots show the mean scores in two groups (didn't study vs study). It looks like there is a difference, as 1 dot is lower than the other. We want to know if chance could produce a difference this big. At the beginning of the animation, the light green and red dots show the individual scores from each of 10 subjects in the design (the purple dots are the means of these original scores). Now, during the randomizations, we randomly shuffle the original scores between the groups. You can see this happening throughout the animation, as the green and red dots appear in different random combinations. The moving yellow dots show you the new means for each group after the randomization. The differences between the yellow dots show you the range of differences that chance could produce.

We are engaging in some visual statistical inference. By looking at the range of motion of the yellow dots, we are watching what kind of differences chance can produce. In this animation, the purple dots, representing the original difference, are generally outside of the range of chance. The yellow dots don't move past the purple dots, as a result chance is an unlikely explanation of the difference.

If the purple dots were inside the range of the yellow dots, then when would know that chance is capable of producing the difference we observed, and that it does so fairly often. As a result, we should not conclude the manipulation caused the difference, because it could have easily occurred by chance.

Let's return to the gum example. After we randomize our scores many times, and computed the new means, and the mean differences, we will have loads of mean differences to look at, which we can plot in a histogram. The histogram gives a picture of **what could have happened**. Then, we can compare **what did happen** with **what could have happened**.

Here's the histogram of the mean differences from the randomization test. For this simulation, we randomized the results from the original experiment 1000 times. This is what could have happened. The blue line in Figure 5.22 shows where the observed difference lies on the x-axis.

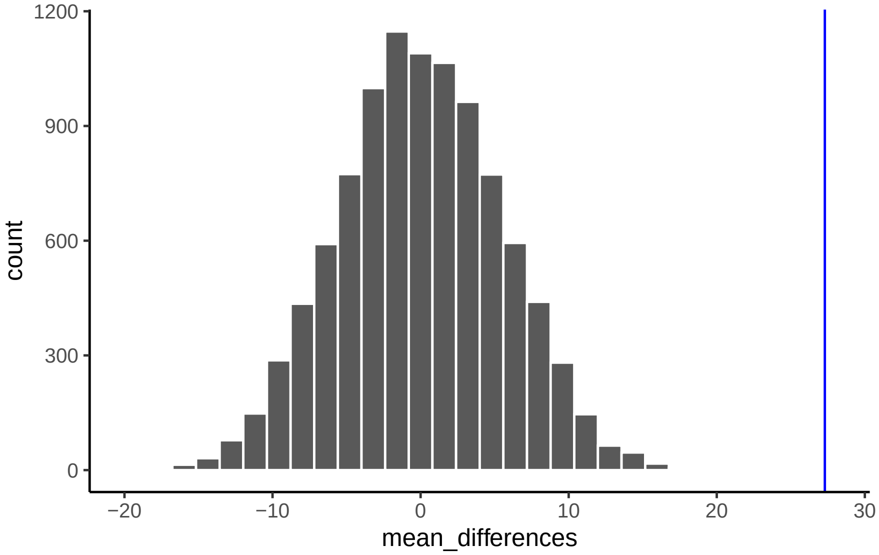


Figure 5.22: A histogram of simulated mean differences for a randomization test

What do you think? Could the difference represented by the blue line have been caused by chance? My answer is probably not. The histogram shows us the window of chance. The blue line is not inside the window. This means we can be pretty confident that the difference we observed was not due to chance.

We are looking at another window of chance. We are seeing a histogram of the kinds of mean differences that could have occurred in our experiment, if we had assigned our subjects to the gum and no gum groups differently. As you can see, the mean differences range from negative to positive. The most frequent difference is 0. Also, the distribution appears to be symmetrical about zero, which shows we had roughly same the chances of getting a positive or negative difference. Also, notice that as the differences get larger (in the positive or negative direction, they become less frequent). The blue line shows us **the observed difference**, this is the one we found in our fake experiment. Where is it? It's way out to the right. It is is well outside the histogram. In other words, when we look at **what could have happened**, we see that **what did happen** doesn't occur very often.

IMPORTANT: In this case, when we speak of **what could have happened**. We are talking about what could have happened **by chance**. When we compare what did happen to what chance could have done, we can get a better idea of whether our result was caused by chance.

OK, let's pretend we got a much smaller mean difference when we first ran the experiment. We can draw new lines (blue and red) to represent a smaller mean that we might have found.

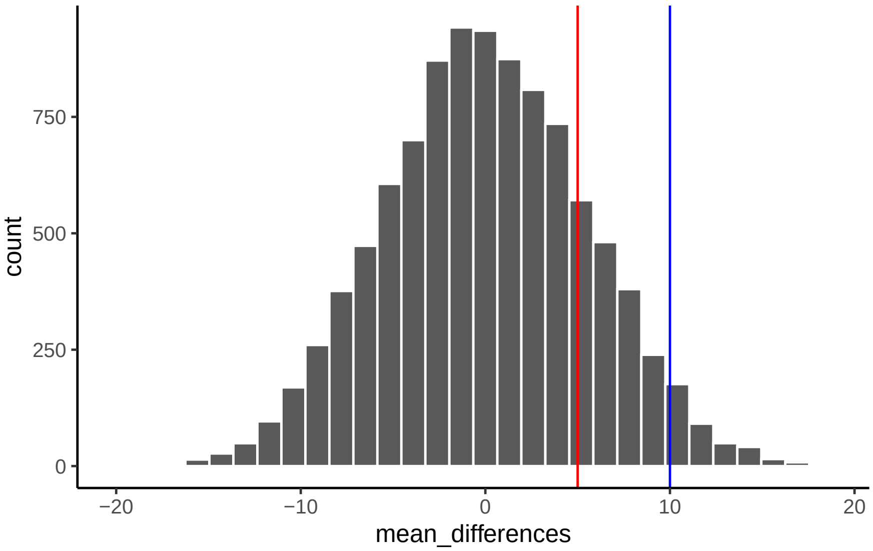


Figure 5.23: Would you expect a mean difference represented by the blue line to occur more or less often by chance compared to the mean difference represented by the red line?

Look at the blue line in Figure 5.23. If you found a mean difference of 10, would you be convinced that your difference was not caused by chance? As you can see, the blue line is inside the chance window. Notably, differences of +10 don't very often. You might infer that your difference was not likely to be due to chance (but you might be a little bit skeptical, because it could have been). How about the red line? The red line represents a difference of +5. If you found a difference of +5 here, would you be confident that your difference was not caused by chance? I wouldn't be. The red line is totally inside the chance window, this kind of difference happens fairly often. I'd need some more evidence to consider the claim the some independent variable actually caused the difference. I'd be much more comfortable assuming that sampling error probably caused the difference.

5.6.2 Take homes so far

Have you noticed that we haven't used any formulas yet, but we have been able to accomplish inferential statistics. We will see some formulas as we progress, but these aren't as the idea behind the formulas.

Inferential statistics is an attempt to solve the problem: **where did my data from?**. In the randomization test example, our question was: **where did the differences between the means in my data come from?**. We know that the differences could be produced by chance alone. We simulated what chance can due using randomization. Then we plotted what chance can do using a histogram. Then, we used to picture to help us make an inference. Did our observed difference come from the distribution, or not? When the observed difference is clearly inside the chance distribution, then we can infer that our difference **could have been produced by chance**. When the observed difference is not clearly inside the chance distribution, then we can infer that our difference was **probably not produced by chance**.

In my opinion, these pictures are very, very helpful. If one of our goals is to help ourselves summarize a bunch of complicated numbers to arrive at an inference, then the pictures do a great job. We don't even need a summary number, we just need to look at the picture and see if the observed difference is inside or outside of the window. This is what it is all about. Creating intuitive and meaningful ways to make inferences from our data. As we move forward, the main thing that we will do is formalize our process, and talk more about "standard" inferential statistics. For example, rather than looking at a picture (which is a good thing to do), we will create some helpful numbers. For example, what if you wanted to the probability that your difference could have been produced by chance? That could be a single number, like 95%. If there was a 95% probability that chance can produce the difference you observed, you might not be very confident that something like your experimental manipulation was causing the difference. If there was only 1% probability that chance could produce your difference, then you might be more confident that **chance did not** produce the difference; and, you might instead be comfortable with the possibility that your experimental manipulation actually caused the difference. So, how can we arrive at those numbers? In order to get there we will introduce you to some more foundational tools for statistical inference.

5.7 Videos

5.7.1 Null and Alternate Hypotheses

5.7.2 Types of Errors

6 Hypothesis Testing

6.1 Hypothesis Testing - The Nuts & Bolts

Hypothesis testing helps us figure out if what we believe about a whole group is likely true, just by looking at a small part of it (a sample).

6.1.1 Clarifying Alpha, P-value, and Confidence Level

Before diving deep, let's clear up some terms you'll come across often.

Alpha (α)

Alpha (α) is the significance level of a statistical test, and it quantifies the risk of committing a Type I error. A Type I error happens when we incorrectly reject a true null hypothesis. The standard value for alpha is often set at 0.05, implying a 5% chance of making a Type I error. In other words, we are willing to accept a 5% risk of concluding that a difference exists when there is no actual difference.

P-value

The p-value is another crucial concept in hypothesis testing. It represents the probability of observing the obtained results, or something more extreme, assuming that the null hypothesis is true. A small p-value (usually 0.05) suggests that the observed data is inconsistent with the null hypothesis, and thus, you have evidence to reject it.

Confidence Level

The confidence level is related but distinct from alpha and p-value. While alpha quantifies the risk of a Type I error, the confidence level indicates how confident we are in our statistical estimates. The confidence level is calculated as the complement of alpha:

$$\text{Confidence Level} = 1 - \alpha$$

For example, if α is 0.05, the confidence level would be (1 - 0.05 = 0.95) or 95%. This means we are 95% confident that our results fall within a specific range.

Bringing It All Together

- **Alpha** (α): Risk of Type I error (usually 5%)
- **P-value**: Probability of observed data given the null is true
- **Confidence Level**: Confidence in the range of our estimates (usually 95%)

Grasping how these three terms connect and differ is key to making sense of the stats we'll discuss.

6.1.2 The Steps of Hypothesis Testing Applied to an Example

Let's say we want to know if the average pollution in a set of water samples is above the legal limit. Or if young deer in a region are, on average, healthy.

Step 1: Define Your Hypotheses: First, we need to define two hypotheses: the **research hypothesis** and the **null hypothesis**.

- **Research Hypothesis ($\mathrm{H_a}$)**: This is what we aim to support. **Keep in mind, we can't exactly "prove" $\mathrm{H_a}$ is correct, we can only say that $\mathrm{H_0}$ isn't likely**. It can take a few forms based on the question:
  - $\mathrm{H_a}$: average pollution > legal limit (pollution is too high)
  - $\mathrm{H_a}$: average pollution < legal limit (pollution is too low)
  - $\mathrm{H_a}$: average pollution   legal limit (pollution is just different)
- **Null Hypothesis ($\mathrm{H_0}$)**: This is the default or 'no change' scenario. It's opposite to the research hypothesis.
  - $\mathrm{H_0}$: average pollution   legal limit (for the first $\mathrm{H_a}$)
  - $\mathrm{H_0}$: average pollution   legal limit (for the second $\mathrm{H_a}$)
  - $\mathrm{H_0}$: average pollution = legal limit (for the third $\mathrm{H_a}$)

Step 2: Choose Your Test Statistic: Based on the data, we'll compute a **test statistic**. This number will help us decide which hypothesis seems more likely.

Step 3: Determine the Rejection Region: Before running the test, we decide on a **rejection region**. If our test statistic falls in this region, we'll reject the null hypothesis.

Step 4: Check Assumptions: Before drawing conclusions, ensure that the test's conditions and assumptions are satisfied.

Step 5: Draw Conclusions: Finally, based on the test statistic and the rejection region, decide whether to reject the null hypothesis.

6.1.3 Errors in Hypothesis Testing

Sometimes, even with the best methods, we make incorrect decisions.

- **Type I Error** (α): This happens when we mistakenly reject the true null hypothesis. Imagine sending an innocent person to jail. Typically, α is set at 0.05 (5%).
- **Type II Error** (β): Here, we mistakenly accept a false null hypothesis. Think of it as letting a guilty person go free.

Decision	If the null hypothesis is True	If the null hypothesis is False
Reject $\mathbf{H_0}$	Type I error (prob = α)	Correct (prob = 1 - β)
Fail to reject $\mathbf{H_0}$	Correct (prob = 1 - α)	Type II error (prob = β)

Key Takeaway: As α gets smaller, β gets bigger, and vice-versa.

6.1.4 Deciphering Significance with P-values

The p-value is like a reality-check. It tells us how weird our results are if we assume the starting belief (null hypothesis) is spot on.

- **One-Tailed Test**: The p-value shows the likelihood of observing an average as extreme as our sample's if the null hypothesis stands.
- **Two-Tailed Test**: This p-value represents the odds of spotting an average as different from the null value as our sample's.

Rule of Thumb: If the p-value is less than α, we opt to reject the null hypothesis.

6.2 Graphical Review

6.2.1 Key Players in Hypothesis Testing Visualization

We define and visualize the core components essential to understanding the graphical representations of hypothesis testing:

1. **Null Distribution** - The hypothesized parent distribution under the assumption that the null hypothesis H_0 is true.
2. **Inferred Parent Distribution** - The parent distribution inferred from our sample data. This is what we conceptualize as the distribution of H_a.
3. **True Parent Distribution** - The actual distribution from which our sample originates.
4. **Sampling Distribution of the Sample Mean** - Represents the distribution of sample means if we were to draw multiple samples from the parent distribution. This is crucial for making inferences about the **Inferred Parent Distribution**.

In the figure below, we've outlined the various elements crucial for hypothesis testing. Think of this section as a handy guide. Whenever you come across detailed graphs later in this chapter, you can circle back here for clarity.

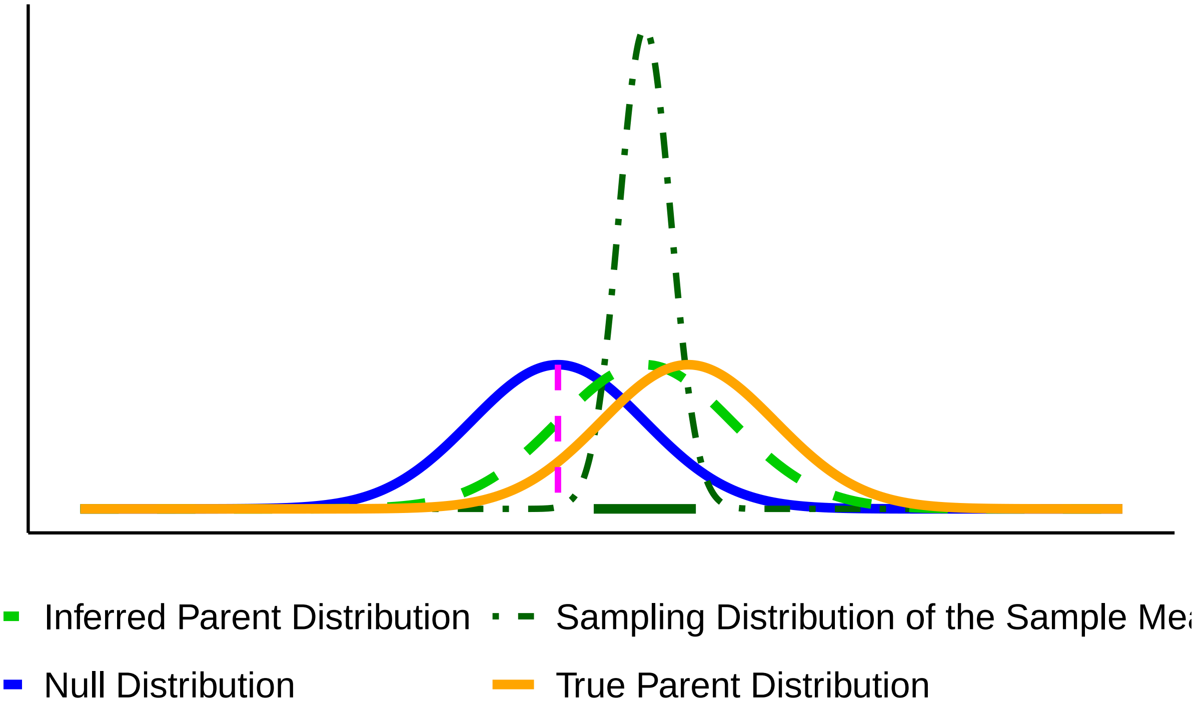


Figure 6.1: Comparison of four distributions essential for hypothesis testing: Null, True Parent, Inferred Parent, and Sampling Distribution of the Sample Mean.

For a two-tailed alternative, we are interested in the possibility that a sample comes from a parent distribution that may have a lower or higher location than the null.

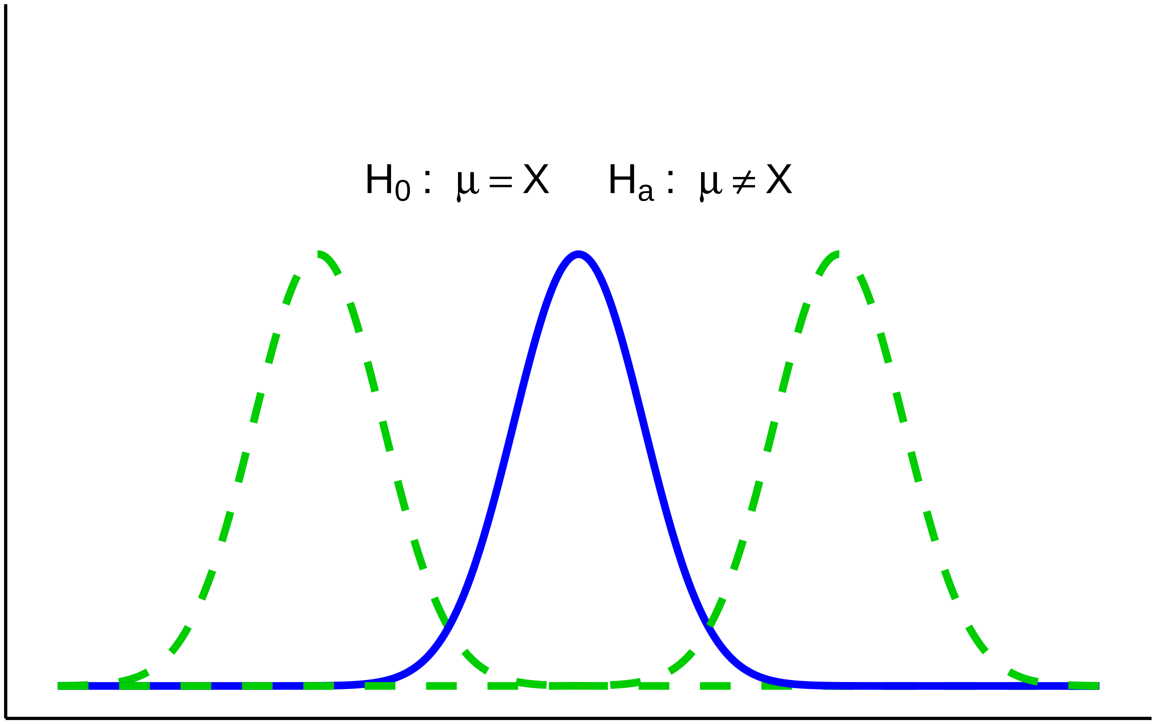


Figure 6.2: Two-tailed distribution

In a one-tailed t-test, we're examining if our sample originates from a parent distribution that's situated either below or above the null hypothesis. Unlike a two-tailed test, we're only interested in one of these directions, not both.

In a "perfect" world in which the null hypothesis is true, the sample's parent distribution (solid, orange) is exactly the same parent distribution described by the null hypothesis (solid, blue).

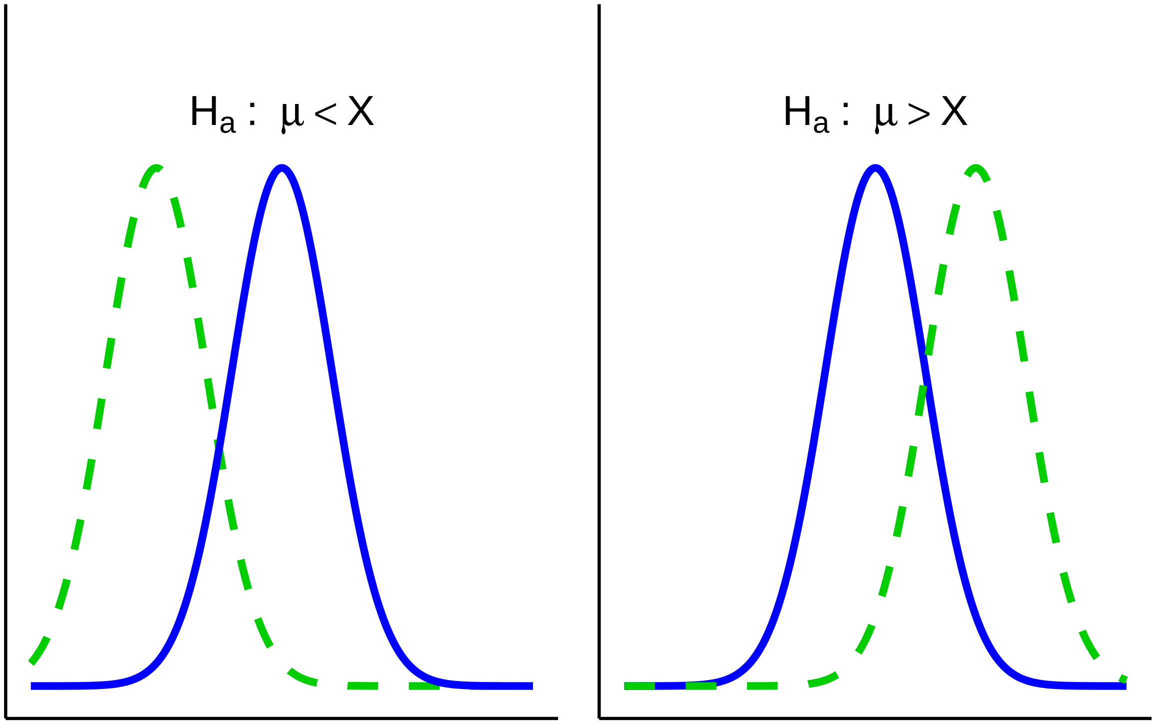


Figure 6.3: Side-by-side comparison of one-tailed t-test scenarios: exploring if our sample comes from a distribution either below or above the null hypothesis.

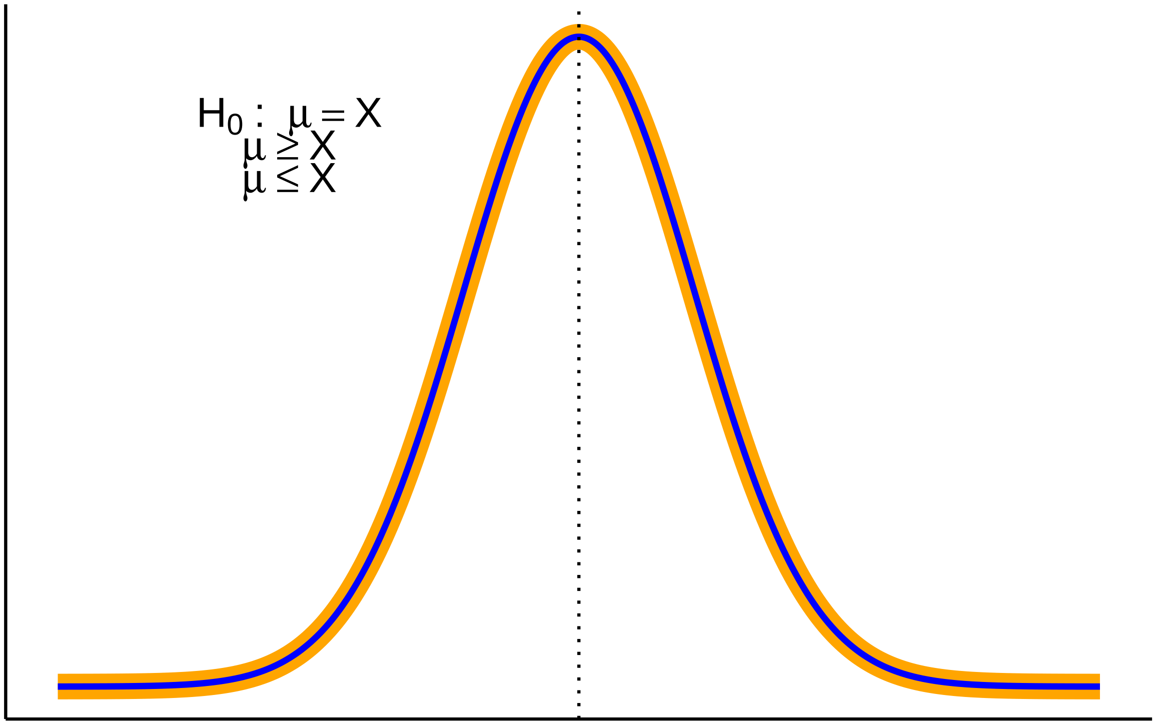


Figure 6.4: True parent distribution & null distribution are the same

We never know the true parent distribution of the sample – we infer it from the sample. Here, the tall dash-dotted line shows the sampling distribution of the mean, from which we infer the parent distribution (green3, dashed).

In this even more perfect world, that parent distribution is the same as the parent distribution described by the null hypothesis and we have taken a perfectly representative sample, so all 3 curves line up perfectly on the same mean. The thick, short, flat dark green line is the confidence interval for the sample mean.

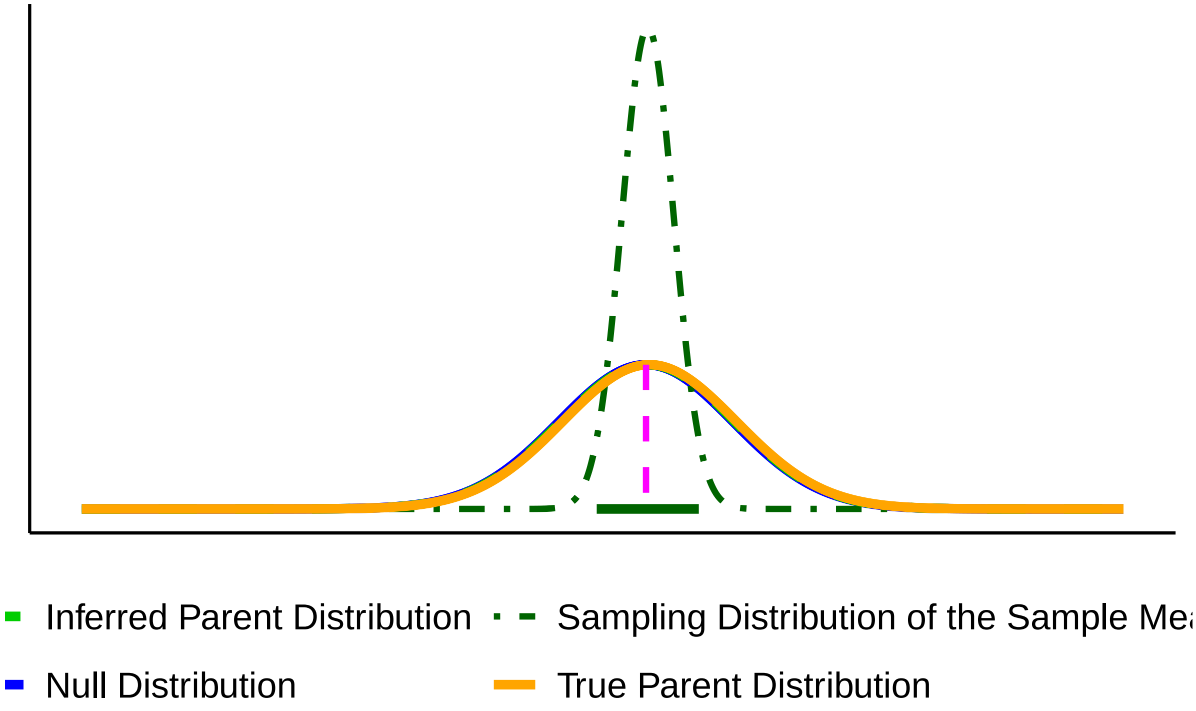


Figure 6.5: Here, we have a perfectly representative sample

In an imperfect but convenient world, the sample is not a perfect representation of the parent population, but is fairly close. The sample mean is close to hypothesized mean, and (in the 2-tailed case) the confidence interval for the sample mean "catches" the mean of the null hypothesis (pink dashed line). A hypothesis test will correctly determine that there is not a significant difference between the sample mean and the mean of the null hypothesis.

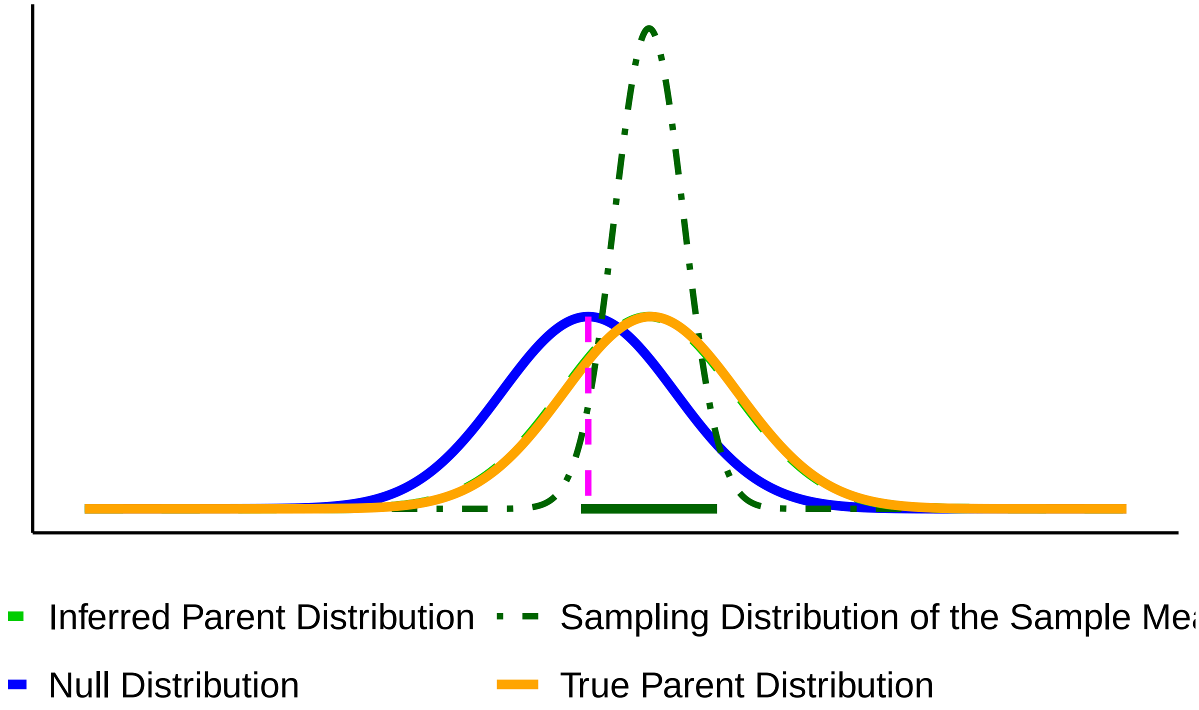


Figure 6.6: Imperfect, but convenient

In an imperfect and inconvenient world, the random sample is, by chance, sufficiently imperfect that the apparent (inferred) parent distribution is far from the true parent distribution and (in the 2-tailed case) the confidence interval for the sample mean no longer “catches” the mean of the null hypothesis. A hypothesis test will now find a significant difference between the sample mean and the mean of the null hypothesis. **This is a type I error**.

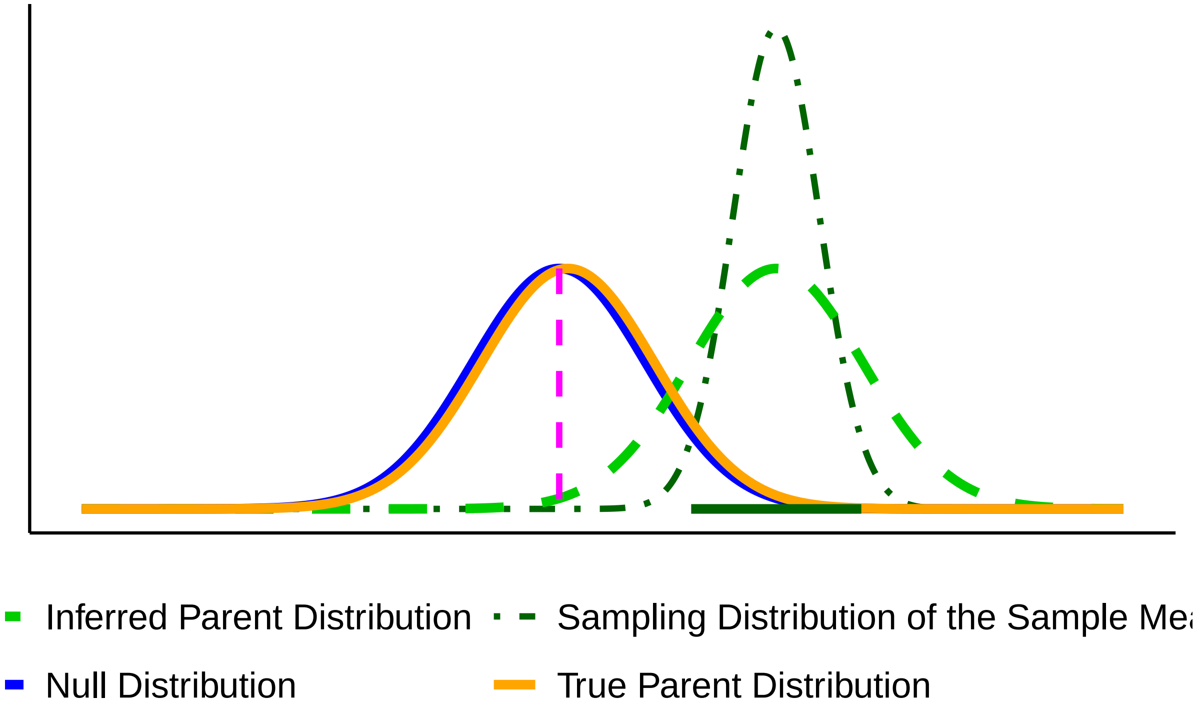


Figure 6.7: Type 1 Error

In another imperfect and inconvenient world, the sample (dashed dark green lines) really is drawn from the alternative distribution (the sample’s true parent distribution; orange), but is unrepresentative of its parent and similar to the null (solid blue line). The confidence interval (in the 2-tailed case) of the sample “catches” the mean of the null hypothesis although it is far from the mean of the true parent of the sample. A hypothesis test will find no significant difference between the sample mean and the mean of the null hypothesis. **This is a type II error.**

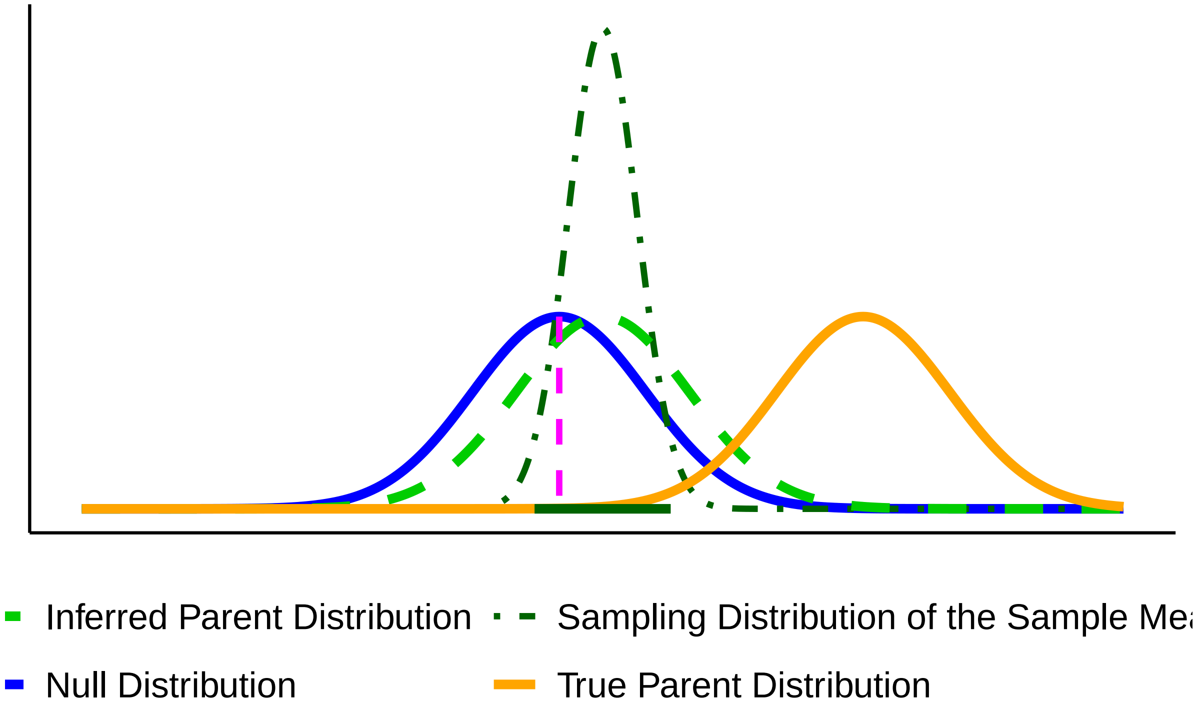


Figure 6.8: Type 2 Error

6.3 Graphical Review of Test Outcomes that are Not in Error

As you review hypothesis testing, it's essential to remember that we don't *accept* the null hypothesis. The possibility of a Type I error means our conclusion might be flawed. Instead of accepting the null hypothesis, we *fail to reject* H_0. The scarcity of data with small sample sizes can lead to significant differences between the sample mean and the null mean (_0). While it's tempting to gather more data to be more certain, in the meantime, the best we can do is fail to reject H_0.

In the figures below, as in the figures above, the blue lines represent the null parent distribution (defined by the null mean and the sample's standard deviation).

The green solid lines denote the apparent parent distribution of our sample:

- **Solid lighter green line**: Represents the distribution described by our sample mean and standard deviation.
- ***Dashed dark green line**: Shows the sampling distribution of the sample mean, described by our sample mean and the standard error (SE).

6.3.1 Graphical Descriptions:

1. **Fail to Reject the Null Hypothesis** - *Sample Mean Supports the Null Hypothesis*: The means are far apart, but **not in our direction of interest**. For a one-tailed test, only data on one side of the rejection region can support the null hypothesis. Question to ponder: If we gather more data and obtain the same sample mean, could our conclusion change?

2. **Fail to Reject the Null Hypothesis**: *The sample mean supports the alternate hypothesis (it is on the appropriate side of the rejection region), but the sample size is too small.* The sample mean is only about 1SE from the null mean, making it too close to be significant. Hypothetical situation: With more data and the same sample mean, could our conclusion differ?

3. **Reject the Null Hypothesis**: The sample mean is on the appropriate side of the rejection region. It's significantly distant from the null mean, over 3 SE, which is typically considered significant for most standard values of .

4. **Reject the Null Hypothesis (Two-Tailed Test)**: Reject the null hypothesis for the same reasons as the previous example. This case is two-tailed, but nothing else has changed.

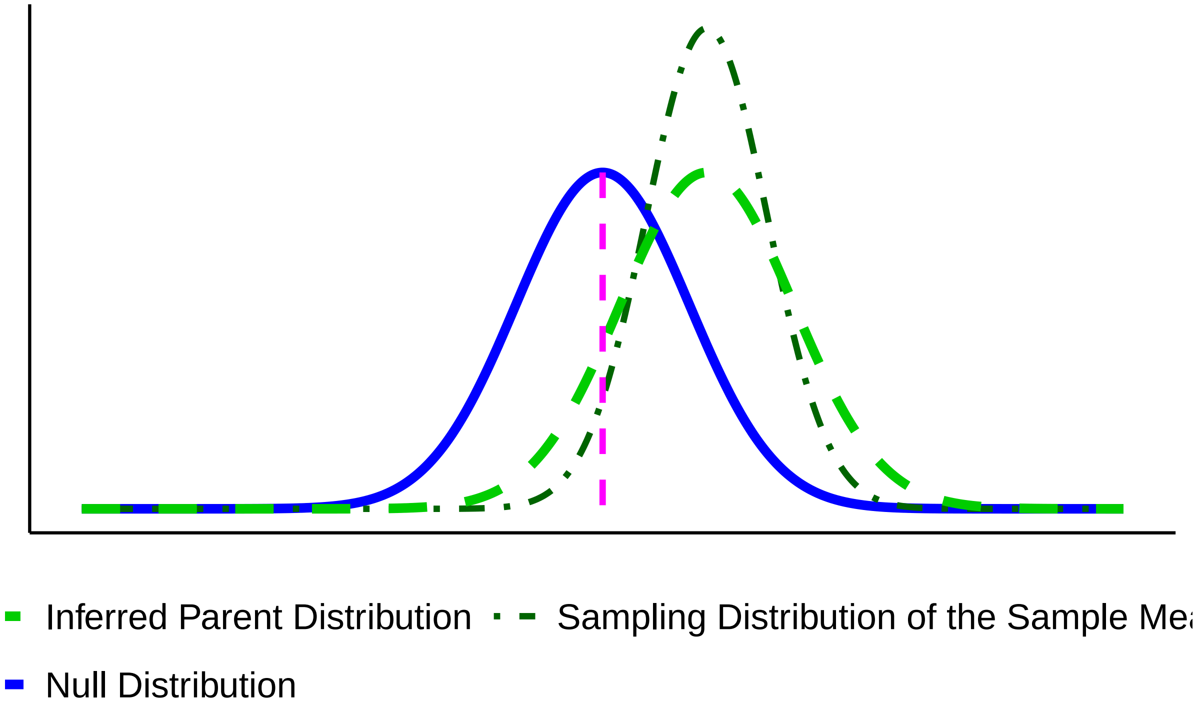


Figure 6.9: Ha: < X

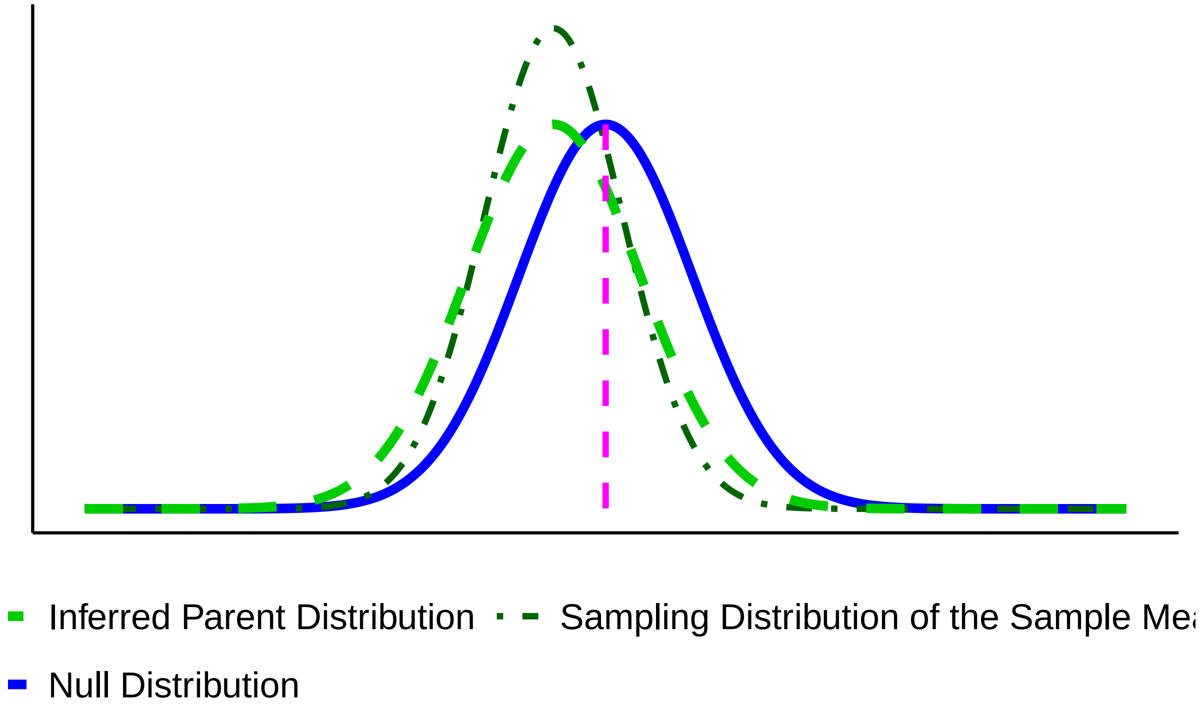


Figure 6.10: Ha: < X

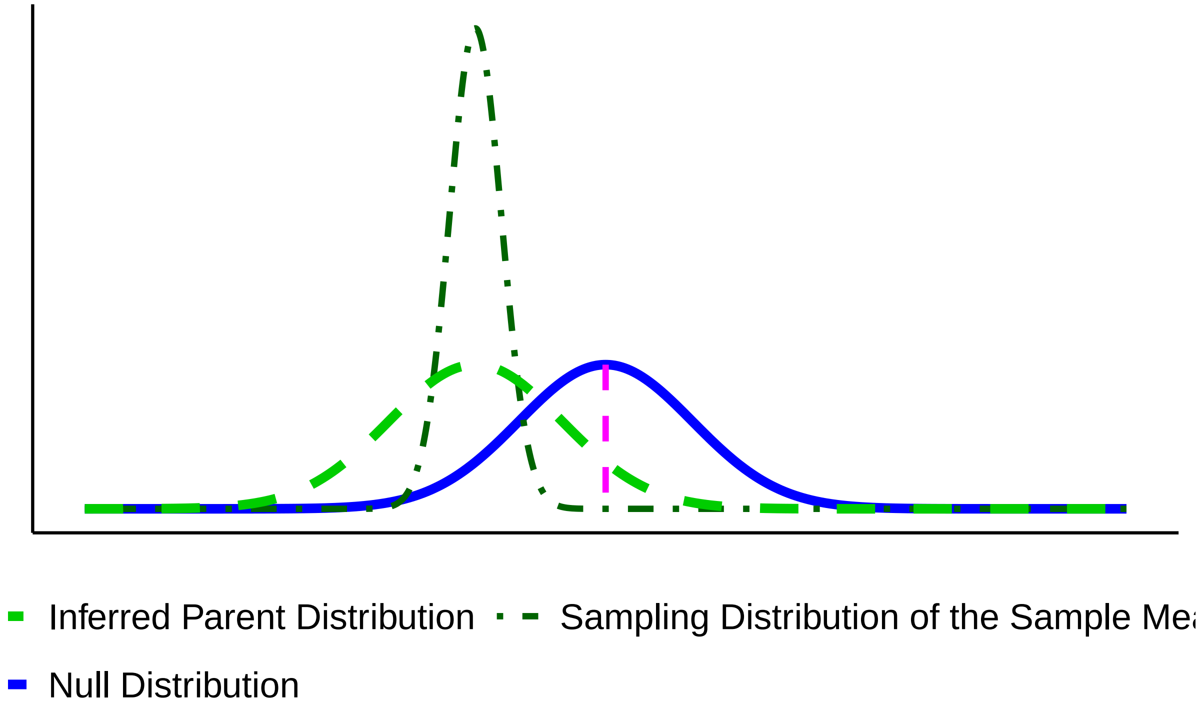


Figure 6.11: Ha: < X

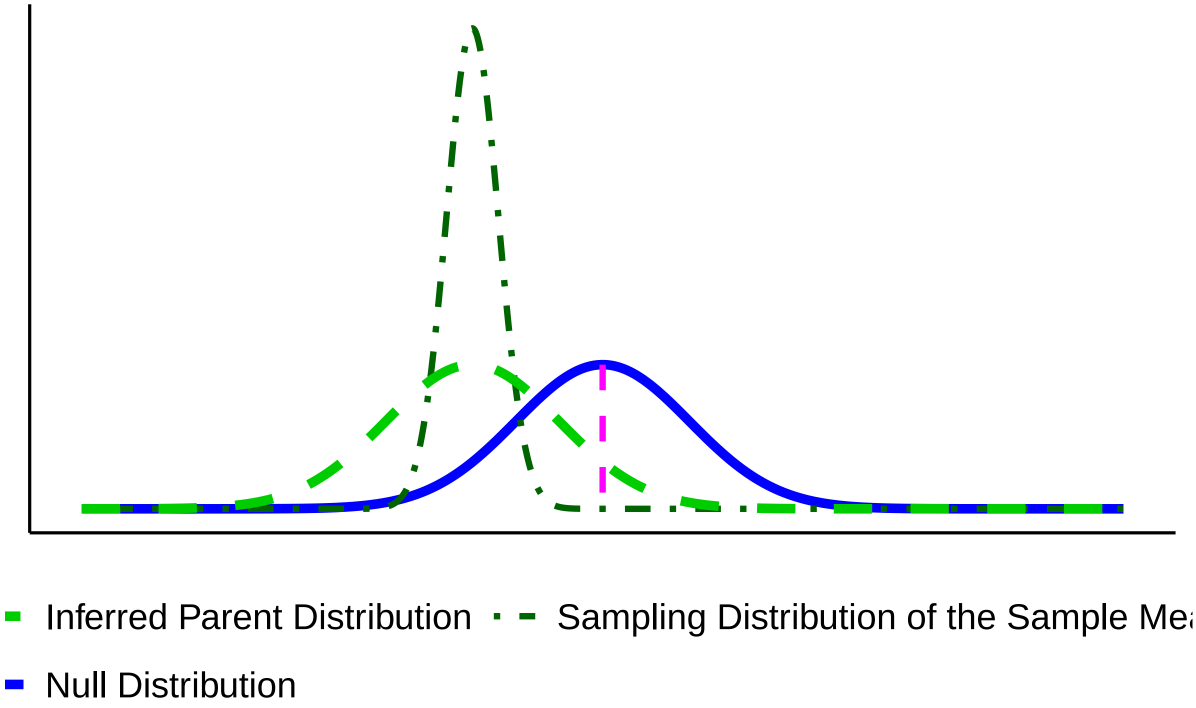


Figure 6.12: Ha: X

6.4 Graphical Review of Sample Size Effect when Test Outcomes are in Error

It's a given that we never truly grasp the actual parent distribution of a sample. An unrepresentative sample can lead either to a Type I or a Type II error. The term *sampling error* is sometimes invoked to depict such unrepresentative samples, but it's imperative to understand that the researcher hasn't committed any mistakes.

6.4.1 Graphical Descriptions:

Type I Error: Here, the green curves depict the sampling distribution (dark green) and the apparent parent distribution (lighter green) of our sample. But in reality, the sample is a product of the null distribution (blue). Question to ponder: How would the representation look if we had utilized a smaller sample size?

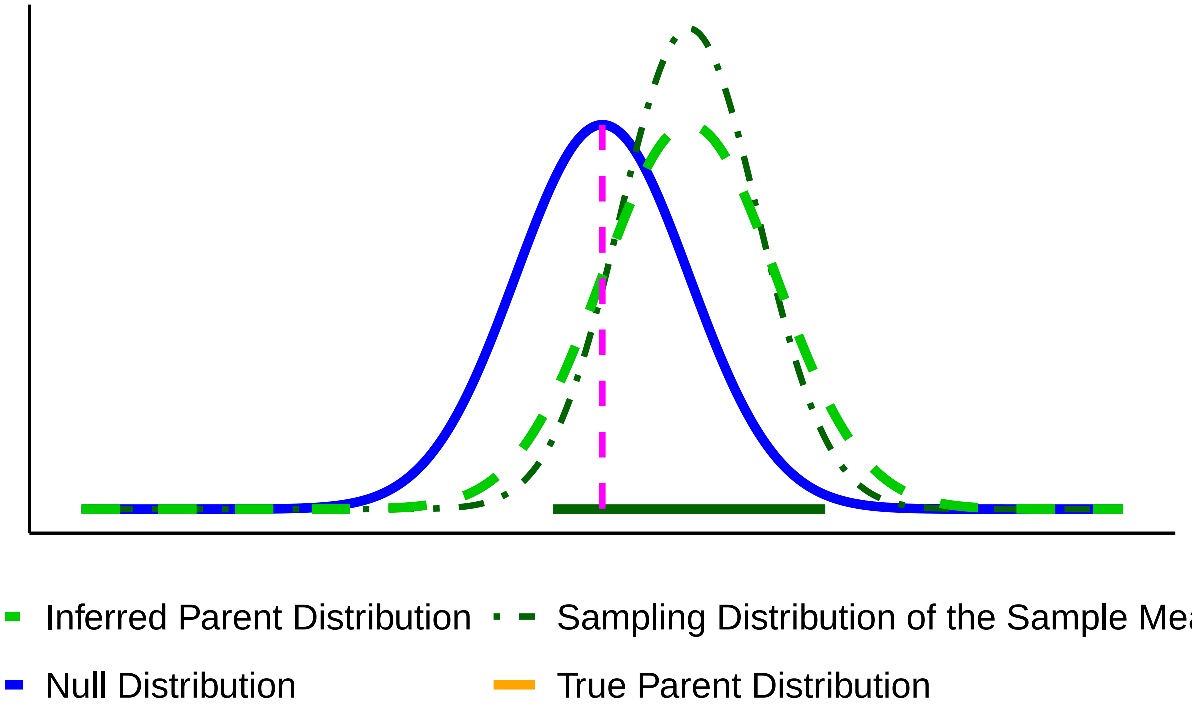


Figure 6.13

The main thing that would change with a larger sample size is that the sampling distribution of sample means becomes much tighter, thus making the confidence interval smaller. So here, are we more or less likely to have a type 1 error with the larger sample size?

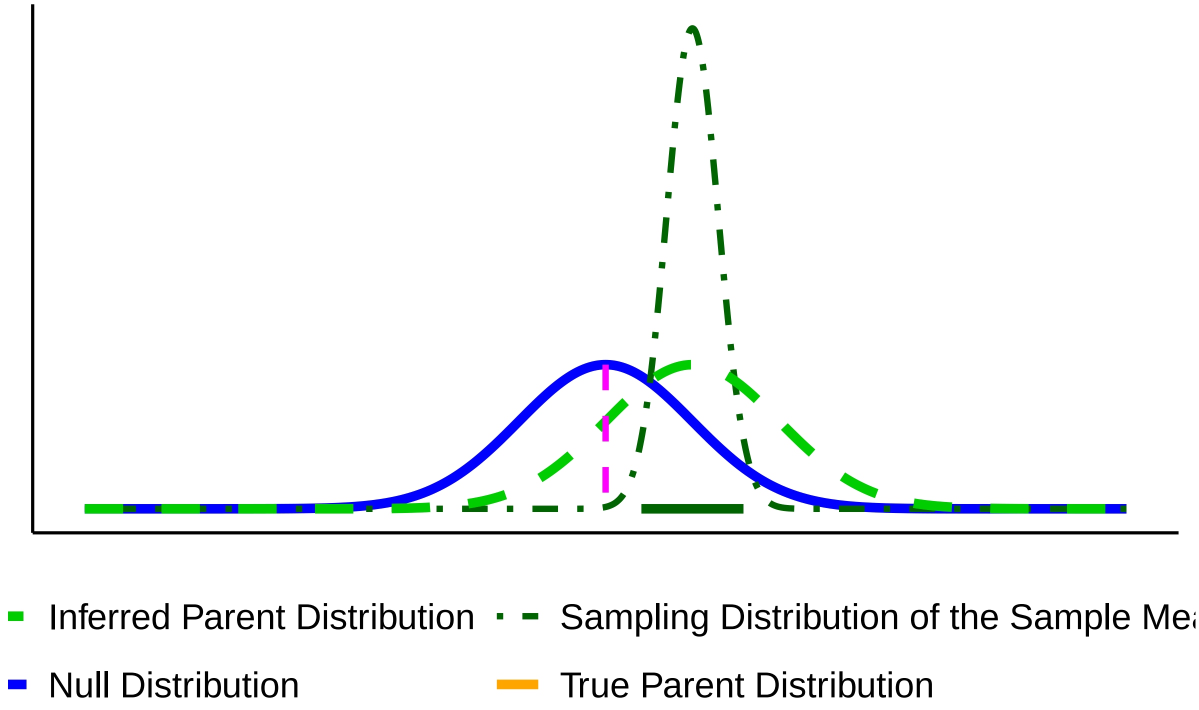

Inferred Parent Distribution
Sampling Distribution of the Sample Me
Null Distribution
True Parent Distribution

Figure 6.14

Type II Error: The sample genuinely hails from the solid orange parent population. However, it was misleading enough (as depicted by the dashed green line) to seem analogous to the null distribution (blue). Query to reflect upon: How would this representation transform if the sample size was substantially larger?

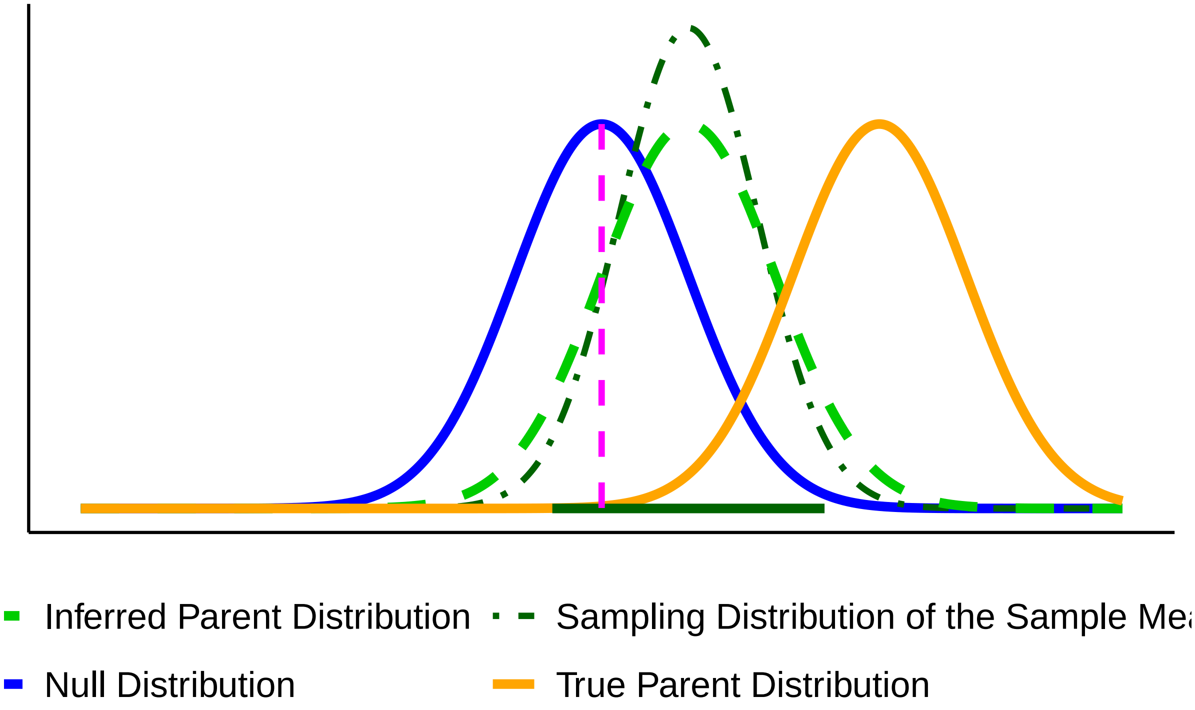


Figure 6.15

The main thing that would change with a larger sample size is that the sampling distribution of sample means becomes much tighter, thus making the confidence interval smaller. So here, are we more or less likely to have a type 2 error with the larger sample size?

6.4.2 How Significant is 'Significant' – Interpreting p-values

When we use a rejection region to test a hypothesis, we get a yes-or-no answer. For a two-tailed test, if we ask whether the confidence interval "captures" the null mean, we get a yes-or-no answer as well.

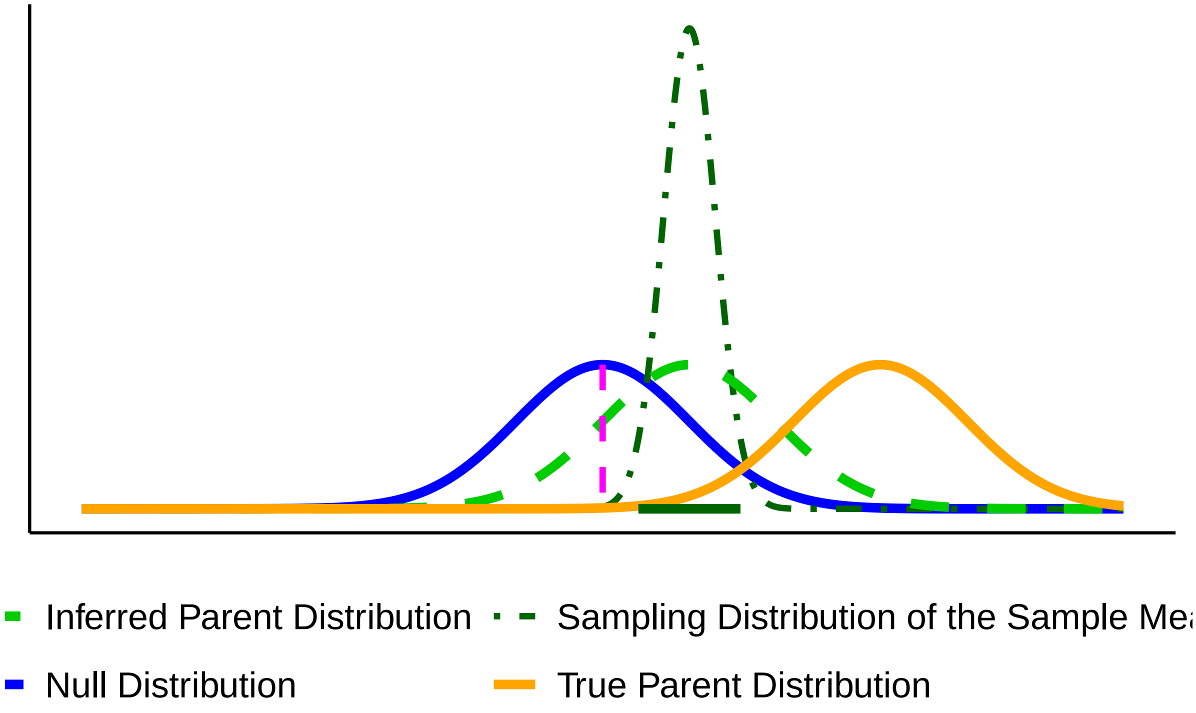

Inferred Parent Distribution
Sampling Distribution of the Sample Mea
Null Distribution
True Parent Distribution

Figure 6.16

6.4.2.1 Calculating p-values

We can do better – we can get an actual probability value. The blue box for the z-test tells us how to calculate our p-value if our mean is in the area of interest for the test.

1. **Calculate Z-value**: First, we calculate how many standard errors our mean is from the null mean. This is the z-value for our mean in the world of the null hypothesis.
    - For ($H_0 < X$), we ask about the upper tail probability of our mean.
    - For ($H_0 > X$), we ask about the lower tail probability of our mean.
    - For ($H_0 = X$), we calculate twice the tail probability.

6.4.2.2 One-Tailed vs Two-Tailed Tests

- **One-Tailed Test**: The p-value tells us the probability of a mean at least as much greater than (H_0) as our mean, when the null hypothesis is true or as much less than (H_0).
- **Two-Tailed Test**: The p-value tells us the probability of a mean at least as different from (H_0) as our mean, when the null hypothesis is true.

6.4.2.3 Rejecting the Null Hypothesis

We reject (H_0) when (p)-value ($<$). At that point, our data are too unusual when the null hypothesis is true for us to believe that the null hypothesis is true.

- **Small p-value**: When (p) is small, our data provide weak support for (H_0), and we are more sure that (H_0) is not true, and that (H_a) is more likely.

6.5 Review of Ways to Test (H_0)

1. **Confidence Interval**: If the alternative is two-tailed, build a 1-( ) confidence interval. If the CI catches (H_0) then fail to reject.
2. **Test Statistic**: Calculate the test statistic and compare to the rejection region of size ( ). If the test statistic is in the rejection region, reject (H_0).
3. **Probability**: Determine the probability of your test statistic. If (p $<$) then reject (H_0).

Note: The first two methods give you a yes-or-no answer. The third method gives you some additional information.

6.5.1 Results Statement

Now that we have started doing statistical tests, we have also started to think about results. A results statement provides an English language version of what we discovered, as well as the statistical results. For a z-test, a results sentence might say:

> The average level of mercury in the ponds within 10 km of the smelter is significantly higher than the legal limit (z = 2.85, n = 32, p = 0.004).

The information in the parentheses, for a one-sample test, is, in this order,

1) the value of the test statistic, in this case, (z);
2) the sample size or degrees of freedom; and
3) the probability of the test statistic when the null hypothesis is true.

Most problems that include a test will require a results sentence.

6.5.2 Beyond the 0.5 cutoff: Effect-size and power

You've probably heard me mention that the 0.5 cutoff for statistical significance is somewhat arbitrary. So, what's the alternative? Enter effect size and statistical power. These aren't just buzzwords; they're foundational elements for conducting meaningful environmental research. Many scientific journals even have guidelines on how to report them. Ideally, you should be thinking about these factors before you collect your first data point. Given their importance, it's time we delve into what these concepts really mean and why they're crucial for research.

6.5.3 The importance of knowing what you're doing

Effect size and power analyses are more than just boxes to tick; they're essential tools in your research toolkit for understanding environmental data. Rather than using them simply because you were advised to, see them as integral to designing meaningful studies. These tools help you filter out statistical "noise," revealing actionable insights that can address real-world environmental issues. They shouldn't be applied blindly but should be part of a thoughtful research strategy aimed at making your data work for you.

6.5.4 Chance vs. Real Effects: The Playground, the Superpower, and the Impact Scale

In environmental research, the goal is often to identify meaningful changes—like the improvement of air quality due to reduced pollution. However, researchers sometimes find themselves grappling with statistical "noise" rather than detecting genuine effects. To navigate this complex landscape, let's use some analogies.

The Playground and the Mischievous Kid

First, consider your sample size as a playground and chance as a mischievous kid running around in it. The smaller the playground, the more room this kid has to create chaos, leading to random variations in your data. On the flip side, a larger playground restricts the kid's antics, minimizing the influence of chance. So, your first task is to design your study like an ultimate playground—spacious and well-planned to keep chance at bay.

The Balls: Different Sizes, Different Impacts

Next, let's focus on the "stuff" being thrown around on this playground. Think of different types of balls—soccer balls, tennis balls, ping pong balls, and marbles—as representing different effect sizes:

- **Soccer Ball (Strong Effect)**: It's big and noticeable. When it lands, you know something significant has happened.
- **Tennis Ball (Medium Effect)**: Still impactful but not as game-changing as a soccer ball.
- **Ping Pong Ball (Small Effect)**: It might bounce around, but it's not going to change the landscape.
- **Marble (Very Small Effect)**: Almost negligible amid the other activities.

The Superhero: Statistical Power

Here's where statistical power comes into play. It's your research superhero, capable of discerning whether the changes you're observing are due to chance, the size of your playground, or the type of ball being thrown (Effect Size). Imagine it as a keen-eyed playground supervisor who can tell the difference between a random bounce and a meaningful impact.

The Takeaway

1. **Plan Well**: Design your study like you're building the ultimate playground—spacious and well-planned to minimize the role of chance.
2. **Know Your Ball**: Understand the potential impact (effect size) of what you're introducing into your study. This helps you make meaningful conclusions.

3. **Power Up**: Conduct a power analysis to ensure your study is equipped to distinguish between meaningful impacts and random noise.

By focusing on these three elements—chance, effect size, and statistical power—you're not just adhering to research best practices; you're elevating the quality and impact of your work.

6.5.5 Effect size: concrete vs. abstract notions

Generally speaking, the big concept of effect size, is simply how big the differences are, that's it. However, the biggness or smallness of effects quickly becomes a little bit complicated. On the one hand, the raw difference in the means can be very meaningful. Let's say we are measuring performance on a final exam, and we are testing whether or not a miracle drug can make you do better on the test. Let's say taking the drug makes you do 5% better on the test, compared to not taking the drug. You know what 5% means, that's basically a whole letter grade. Pretty good. An effect-size of 25% would be even better, right? Lots of measures have a concrete quality to them, and we often want to the size of the effect expressed in terms of the original measure.

Let's talk about concrete measures some more. How about learning a musical instrument. Let's say it takes 10,000 hours to become an expert piano, violin, or guitar player. And, let's say you found something online that says that using their method, you will learn the instrument in less time than normal. That is a claim about the effect size of their method. You would want to know how big the effect is right? For example, the effect-size could be 10 hours. That would mean it would take you 9,980 hours to become an expert (that's a whole 10 hours less). If I knew the effect-size was so tiny, I wouldn't bother with their new method. But, if the effect size was say 1,000 hours, that's a pretty big deal, that's 10% less (still doesn't seem like much, but saving 1,000 hours seems like a lot).

In environmental science, we often encounter measures that are not as straightforward as, say, temperature or pH levels. Take biodiversity indices as an example. These indices can give us a numerical value representing the variety of life in a particular ecosystem, but interpreting what these numbers mean can be challenging.

Imagine you're assessing the impact of a reforestation project. Your biodiversity index might read 3 before the project and 4 after. That's a difference of only 1 unit, but what does that actually signify? Is it a significant improvement, or just a minor change? The raw numbers alone don't provide enough context.

To make these abstract measures more interpretable, we often turn to standardized metrics, like z-scores. If that 1-unit difference in biodiversity corresponds to a shift of one standard deviation, that's a substantial change worth noting. On the other hand, if the shift is only 0.1 in terms of standard deviation, then the 11-unit difference might not be as impactful as it first seemed. Standardized measures like Cohen's d can further help us understand the practical significance of our findings.

6.5.6 Cohen's d

Let's look a few distributions to firm up some ideas about effect-size. Figure 6.17 has four panels. The first panel (0) represents the null distribution of no differences. This is the idea that your manipulation (A vs. B) doesn't do anything at all, as a result when you measure scores in conditions A and B, you are effectively sampling scores from the very same overall distribution. The panel shows the distribution as green for condition B, but the red one for condition A is identical and drawn underneath (it's invisible). There is 0 difference between these distributions, so it represent a null effect.

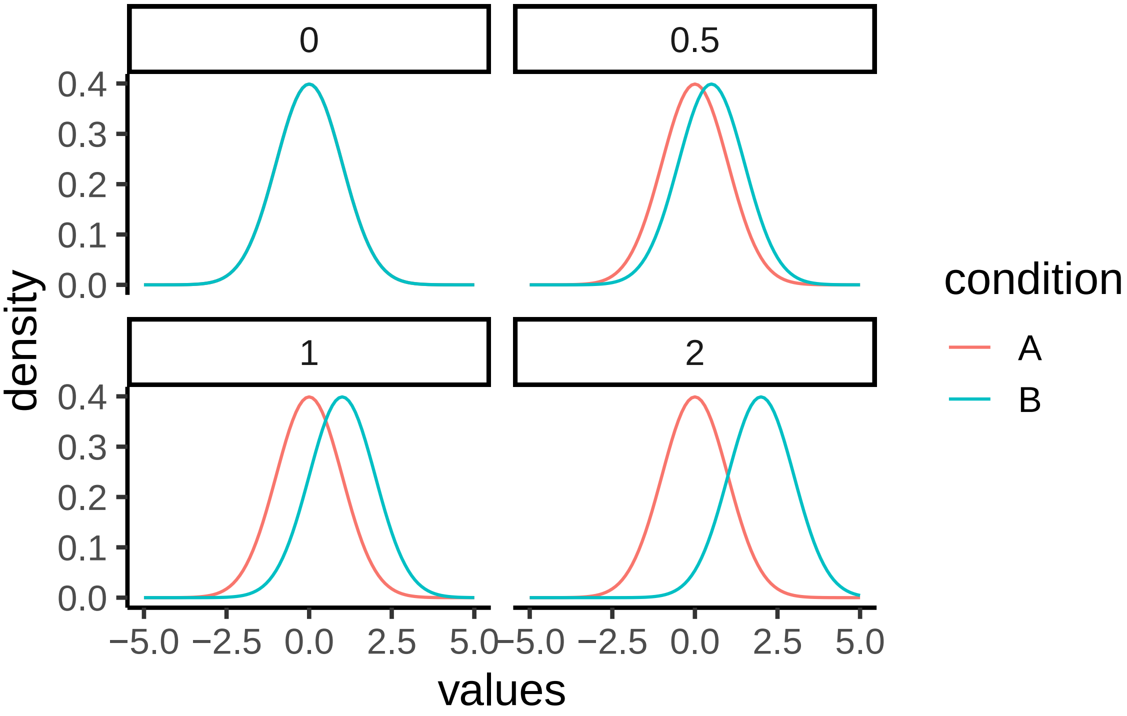


Figure 6.17: Each panel shows hypothetical distributions for two conditions. As the effect-size increases, the difference between the distributions become larger.

The remaining panels are hypothetical examples of what a true effect could look like, when your manipulation actually causes a difference. For example, if condition A is a control group, and condition B is a treatment group, we are looking at three cases where the treatment manipulation causes a positive shift in the mean of distribution. We are using normal curves with mean =0 and sd =1 for this demonstration, so a shift of .5 is a shift of half of a standard deviation. A shift of 1 is a shift of 1 standard deviation, and a shift of 2 is a shift of 2 standard deviations. We could draw many more examples showing even bigger shifts, or shifts that go in the other direction.

Let's look at another example, but this time we'll use some concrete measurements. Let's say we are looking at final exam performance, so our numbers are grade percentages. Let's also say that we know the mean on the test is 65%, with a standard deviation of 5%. Group A could be a control that just takes the test, Group B could receive some "educational" manipulation designed to improve the test score. These graphs then show us some hypotheses about what the manipulation may or may not be doing.

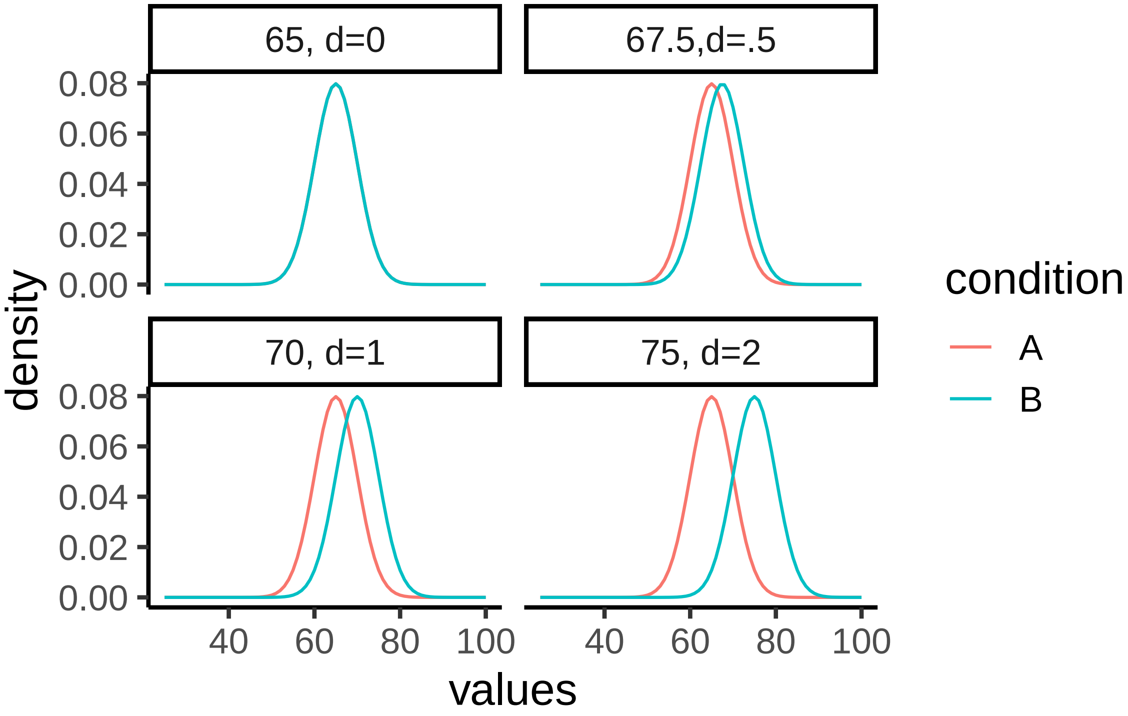


Figure 6.18: Each panel shows hypothetical distributions for two conditions. As the effect-size increases, the difference between the distributions become larger.

The first panel shows that both condition A and B will sample test scores from the same distribution (mean =65, with 0 effect). The other panels show shifted mean for condition B (the treatment that is supposed to increase test performance). So, the treatment could increase the test performance by 2.5% (mean 67.5, .5 sd shift), or by 5% (mean 70, 1 sd shift), or by 10% (mean 75%, 2 sd shift), or by any other amount. In terms of our previous metaphor, a shift of 2 standard deviations is more like jack-hammer in terms of size, and a shift of .5 standard deviations is more like using a pencil. The thing about research, is we often have no clue about whether our manipulation will produce a big or small effect, that's why we are conducting the research.

You might have noticed that the letter *d* appears in the above figure. Why is that? Jacob Cohen (Cohen 1988) used the letter *d* in defining the effect-size for this situation, and now everyone calls it Cohen's d. The formula for Cohen's d is:

$d = \frac{\text{mean for condition 1}-\text{mean for condition 2}}{\text{population standard deviation}}$

If you notice, this is just a kind of z-score. It is a way to standardize the mean difference in terms of the population standard deviation.

It is also worth noting again that this measure of effect-size is entirely hypothetical for most purposes. In general, researchers do not know the population standard deviation, they can only guess at it, or estimate it from the sample. The same goes for means, in the formula these are hypothetical mean differences in two population distributions. In practice, researchers do not know these values, they guess at them from their samples.

Before discussing why the concept of effect-size can be useful, we note that Cohen's d is useful for understanding abstract measures. For example, when you don't know what a difference of 10 or 20 means as a raw score, you can standardize the difference by the sample standard deviation, then you know roughly how big the effect is in terms of standard units. If you thought a 20 was big, but it turned out to be only 1/10th of a standard deviation, then you would know the effect is actually quite small with respect to the overall variability in the data.

6.6 Power

When there is a true effect out there to measure, you want to make sure your design is sensitive enough to detect the effect, otherwise what's the point. We've already talked about the idea that an effect can have different sizes. The next idea is that your design can be more less sensitive in its ability to reliabily measure the effect. We have discussed this general idea many times already in the textbook, for example we know that we will be more likely to detect "significant" effects (when there are real differences) when we increase our sample-size. Here, we will talk about the idea of design sensitivity in terms of the concept of power. Interestingly, the concept of power is a somewhat limited concept, in that it only exists as a concept within some philosophies of statistics.

6.6.1 A digresssion about hypothesis testing

In particular, the concept of power falls out of the Neyman-Pearson concept of null vs. alternative hypothesis testing. Neyman-Pearson ideas are by now the most common and widespread, and in the opinion of some of us, they are also the most widely misunderstood and abused idea.

What we have been mainly doing is talking about hypothesis testing from the Fisherian (Sir Ronald Fisher, the ANOVA guy) perspective. This is a basic perspective that can't be easily ignored. It is also quite limited. The basic idea is this:

1. We know that chance can cause some differences when we measure something between experimental conditions.
2. We want to rule out the possibility that the difference that we observed can not be due to chance
3. We construct large N designs that permit us to do this when a real effect is observed, such that we can confidently say that big differences that we find are so big (well outside the chance window) that it is highly implausible that chance alone could have produced.
4. The final conclusion is that chance was extremely unlikely to have produced the differences. We then infer that something else, like the manipulation, must have caused the difference.
5. We don't say anything else about the something else.
6. We either reject the null distribution as an explanation (that chance couldn't have done it), or retain the null (admit that chance could have done it, and if it did we couldn't tell the difference between what we found and what chance could do)

Neyman and Pearson introduced one more idea to this mix, the idea of an alternative hypothesis. The alternative hypothesis is the idea that if there is a true effect, then the data sampled into each condition of the experiment must have come from two different distributions. Remember, when there is no effect we assume all of the data cam from the same distribution (which by definition can't produce true differences in the long run, because all of the numbers are coming from the same distribution). The graphs of effect-sizes from before show examples of these alternative distributions, with samples for condition A coming from one distribution, and samples from condition B coming from a shifted distribution with a different mean.

So, under the Neyman-Pearson tradition, when a researcher find a signifcant effect they do more than one things. First, they reject the null-hypothesis of no differences, and they accept the alternative hypothesis that there was differences. This seems like a sensible thing to do. And, because the researcher is actually interested in the properties of the real effect, they might be interested in learning more about the actual alternative hypothesis, that is they might want to know if their data come from two different distributions that were separated by some amount…in other words, they would want to know the size of the effect that they were measuring.

6.6.2 Back to power

We have now discussed enough ideas to formalize the concept of statistical power. For this concept to exist we need to do a couple things.

1. Agree to set an alpha criterion. When the p-value for our test-statistic is below this value we will call our finding statistically significant, and agree to reject the null hypothesis and accept the "alternative" hypothesis (sidenote, usually it isn't very clear which specific alternative hypothesis was accepted)

2. In advance of conducting the study, figure out what kinds of effect-sizes our design is capable of detecting with particular probabilites.

The power of a study is determined by the relationship between

1. The sample-size of the study
2. The effect-size of the manipulation
3. The alpha value set by the researcher.

To see this in practice let's do a simulation. We will do a t-test on a between-groups design 10 subjects in each group. Group A will be a control group with scores sampled from a normal distribution with mean of 10, and standard deviation of 5. Group B will be a treatment group, we will say the treatment has an effect-size of Cohen's $d = .5$, that's a standard deviation shift of .5, so the scores with come from a normal distribution with mean =12.5 and standard deivation of 5. Remember 1 standard deviation here is 5, so half of a standard deviation is 2.5.

The following R script runs this simulated experiment 1000 times. We set the alpha criterion to .05, this means we will reject the null whenever the *p*-value is less than .05. With this specific design, how many times out of of 1000 do we reject the null, and accept the alternative hypothesis?

```
#> [1] 191
```

The answer is that we reject the null, and accept the alternative 191 times out of 1000. In other words our experiment succesfully accepts the alternative hypothesis 19.1 percent of the time, this is known as the power of the study. Power is the probability that a design will succesfully detect an effect of a specific size.

Importantly, power is completely abstract idea that is completely determined by many assumptions including N, effect-size, and alpha. As a result, it is best not to think of power as a single number, but instead as a family of numbers.

For example, power is different when we change N. If we increase N, our samples will more precisely estimate the true distributions that they came from. Increasing N reduces sampling error, and shrinks the range of differences that can be produced by chance. Lets' increase our N in this simulation from 10 to 20 in each group and see what happens.

```
#> [1] 345
```

Now the number of significant experiments i 345 out of 1000, or a power of 34.5 percent. That's roughly doubled from before. We have made the design more sensitive to the effect by increasing N.

We can change the power of the design by changing the alpha-value, which tells us how much evidence we need to reject the null. For example, if we set the alpha criterion to 0.01, then we will be more conservative, only rejecting the null when chance can produce the observed difference 1% of the time. In our example, this will have the effect of reducing power. Let's keep N at 20, but reduce the alpha to 0.01 and see what happens:

```
#> [1] 140
```

Now only 140 out of 1000 experiments are significant, that's 14 power.

Finally, the power of the design depends on the actual size of the effect caused by the manipulation. In our example, we hypothesized that the effect caused a shift of .5 standard deviations. What if the effect causes a bigger shift? Say, a shift of 2 standard deviations. Let's keep N= 20, and alpha < .01, but change the effect-size to two standard deviations. When the effect in the real-world is bigger, it should be easier to measure, so our power will increase.

```
#> [1] 1000
```

Neat, if the effect-size is actually huge (2 standard deviation shift), then we have power 100 percent to detect the true effect.

6.6.3 Power curves

We mentioned that it is best to think of power as a family of numbers, rather than as a single number. To elaborate on this consider the power curve below. This is the power curve for a specific design: a between groups experiments with two levels, that uses an independent samples t-test to test whether an observed difference is due to chance. Critically, N is set to 10 in each group, and alpha is set to .05

In Figure 6.19 power (as a proportion, not a percentage) is plotted on the y-axis, and effect-size (Cohen's d) in standard deviation units is plotted on the x-axis.

A power curve like this one is very helpful to understand the sensitivity of a particular design. For example, we can see that a between subjects design with N=10 in both groups, will detect an effect of d=.5 (half a standard deviation shift) about 20% of the time, will detect an effect of d=.8 about 50% of the time, and will detect an effect of d=2 about 100% of the time. All of the percentages reflect the power of the design, which is the percentage of times the design would be expected to find a $p < 0.05$.

Let's imagine that based on prior research, the effect you are interested in measuring is fairly small, d=0.2. If you want to run an experiment that will detect an effect of this size a large percentage of the time, how many subjects do you need to have in each group? We know from

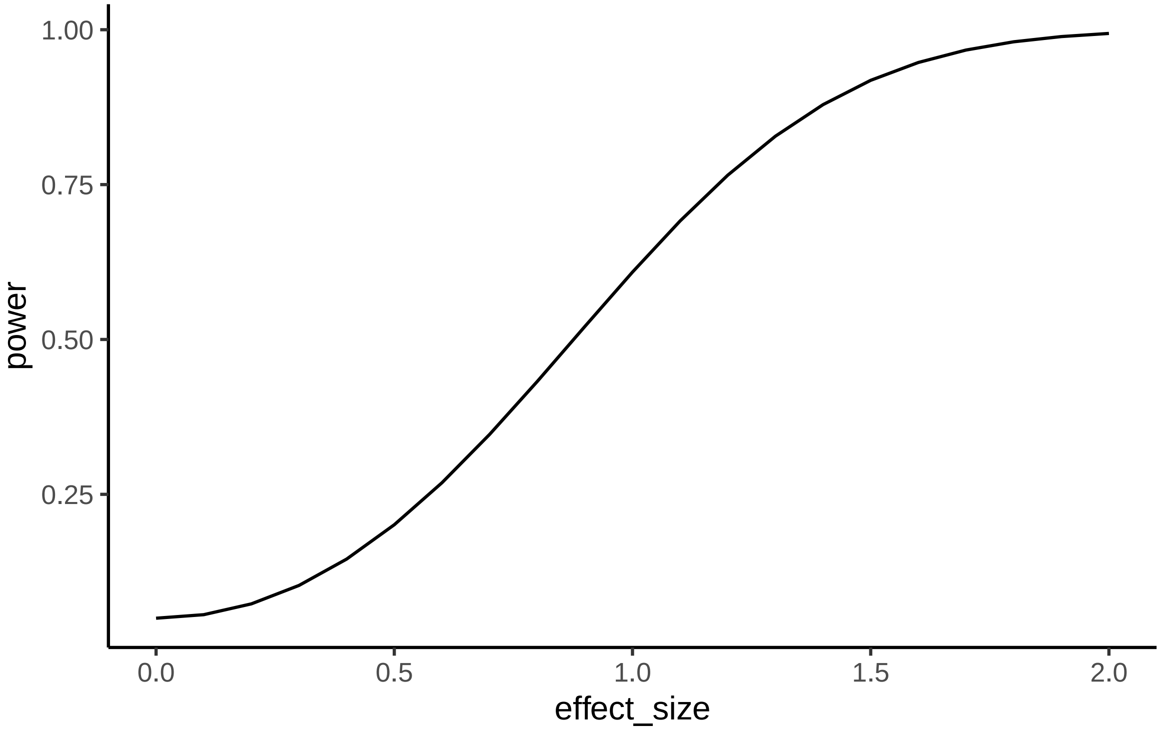


Figure 6.19: This figure shows power as a function of effect-size (Cohen's d) for a between-subjects independent samples t-test, with N=10, and alpha criterion 0.05.

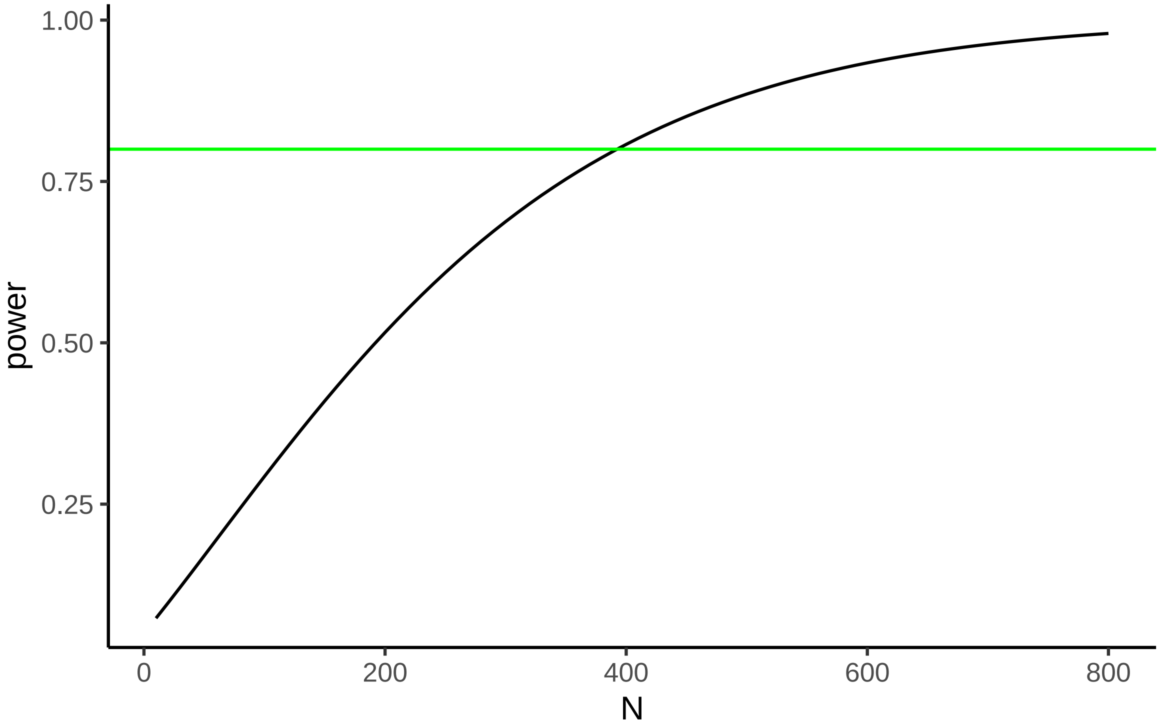


Figure 6.20: This figure shows power as a function of N for a between-subjects independent samples t-test, with d=0.2, and alpha criterion 0.05.

the above graph that with N=10, power is very low to detect an effect of d=0.2. Let's make Figure 6.20 and vary the number of subjects rather than the size of the effect.

The figure plots power to detect an effect of d=0.2, as a function of N. The green line shows where power = .8, or 80%. It looks like we would nee about 380 subjects in each group to measure an effect of d=0.2, with power = .8. This means that 80% of our experiments would succesfully show $p < 0.05$. Often times power of 80% is recommended as a reasonable level of power, however even when your design has power = 80%, your experiment will still fail to find an effect (associated with that level of power) 20% of the time!

6.7 Planning your design

Our discussion of effect size and power highlight the importance of the understanding the statistical limitations of an experimental design. In particular, we have seen the relationship between:

1. Sample-size
2. Effect-size
3. Alpha criterion
4. Power

As a general rule of thumb, small N designs can only reliably detect very large effects, whereas large N designs can reliably detect much smaller effects. As a researcher, it is your responsibility to plan your design accordingly so that it is capable of reliably detecting the kinds of effects it is intended to measure.

6.8 Some considerations

6.8.1 Low powered studies

Consider the following case. A researcher runs a study to detect an effect of interest. There is good reason, from prior research, to believe the effect-size is d=0.5. The researcher uses a design that has 30% power to detect the effect. They run the experiment and find a significant p-value, ($p<.05$). They conclude their manipulation worked, because it was unlikely that their result could have been caused by chance. How would you interpret the results of a study like this? Would you agree with thte researchers that the manipulation likely caused the difference? Would you be skeptical of the result?

The situation above requires thinking about two kinds of probabilities. On the one hand we know that the result observed by the researchers does not occur often by chance (p is less than 0.05). At the same time, we know that the design was underpowered, it only detects results of the expected size 30% of the time. We are face with wondering what kind of luck was driving the difference. The researchers could have gotten unlucky, and the difference really could be due to chance. In this case, they would be making a type I error (saying the result is real when it isn't). If the result was not due to chance, then they would also be lucky, as their design only detects this effect 30% of the time.

Perhaps another way to look at this situation is in terms of the replicability of the result. Replicability refers to whether or not the findings of the study would be the same if the experiment was repeated. Because we know that power is low here (only 30%), we would expect that most replications of this experiment would not find a significant effect. Instead, the experiment would be expected to replicate only 30% of the time.

6.8.2 Large N and small effects

Perhaps you have noticed that there is an intriguing relationship between N (sample-size) and power and effect-size. As N increases, so does power to detect an effect of a particular size. Additionally, as N increases, a design is capable of detecting smaller and smaller effects with greater and greater power. For example, if N was large enough, we would have high power to detect very small effects, say d= 0.01, or even d=0.001. Let's think about what this means.

Imagine a drug company told you that they ran an experiment with 1 billion people to test whether their drug causes a significant change in headache pain. Let's say they found a significant effect (with power =100%), but the effect was very small, it turns out the drug reduces headache pain by less than 1%, let's say 0.01%. For our imaginary study we will also assume that this effect is very real, and not caused by chance.

Clearly the design had enough power to detect the effect, and the effect was there, so the design did detect the effect. However, the issue is that there is little practical value to this effect. Nobody is going to by a drug to reduce their headache pain by 0.01%, even if it was "scientifcally proven" to work. This example brings up two issues. First, increasing N to very large levels will allow designs to detect almost any effect (even very tiny ones) with very high power. Second, sometimes effects are meaningless when they are very small, especially in applied research such as drug studies.

These two issues can lead to interesting suggestions. For example, someone might claim that large N studies aren't very useful, because they can always detect really tiny effects that are practically meaningless. On the other hand, large N studies will also detect larger effects too, and they will give a better estimate of the "true" effect in the population (because we know that larger samples do a better job of estimating population parameters). Additionally, although really small effects are often not interesting in the context of applied research, they can be very important in theoretical research. For example, one theory might predict that manipulating X should have no effect, but another theory might predict that X does have an effect, even if it is a small one. So, detecting a small effect can have theoretical implication that can help rule out false theories. Generally speaking, researchers asking both theoretical and applied questions should think about and establish guidelines for “meaningful” effect-sizes so that they can run designs of appropriate size to detect effects of “meaningful size”.

6.8.3 Small N and Large effects

All other things being equal would you trust the results from a study with small N or large N? This isn’t a trick question, but sometimes people tie themselves into a knot trying to answer it. We already know that large sample-sizes provide better estimates of the distributions the samples come from. As a result, we can safely conclude that we should trust the data from large N studies more than small N studies.

At the same time, you might try to convince yourself otherwise. For example, you know that large N studies can detect very small effects that are practically and possibly even theoretically meaningless. You also know that that small N studies are only capable of reliably detecting very large effects. So, you might reason that a small N study is better than a large N study because if a small N study detects an effect, that effect must be big and meaningful; whereas, a large N study could easily detect an effect that is tiny and meaningless.

This line of thinking needs some improvement. First, just because a large N study can detect small effects, doesn’t mean that it only detects small effects. If the effect is large, a large N study will easily detect it. Large N studies have the power to detect a much wider range of effects, from small to large. Second, just because a small N study detected an effect, does not mean that the effect is real, or that the effect is large. For example, small N studies have more variability, so the estimate of the effect size will have more error. Also, there is 5% (or alpha rate) chance that the effect was spurious. Interestingly, there is a pernicious relationship between effect-size and type I error rate

6.8.4 Type I errors are convincing when N is small

So what is this pernicious relationship between Type I errors and effect-size? Mainly, this relationship is pernicious for small N studies. For example, the following figure illustrates the results of 1000s of simulated experiments, all assuming the null distribution. In other words, for all of these simulations there is no true effect, as the numbers are all sampled from an identical distribution (normal distribution with mean =0, and standard deviation =1). The true effect-size is 0 in all cases.

We know that under the null, researchers will find p values that are less 5% about 5% of the time, remember that is the definition. So, if a researcher happened to be in this situation (where there manipulation did absolutely nothing), they would make a type I error 5% of the time, or if they conducted 100 experiments, they would expect to find a significant result for 5 of them.

Figure 6.21 reports the findings from only the type I errors, where the simulated study did produce $p < 0.05$. For each type I error, we calculated the exact p-value, as well as the effect-size (cohen's D) (mean difference divided by standard deviation). We already know that the true effect-size is zero, however take a look at this graph, and pay close attention to the smaller sample-sizes.

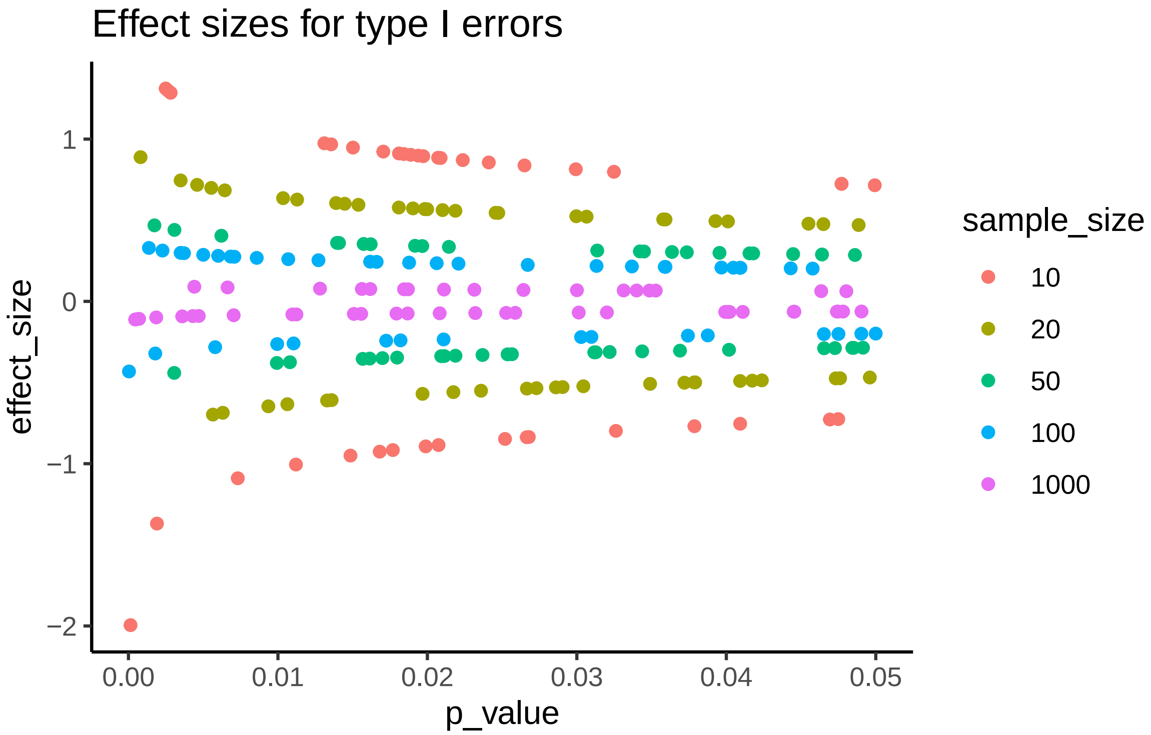


Figure 6.21: Effect size as a function of p-values for type 1 Errors under the null, for a paired samples t-test.

For example, look at the red dots, when sample size is 10. Here we see that the effect-sizes are quite large. When p is near 0.05 the effect-size is around .8, and it goes up and up as when p gets smaller and smaller. What does this mean? It means that when you get unlucky with a small N design, and your manipulation does not work, but you by chance find a "significant" effect, the effect-size measurement will show you a "big effect". This is the pernicious aspect. When you make a type I error for small N, your data will make you think there is no way it could be a type I error because the effect is just so big!. Notice that when N is very large, like 1000, the measure of effect-size approaches 0 (which is the true effect-size in the simulation shown in Figure 6.22).

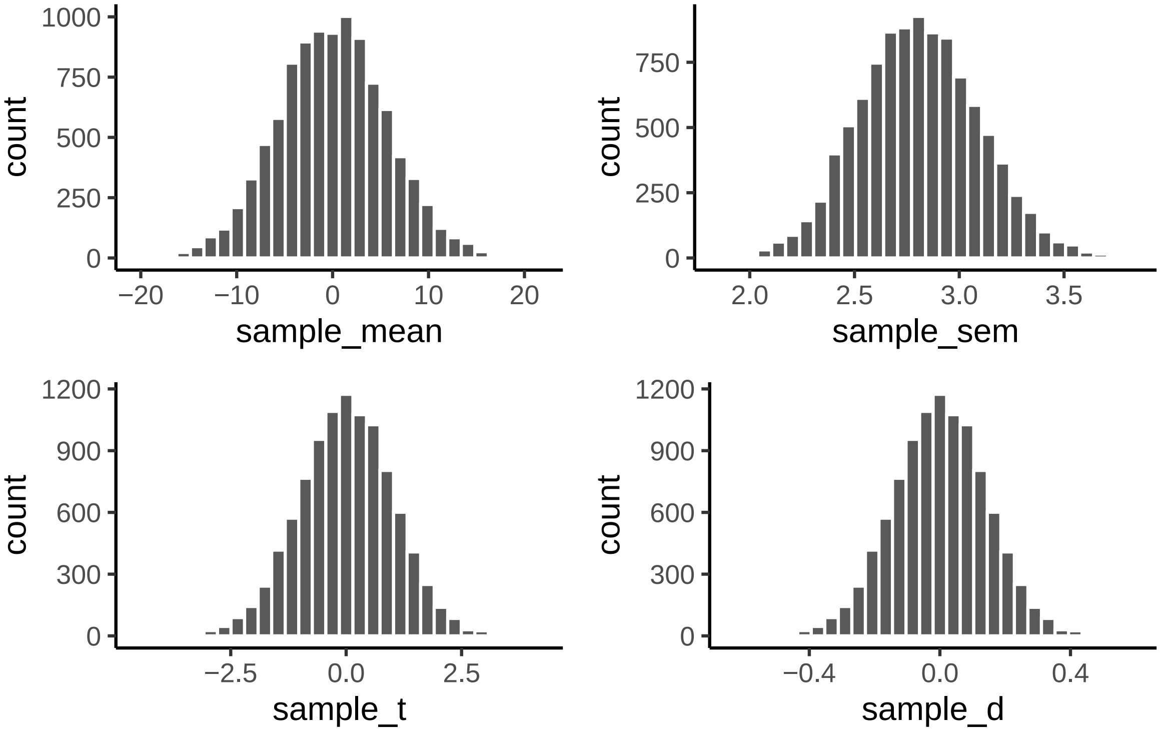


Figure 6.22: Each panel shows a histogram of a different sampling statistic.

7 t-tests

Back in the day, William Sealy Gosset got a job working for Guinness Breweries. They make the famous Irish stout called Guinness. What happens next went something like this (total fabrication, but mostly on point).

Guinness wanted all of their beers to be the best beers. No mistakes, no bad beers. They wanted to improve their quality control so that when Guinness was poured anywhere in the world, it would always comes out fantastic: 5 stars out of 5 every time, the best.

Guinness had some beer tasters, who were super-experts. Every time they tasted a Guinness from the factory that wasn't 5 out of 5, they knew right away.

But, Guinness had a big problem. They would make a keg of beer, and they would want to know if every single pint that would come out would be a 5 out of 5. So, the beer tasters drank pint after pint out of the keg, until it was gone. Some kegs were all 5 out of 5s. Some weren't, Guinness needed to fix that. But, the biggest problem was that, after the testing, there **was no beer left to sell**, the testers drank it all (remember I'm making this part up to illustrate a point, they probably still had beer left to sell).

Guinness had a sampling and population problem. They wanted to know that the entire population of the beers they made were all 5 out of 5 stars. But, if they sampled the entire population, they would drink all of their beer, and wouldn't have any left to sell.

Enter William Sealy Gosset. Gosset figured out the solution to the problem. He asked questions like this:

1. How many samples do I need to take to know the whole population is 5 out of 5?
2. What's the fewest amount of samples I need to take to know the above, that would mean Guinness could test fewer beers for quality, sell more beers for profit, and make the product testing time shorter.

Gosset solved those questions, and he invented something called the *Student's t-test*. Gosset was working for Guinness, and could be fired for releasing trade-secrets that he invented (the t-test). But, Gosset published the work anyways, under a pseudonym (Student 1908). He called himself Student, hence Student's t-test. Now you know the rest of the story.

It turns out this was a very nice thing for Gosset to have done. t-tests are used all the time, and they are useful, that's why they are used. In this chapter we learn how they work.

You'll be surprised to learn that what we've already talked about, (the Crump Test, and the Randomization Test), are both very very similar to the t-test. So, in general, you have already been thinking about the things you need to think about to understand t-tests. You're probably wondering what is this t, what does t mean? We will tell you. Before we tell what it means, we first tell you about one more idea.

7.1 Check your confidence in your mean

We've talked about getting a sample of data. We know we can find the mean, we know we can find the standard deviation. We know we can look at the data in a histogram. These are all useful things to do for us to learn something about the properties of our data.

You might be thinking of the mean and standard deviation as very different things that we would not put together. The mean is about central tendency (where most of the data is), and the standard deviation is about variance (where most of the data isn't). Yes, they are different things, but we can use them together to create useful new things.

What if I told you my sample mean was 50, and I told you nothing else about my sample. Would you be confident that most of the numbers were near 50? Would you wonder if there was a lot of variability in the sample, and many of the numbers were very different from 50. You should wonder all of those things. The mean alone, just by itself, doesn't tell you anything about well the mean represents all of the numbers in the sample.

It could be a representative number, when the standard deviation is very small, and all the numbers are close to 50. It could be a non-representative number, when the standard deviation is large, and many of the numbers are not near 50. You need to know the standard deviation in order to be confident in how well the mean represents the data.

How can we put the mean and the standard deviation together, to give us a new number that tells us about confidence in the mean?

We can do this using a ratio:

$\frac{mean}{\text{standard deviation}}$

Think about what happens here. We are dividing a number by a number. Look at what happens:

$\frac{number}{\text{same number}} = 1$

$\frac{number}{\text{smaller number}} = \text{big number}$

compared to:

$\frac{number}{\text{bigger number}} = \text{smaller number}$

Imagine we have a mean of 50, and a truly small standard deviation of 1. What do we get with our formula?

$\frac{50}{1} = 50$

Imagine we have a mean of 50, and a big standard deviation of 100. What do we get with our formula?

$\frac{50}{100} = 0.5$

Notice, when we have a mean paired with a small standard deviation, our formula gives us a big number, like 50. When we have a mean paired with a large standard deviation, our formula gives us a small number, like 0.5. These numbers can tell us something about confidence in our mean, in a general way. We can be 50 confident in our mean in the first case, and only 0.5 (not at a lot) confident in the second case.

What did we do here? We created a descriptive statistic by dividing the mean by the standard deviation. And, we have a sense of how to interpret this number, when it's big we're more confident that the mean represents all of the numbers, when it's small we are less confident. This is a useful kind of number, a ratio between what we think about our sample (the mean), and the variability in our sample (the standard deviation). Get used to this idea. Almost everything that follows in this textbook is based on this kind of ratio. We will see that our ratio turns into different kinds of "statistics", and the ratios will look like this in general:

$\text{name of statistic} = \frac{\text{measure of what we know}}{\text{measure of what we don't know}}$

or, to say it using different words:

$\text{name of statistic} = \frac{\text{measure of effect}}{\text{measure of error}}$

In fact, this is the general formula for the t-test. Big surprise!

7.2 One-sample t-test: A new t-test

Now we are ready to talk about t-test. We will talk about three of them. We start with the one-sample t-test.

Commonly, the one-sample t-test is used to estimate the chances that your sample came from a particular population. Specifically, you might want to know whether the mean that you found from your sample, could have come from a particular population having a particular mean.

Straight away, the one-sample t-test becomes a little confusing (and I haven't even described it yet). Officially, it uses known parameters from the population, like the mean of the population and the standard deviation of the population. However, most times you don't know those parameters of the population! So, you have to estimate them from your sample. Remember from the chapters on descriptive statistics and sampling, our sample mean is an unbiased estimate of the population mean. And, our sample standard deviation (the one where we divide by n-1) is an unbiased estimate of the population standard deviation. When Gosset developed the t-test, he recognized that he could use these estimates from his samples, to make the t-test. Here is the formula for the one sample t-test, we first use words, and then become more specific:

7.2.1 Formulas for one-sample t-test

name of statistic $= \frac{\text{measure of effect}}{\text{measure of error}}$

t $= \frac{\text{measure of effect}}{\text{measure of error}}$

t $= \frac{\text{Mean difference}}{\text{standard error}}$

t $= \frac{\bar{X}-u}{S_{\bar{X}}}$

t $= \frac{\text{Sample Mean - Population Mean}}{\text{Sample Standard Error}}$

Estimated Standard Error $=$ Standard Error of Sample $= \frac{s}{\sqrt{N}}$

Where, s is the sample standard deviation.

Some of you may have gone cross-eyed looking at all of this. Remember, we've seen it before when we divided our mean by the standard deviation in the first bit. The t-test is just a measure of a sample mean, divided by the standard error of the sample mean. That is it.

7.2.2 What does t represent?

t gives us a measure of confidence, just like our previous ratio for dividing the mean by a standard deviations. The only difference with *t*, is that we divide by the standard error of mean (remember, this is also a standard deviation, it is the standard deviation of the sampling distribution of the mean)

> ℹ Note
>
> What does the t in t-test stand for? Apparently nothing. Gosset originally labelled it z. And, Fisher later called it t, perhaps because t comes after s, which is often used for the sample standard deviation.

t is a property of the data that you collect. You compute it with a sample mean, and a sample standard error (there's one more thing in the one-sample formula, the population mean, which we get to in a moment). This is why we call *t*, a sample-statistic. It's a statistic we compute from the sample.

What kinds of numbers should we expect to find for these *ts*? How could we figure that out?

Let's start small and work through some examples. Imagine your sample mean is 5. You want to know if it came from a population that also has a mean of 5. In this case, what would t be? It would be zero: we first subtract the sample mean from the population mean, $5 - 5 = 0$. Because the numerator is 0, t will be zero. So, $t = 0$, occurs, when there is no difference.

Let's say you take another sample, do you think the mean will be 5 every time, probably not. Let's say the mean is 6. So, what can t be here? It will be a positive number, because 6-5= +1. But, will t be +1? That depends on the standard error of the sample. If the standard error of the sample is 1, then t could be 1, because 1/1 = 1.

If the sample standard error is smaller than 1, what happens to t? It get's bigger right? For example, 1 divided by 0.5 = 2. If the sample standard error was 0.5, t would be 2. And, what could we do with this information? Well, it be like a measure of confidence. As t get's bigger we could be more confident in the mean difference we are measuring.

Can t be smaller than 1? Sure, it can. If the sample standard error is big, say like 2, then t will be smaller than one (in our case), e.g., 1/2 = .5. The direction of the difference between the sample mean and population mean, can also make the t become negative. What if our sample mean was 4. Well, then t will be negative, because the mean difference in the numerator will be negative, and the number in the bottom (denominator) will always be positive (remember why, it's the standard error, computed from the sample standard deviation, which is always positive because of the squaring that we did.).

So, that is some intuitions about what the kinds of values t can take. t can be positive or negative, and big or small.

Let's do one more thing to build our intuitions about what t can look like. How about we sample some numbers and then measure the sample mean **and** the standard error of the mean, and then plot those two things against each each. This will show us how a sample mean typically varies with respect to the standard error of the mean.

In Figure 7.1, I pulled 1,000 samples of $N = 10$ from a normal distribution (mean = 0, sd = 1). Each time I measured the mean and standard error of the sample. That gave two descriptive statistics for each sample, letting us plot each sample as dot in a scatter plot.

What we get is a cloud of dots. You might notice the cloud has a circular quality. There's more dots in the middle, and fewer dots as they radiate out from the middle. The dot cloud shows us the general range of the sample mean, for example most of the dots are in between -1 and 1. Similarly, the range for the sample standard error is roughly between .2 and .5. Remember, each dot represents one sample.

We can look at the same data a different way. For example, rather than using a scatter plot, we can divide the mean for each dot by the standard error for each dot. Figure 7.2 shows the result in a histogram.

Interesting, we can see the histogram is shaped like a normal curve. It is centered on 0, which is the most common value. As values become more extreme, they become less common. If you

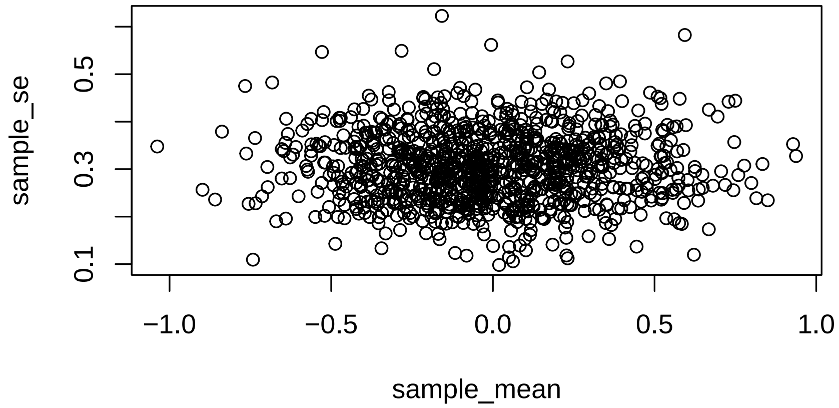


Figure 7.1: A scatter plot with sample mean on the x-axis, and standard error of the mean on the y-axis

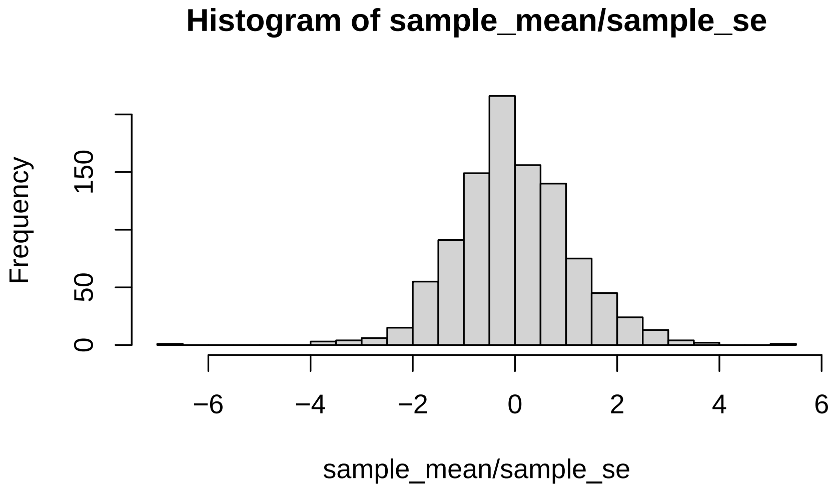


Figure 7.2: A histogram of the sample means divided by the sample standard errors, this is a t-distribution.

remember, our formula for t, was the mean divided by the standard error of the mean. That's what we did here. This histogram is showing you a t-distribution.

7.2.3 Calculating t from data

Let's briefly calculate a t-value from a small sample. Let's say we had 10 students do a true/false quiz with 5 questions on it. There's a 50% chance of getting each answer correct.

Every student completes the 5 questions, we grade them, and then we find their performance (mean percent correct). What we want to know is whether the students were guessing. If they were all guessing, then the sample mean should be about 50%, it shouldn't be different from chance, which is 50%. Let's look at Table 7.1.

Table 7.1: Calculating the t-value for a one-sample test.

students	scores	mean	Difference_from_Mean	Squared_Deviations
1	50	61	-11	121
2	70	61	9	81
3	60	61	-1	1
4	40	61	-21	441
5	80	61	19	361
6	30	61	-31	961
7	90	61	29	841
8	60	61	-1	1
9	70	61	9	81
10	60	61	-1	1
Sums	610	610	0	2890
Means	61	61	0	289
			sd	17.92
			SEM	5.67
			t	1.94003527336861

You can see the `scores` column has all of the test scores for each of the 10 students. We did the things we need to do to compute the standard deviation.

Remember the sample standard deviation is the square root of the sample variance, or:

sample standard deviation = $\sqrt{\frac{\sum_i^n (x_i - \bar{x})^2}{N-1}}$

sd = $\sqrt{\frac{2890}{10-1}} = 17.92$

The standard error of the mean, is the standard deviation divided by the square root of N

SEM = $\frac{s}{\sqrt{N}} = \frac{17.92}{10} = 5.67$

t is the difference between our sample mean (61), and our population mean (50, assuming chance), divided by the standard error of the mean.

t = $\frac{\bar{X}-u}{S_{\bar{X}}} = \frac{\bar{X}-u}{SEM} = \frac{61-50}{5.67} = 1.94$

And, that is you how calculate t, by hand. It's a pain. I was annoyed doing it this way. In the lab, you learn how to calculate t using software, so it will just spit out t. For example in R, all you have to do is this:

```
t.test(scores, mu=50)
#>
#>  One Sample t-test
#>
#> data:  scores
#> t = 1.9412, df = 9, p-value = 0.08415
#> alternative hypothesis: true mean is not equal to 50
#> 95 percent confidence interval:
#>  48.18111 73.81889
#> sample estimates:
#> mean of x
#>        61
```

7.2.4 How does t behave?

If t is just a number that we can compute from our sample (it is), what can we do with it? How can we use t for statistical inference?

Remember back to the chapter on sampling and distributions, that's where we discussed the sampling distribution of the sample mean. Remember, we made a lot of samples, then computed the mean for each sample, then we plotted a histogram of the sample means. Later, in that same section, we mentioned that we could generate sampling distributions for any statistic. For each sample, we could compute the mean, the standard deviation, the standard error, and now even t, if we wanted to. We could generate 10,000 samples, and draw four histograms, one for each sampling distribution for each statistic.

This is exactly what I did, and the results are shown in the four panels of Figure 7.3 below. I used a sample size of 20, and drew random observations for each sample from a normal distribution, with mean = 0, and standard deviation = 1. Let's look at the sampling distributions for each of the statistics. t was computed assuming with the population mean assumed to be 0.

We see four sampling distributions. This is how statistical summaries of these summaries behave. We have used the word chance windows before. These are four chance windows,

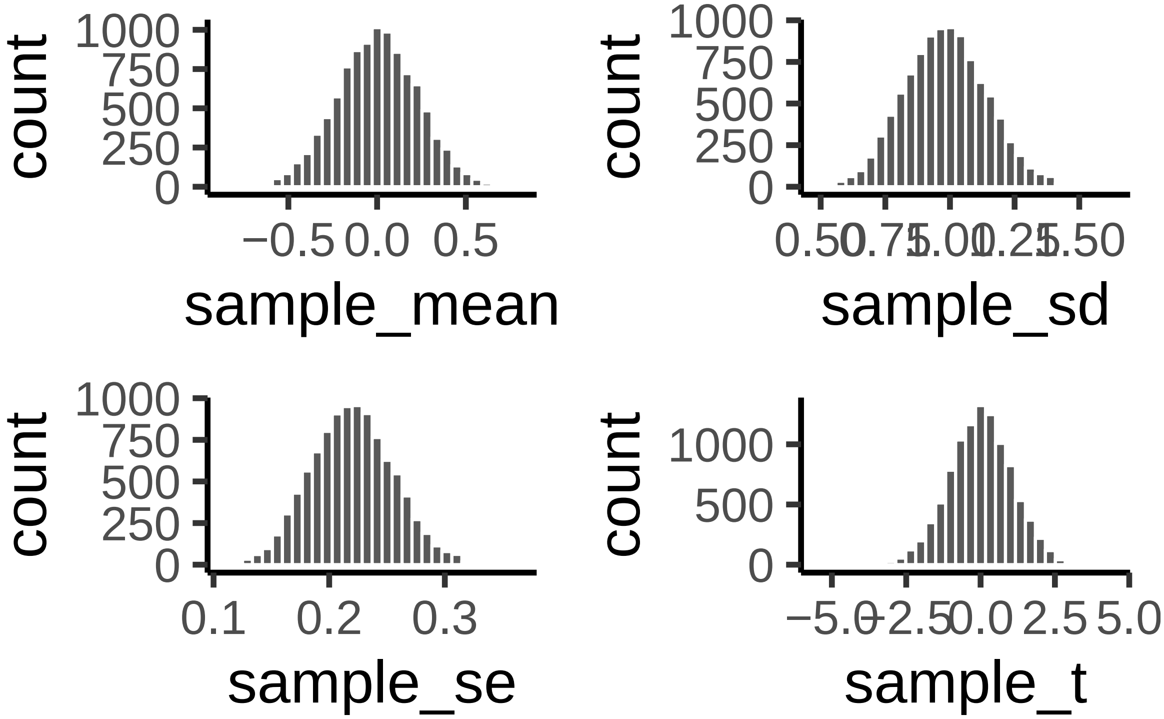


Figure 7.3: Sampling distributions for the mean, standard deviation, standard error of the mean, and t.

measuring different aspects of the sample. In this case, all of the samples came from the same normal distribution. Because of sampling error, each sample is not identical. The means are not identical, the standard deviations are not identical, sample standard error of the means are not identical, and the *t*s of the samples are not identical. They all have some variation, as shown by the histograms. This is how samples of size 20 behave.

We can see straight away, that in this case, we are unlikely to get a sample mean of 2. That's way outside the window. The range for the sampling distribution of the mean is around -.5 to +.5, and is centered on 0 (the population mean, would you believe!).

We are unlikely to get sample standard deviations of between .6 and 1.5, that is a different range, specific to the sample standard deviation.

Same thing with the sample standard error of the mean, the range here is even smaller, mostly between .1, and .3. You would rarely find a sample with a standard error of the mean greater than .3. Virtually never would you find one of say 1 (for this situation).

Now, look at *t*. It's range is basically between -3 and +3 here. 3s barely happen at all. You pretty much never see a 5 or -5 in this situation.

All of these sampling windows are chance windows, and they can all be used in the same way as we have used similar sampling distributions before (e.g., Crump Test, and Randomization Test) for statistical inference. For all of them we would follow the same process:

1. Generate these distributions
2. Look at your sample statistics for the data you have (mean, SD, SEM, and *t*)
3. Find the likelihood of obtaining that value or greater
4. Obtain that probability
5. See if you think your sample statistics were probable or improbable.

We'll formalize this in a second. I just want you to know that what you will be doing is something that you have already done before. For example, in the Crump test and the Randomization test we focused on the distribution of mean differences. We could do that again here, but instead, we will focus on the distribution of *t* values. We then apply the same kinds of decision rules to the *t* distribution, as we did for the other distributions. Below you will see a graph you have already seen, except this time it is a distribution of *t*s, not mean differences:

Remember, if we obtained a single *t* from one sample we collected, we could consult the chance window in Figure 7.4 below to find out whether the *t* we obtained from the sample was likely or unlikely to occur by chance.

7.2.5 Making a decision

From our early example involving the TRUE/FALSE quizzes, we are now ready to make some kind of decision about what happened there. We found a mean difference of 11. We found a *t*

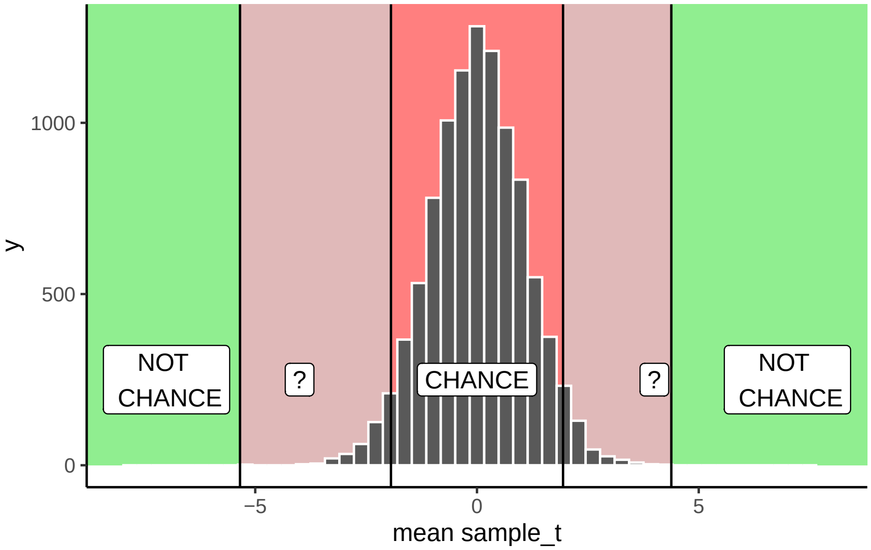


Figure 7.4: Applying decision criteria to the t-distribution. Histogram of ts from samples (n=20) drawn from the same normal distribution (u=0, sd=1)

= 1.9411765. The probability of this t or larger occurring is $p = 0.0841503$. We were testing the idea that our sample mean of 61 could have come from a normal distribution with mean = 50. The t test tells us that the t for our sample, or a larger one, would happen with p = 0.0841503. In other words, chance can do it a kind of small amount of time, but not often. In English, this means that all of the students could have been guessing, but it wasn't that likely that were just guessing.

The next t-test is called a **paired samples t-test**. And, spoiler alert, we will find out that a paired samples t-test is actually a one-sample t-test in disguise (WHAT!), yes it is. If the one-sample t-test didn't make sense to you, read the next section.

7.3 Paired-samples t-test

For me (Crump), many analyses often boil down to a paired samples t-test. It just happens that many things I do reduce down to a test like this.

I am a cognitive psychologist, I conduct research about how people do things like remember, pay attention, and learn skills. There are lots of Psychologists like me, who do very similar things.

We all often conduct the same kinds of experiments. They go like this, and they are called **repeated measures** designs. They are called **repeated measures** designs, because we measure how one person does something more than once, we **repeat** the measure.

So, I might measure somebody doing something in condition A, and measure the same person doing something in Condition B, and then I see that same person does different things in the two conditions. I **repeatedly measure** the same person in both conditions. I am interested in whether the experimental manipulation changes something about how people perform the task in question.

7.3.1 Mehr, Song, and Spelke (2016)

We will introduce the paired-samples t-test with an example using real data, from a real study. Mehr, Song, and Spelke (2016) were interested in whether singing songs to infants helps infants become more sensitive to social cues. For example, infants might need to learn to direct their attention toward people as a part of learning how to interact socially with people. Perhaps singing songs to infants aids this process of directing attention. When an infant hears a familiar song, they might start to pay more attention to the person singing that song, even after they are done singing the song. The person who sang the song might become more socially important to the infant. You will learn more about this study in the lab for this week. This example, prepares you for the lab activities. Here is a brief summary of what they did.

First, parents were trained to sing a song to their infants. After many days of singing this song to the infants, a parent came into the lab with their infant. In the first session, parents sat with their infants on their knees, so the infant could watch two video presentations. There were two videos. Each video involved two unfamiliar new people the infant had never seen before. Each new person in the video (the singers) sang one song to the infant. One singer sang the “familiar” song the infant had learned from their parents. The other singer sang an “unfamiliar” song the infant had not hear before.

There were two really important measurement phases: the baseline phase, and the test phase.

The baseline phase occurred before the infants saw and heard each singer sing a song. During the baseline phase, the infants watched a video of both singers at the same time. The researchers recorded the proportion of time that the infant looked at each singer. The baseline phase was conducted to determine whether infants had a preference to look at either person (who would later sing them a song).

The test phase occurred **after** infants saw and heard each song, sung by each singer. During the test phase, each infant had an opportunity to watch silent videos of both singers. The researchers measured the proportion of time the infants spent looking at each person. The question of interest, was whether the infants would spend a greater proportion of time looking at the singer who sang the familiar song, compared to the singer who sang the unfamiliar song.

There is more than one way to describe the design of this study. We will describe it like this. It was a repeated measures design, with one independent (manipulation) variable called Viewing phase: Baseline versus Test. There was one dependent variable (the measurement), which was proportion looking time (to singer who sung familiar song). This was a repeated measures design because the researchers measured proportion looking time twice (they repeated the measure), once during baseline (before infants heard each singer sing a song), and again during test (after infants head each singer sing a song).

The important question was whether infants would change their looking time, and look more at the singer who sang the familiar song during the test phase, than they did during the baseline phase. This is a question about a change within individual infants. In general, the possible outcomes for the study are:

1. No change: The difference between looking time toward the singer of the familiar song during baseline and test is zero, no difference.
2. Positive change: Infants will look longer toward the singer of the familiar song during the test phase (after they saw and heard the singers), compared to the baseline phase (before they saw and heard the singers). This is a positive difference if we use the formula: Test Phase Looking time - Baseline phase looking time (to familiar song singer).
3. Negative change: Infants will look longer toward the singer of the unfamiliar song during the test phase (after they saw and heard the singers), compared to the baseline phase (before they saw and heard the singers). This is a negative difference if we use the same formula: Test Phase Looking time - Baseline phase looking time (to familiar song singer).

7.3.2 The data

Let's take a look at the data for the first 5 infants in the study. This will help us better understand some properties of the data before we analyze it. We will see that the data is structured in a particular way that we can take advantage of with a paired samples t-test. Note, we look at the first 5 infants to show how the computations work. The results of the paired-samples t-test change when we use all of the data from the study.

Here is a table of the data:

infant	Baseline	Test
1	0.44	0.60
2	0.41	0.68
3	0.75	0.72
4	0.44	0.28
5	0.47	0.50

The table shows proportion looking times toward the singer of the familiar song during the Baseline and Test phases. Notice there are five different infants, (1 to 5). Each infant is measured twice, once during the Baseline phase, and once during the Test phase. To repeat from before, this is a repeated-measures design, because the infants are measured repeatedly (twice in this case). Or, this kind of design is also called a **paired-samples** design. Why? because each participant comes with a pair of samples (two samples), one for each level of the design.

Great, so what are we really interested in here? We want to know if the mean looking time toward the singer of the familiar song for the Test phase is higher than the Baseline phase. We are comparing the two sample means against each other and looking for a difference. We already know that differences could be obtained by chance alone, simply because we took two sets of samples, and we know that samples can be different. So, we are interested in knowing whether chance was likely or unlikely to have produced any difference we might observe.

In other words, we are interested in looking at the difference scores between the baseline and test phase for each infant. The question here is, for each infant, did their proportion looking time to the singer of the familiar song, increase during the test phase as compared to the baseline phase.

7.3.3 The difference scores

Let's add the difference scores to the table of data so it is easier to see what we are talking about. The first step in creating difference scores is to decide how you will take the difference, there are two options:

1. Test phase score - Baseline Phase Score
2. Baseline phase score - Test Phase score

Let's use the first formula. Why? Because it will give us positive differences when the test phase score is higher than the baseline phase score. This makes a positive score meaningful with respect to the study design, we know (because we defined it to be this way), that positive scores will refer to longer proportion looking times (to singer of familiar song) during the test phase compared to the baseline phase.

infant	Baseline	Test	differences
1	0.44	0.60	0.16
2	0.41	0.68	0.27
3	0.75	0.72	-0.03
4	0.44	0.28	-0.16
5	0.47	0.50	0.03

There we have it, the difference scores. The first thing we can do here is look at the difference scores, and ask how many infants showed the effect of interest. Specifically, how many infants showed a positive difference score. We can see that three of five infants showed a positive difference (they looked more at the singer of the familiar song during the test than baseline phase), and two the infants showed the opposite effect (negative difference, they looked more at the singer of the familiar song during baseline than test).

7.3.4 The mean difference

As we have been discussing, the effect of interest in this study is the mean difference between the baseline and test phase proportion looking times. We can calculate the **mean difference**, by finding the **mean of the difference scores**. Let's do that, in fact, for fun let's calculate the mean of the baseline scores, the test scores, and the difference scores.

infant	Baseline	Test	differences
1	0.44	0.6	0.16
2	0.41	0.68	0.27
3	0.75	0.72	-0.03
4	0.44	0.28	-0.16
5	0.47	0.5	0.03
Sums	2.51	2.78	0.27
Means	0.502	0.556	0.054

We can see there was a positive mean difference of 0.054, between the test and baseline phases.

Can we rush to judgment and conclude that infants are more socially attracted to individuals who have sung them a familiar song? I would hope not based on this very small sample. First, the difference in proportion looking isn't very large, and of course we recognize that this difference could have been produced by chance.

We will more formally evaluate whether this difference could have been caused by chance with the paired-samples t-test. But, before we do that, let's again calculate t and discuss what t tells us over and above what our measure of the mean of the difference scores tells us.

7.3.5 Calculate t

OK, so how do we calculate t for a paired-samples t-test? Surprise, we use the one-sample t-test formula that you already learned about! Specifically, we use the one-sample t-test formula on the difference scores. We have one sample of difference scores (you can see they are in one column), so we can use the one-sample t-test on the difference scores. Specifically, we are interested in comparing whether the mean of our difference scores came from a distribution with mean difference = 0. This is a special distribution we refer to as the **null distribution**. It is the distribution no differences. Of course, this **null distribution** can produce differences due to to sampling error, but those differences are not caused by any experimental manipulation, they caused by the random sampling process.

We calculate *t* in a moment. Let's now consider again why we want to calculate *t*? Why don't we just stick with the mean difference we already have?

Remember, the whole concept behind *t*, is that it gives an indication of how confident we should be in our mean. Remember, *t* involves a measure of the mean in the numerator, divided by a measure of variation (standard error of the sample mean) in the denominator. The resulting *t* value is small when the mean difference is small, or when the variation is large. So small *t*-values tell us that we shouldn't be that confident in the estimate of our mean difference. Large *t*-values occur when the mean difference is large and/or when the measure of variation is small. So, large *t*-values tell us that we can be more confident in the estimate of our mean difference. Let's find *t* for the mean difference scores. We use the same formulas as we did last time:

infant	Baseline	Test	differences	diff_from_mean	Squared_differences
1	0.44	0.6	0.16	0.106	0.011236
2	0.41	0.68	0.27	0.216	0.046656
3	0.75	0.72	-0.03	-0.084	0.00705600000000001
4	0.44	0.28	-0.16	-0.214	0.045796
5	0.47	0.5	0.03	-0.024	0.000575999999999999
Sums	2.51	2.78	0.27	0	0.11132
Means	0.502	0.556	0.054	0	0.022264
				sd	0.167
				SEM	0.075
				t	0.72

If we did this test using R, we would obtain almost the same numbers (there is a little bit of rounding in the table).

```
#>
#>  One Sample t-test
#>
#> data:  differences
#> t = 0.72381, df = 4, p-value = 0.5092
#> alternative hypothesis: true mean is not equal to 0
#> 95 percent confidence interval:
#>  -0.1531384  0.2611384
#> sample estimates:
#> mean of x
#>     0.054
```

Here is a quick write up of our t-test results, t(4) = .72, p = .509.

What does all of that tell us? There's a few things we haven't gotten into much yet. For example, the 4 represents degrees of freedom, which we discuss later. The important part, the t value should start to be a little bit more meaningful. We got a kind of small t-value didn't we. It's .72. What can we tell from this value? First, it is positive, so we know the mean difference is positive. The sign of the t-value is always the same as the sign of the mean difference (ours was +0.054). We can also see that the p-value was .509. We've seen p-values before. This tells us that our t value or larger, occurs about 50.9% of the time… Actually it means more than this. And, to understand it, we need to talk about the concept of two-tailed and one-tailed tests.

7.3.6 Interpreting ts

Remember what it is we are doing here. We are evaluating whether our sample data could have come from a particular kind of distribution. The null distribution of no differences. This is the distribution of t-values that would occur for samples of size 5, with a mean difference of 0, and a standard error of the sample mean of .075 (this is the SEM that we calculated from our sample). We can see what this particular null-distribution looks like in Figure 7.5.

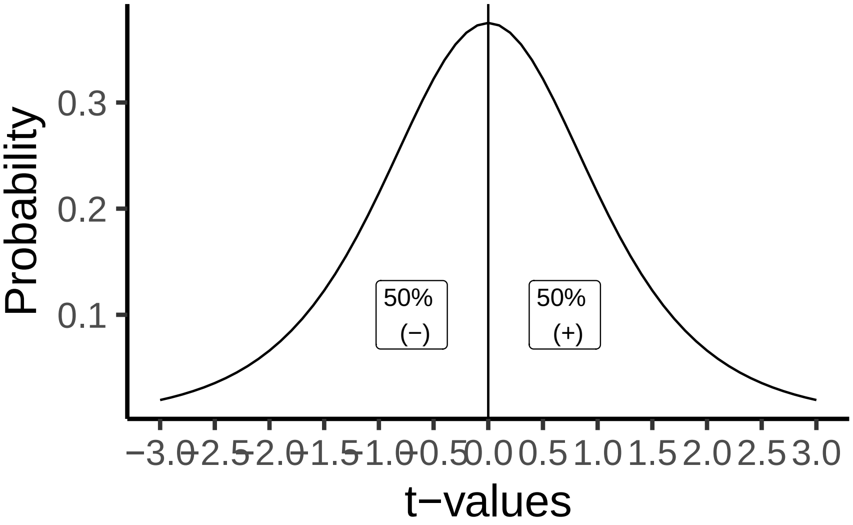


Figure 7.5: A distribution of t-values that can occur by chance alone, when there is no difference between the sample and a population

The t-distribution above shows us the kinds of values t will will take by chance alone, when we measure the mean differences for pairs of 5 samples (like our current). t is most likely to be zero, which is good, because we are looking at the distribution of no-differences, which should most often be 0! But, sometimes, due to sampling error, we can get *t*s that are bigger than 0, either in the positive or negative direction. Notice the distribution is symmetrical, a *t* from the null-distribution will be positive half of the time, and negative half of the time, that is what we would expect by chance.

So, what kind of information do we want know when we find a particular *t* value from our sample? We want to know how likely the *t* value like the one we found occurs just by chance. This is actually a subtly nuanced kind of question. For example, any particular *t* value doesn't have a specific probability of occurring. When we talk about probabilities, we are talking about ranges of probabilities. Let's consider some probabilities. We will use the letter *p*, to talk about the probabilities of particular *t* values.

1. What is the probability that *t* is zero or positive or negative? The answer is p=1, or 100%. We will always have a *t* value that is zero or non-zero…Actually, if we can't compute the t-value, for example when the standard deviation is undefined, I guess then we would have a non-number. But, assuming we can calculate *t*, then it will always be 0 or positive or negative.
2. What is the probability of $t = 0$ or greater than 0? The answer is p=.5, or 50%. 50% of *t*-values are 0 or greater.
3. What is the of $t = 0$ or smaller than 0? The answer is p=.5, or 50%. 50% of *t*-values are 0 or smaller.

We can answer all of those questions just by looking at our t-distribution, and dividing it into two equal regions, the left side (containing 50% of the *t* values), and the right side containing 50% of the *t*-values).

What if we wanted to take a more fine-grained approach, let's say we were interested in regions of 10%. What kinds of *t*s occur 10% of the time. We would apply lines like the following. Notice, the likelihood of bigger numbers (positive or negative) gets smaller, so we have to increase the width of the bars for each of the intervals between the bars to contain 10% of the *t*-values, it looks like Figure 7.6.

Consider the probabilities (*p*) of *t* for the different ranges.

1. *t* <= -1.5 (*t* is less than or equal to -1.5), $p = 10\%$
2. -1.5 >= *t* <= -0.9 (*t* is equal to or between -1.5 and -.9), $p = 10\%$
3. -.9 >= *t* <= -0.6 (*t* is equal to or between -.9 and -.6), $p = 10\%$
4. *t* >= 1.5 (*t* is greater than or equal to 1.5), $p = 10\%$

Notice, that the *p*s are always 10%. *t*s occur in these ranges with 10% probability.

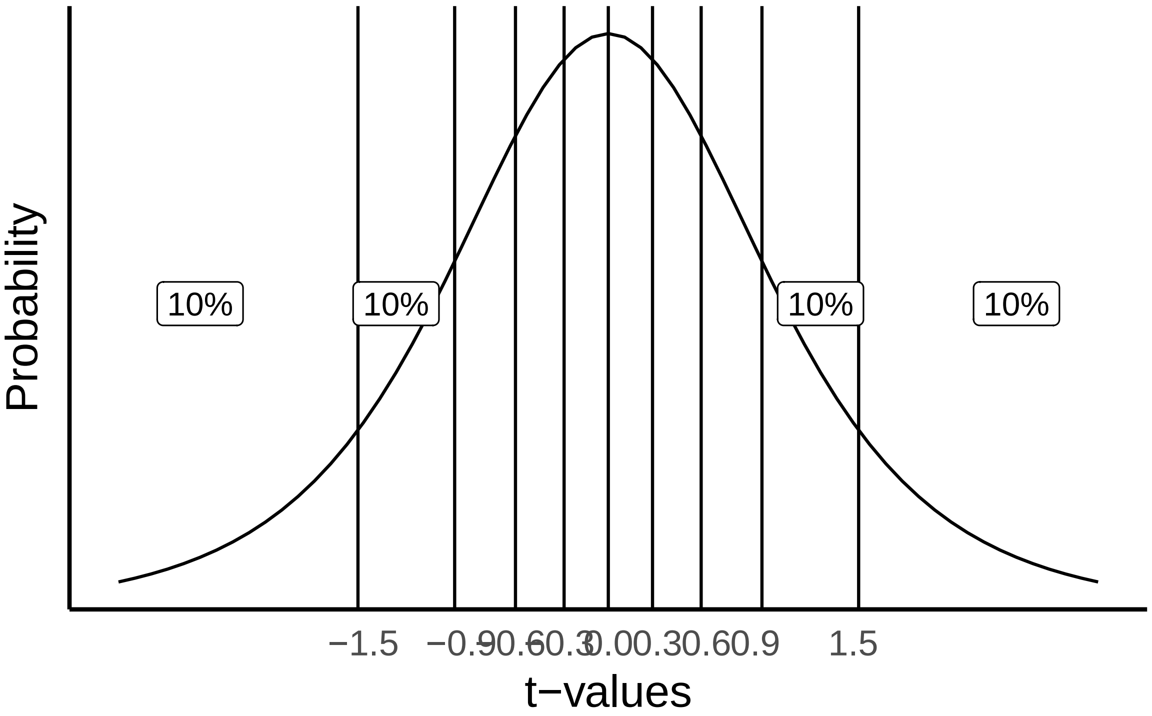


Figure 7.6: Splitting the t distribution up into regions each containing 10% of the *t*-values. The width between the bars narrows as they approach the center of the distribution, where there are more *t*-values.

7.3.7 Getting the p-values for t-values

You might be wondering where I am getting some of these values from. For example, how do I know that 10% of t values (for this null distribution) have a value of approximately 1.5 or greater than 1.5? The answer is I used R to tell me.

In most statistics textbooks the answer would be: there is a table at the back of the book where you can look these things up…This textbook has no such table. We could make one for you. And, we might do that. But, we didn't do that yet…

So, where do these values come from, how can you figure out what they are? The complicated answer is that we are not going to explain the math behind finding these values because, 1) the authors (some of us) admittedly don't know the math well enough to explain it, and 2) it would sidetrack us to much, 3) you will learn how to get these numbers in the lab with software, 4) you will learn how to get these numbers in lab without the math, just by doing a simulation, and 5) you can do it in R, or excel, or you can use an online calculator.

This is all to say that you can find the ts and their associated ps using software. But, the software won't tell you what these values mean. That's we are doing here. You will also see that software wants to know a few more things from you, such as the degrees of freedom for the test, and whether the test is one-tailed or two tailed. We haven't explained any of these things yet. That's what we are going to do now. Note, we explain degrees of freedom last. First, we start with a one-tailed test.

7.3.8 One-tailed tests

A **one-tailed test** is sometimes also called a directional test. It is called a directional test, because a researcher might have a hypothesis in mind suggesting that the difference they observe in their means is going to have a particular direction, either a positive difference, or a negative difference.

Typically, a researcher would set an **alpha criterion**. The alpha criterion describes a line in the sand for the researcher. Often, the alpha criterion is set at $p = .05$. What does this mean? Figure 7.7 shows the t-distribution and the alpha criterion.

The figure shows that t values of +2.13 or greater occur 5% of the time. Because the t-distribution is symmetrical, we also know that t values of -2.13 or smaller also occur 5% of the time. Both of these properties are true under the null distribution of no differences. This means, that when there really are no differences, a researcher can expect to find t values of 2.13 or larger 5% of the time.

Let's review and connect some of the terms:

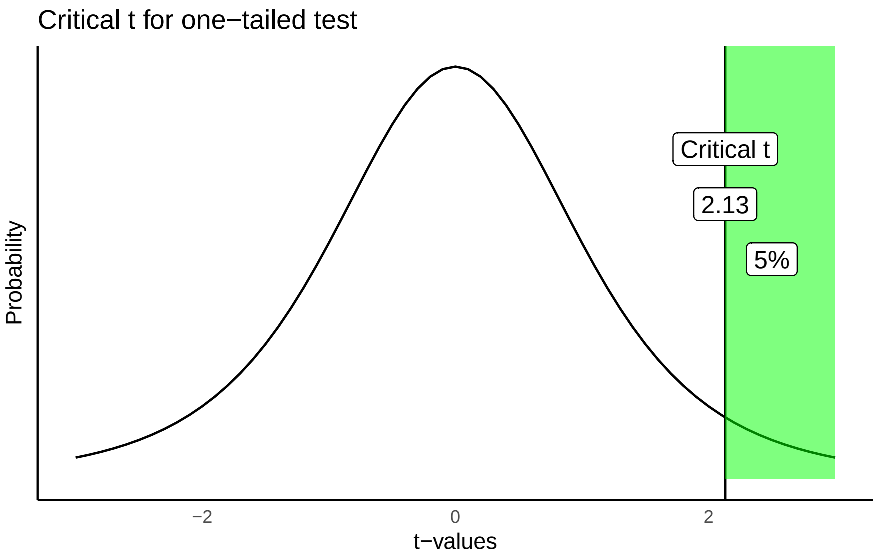


Figure 7.7: The critical value of t for an alpha criterion of 0.05. 5% of all ts are at this value or larger

1. **alpha criterion**: the criterion set by the researcher to make decisions about whether they believe chance did or did not cause the difference. The alpha criterion here is set to $p = .05$.
2. **Critical** ***t***. The critical t is the t-value associated with the alpha-criterion. In this case for a one-tailed test, it is the t value where 5% of all ts are this number or greater. In our example, the critical t is 2.13. 5% of all t values (with degrees of freedom = 4) are +2.13, or greater than +2.13.
3. **Observed** ***t***. The observed t is the one that you calculated from your sample. In our example about the infants, the observed t was t (4) = 0.72.
4. **p-value**. The p-value is the probability of obtaining the observed t value or larger. Now, you could look back at our previous example, and find that the p-value for t (4) = .72, was $p = .509$. HOWEVER, this p-value was not calculated for a one-directional test…(we talk about what .509 means in the next section).

Figure 7.8 shows what the p-value for t (4) = .72 using a one-directional test would would look like:

Let's take this one step at a time. We have located the observed t of .72 on the graph. We shaded the right region all grey. What we see is that the grey region represents .256 or 25.6% of all t values. In other words, 25.6% of t values are .72 or larger than .72. You could expect, by chance alone, to a find a t value of .72 or larger, 25.6% of the time. That's fairly often. We did find a t value of .72. Now that you know this kind of t value or larger occurs 25.6% of the

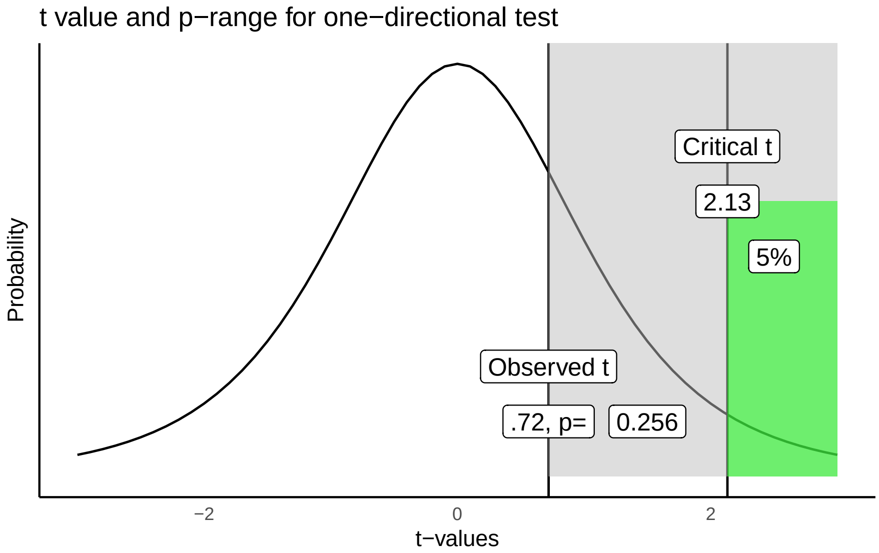


Figure 7.8: A case where the observed value of t is much less than the critical value for a one-directional t-test.

time, would you be confident that the mean difference was not due to chance? Probably not, given that chance can produce this difference fairly often.

Following the "standard" decision making procedure, we would claim that our t value was **not statistically significant**, because it was not large enough. If our observed value was larger than the critical t (larger than 2.13), defined by our alpha criterion, then we would claim that our t value was **statistically significant**. This would be equivalent to saying that we believe it is unlikely that the difference we observed was due to chance. In general, for any observed t value, the associated p-value tells you how likely a t of the observed size or larger would be observed. The p-value **always** refers to a **range** of t-values, never to a single t-value. Researchers use the alpha criterion of .05, as a matter of convenience and convention. There are other ways to interpret these values that do not rely on a strict (significant versus not) dichotomy.

7.3.9 Two-tailed tests

OK, so that was one-tailed tests... What are two tailed tests? The p-value that we originally calculated from our paired-samples t-test was for a 2-tailed test. Often, the default is that the p-value is for a two-tailed test.

The two-tailed test, is asking a more general question about whether a difference is likely to have been produced by chance. The question is: what is probability of any difference. It is also called a **non-directional** test, because here we don't care about the direction or sign of the difference (positive or negative), we just care if there is any kind of difference.

The same basic things as before are involved. We define an alpha criterion ($\alpha = 0.05$). And, we say that any observed t value that has a probability of $p <.05$ (p is less than .05) will be called **statistically significant**, and ones that are more likely ($p >.05$, p is greater than .05) will be called null-results, or not statistically significant. The only difference is how we draw the alpha range. Before it was on the right side of the t distribution (we were conducting a one-sided test remember, so we were only interested in one side).

Figure 7.9 shows what the most extreme 5% of the t-values are when we ignore their sign (whether they are positive or negative).

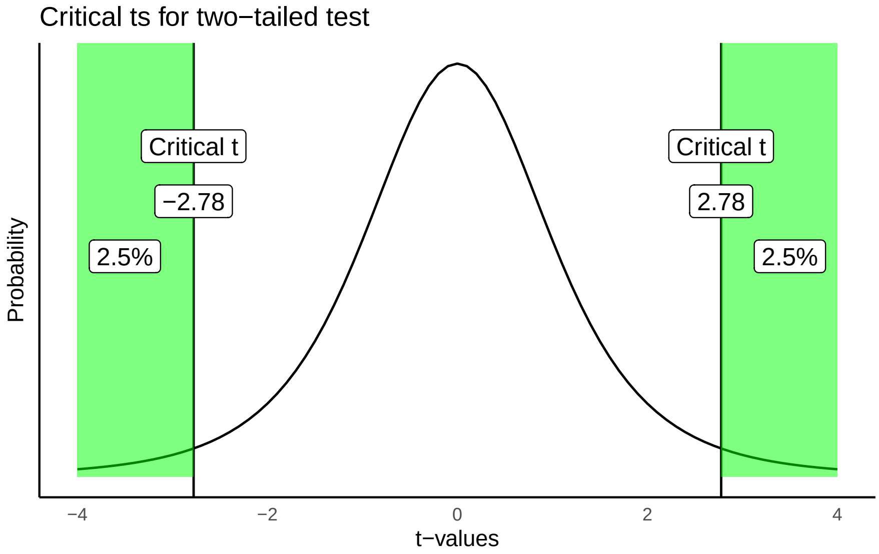


Figure 7.9: Critical values for a two-tailed test. Each line represents the location where 2.5% of all ts are larger or smaller than critical value. The total for both tails is 5%

Here is what we are seeing. A distribution of no differences (the null, which is what we are looking at), will produce ts that are 2.78 or greater 2.5% of the time, and ts that are -2.78 or smaller 2.5% of the time. 2.5% + 2.5% is a total of 5% of the time. We could also say that ts larger than +/- 2.78 occur 5% of the time.

As a result, the critical t value is (+/-) 2.78 for a two-tailed test. As you can see, the two-tailed test is blind to the direction or sign of the difference. Because of this, the critical t value is also higher for a two-tailed test, than for the one-tailed test that we did earlier. Hopefully, now you can see why it is called a two-tailed test. There are two tails of the distribution, one on the left and right, both shaded in green.

7.3.10 One or two tailed, which one?

Now that you know there are two kinds of tests, one-tailed, and two-tailed, which one should you use? There is some conventional wisdom on this, but also some debate. In the end, it is up to you to be able to justify your choice and why it is appropriate for you data. That is the real answer.

The conventional answer is that you use a one-tailed test when you have a theory or hypothesis that is making a directional prediction (the theory predicts that the difference will be positive, or negative). Similarly, use a two-tailed test when you are looking for any difference, and you don't have a theory that makes a directional prediction (it just makes the prediction that there will be a difference, either positive or negative).

Also, people appear to choose one or two-tailed tests based on how risky they are as researchers. If you always ran one-tailed tests, your critical t values for your set alpha criterion would always be smaller than the critical ts for a two-tailed test. Over the long run, you would make more type I errors, because the criterion to detect an effect is a lower bar for one than two tailed tests.

> Remember type 1 errors occur when you reject the idea that chance could have caused your difference. You often never know when you make this error. It happens anytime that sampling error was the actual cause of the difference, but a researcher dismisses that possibility and concludes that their manipulation caused the difference.

Similarly, if you always ran two-tailed tests, even when you had a directional prediction, you would make fewer type I errors over the long run, because the t for a two-tailed test is higher than the t for a one-tailed test. It seems quite common for researchers to use a more conservative two-tailed test, even when they are making a directional prediction based on theory. In practice, researchers tend to adopt a standard for reporting that is common in their field. Whether or not the practice is justifiable can sometimes be an open question. The important task for any researcher, or student learning statistics, is to be able to justify their choice of test.

7.3.11 Degrees of freedom

Before we finish up with paired-samples t-tests, we should talk about degrees of freedom. Our sense is that students don't really understand degrees of freedom very well. If you are reading this textbook, you are probably still wondering what is degrees of freedom, seeing as we haven't really talked about it all.

For the t-test, there is a formula for degrees of freedom. For the one-sample and paired sample t-tests, the formula is:

Degrees of Freedom = df = $n - 1$. Where n is the number of samples in the test.

In our paired t-test example, there were 5 infants. Therefore, degrees of freedom = 5-1 = 4.

OK, that's a formula. Who cares about degrees of freedom, what does the number mean? And why do we report it when we report a t-test… you've probably noticed the number in parentheses e.g., $t(4)$=.72, the 4 is the df, or degrees of freedom.

Degrees of freedom is both a concept, and a correction. The concept is that if you estimate a property of the numbers, and you use this estimate, you will be forcing some constraints on your numbers.

Consider the numbers: 1, 2, 3. The mean of these numbers is 2. Now, let's say I told you that the mean of three numbers is 2. Then, how many of these three numbers have freedom? Funny question right. What we mean is, how many of the three numbers could be any number, or have the freedom to be any number.

The first two numbers could be any number. But, once those two numbers are set, the final number (the third number), MUST be a particular number that makes the mean 2. The first two numbers have freedom. The third number has no freedom.

To illustrate. Let's freely pick two numbers: 51 and -3. I used my personal freedom to pick those two numbers. Now, if our three numbers are 51, -3, and x, and the mean of these three numbers is 2. There is only one solution, x has to be -42, otherwise the mean won't be 2. This is one way to think about degrees of freedom. The degrees of freedom for these three numbers is n-1 = 3-1= 2, because 2 of the numbers can be free, but the last number has no freedom, it becomes fixed after the first two are decided.

Now, statisticians often apply degrees of freedom to their calculations, especially when a second calculation relies on an estimated value. For example, when we calculate the standard deviation of a sample, we first calculate the mean of the sample right! By estimating the mean, we are fixing an aspect of our sample, and so, our sample now has n-1 degrees of freedom when we calculate the standard deviation (remember for the sample standard deviation, we divide by n-1…there's that n-1 again.)

7.3.11.1 Simulating how degrees of freedom affects the t distribution

There are at least two ways to think the degrees of freedom for a t-test. For example, if you want to use math to compute aspects of the t distribution, then you need the degrees of freedom to plug in to the formula… If you want to see the formulas I'm talking about, scroll down on the t-test wikipedia page and look for the probability density or cumulative distribution functions…We think that is quite scary for most people, and one reason why degrees of freedom are not well-understood.

If we wanted to simulate the t distribution we could more easily see what influence degrees of freedom has on the shape of the distribution. Remember, t is a sample statistic, it is something we measure from the sample. So, we could simulate the process of measuring t from many different samples, then plot the histogram of t to show us the simulated t distribution.

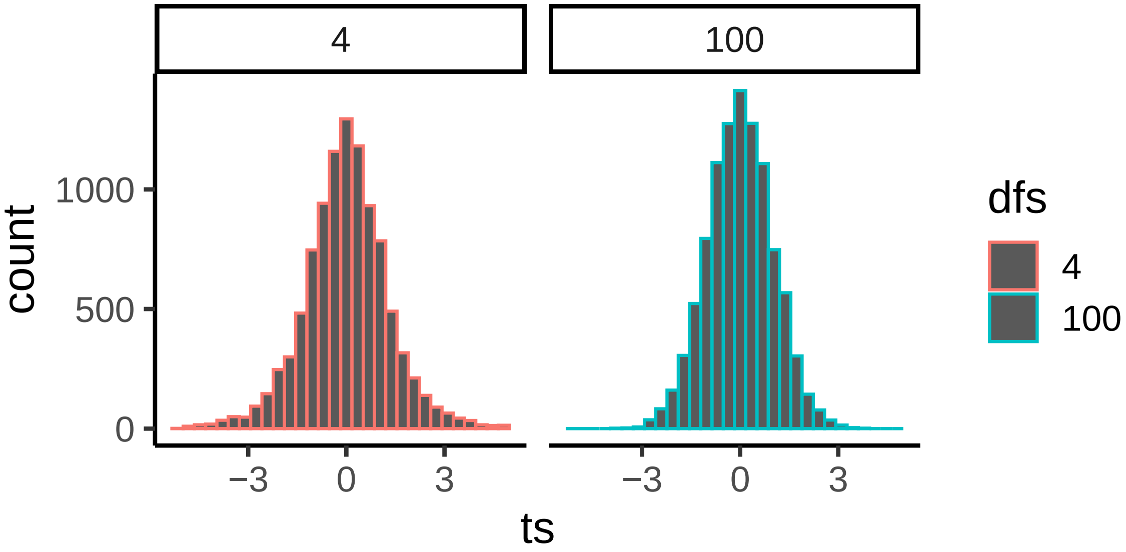


Figure 7.10: The width of the t distribution shrinks as sample size and degrees of freedom (from 4 to 100) increases.

In Figure 7.10 notice that the red distribution for $df = 4$, is a little bit shorter, and a little bit wider than the bluey-green distribution for $df = 100$. As degrees of freedom increase the t distribution gets taller (in the middle), and narrower in the range. It get's more peaky. Can you guess the reason for this? Remember, we are estimating a sample statistic, and degrees of freedom is really just a number that refers to the number of subjects (well minus one). And, we already know that as we increase n, our sample statistics become better estimates (less variance) of the distributional parameters they are estimating. So, t becomes a better estimate of it's "true" value as sample size increase, resulting in a more narrow distribution of ts.

There is a slightly different t distribution for every degrees of freedom, and the critical regions associated with 5% of the extreme values are thus slightly different every time. This is why we report the degrees of freedom for each t-test, they define the distribution of t values for the sample-size in question. Why do we use n-1 and not n? Well, we calculate t using the sample standard deviation to estimate the standard error or the mean, that estimate uses n-1 in the denominator, so our t distribution is built assuming n-1. That's enough for degrees of freedom…

7.4 The paired samples t-test strikes back

You must be wondering if we will ever be finished talking about paired samples t-tests... why are we doing round 2, oh no! Don't worry, we're just going to 1) remind you about what we were doing with the infant study, and 2) do a paired samples t-test on the entire data set and discuss.

Remember, we were wondering if the infants would look longer toward the singer who sang the familiar song during the test phase compared to the baseline phase. We showed you data from 5 infants, and walked through the computations for the *t*-test. As a reminder, it looked like this:

infant	Baseline	Test	differences	diff_from_mean	Squared_differences
1	0.44	0.6	0.16	0.106	0.011236
2	0.41	0.68	0.27	0.216	0.046656
3	0.75	0.72	-0.03	-0.084	0.00705600000000001
4	0.44	0.28	-0.16	-0.214	0.045796
5	0.47	0.5	0.03	-0.024	0.000575999999999999
Sums	2.51	2.78	0.27	0	0.11132
Means	0.502	0.556	0.054	0	0.022264
				sd	0.167
				SEM	0.075
				t	0.72

```
#>
#>  One Sample t-test
#>
#> data:  round(differences, digits = 2)
#> t = 0.72381, df = 4, p-value = 0.5092
#> alternative hypothesis: true mean is not equal to 0
#> 95 percent confidence interval:
#>  -0.1531384  0.2611384
#> sample estimates:
#> mean of x
#>     0.054
```

Let's write down the finding one more time: The mean difference was 0.054, $t(4) = .72$, p =.509. We can also now confirm, that the p-value was from a two-tailed test. So, what does this all really mean.

We can say that a t value with an absolute of .72 or larger occurs 50.9% of the time. More precisely, the distribution of no differences (the null), will produce a t value this large or larger

50.9% of the time. In other words, chance alone good have easily produced the *t* value from our sample, and the mean difference we observed or .054, could easily have been a result of chance.

Let's quickly put all of the data in the *t*-test, and re-run the test using all of the infant subjects.

```
#>
#>  One Sample t-test
#>
#> data:  differences
#> t = 2.4388, df = 31, p-value = 0.02066
#> alternative hypothesis: true mean is not equal to 0
#> 95 percent confidence interval:
#>  0.01192088 0.13370412
#> sample estimates:
#> mean of x
#> 0.0728125
```

Now we get a very different answer. We would summarize the results saying the mean difference was .073, t(31) = 2.44, p = 0.020. How many total infants were their? Well the degrees of freedom was 31, so there must have been 32 infants in the study. Now we see a much smaller *p*-value. This was also a two-tailed test, so we that observing a *t* value of 2.4 or greater (absolute value) only occurs 2% of the time. In other words, the distribution of no differences will produce the observed t-value very rarely. So, it is unlikely that the observed mean difference of .073 was due to chance (it could have been due to chance, but that is very unlikely). As a result, we can be somewhat confident in concluding that something about seeing and hearing a unfamiliar person sing a familiar song, causes an infant to draw their attention toward the singer, and this potentially benefits social learning on the part of the infant.

7.5 Independent samples t-test: The return of the t-test?

If you've been following the Star Wars references, we are on last movie (of the original trilogy)… the independent t-test. This is were basically the same story plays out as before, only slightly different.

Remember there are different *t*-tests for different kinds of research designs. When your design is a **between-subjects** design, you use an **independent samples t-test**. Between-subjects design involve different people or subjects in each experimental condition. If there are two conditions, and 10 people in each, then there are 20 total people. And, there are no paired scores, because every single person is measured once, not twice, no repeated measures. Because there are no repeated measures we can't look at the difference scores between conditions one and two. The scores are not paired in any meaningful way, to it doesn't make sense to subtract them. So what do we do?

The logic of the independent samples t-test is the very same as the other *t*-tests. We calculated the means for each group, then we find the difference. That goes into the numerator of the t formula. Then we get an estimate of the variation for the denominator. We divide the mean difference by the estimate of the variation, and we get *t*. It's the same as before.

The only wrinkle here is what goes into the denominator? How should we calculate the estimate of the variance? It would be nice if we could do something very straightforward like this, say for an experiment with two groups A and B:

$t = \frac{\bar{A}-\bar{B}}{(\frac{SEM_A+SEM_B}{2})}$

In plain language, this is just:

1. Find the mean difference for the top part
2. Compute the SEM (standard error of the mean) for each group, and average them together to make a single estimate, pooling over both samples.

This would be nice, but unfortunately, it turns out that finding the average of two standard errors of the mean is not the best way to do it. This would create a biased estimator of the variation for the hypothesized distribution of no differences. We won't go into the math here, but instead of the above formula, we an use a different one that gives as an **unbiased estimate of the pooled standard error of the sample mean**. Our new and improved *t* formula would look like this:

$t = \frac{\bar{X_A}-\bar{X_B}}{s_p * \sqrt{\frac{1}{n_A}+\frac{1}{n_B}}}$

and, s_p, which is the pooled sample standard deviation is defined as, note the $s$es in the formula are variances:

$s_p = \sqrt{\frac{(n_A-1)s_A^2+(n_B-1)s_B^2}{n_A+n_B-2}}$

Believe you me, that is so much more formula than I wanted to type out. Shall we do one independent *t*-test example by hand, just to see the computations? Let's do it…but in a slightly different way than you expect. I show the steps using R. I made some fake scores for groups A and B. Then, I followed all of the steps from the formula, but made R do each of the calculations. This shows you the needed steps by following the code. At the end, I print the *t*-test values I computed "by hand", and then the *t*-test value that the R software outputs using the *t*-test function. You should be able to get the same values for *t*, if you were brave enough to compute *t* by hand.

```
## By "hand" using R r code
a <- c(1,2,3,4,5)
b <- c(3,5,4,7,9)

mean_difference <- mean(a)-mean(b) # compute mean difference

variance_a <- var(a) # compute variance for A
variance_b <- var(b) # compute variance for B

# Compute top part and bottom part of sp formula

sp_numerator <- (4*variance_a + 4* variance_b)
sp_denominator <- 5+5-2
sp <- sqrt(sp_numerator/sp_denominator) # compute sp


# compute t following formulat

t <- mean_difference / ( sp * sqrt( (1/5) +(1/5) ) )

t # print results
#> [1] -2.017991


# using the R function t.test
t.test(a,b, paired=FALSE, var.equal = TRUE)
#>
#>  Two Sample t-test
#>
#> data:  a and b
#> t = -2.018, df = 8, p-value = 0.0783
#> alternative hypothesis: true difference in means is not equal to 0
#> 95 percent confidence interval:
#>  -5.5710785  0.3710785
#> sample estimates:
#> mean of x mean of y
#>       3.0       5.6
```

7.6 Simulating data for t-tests

An “advanced” topic for *t*-tests is the idea of using R to conduct simulations for *t*-tests.

If you recall, *t* is a property of a sample. We calculate *t* from our sample. The *t* distribution is the hypothetical behavior of our sample. That is, if we had taken thousands upon thousands of samples, and calculated *t* for each one, and then looked at the distribution of those *t*’s, we would have the sampling distribution of *t*!

It can be very useful to get in the habit of using R to simulate data under certain conditions, to see how your sample data, and things like *t* behave. Why is this useful? It mainly prepares you with some intuitions about how sampling error (random chance) can influence your results, given specific parameters of your design, such as sample-size, the size of the mean difference you expect to find in your data, and the amount of variation you might find. These methods can be used formally to conduct power-analyses. Or more informally for data sense.

7.6.1 Simulating a one-sample t-test

Here are the steps you might follow to simulate data for a one sample *t*-test.

1. Make some assumptions about what your sample (that you might be planning to collect) might look like. For example, you might be planning to collect 30 subjects worth of data. The scores of those data points might come from a normal distribution (mean = 50, sd = 10).
2. sample simulated numbers from the distribution, then conduct a *t*-test on the simulated numbers. Save the statistics you want (such as *t*s and *p*s), and then see how things behave.

Let’s do this a couple different times. First, let’s simulate samples with N = 30, taken from a normal (mean= 50, sd = 25). We’ll do a simulation with 1000 simulations. For each simulation, we will compare the sample mean with a population mean of 50. There should be no difference on average here. Figure 7.11 is the null distribution that we are simulating.

Neat. We see both a *t* distribution, that looks like *t* distribution as it should. And we see the *p* distribution. This shows us how often we get *t* values of particular sizes. You may find it interesting that the *p*-distribution is flat under the null, which we are simulating here. This means that you have the same chances of a getting a *t* with a p-value between 0 and 0.05, as you would for getting a *t* with a p-value between .90 and .95. Those ranges are both ranges of 5%, so there are an equal amount of *t* values in them by definition.

Here’s another way to do the same simulation in R, using the replicate function, instead a for loop:

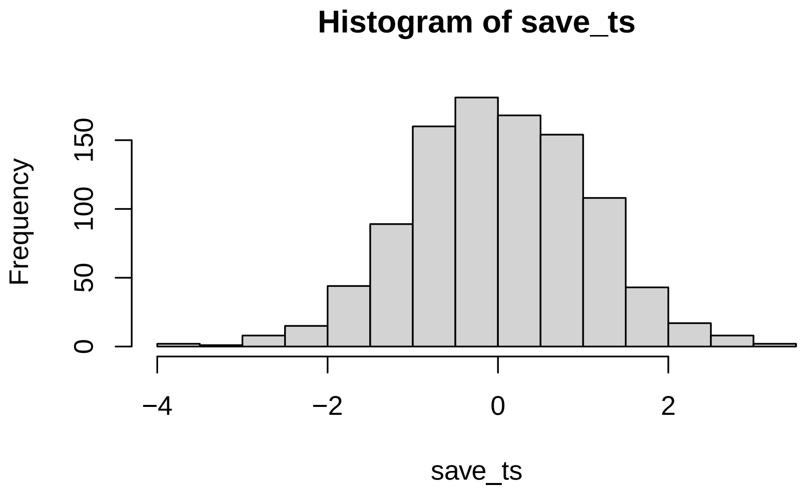


Figure 7.11: The distribution of t-values under the null. These are the t values that are produced by chance alone.

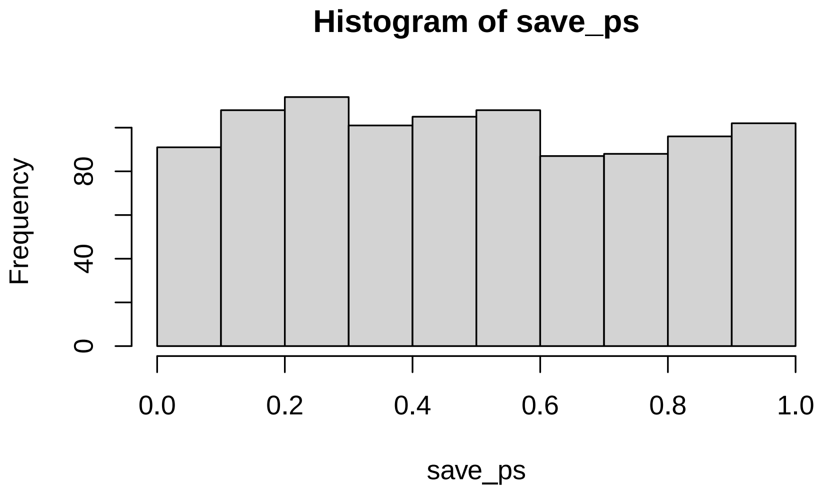


Figure 7.12: The distribution of p-values that are observed is flat under the null.

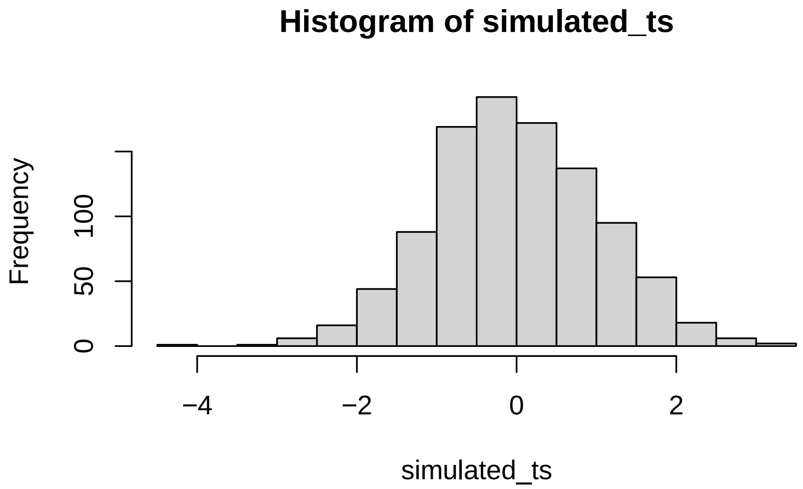


Figure 7.13: Simulating ts in R.

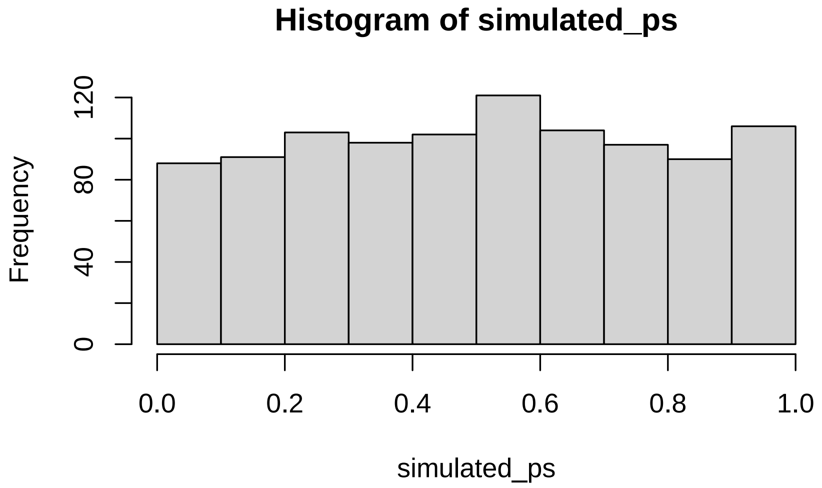


Figure 7.14: Simulating ps in R.

7.6.2 Simulating a paired samples t-test

The code below is set up to sample 10 scores for condition A and B from the same normal distribution. The simulation is conducted 1000 times, and the *t*s and *p*s are saved and plotted for each.

```
save_ps <- length(1000)
save_ts <- length(1000)
for ( i in 1:1000 ){
  condition_A <- rnorm(10,10,5)
  condition_B <- rnorm(10,10,5)
  differences <- condition_A - condition_B
  t_test <- t.test(differences, mu=0)
  save_ps[i] <- t_test$p.value
  save_ts[i] <- t_test$statistic
}
```

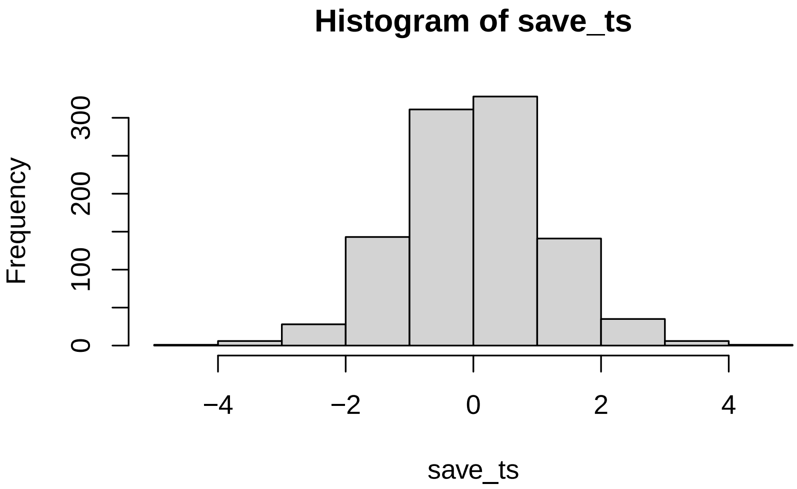


Figure 7.15: 1000 simulated ts from the null distribution

According to the simulation. When there are no differences between the conditions, and the samples are being pulled from the very same distribution, you get these two distributions for t and p. These again show how the null distribution of no differences behaves.

For any of these simulations, if you rejected the null-hypothesis (that your difference was only due to chance), you would be making a type I error. If you set your alpha criteria to $\alpha = .05$, we can ask how many type I errors were made in these 1000 simulations. The answer is:

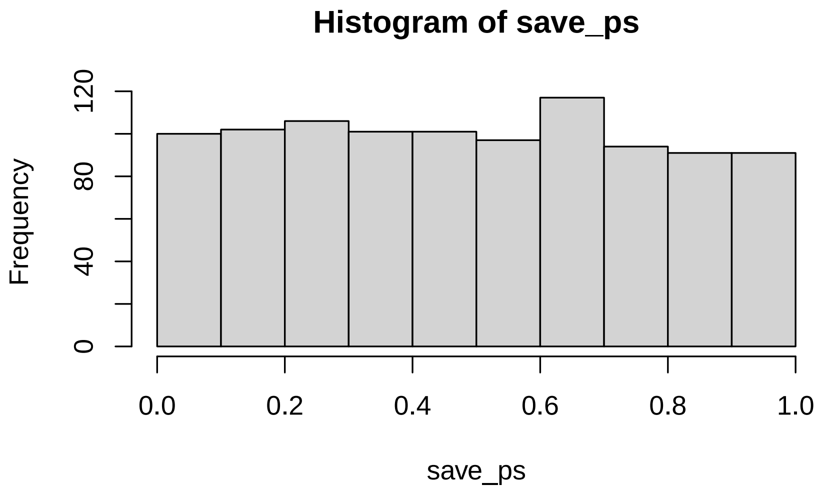


Figure 7.16: 1000 simulated ps from the null distribution

```
length(save_ps[save_ps<.05])
#> [1] 48
length(save_ps[save_ps<.05])/1000
#> [1] 0.048
```

We happened to make 48. The expectation over the long run is 5% type I error rates (if your alpha is .05).

What happens if there actually is a difference in the simulated data, let's set one condition to have a larger mean than the other:

```
save_ps <- length(1000)
save_ts <- length(1000)
for ( i in 1:1000 ){
  condition_A <- rnorm(10,10,5)
  condition_B <- rnorm(10,13,5)
  differences <- condition_A - condition_B
  t_test <- t.test(differences, mu=0)
  save_ps[i] <- t_test$p.value
  save_ts[i] <- t_test$statistic
}
```

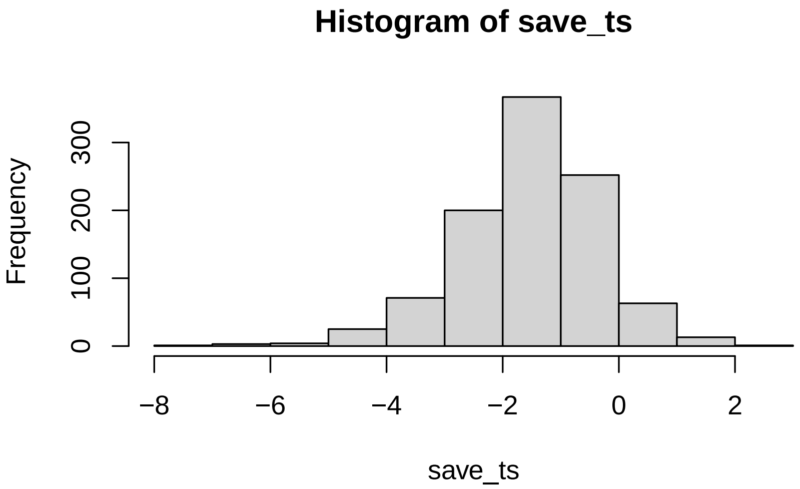


Figure 7.17: 1000 ts when there is a true difference

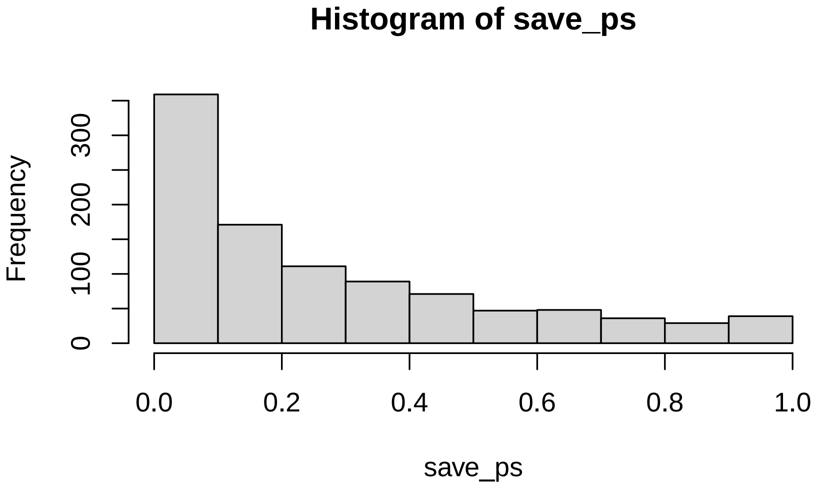


Figure 7.18: 1000 ps when there is a true difference

Now you can see that the p-value distribution is skewed to the left. This is because when there is a true effect, you will get p-values that are less than .05 more often. Or, rather, you get larger t values than you normally would if there were no differences.

In this case, we wouldn't be making a type I error if we rejected the null when p was smaller than .05. How many times would we do that out of our 1000 experiments?

```r
length(save_ps[save_ps<.05])
#> [1] 231
length(save_ps[save_ps<.05])/1000
#> [1] 0.231
```

We happened to get 231 simulations where p was less than .05, that's only 0.231 experiments. If you were the researcher, would you want to run an experiment that would be successful only 0.231 of the time? I wouldn't. I would run a better experiment.

How would you run a better simulated experiment? Well, you could increase n, the number of subjects in the experiment. Let's increase n from 10 to 100, and see what happens to the number of "significant" simulated experiments.

```r
save_ps <- length(1000)
save_ts <- length(1000)
for ( i in 1:1000 ){
  condition_A <- rnorm(100,10,5)
  condition_B <- rnorm(100,13,5)
  differences <- condition_A - condition_B
  t_test <- t.test(differences, mu=0)
  save_ps[i] <- t_test$p.value
  save_ts[i] <- t_test$statistic
}
```

```
#> [1] 988
#> [1] 0.988
```

Cool, now almost all of the experiments show a p-value of less than .05 (using a two-tailed test, that's the default in R). See, you could use this simulation process to determine how many subjects you need to reliably find your effect.

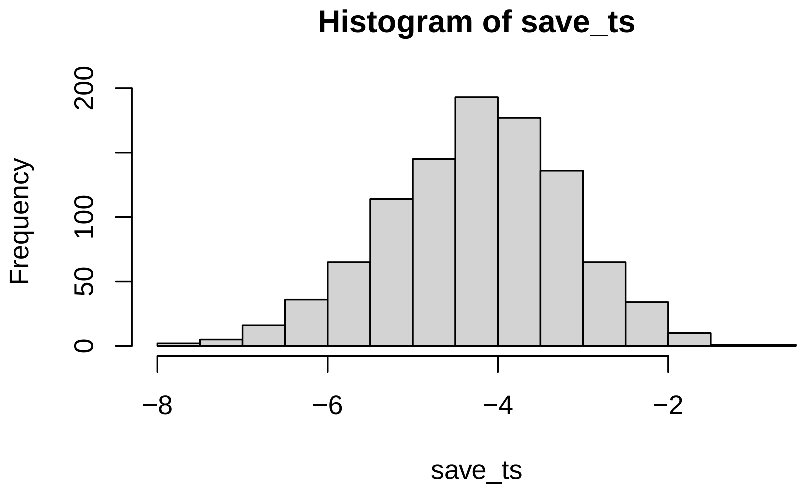


Figure 7.19: 1000 ts for n =100, when there is a true effect

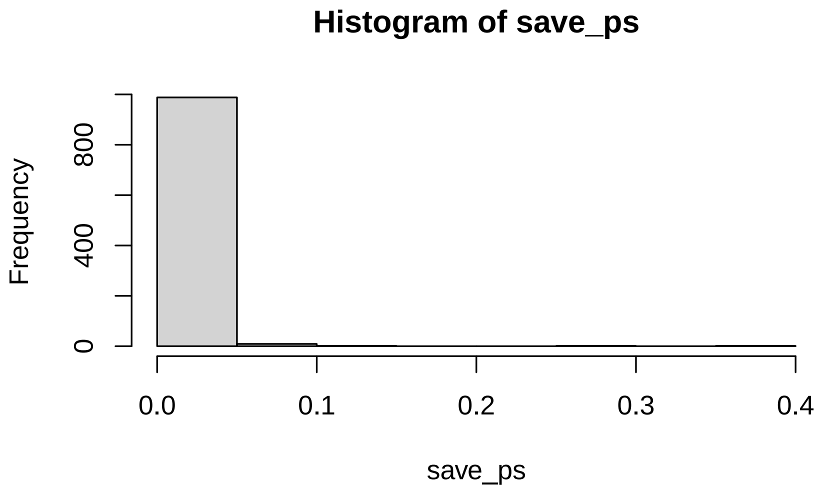


Figure 7.20: 1000 ps for n =100, when there is a true effect

7.6.3 Simulating an independent samples t.test

Just change the t.test function like so... this is for the null, assuming no difference between groups.

```r
save_ps <- length(1000)
save_ts <- length(1000)
for ( i in 1:1000 ){
  group_A <- rnorm(10,10,5)
  group_B <- rnorm(10,10,5)
  t_test <- t.test(group_A, group_B, paired=FALSE, var.equal=TRUE)
  save_ps[i] <- t_test$p.value
  save_ts[i] <- t_test$statistic
}
```

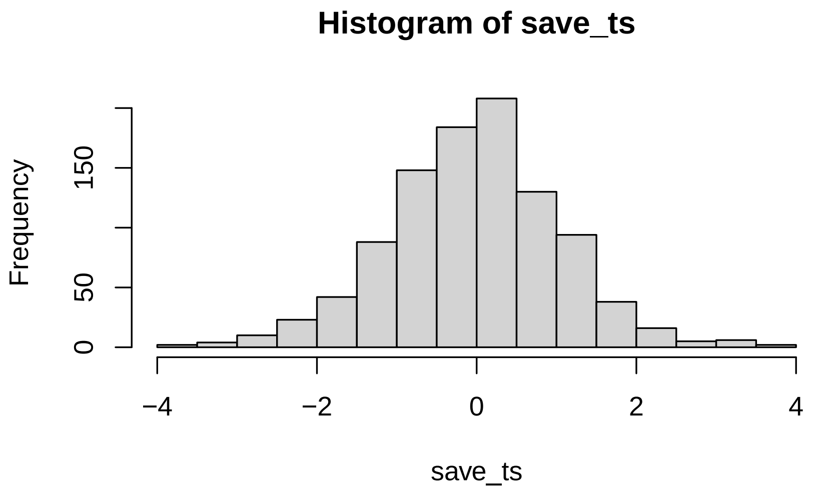


Figure 7.21: 1000 ts for n =100, when there is a true effect

```
#> [1] 55
#> [1] 0.055
```

7.7 Videos

7.7.1 One or Two tailed tests

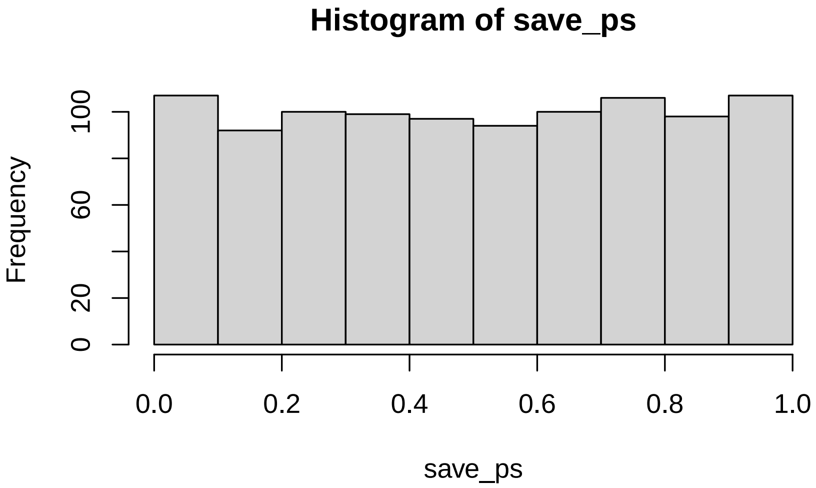


Figure 7.22: 1000 ps for n =100, when there is a true effect

8 ANOVA

A fun bit of stats history (Salsburg 2001). Sir Ronald Fisher invented the ANOVA, which we learn about in this section. He wanted to publish his new test in the journal Biometrika. The editor at the time was Karl Pearson (remember Pearson's r for correlation?). Pearson and Fisher were apparently not on good terms, they didn't like each other. Pearson refused to publish Fisher's new test. So, Fisher eventually published his work in the Journal of Agricultural Science. Funnily enough, the feud continued onto the next generation. Years after Fisher published his ANOVA, Karl Pearson's son Egon Pearson, and Jersey Neyman revamped Fisher's ideas, and re-cast them into what is commonly known as null vs. alternative hypothesis testing. Fisher didn't like this very much.

We present the ANOVA in the Fisherian sense, and at the end describe the Neyman-Pearson approach that invokes the concept of null vs. alternative hypotheses.

8.1 ANOVA is Analysis of Variance

ANOVA stands for Analysis Of Variance. It is a widely used technique for assessing the likelihood that differences found between means in sample data could be produced by chance. You might be thinking, well don't we have t-tests for that? Why do we need the ANOVA, what do we get that's new that we didn't have before?

What's new with the ANOVA, is the ability to test a wider range of means beyond just two. In all of the t-test examples we were always comparing two things. For example, we might ask whether the difference between two sample means could have been produced by chance. What if our experiment had more than two conditions or groups? We would have more than 2 means. We would have one mean for each group or condition. That could be a lot depending on the experiment. How would we compare all of those means? What should we do, run a lot of t-tests, comparing every possible combination of means? Actually, you could do that. Or, you could do an ANOVA.

In practice, we will combine both the ANOVA test and t-tests when analyzing data with many sample means (from more than two groups or conditions). Just like the t-test, there are different kinds of ANOVAs for different research designs. There is one for between-subjects designs, and a slightly different one for repeated measures designs. We talk about both, beginning with the ANOVA for between-subjects designs.

8.2 One-factor ANOVA

The one-factor ANOVA is sometimes also called a between-subjects ANOVA, an independent factor ANOVA, or a one-way ANOVA (which is a bit of a misnomer as we discuss later). The critical ingredient for a one-factor, between-subjects ANOVA, is that you have one independent variable, with at least two-levels. When you have one IV with two levels, you can run a t-test. You can also run an ANOVA. Interestingly, they give you almost the exact same results. You will get a p-value from both tests that is identical (they are really doing the same thing under the hood). The t-test gives a t-value as the important sample statistic. The ANOVA gives you the F-value (for Fisher, the inventor of the test) as the important sample statistic. It turns out that t^2 equals F, when there are only two groups in the design. They are the same test. Side-note, it turns out they are all related to Pearson's r too (but we haven't written about this relationship yet in this textbook).

Remember that t is computed directly from the data. It's like a mean and standard error that we measure from the sample. In fact it's the mean difference divided by the standard error of the sample. It's just another descriptive statistic isn't it.

The same thing is true about F. F is computed directly from the data. In fact, the idea behind F is the same basic idea that goes into making t. Here is the general idea behind the formula, it is again a ratio of the effect we are measuring (in the numerator), and the variation associated with the effect (in the denominator).

$\text{name of statistic} = \frac{\text{measure of effect}}{\text{measure of error}}$

$\text{F} = \frac{\text{measure of effect}}{\text{measure of error}}$

The difference with F, is that we use variances to describe both the measure of the effect and the measure of error. So, F is a ratio of two variances.

Remember what we said about how these ratios work. When the variance associated with the effect is the same size as the variance associated with sampling error, we will get two of the same numbers, this will result in an F-value of 1. When the variance due to the effect is larger than the variance associated with sampling error, then F will be greater than 1. When the variance associated with the effect is smaller than the variance associated with sampling error, F will be less than one.

Let's rewrite in plainer English. We are talking about two concepts that we would like to measure from our data. 1) A measure of what we can explain, and 2) a measure of error, or stuff about our data we can't explain. So, the F formula looks like this:

$\text{F} = \frac{\text{Can Explain}}{\text{Can't Explain}}$

When we can explain as much as we can't explain, $F = 1$. This isn't that great of a situation for us to be in. It means we have a lot of uncertainty. When we can explain much more than we can't we are doing a good job, F will be greater than 1. When we can explain less than what we can't, we really can't explain very much, F will be less than 1. That's the concept behind making F.

If you saw an F in the wild, and it was .6. Then you would automatically know the researchers couldn't explain much of their data. If you saw an F of 5, then you would know the researchers could explain 5 times more than the couldn't, that's pretty good. And the point of this is to give you an intuition about the meaning of an F-value, even before you know how to compute it.

8.2.1 Computing the F-value

Fisher's ANOVA is very elegant in my opinion. It starts us off with a big problem we always have with data. We have a lot of numbers, and there is a lot of variation in the numbers, what to do? Wouldn't it be nice to split up the variation into to kinds, or sources. If we could know what parts of the variation were being caused by our experimental manipulation, and what parts were being caused by sampling error, we would be making really good progress. We would be able to know if our experimental manipulation was causing more change in the data than sampling error, or chance alone. If we could measure those two parts of the total variation, we could make a ratio, and then we would have an F value. This is what the ANOVA does. It splits the total variation in the data into two parts. The formula is:

Total Variation = Variation due to Manipulation + Variation due to sampling error

This is a nice idea, but it is also vague. We haven't specified our measure of variation. What should we use?

Remember the sums of squares that we used to make the variance and the standard deviation? That's what we'll use. Let's take another look at the formula, using sums of squares for the measure of variation:

$SS_{\text{total}} = SS_{\text{Effect}} + SS_{\text{Error}}$

8.2.2 SS Total

The total sums of squares, or SSTotal is a way of thinking about all of the variation in a set of data. It's pretty straightforward to measure. No tricky business. All we do is find the difference between each score and the grand mean, then we square the differences and add them all up.

Let's imagine we had some data in three groups, A, B, and C. For example, we might have 3 scores in each group. The data could look like this:

groups	scores	diff	diff_squared
A	20	13	169
A	11	4	16
A	2	-5	25
B	6	-1	1
B	2	-5	25
B	7	0	0
C	2	-5	25
C	11	4	16
C	2	-5	25
Sums	63	0	302
Means	7	0	33.5555555555556

The data is organized in long format, so that each row is a single score. There are three scores for the A, B, and C groups. The mean of all of the scores is called the **Grand Mean**. It's calculated in the table, the Grand Mean = 7.

We also calculated all of the difference scores **from the Grand Mean**. The difference scores are in the column titled `diff`. Next, we squared the difference scores, and those are in the next column called `diff_squared`.

Remember, the difference scores are a way of measuring variation. They represent how far each number is from the Grand Mean. If the Grand Mean represents our best guess at summarizing the data, the difference scores represent the error between the guess and each actual data point. The only problem with the difference scores is that they sum to zero (because the mean is the balancing point in the data). So, it is convenient to square the difference scores, this turns all of them into positive numbers. The size of the squared difference scores still represents error between the mean and each score. And, the squaring operation exacerbates the differences as the error grows larger (squaring a big number makes a really big number, squaring a small number still makes a smallish number).

OK fine! We have the squared deviations from the grand mean, we know that they represent the error between the grand mean and each score. What next? SUM THEM UP!

When you add up all of the individual squared deviations (difference scores) you get the sums of squares. That's why it's called the sums of squares (SS).

Now, we have the first part of our answer:

$SS_\text{total} = SS_\text{Effect} + SS_\text{Error}$

$SS_\text{total} = 302$ and

$302 = SS_\text{Effect} + SS_\text{Error}$

What next? If you think back to what you learned about algebra, and solving for X, you might notice that we don't really need to find the answers to both missing parts of the equation. We only need one, and we can solve for the other. For example, if we found SS_{Effect}, then we could solve for SS_{Error}.

8.2.3 SS Effect

SS_{Total} gave us a number representing all of the change in our data, how all the scores are different from the grand mean.

What we want to do next is estimate how much of the total change in the data might be due to the experimental manipulation. For example, if we ran an experiment that causes causes change in the measurement, then the means for each group will be different from other. As a result, the manipulation forces change onto the numbers, and this will naturally mean that some part of the total variation in the numbers is caused by the manipulation.

The way to isolate the variation due to the manipulation (also called effect) is to look at the means in each group, and calculate the difference scores between each group mean and the grand mean, and then sum the squared deviations to find SS_{Effect}.

Consider this table, showing the calculations for SS_{Effect}.

groups	scores	means	diff	diff_squared
A	20	11	4	16
A	11	11	4	16
A	2	11	4	16
B	6	5	-2	4
B	2	5	-2	4
B	7	5	-2	4
C	2	5	-2	4
C	11	5	-2	4
C	2	5	-2	4
Sums	63	63	0	72
Means	7	7	0	8

Notice we created a new column called `means`. For example, the mean for group A was 11. You can see there are three 11s, one for each observation in row A. The means for group B and C happen to both be 5. So, the rest of the numbers in the means column are 5s.

What we are doing here is thinking of each score in the data from the viewpoint of the group means. The group means are our best attempt to summarize the data in those groups. From the point of view of the mean, all of the numbers are treated as the same. The mean doesn't know how far off it is from each score, it just knows that all of the scores are centered on the mean.

> Let's pretend you are the mean for group A. That means you are an 11. Someone asks you "hey, what's the score for the first data point in group A?". Because you are the mean, you say, I know that, it's 11. "What about the second score?"…it's 11… they're all 11, so far as I can tell…"Am I missing something…", asked the mean.

Now that we have converted each score to it's mean value we can find the differences between each mean score and the grand mean, then square them, then sum them up. We did that, and found that the $SS_{\text{Effect}} = 72$.

SS_{Effect} represents the amount of variation that is caused by differences between the means. I also refer to this as the amount of variation that the researcher can explain (by the means, which represent differences between groups or conditions that were manipulated by the researcher).

Notice also that $SS_{\text{Effect}} = 72$, and that 72 is smaller than $SS_{\text{total}} = 302$. That is very important. SS_{Effect} by definition can never be larger than SS_{total}.

8.2.4 SS Error

Great, we made it to SS Error. We already found SS Total, and SS Effect, so now we can solve for SS Error just like this:

$SS_{\text{total}} = SS_{\text{Effect}} + SS_{\text{Error}}$

switching around:

$ SS_Error = SS_total - SS_Effect $

$ SS_Error = 302 - 72 = 230 $

We could stop here and show you the rest of the ANOVA, we're almost there. But, the next step might not make sense unless we show you how to calculate SS_{Error} directly from the data, rather than just solving for it. We should do this just to double-check our work anyway.

groups	scores	means	diff	diff_squared
A	20	11	-9	81
A	11	11	0	0
A	2	11	9	81
B	6	5	-1	1
B	2	5	3	9
B	7	5	-2	4
C	2	5	3	9

groups	scores	means	diff	diff_squared
C	11	5	-6	36
C	2	5	3	9
Sums	63	63	0	230
Means	7	7	0	25.5555555555556

Alright, we did almost the same thing as we did to find SS_{Effect}. Can you spot the difference? This time for each score we first found the group mean, then we found the error in the group mean estimate for each score. In other words, the values in the $diff$ column are the differences between each score and it's group mean. The values in the `diff_squared` column are the squared deviations. When we sum up the squared deviations, we get another Sums of Squares, this time it's the SS_{Error}. This is an appropriate name, because these deviations are the ones that the group means can't explain!

8.2.5 Degrees of freedom

Degrees of freedom come into play again with ANOVA. This time, their purpose is a little bit more clear. Dfs can be fairly simple when we are doing a relatively simple ANOVA like this one, but they can become complicated when designs get more complicated.

Let's talk about the degrees of freedom for the SS_{Effect} and SS_{Error}.

The formula for the degrees of freedom for SS_{Effect} is

$df_{\text{Effect}} = \text{Groups} - 1$, where Groups is the number of groups in the design.

In our example, there are 3 groups, so the df is 3-1 = 2. You can think of the df for the effect this way. When we estimate the grand mean (the overall mean), we are taking away a degree of freedom for the group means. Two of the group means can be anything they want (they have complete freedom), but in order for all three to be consistent with the Grand Mean, the last group mean has to be fixed.

The formula for the degrees of freedom for SS_{Error} is

$df_{\text{Error}} = \text{scores} - \text{groups}$, or the number of scores minus the number of groups. We have 9 scores and 3 groups, so our df for the error term is 9-3 = 6. Remember, when we computed the difference score between each score and its group mean, we had to compute three means (one for each group) to do that. So, that reduces the degrees of freedom by 3. 6 of the difference scores could be anything they want, but the last 3 have to be fixed to match the means from the groups.

8.2.6 Mean Squared Error

OK, so we have the degrees of freedom. What's next? There are two steps left. First we divide the SSes by their respective degrees of freedom to create something new called Mean Squared Error. Let's talk about why we do this.

First of all, remember we are trying to accomplish this goal:

$\text{F} = \frac{\text{measure of effect}}{\text{measure of error}}$

We want to build a ratio that divides a measure of an effect by a measure of error. Perhaps you noticed that we already have a measure of an effect and error! How about the SS_{Effect} and SS_{Error}. They both represent the variation due to the effect, and the leftover variation that is unexplained. Why don't we just do this?

$\frac{SS_{\text{Effect}}}{SS_{\text{Error}}}$

Well, of course you could do that. What would happen is you can get some really big and small numbers for your inferential statistic. And, the kind of number you would get wouldn't be readily interpretable like a t value or a z score.

The solution is to **normalize** the SS terms. Don't worry, normalize is just a fancy word for taking the average, or finding the mean. Remember, the SS terms are all sums. And, each sum represents a different number of underlying properties.

For example, the SS_Effect represents the sum of variation for three means in our study. We might ask the question, well, what is the average amount of variation for each mean...You might think to divide SS_Effect by 3, because there are three means, but because we are estimating this property, we divide by the degrees of freedom instead (# groups - 1 = 3-1 = 2). Now we have created something new, it's called the MSE_{Effect}.

$MSE_{\text{Effect}} = \frac{SS_{\text{Effect}}}{df_{\text{Effect}}}$

$MSE_{\text{Effect}} = \frac{72}{2} = 36$

This might look alien and seem a bit complicated. But, it's just another mean. It's the mean of the sums of squares for the effect. If this reminds you of the formula for the variance, good memory. The SME_{Effect} is a measure variance for the change in the data due to changes in the means (which are tied to the experimental conditions).

The SS_{Error} represents the sum of variation for nine scores in our study. That's a lot more scores, so the SS_{Error} is often way bigger than than SS_{Effect}. If we left our SSes this way and divided them, we would almost always get numbers less than one, because the SS_{Error} is so big. What we need to do is bring it down to the average size. So, we might want to divide our SS_{Error} by 9, after all there were nine scores. However, because we are estimating this property, we divide by the degrees of freedom instead (scores-groups) = 9-3 = 6). Now we have created something new, it's called the MSE_{Error}.

$MSE_{\text{Error}} = \frac{SS_{\text{Error}}}{df_{\text{Error}}}$

$MSE_{\text{Error}} = \frac{230}{6} = 38.33$

8.2.7 Calculate F

Now that we have done all of the hard work, calculating F is easy:

$\text{F} = \frac{\text{measure of effect}}{\text{measure of error}}$

$\text{F} = \frac{MSE_{\text{Effect}}}{MSE_{\text{Error}}}$

$\text{F} = \frac{36}{38.33} = .939$

Done!

8.2.8 The ANOVA TABLE

You might suspect we aren't totally done here. We've walked through the steps of computing F. Remember, F is a sample statistic, we computed F directly from the data. There were a whole bunch of pieces we needed, the dfs, the SSes, the MSEs, and then finally the F.

All of these little pieces are conveniently organized by ANOVA tables. ANOVA tables look like this:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
groups	2	72	36.00000	0.9391304	0.4417359
Residuals	6	230	38.33333	NA	NA

You are looking at the print-out of an ANOVA summary table from R. Notice, it had columns for Df, SS (Sum Sq), MSE (Mean Sq), F, and a p-value. There are two rows. The `groups` row is for the Effect (what our means can explain). The `Residuals` row is for the Error (what our means can't explain). Different programs give slightly different labels, but they are all attempting to present the same information in the ANOVA table. There isn't anything special about the ANOVA table, it's just a way of organizing all the pieces. Notice, the MSE for the effect (36) is placed above the MSE for the error (38.333), and this seems natural because we divide 36/38.33 in or to get the F-value!

8.3 What does F mean?

We've just noted that the ANOVA has a bunch of numbers that we calculated straight from the data. All except one, the p-value. We did not calculate the p-value from the data. Where did it come from, what does it mean? How do we use this for statistical inference. Just so you don't get too worried, the p-value for the ANOVA has the very same general meaning as the p-value for the t-test, or the p-value for any sample statistic. It tells us that the probability that we would observe our test statistic or larger, under the distribution of no differences (the null).

As we keep saying, F is a sample statistic. Can you guess what we do with sample statistics in this textbook? We did it for the Crump Test, the Randomization Test, and the t-test… We make fake data, we simulate it, we compute the sample statistic we are interested in, then we see how it behaves over many replications or simulations.

Let's do that for F. This will help you understand what F really is, and how it behaves. We are going to created the sampling distribution of F. Once we have that you will be able to see where the p-values come from. It's the same basic process that we followed for the t tests, except we are measuring F instead of t.

Here is the set-up, we are going to run an experiment with three levels. In our imaginary experiment we are going to test whether a new magic pill can make you smarter. The independent variable is the number of magic pills you take: 1, 2, or 3. We will measure your smartness using a smartness test. We will assume the smartness test has some known properties, the mean score on the test is 100, with a standard deviation of 10 (and the distribution is normal).

The only catch is that our magic pill does NOTHING AT ALL. The fake people in our fake experiment will all take sugar pills that do absolutely nothing to their smartness. Why would we want to simulate such a bunch of nonsense? The answer is that this kind of simulation is critical for making inferences about chance if you were to conduct a real experiment.

Here are some more details for the experiment. Each group will have 10 different subjects, so there will be a total of 30 subjects. We are going to run this experiment 10,000 times. Each time drawing numbers randomly from the very same normal distribution. We are going to calculate F from our sample data every time, and then we are going to draw the histogram of F-values. Figure 8.1 shows the sampling distribution of F for our situation.

Let's note a couple things about the F distribution. 1) The smallest value is 0, and there are no negative values. Does this make sense? F can never be negative because it is the ratio of two variances, and variances are always positive because of the squaring operation. So, yes, it makes sense that the sampling distribution of F is always 0 or greater. 2) it does not look normal. No it does not. F can have many different looking shapes, depending on the degrees of freedom in the numerator and denominator. However, these aspects are too important for now.

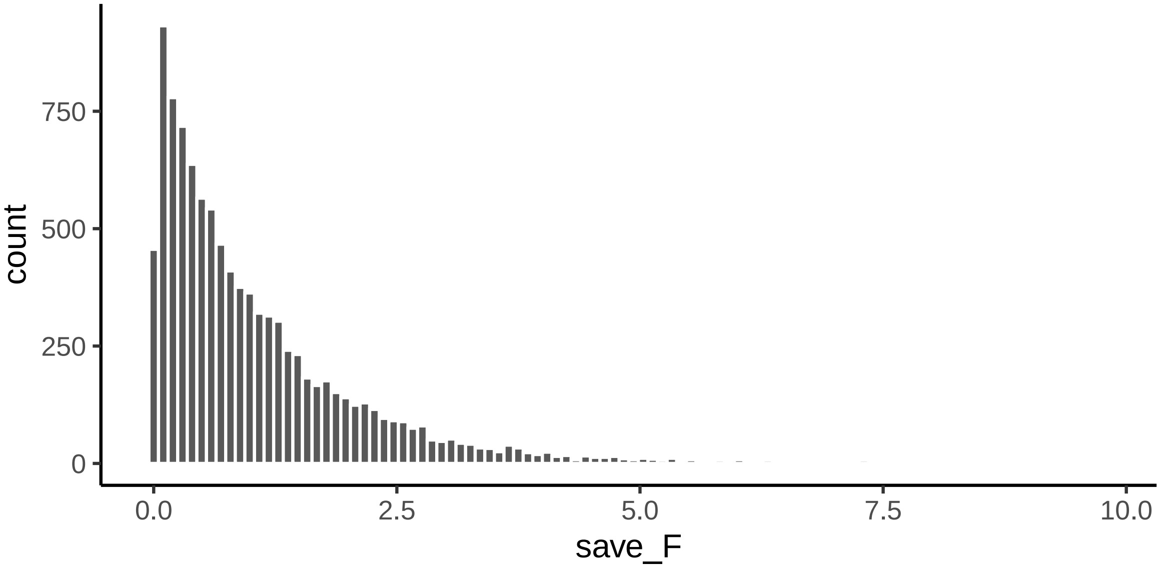


Figure 8.1: A simulation of 10,000 experiments from a null distribution where there is no differences. The histogram shows 10,000 F-values, one for each simulation. These are values that F can take in this situation. All of these F-values were produced by random sampling error.

Remember, before we talked about some intuitive ideas for understanding F, based on the idea that F is a ratio of what we can explain (variance due to mean differences), divided by what we can't explain (the error variance). When the error variance is higher than the effect variance, then we will always get an F-value less than one. You can see that we often got F-values less than one in the simulation. This is sensible, after all we were simulating samples coming from the very same distribution. On average there should be no differences between the means. So, on average the part of the total variance that is explained by the means should be less than one, or around one, because it should be roughly the same as the amount of error variance (remember, we are simulating no differences).

At the same time, we do see that some F-values are larger than 1. There are little bars that we can see going all the way up to about 5. If you were to get an F-value of 5, you might automatically think, that's a pretty big F-value. Indeed it kind of is, it means that you can explain 5 times more of variance than you can't explain. That seems like a lot. You can also see that larger F-values don't occur very often. As a final reminder, what you are looking at is how the F-statistic (measured from each of 10,000 simulated experiments) behaves when the only thing that can cause differences in the means is random sampling error. Just by chance sometimes the means will be different. You are looking at another chance window. These are the Fs that chance can produce.

8.3.1 Making Decisions

We can use the sampling distribution of F (for the null) to make decisions about the role of chance in a real experiment. For example, we could do the following.

1. Set an alpha criterion of $p = 0.05$
2. Find out the critical value for F, for our particular situation (with our dfs for the numerator and denominator).

Let's do that. I've drawn the line for the critical value onto the histogram in Figure 8.2:

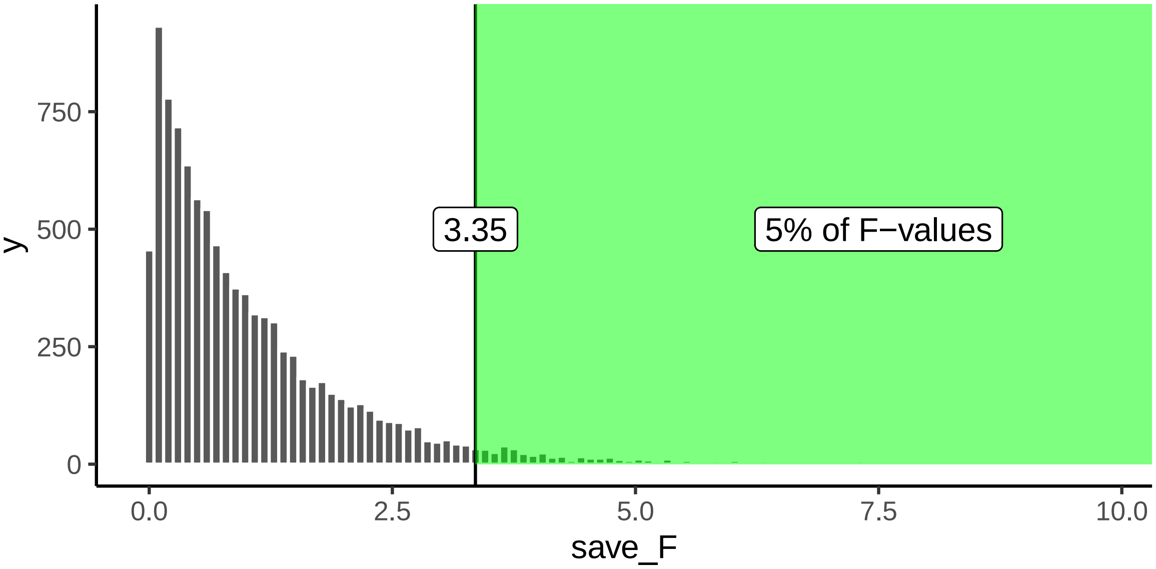


Figure 8.2: The critical value for F where 5% of all F-values lie beyond this point

Alright, now we can see that only 5% of all F-values from from this sampling distribution will be 3.35 or larger. We can use this information.

How would we use it? Imagine we ran a real version of this experiment. And, we really used some pills that just might change smartness. If we ran the exact same design, with 30 people in total (10 in each group), we could set an F criterion of 3.35 for determining whether any of our results reflected a causal change in smartness due to the pills, and not due to random chance. For example, if we found an F-value of 3.34, which happens, just less than 5% of the time, we might conclude that random sampling error did not produce the differences between our means. Instead, we might be more confident that the pills actually did something, after all an F-value of 3.34 doesn't happen very often, it is unlikely (only 5 times out of 100) to occur by chance.

8.3.2 Fs and means

Up to here we have been building your intuition for understanding F. We went through the calculation of F from sample data. We went through the process of simulating thousands of Fs to show you the null distribution. We have not talked so much about what researchers really care about…The MEANS! The actual results from the experiment. Were the means different? that's often what people want to know. So, now we will talk about the means, and F, together.

Notice, if I told you I ran an experiment with three groups, testing whether some manipulation changes the behavior of the groups, and I told you that I found a big F!, say an F of 6!. And, that the F of 6 had a p-value of .001. What would you know based on that information alone? You would only know that Fs of 6 don't happen very often by chance. In fact they only happen 0.1% of the time, that's hardly at all. If someone told me those values, I would believe that the results they found in their experiment were not likely due to chance. However, I still would not know what the results of the experiment were! Nobody told us what the means were in the different groups, we don't know what happened!

IMPORTANT: even though we don't know what the means were, we do know something about them, whenever we get F-values and p-values like that (big Fs, and very small associated ps)… Can you guess what we know? I'll tell you. We automatically know that there **must have been some differences between the means**. If there was no differences between the means, then the variance explained by the means (the numerator for F) would not be very large. So, we know that there must be some differences, we just don't know what they are. Of course, if we had the data, all we would need to do is look at the means for the groups (the ANOVA table doesn't report this, we need to do it as a separate step).

8.3.2.1 ANOVA is an omnibus test

This property of the ANOVA is why the ANOVA is sometimes called the **omnibus test**. Omnibus is a fun word, it sounds like a bus I'd like to ride. The meaning of omnibus, according to the dictionary, is "comprising several items". The ANOVA is, in a way, one omnibus test, comprising several little tests.

For example, if you had three groups, A, B, and C. You get could differences between

1. A and B
2. B and C
3. A and C

That's three possible differences you could get. You could run separate t-tests, to test whether each of those differences you might have found could have been produced by chance. Or, you could run an ANOVA, like what we have been doing, to ask one more general question about the differences. Here is one way to think about what the omnibus test is testing:

Hypothesis of no differences anywhere: $ A = B = C $

Any differences anywhere:

a. $ A B = C $
b. $ A = B C $
c. $ A C = B $

The $\neq$ symbol means "does not equal", it's an equal sign with a cross through it (no equals allowed!).

How do we put all of this together. Generally, when we get a small *F*-value, with a large *p*-value, we will not reject the hypothesis of no differences. We will say that we do not have evidence that the means of the three groups are in any way different, and the differences that are there could easily have been produced by chance. When we get a large F with a small *p*-value (one that is below our alpha criterion), we will generally reject the hypothesis of no differences. We would then assume that at least one group mean is not equal to one of the others. That is the omnibus test. Rejecting the null in this way is rejecting the idea there are no differences. But, the *F* test still does not tell you which of the possible group differences are the ones that are different.

8.3.2.2 Looking at a bunch of group means

We just ran 10,000 experiments and we didn't even once look at the group means for any of the experiments. Different patterns of group means under the null are shown in Figure 8.3 for a subset of 10 random simulations.

Whoa, that's a lot to look at. What is going on here? Each little box represents the outcome of a simulated experiment. The dots are the means for each group (whether subjects took 1 , 2, or 3 magic pills). The y-axis shows the mean smartness for each group. The error bars are standard errors of the mean.

You can see that each of the 10 experiments turn out different. Remember, we sampled 10 numbers for each group from the **same** normal distribution with mean = 100, and sd = 10. So, we know that the **correct** means for each sample should actually be 100 every single time. However, they are not 100 every single time because of?…**sampling error** (Our good friend that we talk about all the time).

For most of the simulations the error bars are all overlapping, this suggests visually that the means are not different. However, some of them look like they are not overlapping so much, and this would suggest that they are different. This is the siren song of chance (sirens lured sailors to their deaths at sea…beware of the siren call of chance). If we concluded that any of these sets of means had a true difference, we would be committing a type I error. Because we made the simulation, we know that none of these means are actually different. But, when you are running a real experiment, you don't get to know this for sure.

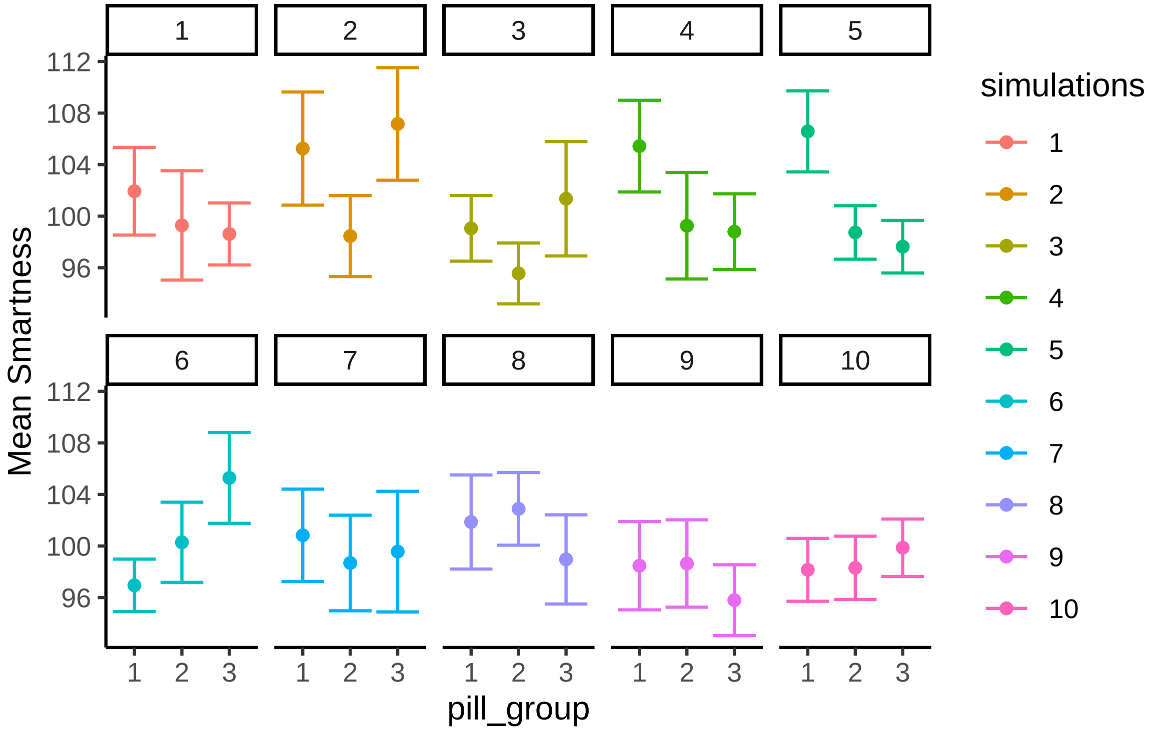


Figure 8.3: Different patterns of group means under the null (all scores for each group sampled from the same distribution).

8.3.2.3 Looking at bar graphs

Let's look at the exact same graph as above, but this time use bars to visually illustrate the means, instead of dots. We'll re-do our simulation of 10 experiments, so the pattern will be a little bit different:

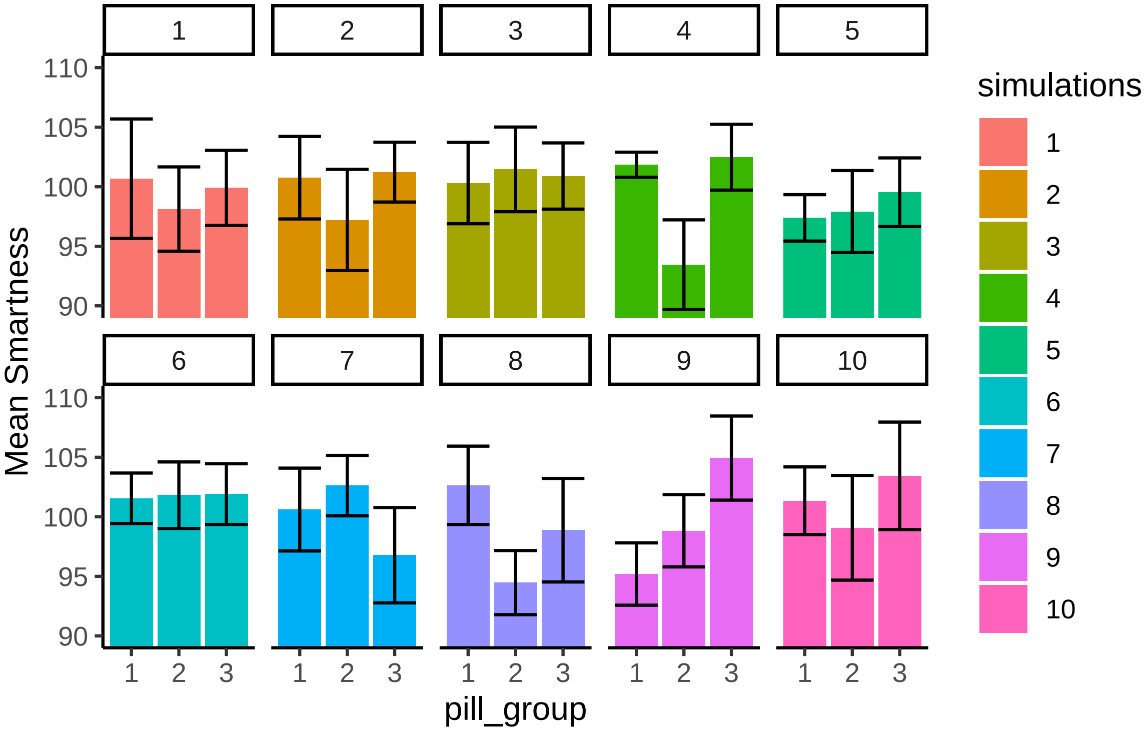


Figure 8.4: Different patterns of group means under the null (all scores for each group sampled from the same distribution).

In Figure 8.4 the heights of the bars display the means for each pill group. The pattern across simulations is generally the same. Some of the fake experiments look like there might be differences, and some of them don't.

8.3.2.4 What mean differences look like when F is less than 1

We are now giving you some visual experience looking at what means look like from a particular experiment. This is for your stats intuition. We're trying to improve your data senses.

What we are going to do now is similar to what we did before. Except this time we are going to look at 10 simulated experiments, where all of the F-values were less than 1. All of these F-values would also be associated with fairly large p-values. When F is less than one, we would not reject the hypothesis of no differences. So, when we look at patterns of means when F is less than 1, we should see mostly the same means, and no big differences.

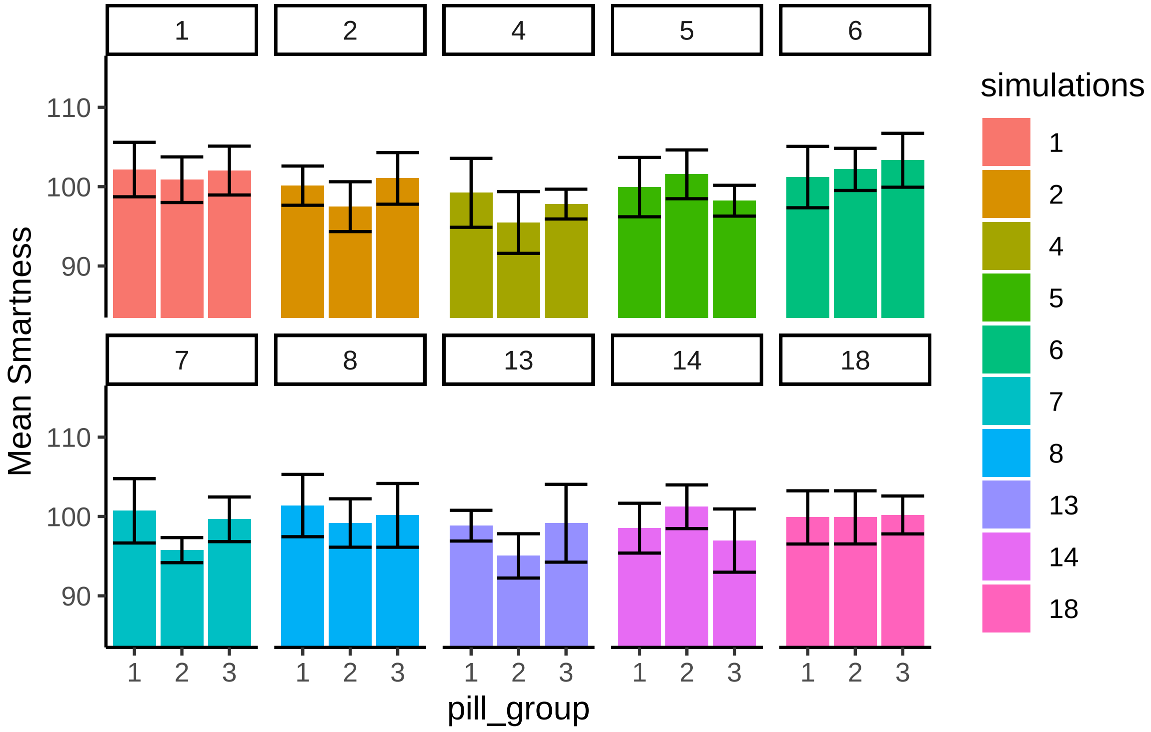


Figure 8.5: Different patterns of group means under the null (sampled from same distribution) when F is less than 1.

In Figure 8.5 the numbers in the panels now tell us which simulations actually produced Fs of less than 1.

We see here that all the bars aren't perfectly flat, that's OK. What's more important is that for each panel, the error bars for each mean are totally overlapping with all the other error bars. We can see visually that our estimate of the mean for each sample is about the same for all of the bars. That's good, we wouldn't make any type I errors here.

8.3.2.5 What mean differences look like when F $>$ 3.35

Earlier we found that the critical value for F in our situation was 3.35, this was the location on the F distribution where only 5% of Fs were 3.35 or greater. We would reject the hypothesis of no differences whenever F was greater than 3.35. In this case, whenever we did that, we would be making a type I error. That is because we are simulating the distribution of no differences (remember all of our sample means are coming from the exact same distribution).

So, now we can take a look at what type I errors look like. In other words, we can run some simulations and look at the pattern in the means, only when F happens to be 3.35 or greater (this only happens 5% of the time, so we might have to let the computer simulate for a while). Let's see what that looks like:

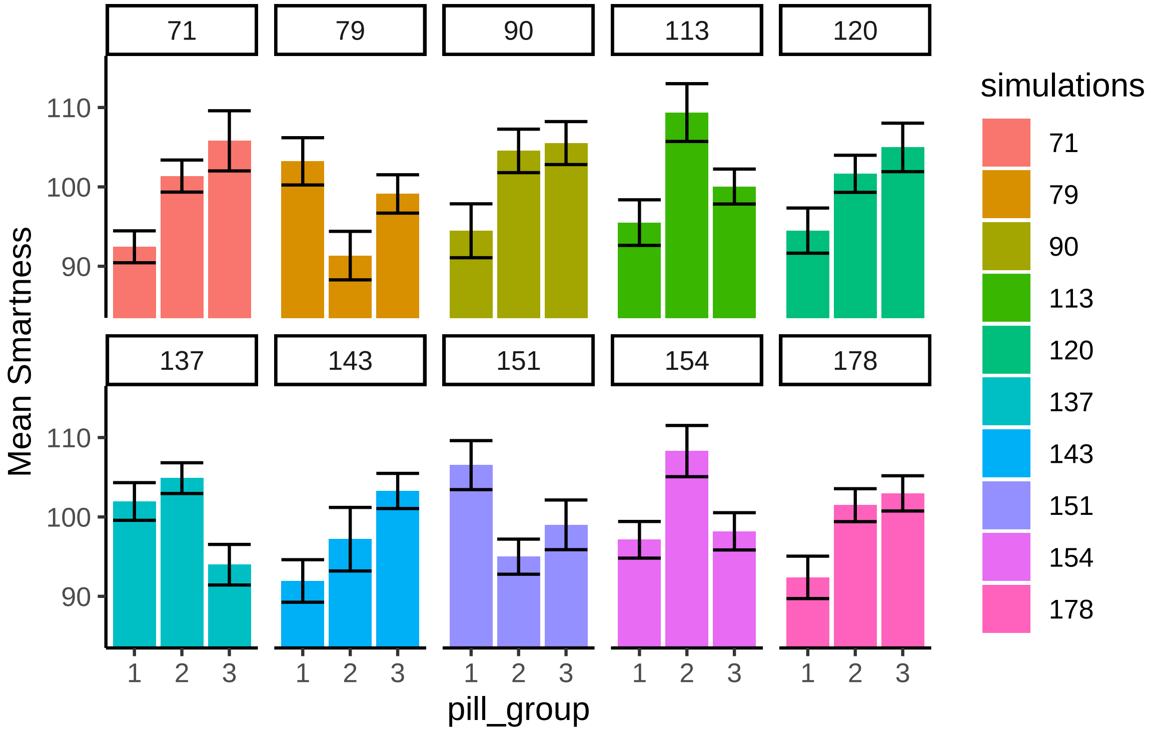


Figure 8.6: Different patterns of group means under the null when F is above critical value (these are all type I Errors).

The numbers in the panels now tell us which simulations actually produced Fs that were greater than 3.35

What do you notice about the pattern of means inside each panel of Figure 8.6? Now, every the panels show at least one mean that is different from the others. Specifically, the error bars for one mean do not overlap with the error bars for one or another mean. This is what mistakes looks like. These are all type I errors. They are insidious. When they happen to you by chance, the data really does appear to show a strong pattern, your F-value is large, and your p-value is small! It is easy to be convinced by a type I error (it's the siren song of chance).

8.4 ANOVA on Real Data

We've covered many fundamentals about the ANOVA, how to calculate the necessary values to obtain an F-statistic, and how to interpret the F-statistic along with it's associate p-value once we have one. In general, you will be conducting ANOVAs and playing with Fs and ps using software that will automatically spit out the numbers for you. It's important that you understand what the numbers mean, that's why we've spent time on the concepts. We also recommend that you try to compute an ANOVA by hand at least once. It builds character, and let's you know that you know what you are doing with the numbers.

But, we've probably also lost the real thread of all this. The core thread is that when we run an experiment we use our inferential statistics, like ANOVA, to help us determine whether the differences we found are likely due to chance or not. In general, we like to find out that the differences that we find are not due to chance, but instead to due to our manipulation.

So, we return to the application of the ANOVA to a real data set with a real question. This is the same one that you will be learning about in the lab. We give you a brief overview here so you know what to expect.

8.4.1 Tetris and bad memories

Yup, you read that right. The research you will learn about tests whether playing Tetris after watching a scary movie can help prevent you from having bad memories from the movie (James et al. 2015). Sometimes in life people have intrusive memories, and they think about things they'd rather not have to think about. This research looks at one method that could reduce the frequency of intrusive memories.

Here's what they did. Subjects watched a scary movie, then at the end of the week they reported how many intrusive memories about the movie they had. The mean number of intrusive memories was the measurement (the dependent variable). This was a between-subjects experiment with four groups. Each group of subjects received a different treatment following the scary movie. The question was whether any of these treatments would reduce the number of intrusive memories. All of these treatments occurred after watching the scary movie:

1. No-task control: These participants completed a 10-minute music filler task after watching the scary movie.
2. Reactivation + Tetris: These participants were shown a series of images from the trauma film to reactivate the traumatic memories (i.e., reactivation task). Then, participants played the video game Tetris for 12 minutes.
3. Tetris Only: These participants played Tetris for 12 minutes, but did not complete the reactivation task.

4. Reactivation Only: These participants completed the reactivation task, but did not play Tetris.

For reasons we elaborate on in the lab, the researchers hypothesized that the `Reactivation+Tetris` group would have fewer intrusive memories over the week than the other groups.

Let's look at the findings. Note you will learn how to do all of these steps in the lab. For now, we just show the findings and the ANOVA table. Then we walk through how to interpret it.

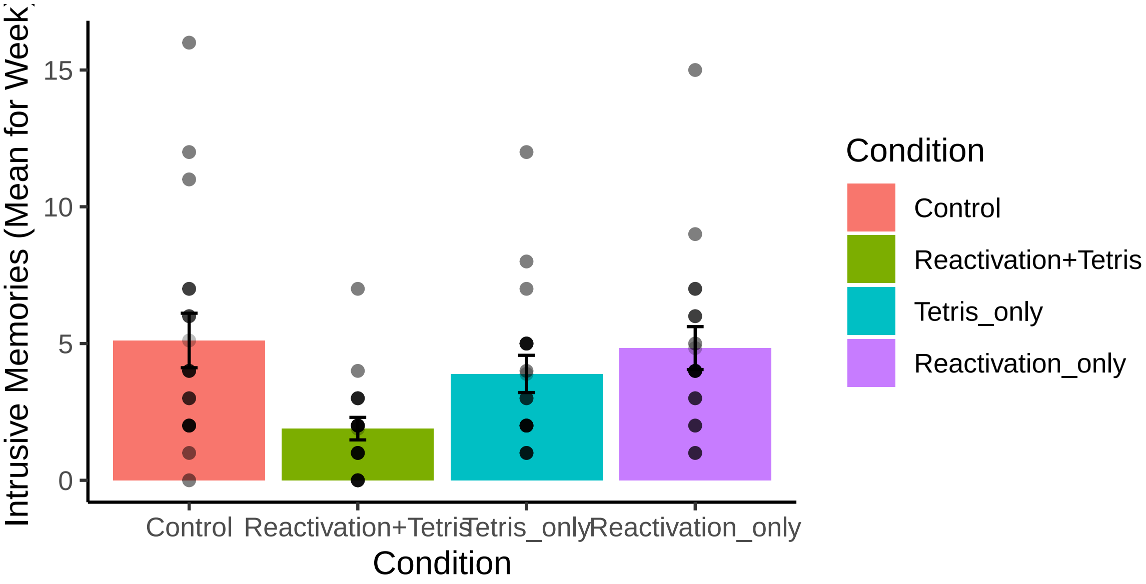


Figure 8.7: Mean number of intrusive memories per week as a function of experimental treatments.

OOooh, look at that. We did something fancy. Figure 8.7 shows the data from the four groups. The height of each bar shows the mean intrusive memories for the week. The dots show the individual scores for each subject in each group (useful to to the spread of the data). The error bars show the standard errors of the mean.

What can we see here? Right away it looks like there is some support for the research hypothesis. The green bar, for the Reactivation + Tetris group had the lowest mean number of intrusive memories. Also, the error bar is not overlapping with any of the other error bars. This implies that the mean for the Reactivation + Tetris group is different from the means for the other groups. And, this difference is probably not very likely by chance.

We can now conduct the ANOVA on the data to ask the omnibus question. If we get a an F-value with an associated p-value of less than .05 (the alpha criterion set by the authors), then we can reject the hypothesis of no differences. Let's see what happens:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Condition	3	114.8194	38.27315	3.794762	0.0140858
Residuals	68	685.8333	10.08578	NA	NA

We see the ANOVA table, it's up there. We could report the results from the ANOVA table like this:

> There was a significant main effect of treatment condition, F(3, 68) = 3.79, MSE = 10.08, p=0.014.

We called this a significant effect because the *p*-value was less than 0.05. In other words, the *F*-value of 3.79 only happens 1.4% of the time when the null is true. Or, the differences we observed in the means only occur by random chance (sampling error) 1.4% of the time. Because chance rarely produces this kind of result, the researchers made the inference that chance DID NOT produce their differences, instead, they were inclined to conclude that the Reactivation + Tetris treatment really did cause a reduction in intrusive memories. That's pretty neat.

8.4.2 Comparing means after the ANOVA

Remember that the ANOVA is an omnibus test, it just tells us whether we can reject the idea that all of the means are the same. The F-test (synonym for ANOVA) that we just conducted suggested we could reject the hypothesis of no differences. As we discussed before, that must mean that there are some differences in the pattern of means.

Generally after conducting an ANOVA, researchers will conduct follow-up tests to compare differences between specific means. We will talk more about this practice throughout the textbook. There are many recommended practices for follow-up tests, and there is a lot of debate about what you should do. We are not going to wade into this debate right now. Instead we are going to point out that **you need to do something** to compare the means of interest after you conduct the ANOVA, because the ANOVA is just the beginning…It usually doesn't tell you want you want to know. You might wonder why bother conducting the ANOVA in the first place…Not a terrible question at all. A good question. You will see as we talk about more complicated designs, why ANOVAs are so useful. In the present example, they are just a common first step. There are required next steps, such as what we do next.

How can you compare the difference between two means, from a between-subjects design, to determine whether or not the difference you observed is likely or unlikely to be produced by chance? We covered this one already, it's the independent *t*-test. We'll do a couple *t*-tests, showing the process.

8.4.2.1 Control vs. Reactivation+Tetris

What we really want to know is if Reactivation+Tetris caused fewer intrusive memories…but compared to what? Well, if it did something, the Reactivation+Tetris group should have a smaller mean than the Control group. So, let's do that comparison:

```
#> 
#>  Two Sample t-test
#> 
#> data:  Days_One_to_Seven_Number_of_Intrusions by Condition
#> t = 2.9893, df = 34, p-value = 0.005167
#> alternative hypothesis: true difference in means between group Control and group Reactiva
#> 95 percent confidence interval:
#>  1.031592 5.412852
#> sample estimates:
#>             mean in group Control mean in group Reactivation+Tetris
#>                          5.111111                          1.888889
```

We found that there was a significant difference between the control group (M=5.11) and Reactivation + Tetris group (M=1.89), t(34) = 2.99, p=0.005.

Above you just saw an example of reporting another *t*-test. This sentences does an OK job of telling the reader everything they want to know. It has the means for each group, and the important bits from the *t*-test.

More important, as we suspected the difference between the control and Reactivation + Tetris group was likely not due to chance.

8.4.2.2 Control vs. Tetris_only

Now we can really start wondering what caused the difference. Was it just playing Tetris? Does just playing Tetris reduce the number of intrusive memories during the week? Let's compare that to control:

```
#> 
#>  Two Sample t-test
#> 
#> data:  Days_One_to_Seven_Number_of_Intrusions by Condition
#> t = 1.0129, df = 34, p-value = 0.3183
#> alternative hypothesis: true difference in means between group Control and group Tetris_on
#> 95 percent confidence interval:
#>  -1.230036  3.674480
#> sample estimates:
#>     mean in group Control mean in group Tetris_only
#>                  5.111111                  3.888889
```

Here we did not find a significant difference. We found that no significant difference between the control group (M=5.11) and Tetris Only group (M=3.89), t(34) = 2.99, p=0.318.

So, it seems that not all of the differences between our means are large enough to be called statistically significant. In particular, the difference here, or larger, happens by chance 31.8% of the time.

You could go on doing more comparisons, between all of the different pairs of means. Each time conducting a *t*-test, and each time saying something more specific about the patterns across the means than you get to say with the omnibus test provided by the ANOVA.

Usually, it is the pattern of differences across the means that you as a researcher are primarily interested in understanding. Your theories will make predictions about how the pattern turns out (e.g., which specific means should be higher or lower and by how much). So, the practice of doing comparisons after an ANOVA is really important for establishing the patterns in the means.

8.5 ANOVA Summary

We have just finished a rather long introduction to the ANOVA, and the *F*-test. The next couple of chapters continue to explore properties of the ANOVA for different kinds of experimental designs. In general, the process to follow for all of the more complicated designs is very similar to what we did here, which boils down to two steps:

1) conduct the ANOVA on the data
2) conduct follow-up tests, looking at differences between particular means

So what's next…the ANOVA for repeated measures designs. See you in the next chapter.

9 Repeated Measures ANOVA

This chapter introduces you to **repeated measures ANOVA**. Repeated measures ANOVAs are very common in Psychology, because psychologists often use repeated measures designs, and repeated measures ANOVAs are the appropriate test for making inferences about repeated measures designs.

Remember the paired sample *t*-test? We used that test to compare two means from a repeated measures design. Remember what a repeated measures design is? It's also called a within-subjects design. These designs involve measuring the same subject more than once. Specifically, at least once for every experimental condition. In the paired *t*-test example, we discussed a simple experiment with only two experimental conditions. There, each subject would contribute a measurement to level one and level two of the design.

However, paired-samples *t*-tests are limited to comparing two means. What if you had a design that had more than two experimental conditions? For example, perhaps your experiment had 3 levels for the independent variable, and each subject contributed data to each of the three levels?

This is starting to sounds like an ANOVA problem. ANOVAs are capable of evaluating whether there is a difference between any number of means, two or greater. So, we can use an ANOVA for our repeated measures design with three levels for the independent variable.

Great! So, what makes a repeated measures ANOVA different from the ANOVA we just talked about?

9.1 Repeated measures design

Let's use the exact same toy example from the previous chapter, but let's convert it to a repeated measures design.

Last time, we imagined we had some data in three groups, A, B, and C, such as in Table 9.1:

Table 9.1: Example data for three groups.

groups	scores
A	20
A	11

Table 9.1: Example data for three groups.

groups	scores
A	2
B	6
B	2
B	7
C	2
C	11
C	2

The above table represents a between-subject design where each score involves a unique subject.

Let's change things up a tiny bit, and imagine we only had 3 subjects in total in the experiment. And, that each subject contributed data to the three levels of the independent variable, A, B, and C. Before we called the IV `groups`, because there were different groups of subjects. Let's change that to `conditions`, because now the same group of subjects participates in all three conditions. Table 9.2 shows a within-subjects (repeated measures) version of this experiment:

Table 9.2: Example data for a repeated measures design with three conditions, where each subject contributes data in each condition.

subjects	conditions	scores
1	A	20
2	A	11
3	A	2
1	B	6
2	B	2
3	B	7
1	C	2
2	C	11
3	C	2

9.2 Partitioning the Sums of Squares

Time to introduce a new name for an idea you learned about last chapter, it's called **partitioning the sums of squares**. Sometimes an obscure new name can be helpful for your understanding of what is going on. ANOVAs are all about partitioning the sums of squares. We already did some partitioning in the last chapter. What do we mean by partitioning?

Imagine you had a big empty house with no rooms in it. What would happen if you partitioned the house? What would you be doing? One way to partition the house is to split it up into different rooms. You can do this by adding new walls and making little rooms everywhere. That's what partitioning means, to split up.

The act of partitioning, or splitting up, is the core idea of ANOVA. To use the house analogy. Our total sums of squares (SS Total) is our big empty house. We want to split it up into little rooms. Before we partitioned SS Total using this formula:

$SS_\text{TOTAL} = SS_\text{Effect} + SS_\text{Error}$

Remember, the SS_Effect was the variance we could attribute to the means of the different groups, and SS_Error was the leftover variance that we couldn't explain. SS_Effect and SS_Error are the partitions of SS_TOTAL, they are the little rooms.

In the between-subjects case above, we got to split SS_TOTAL into two parts. What is most interesting about the repeated-measures design, is that we get to split SS_TOTAL into three parts, there's one more partition. Can you guess what the new partition is? Hint: whenever we have a new way to calculate means in our design, we can always create a partition for those new means. What are the new means in the repeated measures design?

Here is the new idea for partitioning SS_TOTAL in a repeated-measures design:

$SS_\text{TOTAL} = SS_\text{Effect} + SS_\text{Subjects} + SS_\text{Error}$

We've added SS_Subjects as the new idea in the formula. What's the idea here? Well, because each subject was measured in each condition, we have a new set of means. These are the means for each subject, collapsed across the conditions. For example, subject 1 has a mean (mean of their scores in conditions A, B, and C); subject 2 has a mean (mean of their scores in conditions A, B, and C); and subject 3 has a mean (mean of their scores in conditions A, B, and C). There are three subject means, one for each subject, collapsed across the conditions. And, we can now estimate the portion of the total variance that is explained by these subject means.

We just showed you a "formula" to split up SS_TOTAL into three parts, but we called the formula an idea. We did that because the way we wrote the formula is a little bit misleading, and we need to clear something up. Before we clear the thing up, we will confuse you just a little bit. Be prepared to be confused a little bit.

First, we need to introduce you to some more terms. It turns out that different authors use different words to describe parts of the ANOVA. This can be really confusing. For example, we described the SS formula for a between subjects design like this:

$SS_\text{TOTAL} = SS_\text{Effect} + SS_\text{Error}$

However, the very same formula is often written differently, using the words between and within in place of effect and error, it looks like this:

$SS_{\text{TOTAL}} = SS_{\text{Between}} + SS_{\text{Within}}$

Whoa, hold on a minute. Haven't we switched back to talking about a **between-subjects** ANOVA. YES! Then why are we using the word **within**, what does that mean? YES! We think this is very confusing for people. Here the word **within** has a special meaning. It **does not** refer to a within-subjects design. Let's explain. First, SS_{Between} (which we have been calling SS_{Effect}) refers to variation **between** the group means, that's why it is called SS_{Between}. Second, and most important, SS_{Within} (which we have been calling SS_{Error}), refers to the leftover variation within each group mean. Specifically, it is the variation between each group mean and each score in the group. "AAGGH, you've just used the word between to describe within group variation!". Yes! We feel your pain. Remember, for each group mean, every score is probably off a little bit from the mean. So, the scores within each group have some variation. This is the within group variation, and it is why the leftover error that we can't explain is often called SS_{Within}.

OK. So why did we introduce this new confusing way of talking about things? Why can't we just use SS_{Error} to talk about this instead of SS_{Within}, which you might (we do) find confusing. We're getting there, but perhaps Figure 9.1 will help out.

The figure lines up the partitioning of the Sums of Squares for both between-subjects and repeated-measures designs. In both designs, SS_{Total} is first split up into two pieces $SS_{\text{Effect (between-groups)}}$ and $SS_{\text{Error (within-groups)}}$. At this point, both ANOVAs are the same. In the repeated measures case we split the $SS_{\text{Error (within-groups)}}$ into two more littler parts, which we call $SS_{\text{Subjects (error variation about the subject mean)}}$ and $SS_{\text{Error (left-over variation we can't explain)}}$.

So, when we earlier wrote the formula to split up SS in the repeated-measures design, we were kind of careless in defining what we actually meant by SS_{Error}, this was a little too vague:

$SS_{\text{TOTAL}} = SS_{\text{Effect}} + SS_{\text{Subjects}} + SS_{\text{Error}}$

The critical feature of the repeated-measures ANOVA, is that the SS_{Error} that we will later use to compute the MSE in the denominator for the F-value, is smaller in a repeated-measures design, compared to a between subjects design. This is because the $SS_{\text{Error (within-groups)}}$ is split into two parts, $SS_{\text{Subjects (error variation about the subject mean)}}$ and $SS_{\text{Error (left-over variation we can't explain)}}$.

To make this more clear, consider Figure 9.2:

As we point out, the $SS_{\text{Error (left-over)}}$ in the green circle will be a smaller number than the $SS_{\text{Error (within-group)}}$. That's because we are able to subtract out the SS_{Subjects} part of the $SS_{\text{Error (within-group)}}$. As we will see shortly, this can have the effect of producing larger F-values when using a repeated-measures design compared to a between-subjects design.

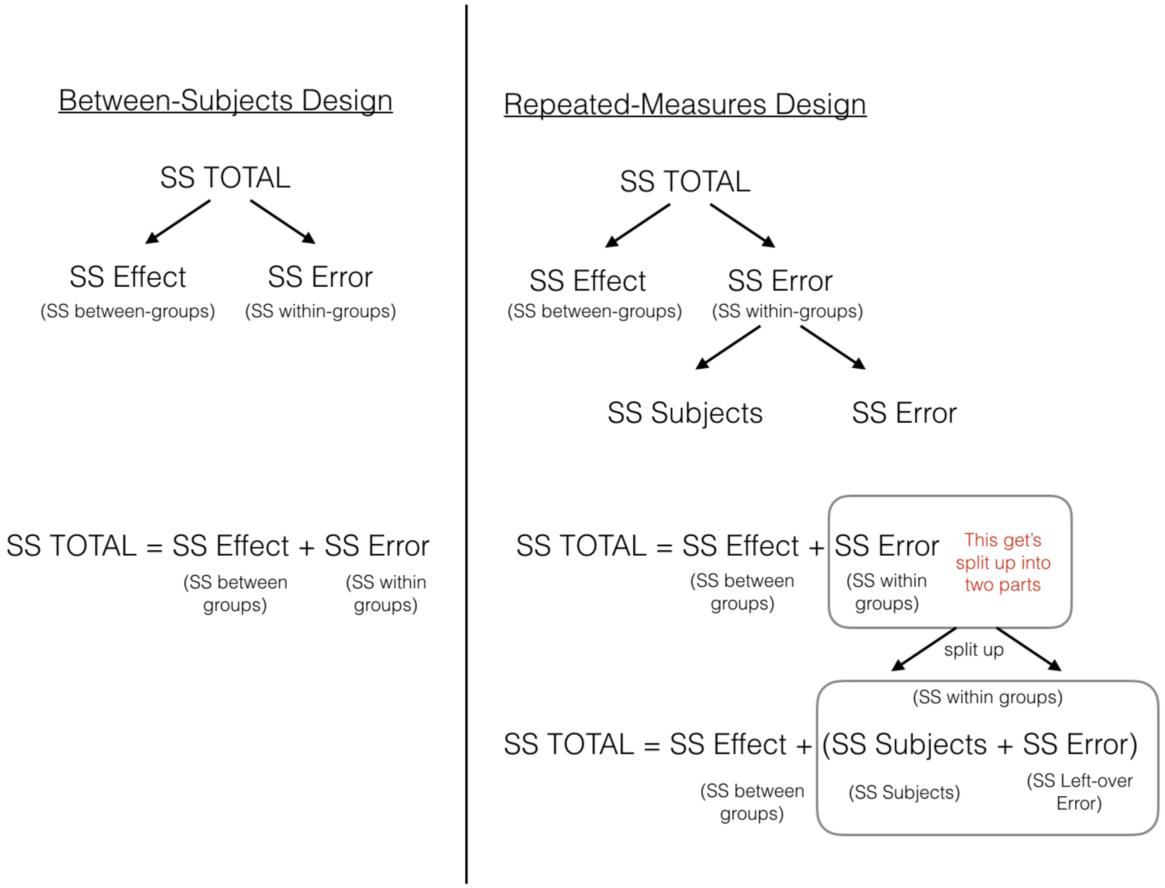


Figure 9.1: Illustration showing how the total sums of squares are partitioned differently for a between versus repeated-measures design

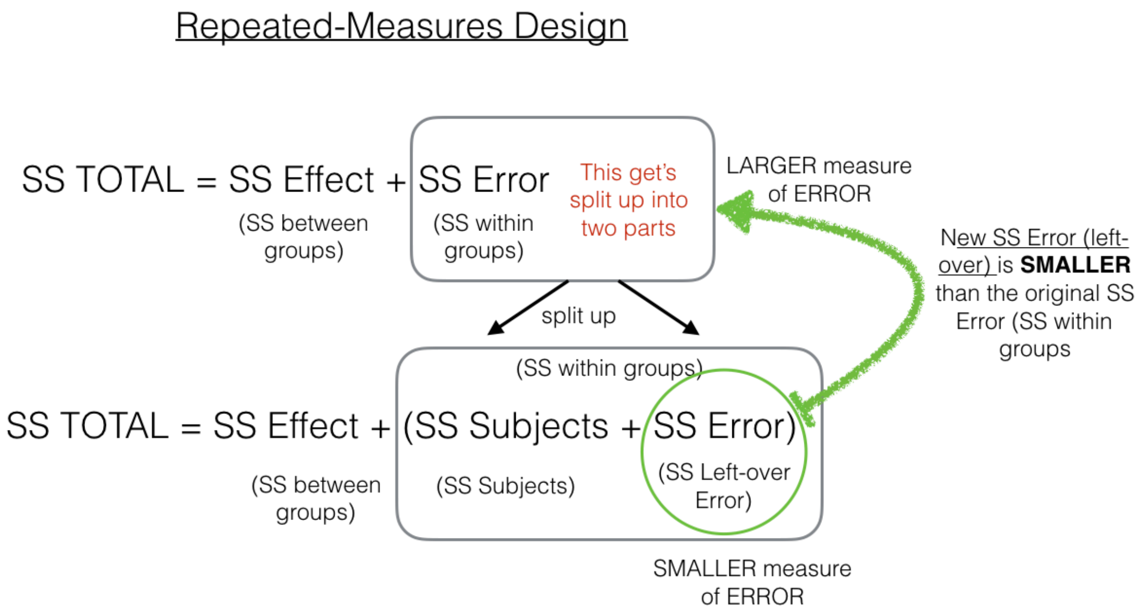
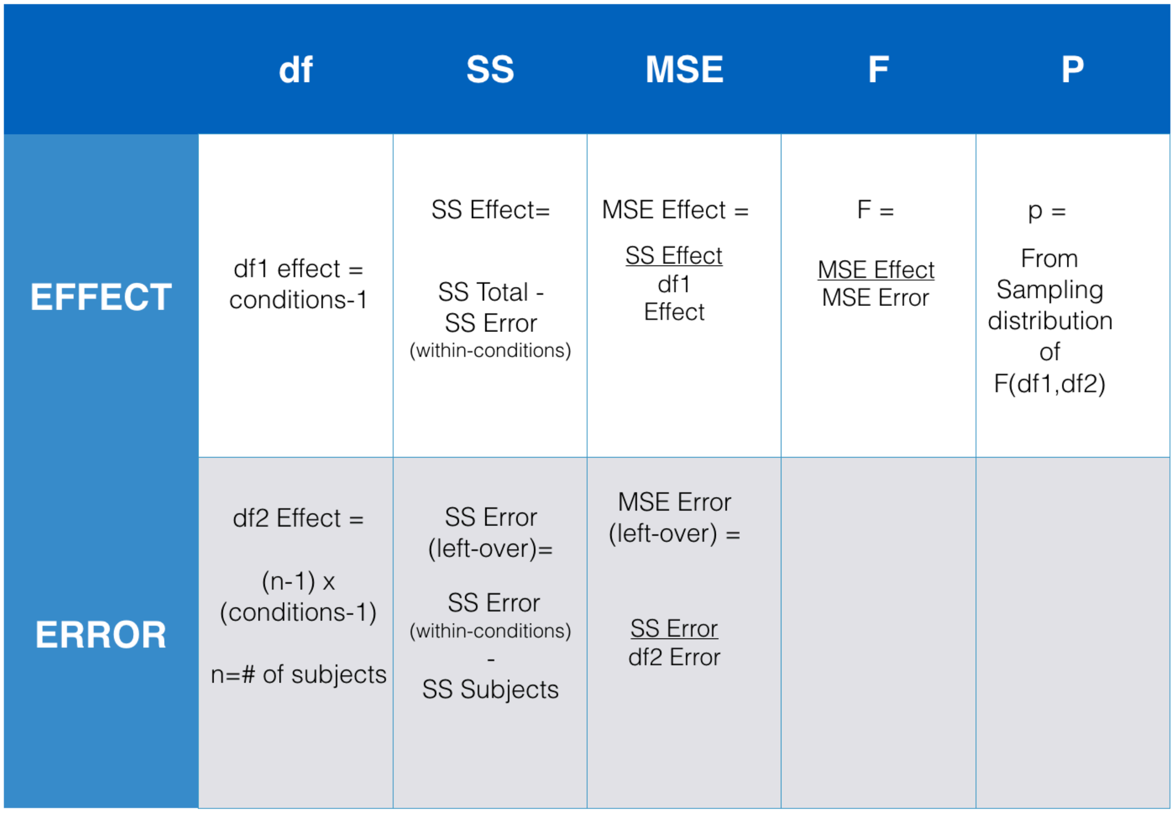

Repeated-Measures Design
SS TOTAL = SS Effect + SS Error
(SS between groups)
(SS within groups)
This get's split up into two parts
LARGER measure of ERROR
New SS Error (left-over) is SMALLER than the original SS Error (SS within groups
split up
(SS within groups)
SS TOTAL = SS Effect + (SS Subjects + SS Error)
(SS between groups)
(SS Subjects)
(SS Left-over Error)
SMALLER measure of ERROR

Figure 9.2: Close-up showing that the Error term is split into two parts in the repeated measures design

9.3 Calculating the RM ANOVA

Now that you are familiar with the concept of an ANOVA table (remember the table from last chapter where we reported all of the parts to calculate the F-value?), we can take a look at the things we need to find out to make the ANOVA table. Figure 9.3 presents an abstract for the repeated-measures ANOVA table. It shows us all the thing we need to calculate to get the F-value for our data.



	df	SS	MSE	F	P
EFFECT	df1 effect = conditions-1	SS Effect= SS Total - SS Error (within-conditions)	MSE Effect = $\frac{\text{SS Effect}}{\text{df1 Effect}}$	F = $\frac{\text{MSE Effect}}{\text{MSE Error}}$	p = From Sampling distribution of F(df1,df2)
ERROR	df2 Effect = (n-1) x (conditions-1) n=# of subjects	SS Error (left-over)= SS Error (within-conditions) - SS Subjects	MSE Error (left-over) = $\frac{\text{SS Error}}{\text{df2 Error}}$		

Figure 9.3: Equations for computing the ANOVA table for a repeated measures design

So, what we need to do is calculate all the SSes that we did before for the between-subjects ANOVA. That means the next three steps are identical to the ones you did before. In fact, I will just basically copy the next three steps to find SS_{TOTAL}, SS_{Effect} , and $SS_{\text{Error (within-conditions)}}$. After that we will talk about splitting up $SS_{\text{Error (within-conditions)}}$ into two parts, this is the new thing for this chapter. Here we go!

9.3.1 SS Total

The total sums of squares, or SSTotal measures the total variation in a set of data. All we do is find the difference between each score and the grand mean, then we square the differences and add them all up.

subjects	conditions	scores	diff	diff_squared
1	A	20	13	169
2	A	11	4	16
3	A	2	-5	25
1	B	6	-1	1
2	B	2	-5	25
3	B	7	0	0
1	C	2	-5	25
2	C	11	4	16
3	C	2	-5	25
Sums		63	0	302
Means		7	0	33.5555555555556

The mean of all of the scores is called the **Grand Mean**. It's calculated in the table, the Grand Mean = 7.

We also calculated all of the difference scores **from the Grand Mean**. The difference scores are in the column titled `diff`. Next, we squared the difference scores, and those are in the next column called `diff_squared`.

When you add up all of the individual squared deviations (difference scores) you get the sums of squares. That's why it's called the sums of squares (SS).

Now, we have the first part of our answer:

$SS_{\text{total}} = SS_{\text{Effect}} + SS_{\text{Error}}$

$SS_{\text{total}} = 302$ and

$302 = SS_{\text{Effect}} + SS_{\text{Error}}$

9.3.2 SS Effect

SS_{Total} gave us a number representing all of the change in our data, how they all are different from the grand mean.

What we want to do next is estimate how much of the total change in the data might be due to the experimental manipulation. For example, if we ran an experiment that causes causes change in the measurement, then the means for each group will be different from other, and the scores in each group will be different from each. As a result, the manipulation forces change onto the numbers, and this will naturally mean that some part of the total variation in the numbers is caused by the manipulation.

The way to isolate the variation due to the manipulation (also called effect) is to look at the means in each group, and the calculate the difference scores between each group mean and the grand mean, and then the squared deviations to find the sum for SS_{Effect}.

Consider this table, showing the calculations for SS_{Effect}.

subjects	conditions	scores	means	diff	diff_squared
1	A	20	11	4	16
2	A	11	11	4	16
3	A	2	11	4	16
1	B	6	5	-2	4
2	B	2	5	-2	4
3	B	7	5	-2	4
1	C	2	5	-2	4
2	C	11	5	-2	4
3	C	2	5	-2	4
Sums		63	63	0	72
Means		7	7	0	8

Notice we created a new column called `means`, these are the means for each condition, A, B, and C.

SS_{Effect} represents the amount of variation that is caused by differences between the means. The `diff` column is the difference between each condition mean and the grand mean, so for the first row, we have 11-7 = 4, and so on.

We found that $SS_{\text{Effect}} = 72$, this is the same as the ANOVA from the previous chapter

9.3.3 SS Error (within-conditions)

Great, we made it to SS Error. We already found SS Total, and SS Effect, so now we can solve for SS Error just like this:

$SS_{\text{total}} = SS_{\text{Effect}} + SS_{\text{Error (within-conditions)}}$

switching around:

$ SS_Error = SS_total - SS_Effect $

$ SS_Error (within conditions) = 302 - 72 = 230 $

Or, we could compute $SS_{\text{Error (within conditions)}}$ directly from the data as we did last time:

subjects	conditions	scores	means	diff	diff_squared
1	A	20	11	-9	81
2	A	11	11	0	0
3	A	2	11	9	81
1	B	6	5	-1	1
2	B	2	5	3	9
3	B	7	5	-2	4
1	C	2	5	3	9
2	C	11	5	-6	36
3	C	2	5	3	9
Sums		63	63	0	230
Means		7	7	0	25.5555555555556

When we compute $SS_{\text{Error (within conditions)}}$ directly, we find the difference between each score and the condition mean for that score. This gives us the remaining error variation around the condition mean, that the condition mean does not explain.

9.3.4 SS Subjects

Now we are ready to calculate new partition, called SS_{Subjects}. We first find the means for each subject. For subject 1, this is the mean of their scores across Conditions A, B, and C. The mean for subject 1 is 9.33 (repeating). Notice there is going to be some rounding error here, that's OK for now.

The `means` column now shows all of the subject means. We then find the difference between each subject mean and the grand mean. These deviations are shown in the `diff` column. Then we square the deviations, and sum them up.

subjects	conditions	scores	means	diff	diff_squared
1	A	20	9.33	2.33	5.4289
2	A	11	8	1	1
3	A	2	3.66	-3.34	11.1556
1	B	6	9.33	2.33	5.4289
2	B	2	8	1	1
3	B	7	3.66	-3.34	11.1556
1	C	2	9.33	2.33	5.4289
2	C	11	8	1	1
3	C	2	3.66	-3.34	11.1556

subjects	conditions	scores	means	diff	diff_squared
Sums		63	62.97	-0.0299999999999994	52.7535
Means		7	6.99666666666667	- 0.00333333333333326	5.8615

We found that the sum of the squared deviations $SS_{\text{Subjects}} = 52.75$. Note again, this has some small rounding error because some of the subject means had repeating decimal places, and did not divide evenly.

We can see the effect of the rounding error if we look at the sum and mean in the `diff` column. We know these should be both zero, because the Grand mean is the balancing point in the data. The sum and mean are both very close to zero, but they are not zero because of rounding error.

9.3.5 SS Error (left-over)

Now we can do the last thing. Remember we wanted to split up the $SS_{\text{Error (within conditions)}}$ into two parts, SS_{Subjects} and $SS_{\text{Error (left-over)}}$. Because we have already calculate $SS_{\text{Error (within conditions)}}$ and SS_{Subjects}, we can solve for $SS_{\text{Error (left-over)}}$:

$SS_{\text{Error (left-over)}} = SS_{\text{Error (within conditions)}} - SS_{\text{Subjects}}$

$SS_{\text{Error (left-over)}} = SS_{\text{Error (within conditions)}} - SS_{\text{Subjects}} = 230 - 52.75 = 177.25$

9.3.6 Check our work

Before we continue to compute the MSEs and F-value for our data, let's quickly check our work. For example, we could have R compute the repeated measures ANOVA for us, and then we could look at the ANOVA table and see if we are on the right track so far.

Table 9.7: Example ANOVA table table reporting the degrees of freedom, sums of squares, mean squares, F value and associated p value.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Residuals	2	52.66667	26.33333	NA	NA
conditions	2	72.00000	36.00000	0.8120301	0.505848
Residuals	4	177.33333	44.33333	NA	NA

Table 9.7 looks good. We found the SS_{Effect} to be 72, and the SS for the conditions (same thing) in the table is also 72. We found the SS_{Subjects} to be 52.75, and the SS for the first

residual (same thing) in the table is also 53.66 repeating. That's close, and our number is off because of rounding error. Finally, we found the $SS_{\text{Error (left-over)}}$ to be 177.25, and the SS for the bottom residuals in the table (same thing) in the table is 177.33 repeating, again close but slightly off due to rounding error.

We have finished our job of computing the sums of squares that we need in order to do the next steps, which include computing the MSEs for the effect and the error term. Once we do that, we can find the F-value, which is the ratio of the two MSEs.

Before we do that, you may have noticed that we solved for $SS_{\text{Error (left-over)}}$, rather than directly computing it from the data. In this chapter we are not going to show you the steps for doing this. We are not trying to hide anything from, instead it turns out these steps are related to another important idea in ANOVA. We discuss this idea, which is called an **interaction** in the next chapter, when we discuss **factorial** designs (designs with more than one independent variable).

9.3.7 Compute the MSEs

Calculating the MSEs (mean squared error) that we need for the F-value involves the same general steps as last time. We divide each SS by the degrees of freedom for the SS.

The degrees of freedom for SS_{Effect} are the same as before, the number of conditions - 1. We have three conditions, so the df is 2. Now we can compute the MSE_{Effect}.

$MSE_{\text{Effect}} = \frac{SS_{\text{Effect}}}{df} = \frac{72}{2} = 36$

The degrees of freedom for $SS_{\text{Error (left-over)}}$ are different than before, they are the (number of subjects - 1) multiplied by the (number of conditions -1). We have 3 subjects and three conditions, so $(3-1)*(3-1) = 2*2 = 4$. You might be wondering why we are multiplying these numbers. Hold that thought for now and wait until the next chapter. Regardless, now we can compute the $MSE_{\text{Error (left-over)}}$.

$MSE_{\text{Error (left-over)}} = \frac{SS_{\text{Error (left-over)}}}{df} = \frac{177.33}{4} = 44.33$

9.3.8 Compute F

We just found the two MSEs that we need to compute F. We went through all of this to compute F for our data, so let's do it:

$F = \frac{MSE_{\text{Effect}}}{MSE_{\text{Error (left-over)}}} = \frac{36}{44.33} = 0.812$

And, there we have it!

9.3.9 p-value

We already conducted the repeated-measures ANOVA using R and reported the ANOVA. Here it is again. The table shows the *p*-value associated with our *F*-value.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Residuals	2	52.66667	26.33333	NA	NA
conditions	2	72.00000	36.00000	0.8120301	0.505848
Residuals	4	177.33333	44.33333	NA	NA

We might write up the results of our experiment and say that the main effect condition was not significant, F(2,4) = 0.812, MSE = 44.33, p = 0.505.

What does this statement mean? Remember, that the *p*-value represents the probability of getting the *F* value we observed or larger under the null (assuming that the samples come from the same distribution, the assumption of no differences). So, we know that an *F*-value of 0.812 or larger happens fairly often by chance (when there are no real differences), in fact it happens 50.5% of the time. As a result, we do not reject the idea that any differences in the means we have observed could have been produced by chance.

9.4 Things worth knowing

Repeated Measures ANOVAs have some special properties that are worth knowing about. The main special property is that the error term used to for the *F*-value (the MSE in the denominator) will always be smaller than the error term used for the *F*-value the ANOVA for a between-subjects design. We discussed this earlier. It is smaller, because we subtract out the error associated with the subject means.

This can have the consequence of generally making *F*-values in repeated measures designs larger than *F*-values in between-subjects designs. When the number in the bottom of the *F* formula is generally smaller, it will generally make the resulting ratio a larger number. That's what happens when you make the number in the bottom smaller.

Because big *F* values usually let us reject the idea that differences in our means are due to chance, the repeated-measures ANOVA becomes a more sensitive test of the differences (its *F*-values are usually larger).

At the same time, there is a trade-off here. The repeated measures ANOVA uses different degrees of freedom for the error term, and these are typically a smaller number of degrees of freedom. So, the *F*-distributions for the repeated measures and between-subjects designs are actually different *F*-distributions, because they have different degrees of freedom.

9.4.1 Repeated vs between-subjects ANOVA

Let's do a couple simulations to see some the differences between the ANOVA for a repeated measures design, and the ANOVA for a between-subjects design.

We will do the following.

1. Simulate a design with three conditions, A, B, and C
2. sample 10 scores into each condition from the same normal distribution (mean = 100, SD = 10)
3. We will include a subject factor for the repeated-measures version. Here there are 10 subjects, each contributing three scores, one each condition
4. For the between-subjects design there are 30 different subjects, each contributing one score in the condition they were assigned to (really the group).

We run 1000 simulated experiments for each design. We calculate the F for each experiment, for both the between and repeated measures designs. Figure 9.4 has the sampling distributions of F for both designs.

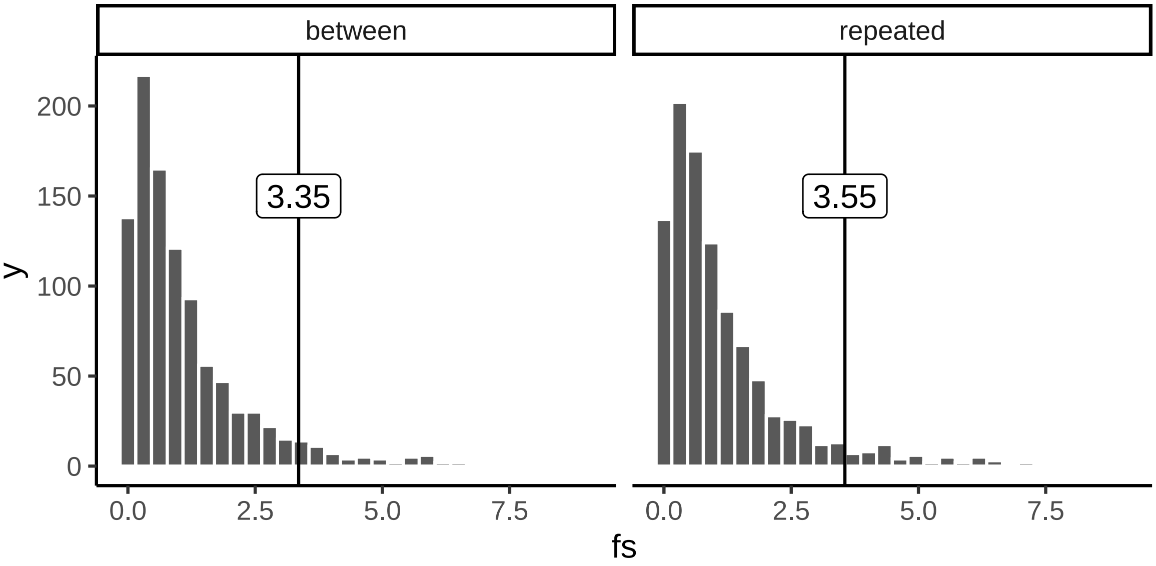


Figure 9.4: Comparing critical F values for a between and repeated measures design

These two F sampling distributions look pretty similar. However, they are subtly different. The between F distribution has degrees of freedom 2, and 27, for the numerator and denominator. There are 3 conditions, so $df$1 = 3-1 = 2. There are 30 subjects, so $df$2 = 30-3 =27. The critical value, assuming an alpha of 0.05 is 3.35. This means F is 3.35 or larger 5% of the time under the null.

The repeated-measures F distribution has degrees of freedom 2, and 18, for the numerator and denominator. There are 3 conditions, so \$df\$1 = 3-1 = 2. There are 10 subjects, so \$df\$2 = (10-1)*(3-1) = 9*2 = 18. The critical value, assuming an alpha of 0.05 is 3.55. This means F is 3.55 or larger 5% of the time under the null.

The critical value for the repeated measures version is slightly higher. This is because when \$df\$2 (the denominator) is smaller, the F-distribution spreads out to the right a little bit. When it is skewed like this, we get some bigger Fs a greater proportion of the time.

So, in order to detect a real difference, you need an F of 3.35 or greater in a between-subjects design, or an F of 3.55 or greater for a repeated-measures design. The catch here is that when there is a real difference between the means, you will detect it more often with the repeated-measures design, even though you need a larger F (to pass the higher critical F-value for the repeated measures design).

9.4.2 repeated measures designs are more sensitive

To illustrate why repeated-measures designs are more sensitive, we will conduct another set of simulations.

We will do something slightly different this time. We will make sure that the scores for condition A, are always a little bit higher than the other scores. In other words, we will program in a real true difference. Specifically, the scores for condition will be sampled from a normal distribution with mean = 105, and SD = 10. This mean is 5 larger than the means for the other two conditions (still set to 100).

With a real difference in the means, we should now reject the hypothesis of no differences more often. We should find F values larger than the critical value more often. And, we should find p-values for each experiment that are smaller than .05 more often, those should occur more than 5% of the time.

To look at this we conduct 1000 experiments for each design, we conduct the ANOVA, then we save the p-value we obtained for each experiment. This is like asking how many times will we find a p-value less than 0.05, when there is a real difference (in this case an average of 5) between some of the means. Figure 9.5 contains histograms of the p-values:

Here we have two distributions of observed p-values for the simulations. The red line shows the location of 0.05. Overall, we can see that for both designs, we got a full range of p-values from 0 to 1. This means that many times we would not have rejected the hypothesis of no differences (even though we know there is a small difference). We would have rejected the null every time the p-value was less than 0.05.

For the between subject design, there were 579 experiments with a p less than 0.05, or 0.579 of experiments were “significant”, with alpha=.05.

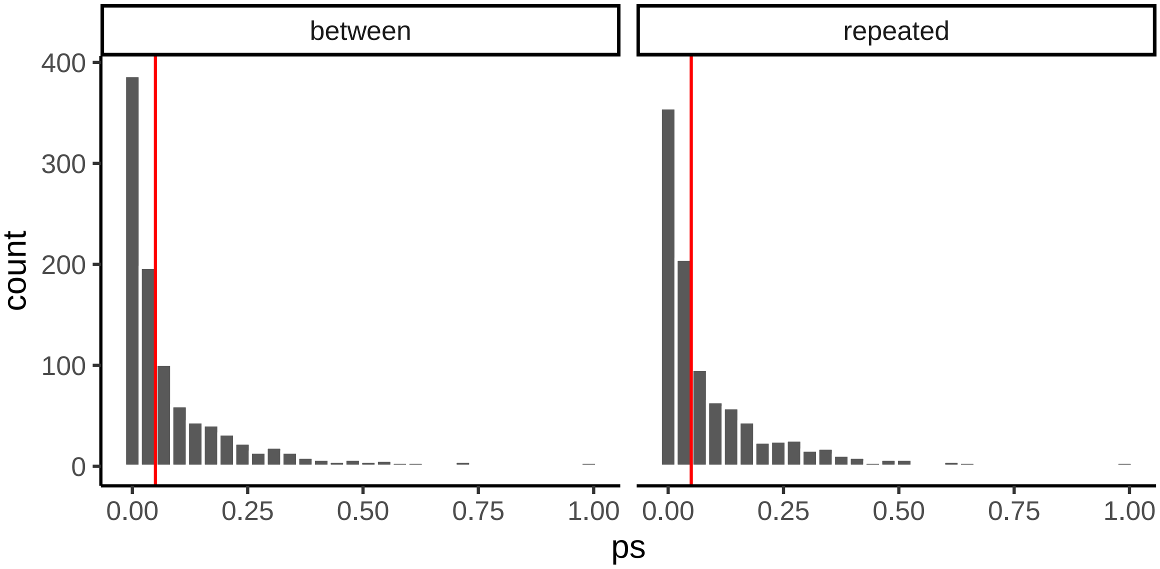


Figure 9.5: p-value distributions for a between and within-subjects ANOVA

For the within subject design, there were 554 experiments with a p less than 0.05, or 0.554 of experiments were "significant", with alpha=.05.

OK, well, you still might not be impressed. In this case, the between-subjects design detected the true effect slightly more often than the repeated measures design. Both them were right around 55% of the time. Based on this, we could say the two designs are pretty comparable in their sensitivity, or ability to detect a true difference when there is one.

However, remember that the between-subjects design uses 30 subjects, and the repeated measures design only uses 10. We had to make a big investment to get our 30 subjects. And, we're kind of unfairly comparing the between design (which is more sensitive because it has more subjects) with the repeated measures design that has fewer subjects.

What do you think would happen if we ran 30 subjects in the repeated measures design? Let's find out. Figure 9.6 re-plots the above, but this time only for the repeated measures design. We increase N from 10 to 30.

Wowsers! Look at that. When we ran 30 subjects in the repeated measures design almost all of the p-values were less than .05. There were 981 experiments with a p less than 0.05, or 0.981 of experiments were "significant", with alpha=.05. That's huge! If we ran the repeated measures design, we would almost always detect the true difference when it is there. This is why the repeated measures design can be more sensitive than the between-subjects design.

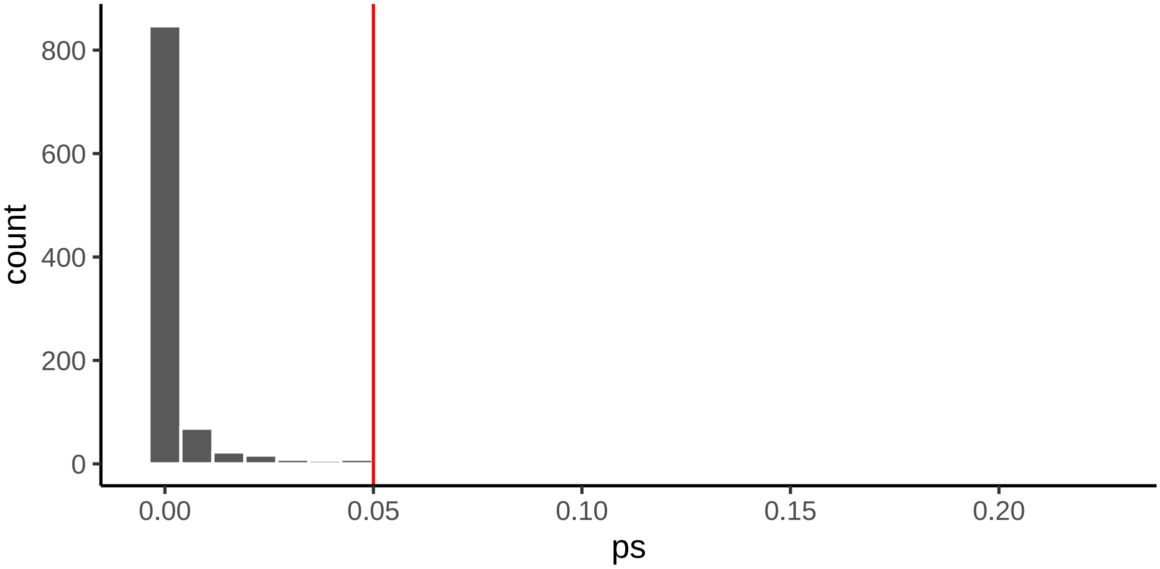


Figure 9.6: p-value distribution for within-subjects design with $n = 30$

9.5 Real Data

Let's look at some real data from a published experiment that uses a repeated measures design. This is the same example that you will be using in the lab for repeated measures ANOVA. The data happen to be taken from a recent study conducted by Lawrence Behmer and myself, at Brooklyn College (Behmer and Crump 2017).

We were interested in how people perform sequences of actions. One question is whether people learn individual parts of actions, or the whole larger pattern of a sequence of actions. We looked at these issues in a computer keyboard typing task. One of our questions was whether we would replicate some well known findings about how people type words and letters.

From prior work we knew that people type words way faster than than random letters, but if you made the random letters a little bit more English-like, then people type those letter strings a little bit faster, but not as slow as random string.

In the study, 38 participants sat in front of a computer and typed 5 letter strings one at a time. Sometimes the 5 letter made a word (Normal condition, TRUCK), sometimes they were completely random (Random Condition, JWYFG), and sometimes they followed patterns like you find in English (Bigram Condition, QUEND), but were not actual words. So, the independent variable for the typing material had three levels. We measured every single keystroke that participants made. This gave us a few different dependent measures. Let's take a look a the reaction times. This is how long it took for participants to start typing the first letter in the string.

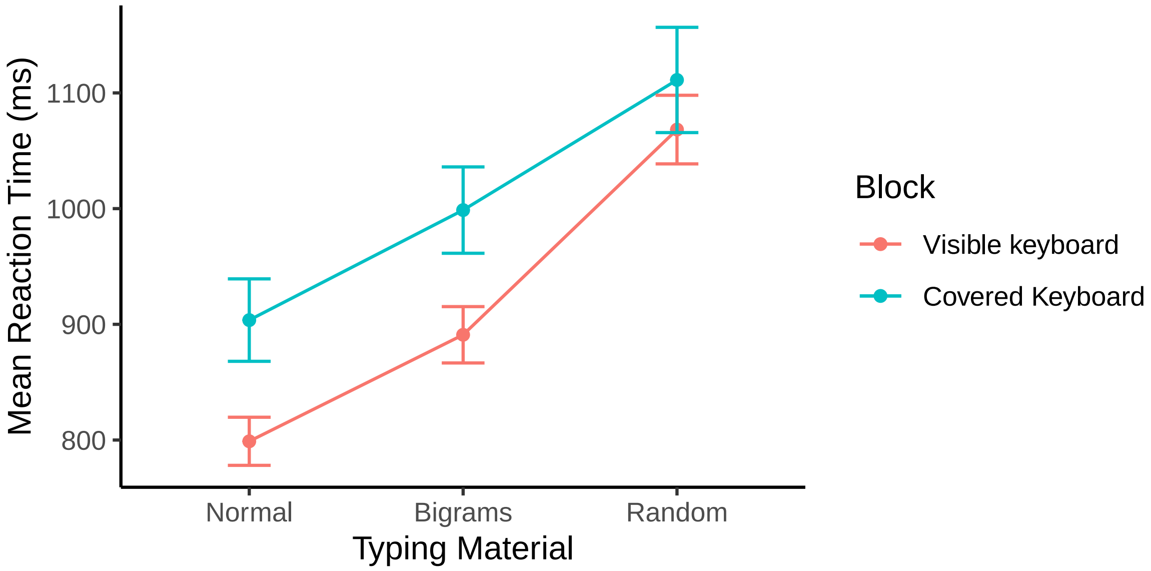


Figure 9.7: Results from Behmer & Crump (2017)

OK, I made a figure showing the mean reaction times for the different typing material conditions. You will notice that there are two sets of lines. That's because there was another manipulation I didn't tell you about. In one block of trials participants got to look at the keyboard while they typed, but in the other condition we covered up the keyboard so people had to type without looking. Finally, the error bars are standard error of the means.

ℹ Note

Note, the use of error bars for repeated-measures designs is not very straightforward. In fact the standard error of the means that we have added here are not very meaningful for judging whether the differences between the means are likely not due to chance. They would be if this was a between-subjects design. We will update this textbook with a longer discussion of this issue, for now we will just live with these error bars.

For the purpose of this example, we will say, it sure looks like the previous finding replicated. For example, people started typing Normal words faster than Bigram strings (English-like), and they started typing random letters the most slowly of all. Just like prior research had found.

Let's focus only on the block of trials where participants were allowed to look at the keyboard while they typed, that's the red line, for the "visible keyboard" block. We can see the means look different. Let's next ask, what is the likelihood that chance (random sampling error) could have produced these mean differences. To do that we run a repeated-measures ANOVA in R. Here is the ANOVA table.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Residuals	37	2452611.9	66286.808	NA	NA
Stimulus	2	1424914.0	712457.010	235.7342	0
Residuals1	74	223649.4	3022.289	NA	NA

Alright, we might report the results like this. There was a significant main effect of Stimulus type, F(2, 74) = 235.73, MSE = 3022.289, p < 0.001.

Notice a couple things. First, this is a huge F-value. It's 253! Notice also that the p-value is listed as 0. That doesn't mean there is zero chance of getting an F-value this big under the null. This is a rounding error. The true p-value is 0.00000000000000… The zeros keep going for a while. This means there is only a vanishingly small probability that these differences could have been produced by sampling error. So, we reject the idea that the differences between our means could be explained by chance. Instead, we are pretty confident, based on this evidence and and previous work showing the same thing, that our experimental manipulation caused the difference. In other words, people really do type normal words faster than random letters, and they type English-like strings somewhere in the middle in terms of speed.

9.6 Summary

In this chapter you were introduced to the repeated-measures ANOVA. This analysis is appropriate for within-subjects or repeated measures designs. The main difference between the independent factor ANOVA and the repeated measures ANOVA, is the ability to partial out variance due to the individual subject means. This can often result in the repeated-measures ANOVA being more sensitive to true effects than the between-subjects ANOVA.

10 Factorial ANOVA

In environmental science, things are rarely simple. Factors like climate, soil type, and water availability all work together, affecting ecosystems in ways that aren't always straightforward. That's where factorial designs come in handy. They help us understand how different factors interact with each other.

So far, we've mostly looked at situations where we change one thing at a time and see what happens. But in real-world ecosystems or when studying climate interactions, there are usually many things changing at once. Factorial designs allow us to study these complex situations more effectively. They help us see not just what each factor does on its own, but also how they influence each other when they're all at play together.

Imagine an investigation into the growth rates of a particular plant species. The dependent variable here would be the growth rate, while the independent variables could range from soil pH to sunlight exposure. In environmental factorial designs, each independent variable, with their respective levels, weaves into a tapestry of possible outcomes, allowing us to observe not just isolated effects, but the symphony of interactions.

Let's explore some factorial design examples tailored to our environmental inquiries:

1. 1 IV (two levels)

A t-test would suffice here, as we are dealing with just two levels of our independent variable.

a. Soil pH (Acidic vs. Neutral): How does soil pH affect plant growth? We have one IV (soil pH), with two levels (Acidic vs. Neutral).

b. Sunlight Exposure (Full Sun vs. Partial Shade): Does the amount of sunlight influence plant growth rates? Here, there's one IV (sunlight exposure), with two levels (Full Sun vs. Partial Shade).

2. 1 IV (three levels):

An ANOVA is appropriate here, given the variable has more than two levels.

a. Water Availability (Low, Medium, High): How does varying water availability impact plant growth? Our single IV (water availability) has three levels (Low, Medium, High).

b. Fertilizer Type (No fertilizer, Organic, Synthetic): What's the effect of different fertilizer types on plant growth? One IV (fertilizer type) has three levels.

3. **2 IVs: IV1 (two levels), IV2 (two levels)**

We would employ a factorial ANOVA for this design, which we haven't discussed yet.

a. IV1 (Soil pH: Acidic vs. Neutral); IV2 (Water Availability: Low vs. High): How do soil pH and water availability together affect plant growth? Here, we have two IVs—soil pH and water availability. Each IV has two levels, leading to a **2x2** Factorial Design. This type of design is fully crossed; each level of one IV is paired with each level of the other IV to observe the combined effect on growth rates.

OK, let's stop here for the moment. The first two designs both had one IV. The third design shows an example of a design with 2 IVs (soil pH and water availability), each with two levels. This is called a **2x2 Factorial Design**. It is called a **factorial** design, because the levels of each independent variable are fully crossed. This means that first each level of one IV, the levels of the other IV are also manipulated.

10.1 Factorial basics

10.1.1 2x2 Designs

We've just started talking about a **2x2 Factorial design**. We said this means the IVs are crossed. To illustrate this, take a look at Figure 10.1. We show an abstract version and a concrete version using soil pH and water availability as the two IVs, each with two levels in the design:

2x2 Design		IV 1	
		IV1: Level 1	IV1: Level 2
IV 2	IV2: Level 1	dv	dv
	IV2: Level 2	dv	dv

2x2 Design		Time of Day	
		Morning	Afternoon
Caffeine	Some Caffeine	dv	dv
	No Caffeine	dv	dv

Figure 10.1: Structure of 2x2 factorial designs

- IV1: Soil Type (Clay vs. Loam)
- IV2: Drought Condition (Yes vs. No)

Each combination of soil type and drought condition creates a unique environment for the plants, leading to four distinct scenarios to measure growth rates. So, we have 2 IVs, each with 2 levels, for a total of 4 conditions. This is why we call it a 2x2 design. 2x2 = 4. The notation tells us how to calculate the total number of conditions.

10.1.2 Factorial Notation

Anytime **all of the levels** of each IV in a design are fully crossed, so that they all occur for each level of every other IV, we can say the design is a **fully factorial** design.

Our notation system succinctly encapsulates the structure of factorial designs. Each IV gets a number representing its levels. Here are a few examples:

- 2x2: Two IVs, each with two levels, yielding four unique conditions.
- 2x3: Two IVs, with the first having two levels and the second three, resulting in six conditions.
- 3x2: Similar to the 2x3, but the first IV has three levels, and the second has two, also giving us six conditions.
- 4x4: Two IVs, each with four levels, resulting in sixteen conditions.

Expanding on our designs, a 2x3 factorial design could investigate the impact of two soil types across three different levels of water availability, equating to six unique conditions to analyze plant growth.

10.2 Purpose of Factorial Designs

Factorial designs let us ask nuanced questions about ecological phenomena. By manipulating multiple variables simultaneously, we gain insights into how different environmental factors might interact to influence a specific outcome, like plant growth or species distribution.

10.2.1 Factorials manipulate an effect of interest

Factorial designs enable researchers to sift through multiple layers of influence to grasp the broader picture. Complicated? Certainly, but it's complexity that mirrors the real world where multiple factors often come into play simultaneously.

Let's make sense of this by contemplating a multifaceted environmental study. Imagine we're environmental scientists, eager to measure the effects of various pollutants on aquatic life. Here's what we could do:

1. **Choose a bioindicator**: Our subjects are the varied aquatic macroinvertebrates, whose species diversity is a critical measure of water health—our dependent variable.

2. **Identify impactful variables**: We've identified the usual suspect – excess Nutrient-loaded agricultural runoff (Pollutant A), notorious for depleting oxygen and threatening our aquatic buddies. We'll add this to our macroinvertebrates' environment and observe the ripple effects on their diversity, using a pristine control (No Pollutant) for comparison.
3. **Measure the impact**: We'll examine the richness and variety of macroinvertebrate species when exposed to each type of pollutant.
4. **Detect the variation**: Variations in macroinvertebrate diversity will illuminate the 'Pollution effect'. The impact of pollutaiton on macroinvertebrate diversity.

The proof is in the visuals. We aim to contrast macroinvertebrate diversity in the shadow of Pollutant A with a pollutant-free scenario. Figure 10.2 shows how the data might look.

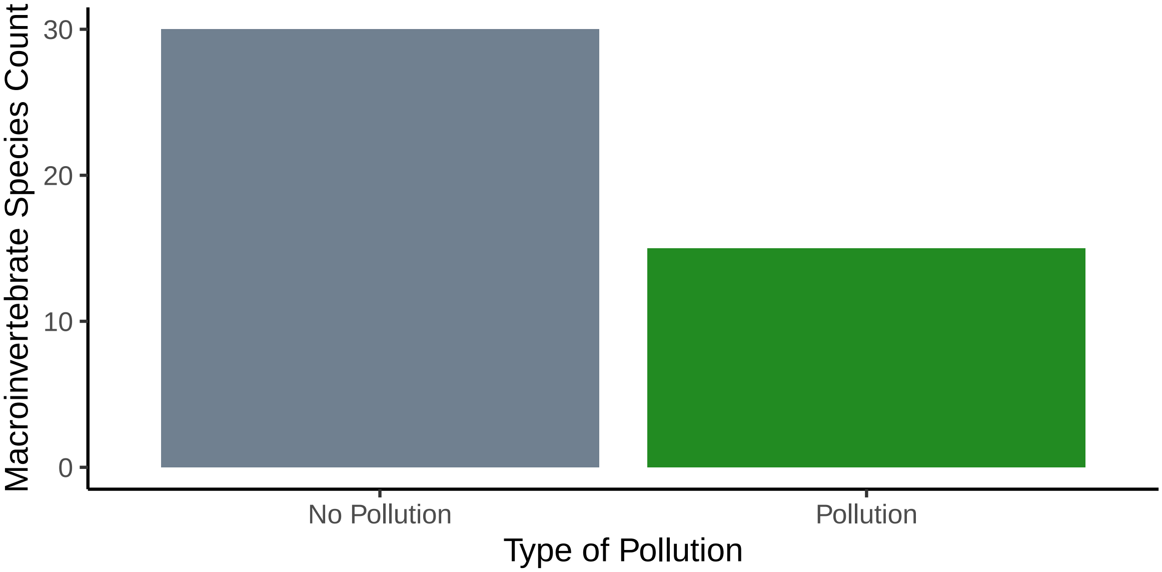


Figure 10.2: Example data illustrating the influence of different pollution types on aquatic macroinvertebrate species diversity

Behold! The macroinvertebrate lineup takes a hit from both pollutants, but they really don't like the industrial waste after-party. In general, it is very common to use the word **effect** to refer to the differences caused by the manipulation. This is what we could call the "Pollution effect".

10.2.2 Manipulating the Pollution effect

This is where factorial designs come in to play. We've already pinpointed a 'Pollution effect'—the change in macroinvertebrate biodiversity when pollutants enter their habitat. This effect serves as our guide, leading us to ask: Where does pollution strike the hardest? What conditions exacerbate its harmful impact?

One possible lever for controlling this Pollution effect could be the timing of fertilizer application, reducing runoff by strategic scheduling post-rainfall. Cover cropping might also act as a shield, protecting our waters from nutrient excess. But now, we introduce a new player: flow rate. Will high flow rate reduce impacts of pollution, since the pollution will scoot on out of there? Or, paradoxically, could it exacerbate the issue, hastening the spread of pollutants that might otherwise degrade if left undisturbed?

Our question evolves: Does flow rate influence the Pollution Effect? We hypothesize that a higher flow rate might reduce the observable Pollution Effect compared to a more stagnant, low flow environment. If our theory holds water, the results could flow out like the data in Figure 10.3.

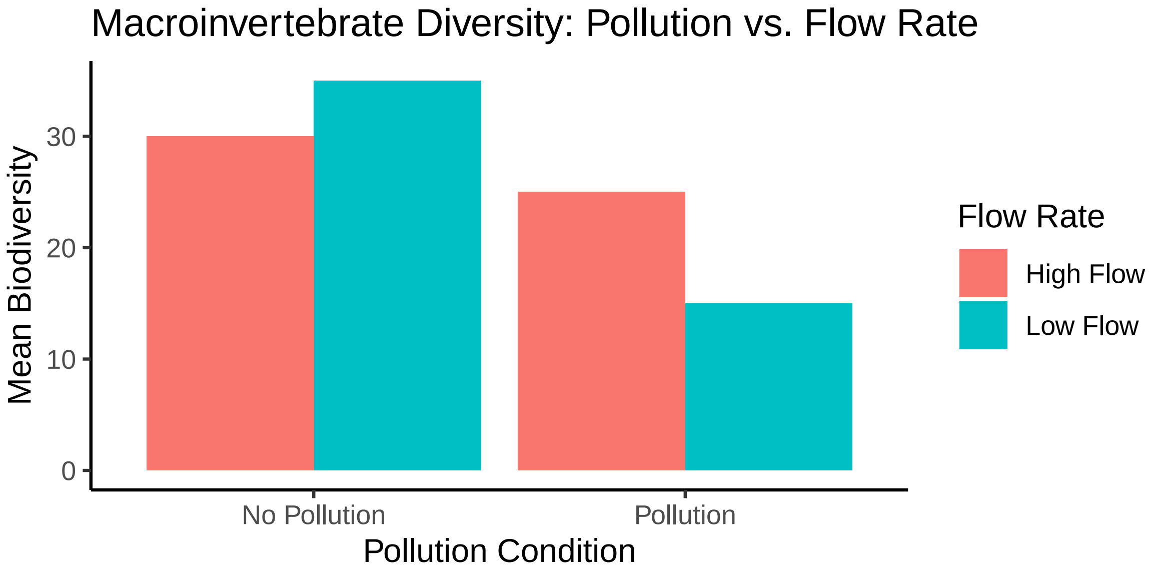


Figure 10.3: Example data showing how the Pollution effect could be modulated by a Flow manipulation. Pollution plotted on the x-axis, makes it more difficult to compare the changes in the pollution effect between flow rate conditions

In this graph, we maintain consistency by placing the Pollution conditions on the x-axis.The bars represent the average macroinvertebrate diversity for each combination of flow and pollution conditions, with colors distinguishing the flow rates. But, it's not as helpful as it could be. We can try to interpret this graph, but **?@fig-10flowB** plots the same data in a way that makes it easier to see what we are talking about.

Here, the x-axis represents the Flow Rate, with the color of the bars indicating the Pollution condition. This graph layout simplifies the comparison of the Pollution effect within each Flow

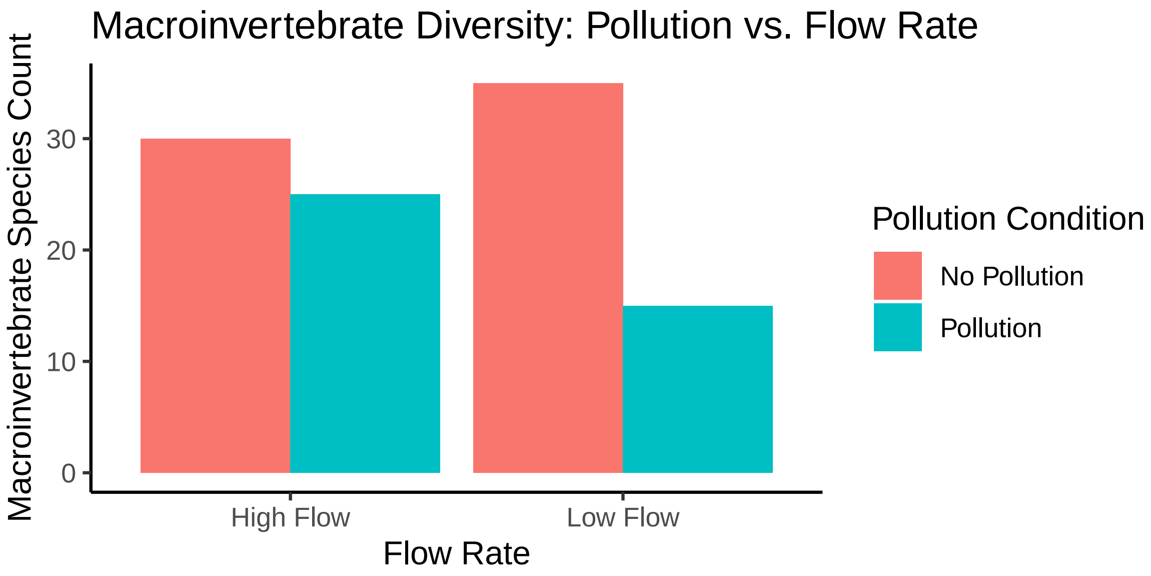


Figure 10.4: Example data showing how the pollution effect could be modulated by a flow rate manipulation. Flow condition plotted on the x-axis, makes it easier to compare the changes in the pollution effect between Flow conditions

Rate condition, making it easier to discern the interaction between these two factors.

Low-Flow condition: In the low flow condition, our macroinvertebrates were observed in both polluted and unpolluted waters. This mirrors the baseline scenario of our study. Consistently with our predictions, the graph would likely show a significant difference: a higher species count in unpolluted waters compared to polluted ones. Thus, we'd observe a Pollution effect, with a specific difference in species count illustrating the impact of pollution under low flow conditions.

Flow condition: In the high flow condition, the macroinvertebrates also faced both polluted and unpolluted environments. However, the dynamic nature of high flow was hypothesized to influence their response to pollution. The expectation was that the swift current might diminish the observable Pollution effect by dispersing pollutants more effectively. If the graph supports our hypothesis, we'd see a smaller difference in species count between the polluted and unpolluted conditions under high flow, indicating a lessened Pollution effect.

Should our research validate these predictions, we could infer that flow rate does indeed modulate the Pollution effect. In a low flow environment, the Pollution effect might be pronounced, as observed by the larger difference in species counts. Conversely, in a high flow scenario, the effect diminishes, with the difference in biodiversity less stark. Therefore, the manipulation of flow rate could potentially alter the Pollution effect by a quantifiable margin, underscored by the variation in species counts between the two flow conditions.v

This is our description of why factorial designs are so useful. They allow researchers to find out what kinds of manipulations can cause changes in the effects they measure. In our environmental study, for instance, we've measured the Pollution effect on macroinvertebrate biodiversity. Then, we introduced the variable of flow rate to see if and how it might alter this effect. If our goal is to untangle the web of ecological dynamics, we'd need to delve into the mechanisms by which flow rate could influence pollutant dispersion and, subsequently, biodiversity. We have the initial evidence suggesting that flow rate can indeed sway the Pollution effect. The next step is to craft hypotheses on the nature of this influence—perhaps proposing that faster flows dilute pollutants more effectively, thus mitigating their negative impact on biodiversity. These hypotheses then become the basis for further experiments, designed to test their validity and expand our understanding of these complex environmental interactions.

10.3 Graphing the means

In our example above we showed you two bar graphs of the very same means for our 2x2 design. Even though the graphs plot identical means, they look different, so they are more or less easy to interpret by looking at them. Results from 2x2 designs are also often plotted with line graphs. Those look different too. There are four different graphs in Figure 10.5, using bars and lines to plot the very same means from before. We are showing you this so that you realize **how you graph your data matters** because it makes it more or less easy for people to understand the results. Also, how the data is plotted matters for what you need to look at to interpret the results.

10.4 Knowing what you want to find out

When you conduct a design with more than one IV, you get more means to look at. As a result, there are more kinds of questions that you can ask of the data. Sometimes it turns out that the questions that you can ask, are not the ones that you want to ask, or have an interest in asking. Because you ran the design with more than one IV, you have the opportunity to ask these kinds of extra questions.

What kinds of extra questions? Let's keep going with our Pollution effect experiment. We have the first IV where we manipulated Pollution. So, we could find the overall means in spot-the difference for the Pollution vs. no-Pollution conditions (that's two means). The second IV was Flow. We could find the overall means in spot-the-difference performance for the Flow vs. no-Flow conditions (that's two more means). We could do what we already did, and look at the means for each combination, that is the mean for Pollution/Flow, Pollution/no-Flow, no-Pollution/Flow, and no-Pollution/no-Flow (that's four more means, if you're counting).

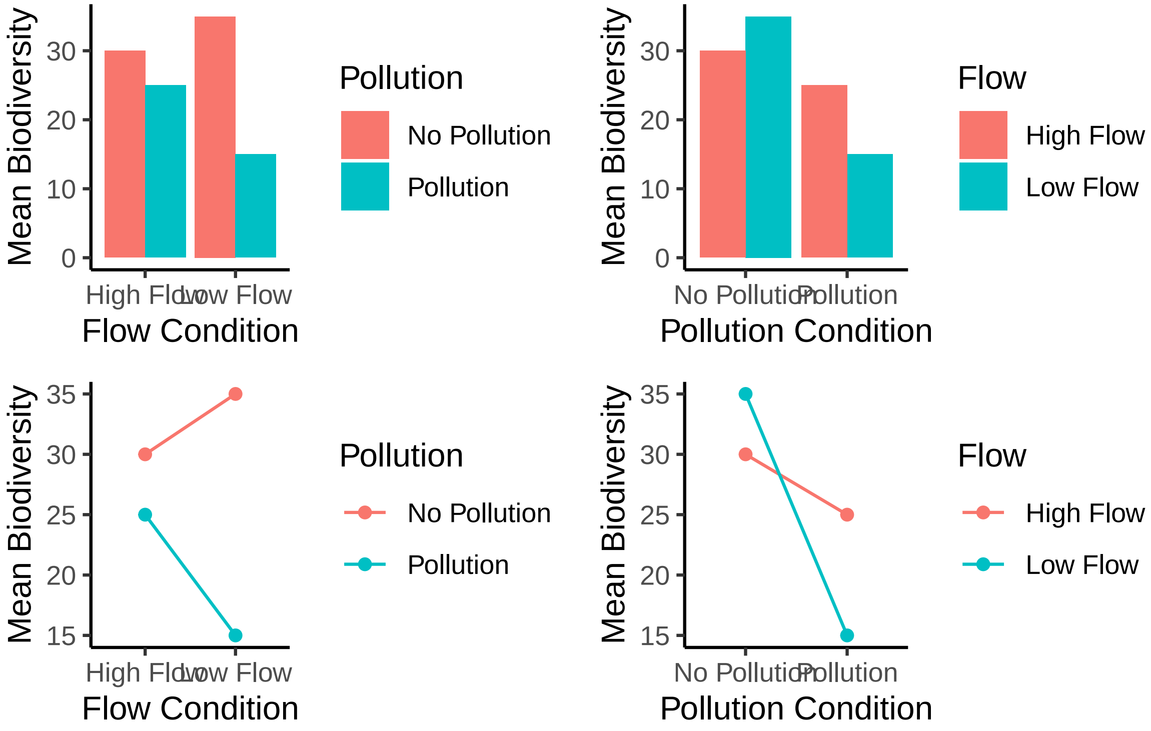


Figure 10.5: same example means plotted using bar graphs or line graphs, and with Pollution or Flow on the x-axis

There's even more. We could look at the mean Pollution effect (the difference between pollution and no pollution) for the low flow condition, and the mean Pollution effect for the high flow condition (that's two more).

Figure 10.6 shows multiple ways of looking at the means across four panels.

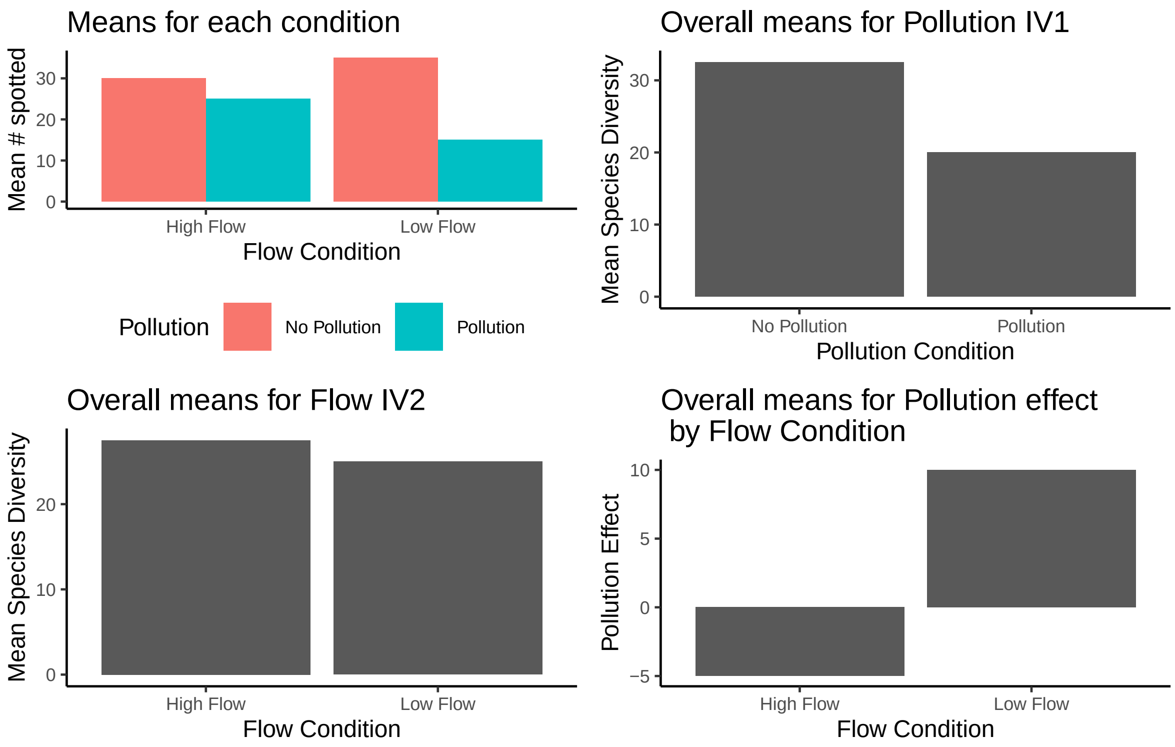


Figure 10.6: Each panel shows the mean for different effects in the design

The top left panel of our graph offers a snapshot of macroinvertebrate populations under different environmental scenarios. To assess the impact of Pollution, we look at the differences in species counts between the unpolluted (aqua bar) and polluted (red bar) conditions within both High Flow and Low Flow contexts. This visual comparison sheds light on how Pollution affects species diversity across varying flow rates.

The top right panel, however, does not delve into the effects of Flow. Rather, it showcases the general impact of Pollution on species diversity, known as the main effect of Pollution. This is a direct comparison of the number of species found in unpolluted versus polluted waters, without the flow rate being a factor.

The bottom left panel similarly isolates another variable, focusing solely on the main effect of Flow. It contrasts species diversity in High Flow versus Low Flow conditions, without the influence of Pollution being considered.

It's in the bottom right panel that we explore the interaction between Flow and Pollution. Here, the y-axis quantifies the Pollution effect, reflecting the change in species diversity from unpolluted to polluted conditions. In the Low-Flow scenario, there's a substantial drop (a difference of 20, from 35 to 15) when Pollution is introduced. In contrast, under High Flow, the introduction of Pollution results in a smaller decrease (a difference of -5, from 30 to 25). The disparity between these two outcomes (20 in Low Flow versus -5 in High Flow) indicates that Flow significantly moderates the effect of Pollution. This differential effect, where the influence of one factor (Pollution) varies according to the level of another (Flow), is what we define as an **interaction**.

> Pro tip
>
> Before you start your environmental detective work, know what you're looking for. What's your question? What clues (means) are relevant? Keep your eyes on the prize, and don't get lost in the sea of data!

10.5 Simplified Analysis of 2x2 Repeated Measures Design

In discussions of factorial designs, the concept of Factorial ANOVAs often comes up. These are complex statistical tests that we'll explore shortly. However, before we delve into Factorial ANOVAs, let's look at how to analyze a 2x2 repeated measures design using paired-samples t-tests. This approach is less common but yields results comparable to those from an ANOVA.

Understanding ANOVA can be challenging, and it gets even trickier with factorial designs. To ease into this complexity, we'll start with t-tests to demonstrate the principles behind Factorial ANOVA. Conducting t-tests requires precise comparisons, which will help you grasp what you're seeking to understand from factorial designs. Once you're clear on your research questions, you can apply ANOVA to uncover the answers and know exactly what to look for in the results. This step-by-step approach makes learning a more enjoyable journey!

First, let's define two key terms: **main effects** and **interactions**. In any factorial design, you have the chance to analyze these elements. The number of **main effects** and **interactions** you can investigate depends on the number of independent variables (IVs) in your design.

10.5.1 Main Effects

A main effect represents the average difference associated with a single independent variable (IV). There's one main effect for each IV. In a 2x2 design, which includes two IVs, there are two main effects. For instance, we might have one main effect for Pollution and another for Flow. A significant main effect suggests that the differences observed are not likely due to random chance.

	No Pollution	Pollution	No Pollution	Pollution
	Low Flow		High Flow	
subject	A	B	C	D
1	10	5	12	9
2	8	4	13	8
3	11	3	14	10
4	9	4	11	11
5	10	2	13	12

In a 2x2x2 design, you'd evaluate three main effects, corresponding to each IV. A 3x3x3 design, despite having more levels, still involves three IVs, so you'd again have three main effects.

10.5.2 Interaction

The concept of interaction can be perplexing in factorial designs. Simply put, an interaction occurs when the effect of one IV on the outcome is influenced by another IV. For example, we observed that the presence of Flow affected the magnitude of the Pollution effect. The Pollution effect was more pronounced without Flow and less so with Flow, indicating an interaction.

Another way to think about interactions is to consider them as the difference between differences. If the Pollution effect (the difference in outcomes between polluted and unpolluted conditions) changes when we introduce different Flow conditions, we're observing an interaction. This concept can seem complex, but don't worry—we'll go through more examples to clarify it.

The number of possible interactions in a design is tied to the number of IVs. A 2x2 design has one interaction: the combined effect of the two IVs. This single interaction examines whether the impact of one IV varies across the levels of the other IV. In designs with more than two IVs, the potential for interactions increases. For example, a design with three IVs (A, B, and C) would have three 2-way interactions (AB, AC, and BC) and one 3-way interaction (ABC).

10.5.3 Looking at the data

Understanding our analysis begins with a clear view of the data. Consider our hypothetical study on attention, with five experimental tanks that each experience all conditions, making this a fully repeated-measures design. The dataset might look something like this:

10.5.4 Main effect of Pollution

To assess the main effect of Pollution, we compare the mean scores across the no-Pollution and Pollution conditions, regardless of the Flow conditions.

	No Pollution		Pollution				
subject	A	B	C	D	Mean_No_Pollution	Mean_Pollution	Pollution_Effect
1	10	5	12	9	11	7	4
2	8	4	13	8	10.5	6	4.5
3	11	3	14	10	12.5	6.5	6
4	9	4	11	11	10	7.5	2.5
5	10	2	13	12	11.5	7	4.5
Means					11.1	6.8	4.3

The mean score in the no-Pollution condition is 11.1, while in the Pollution condition, it is 6.8. The main effect of Pollution, therefore, is 4.3, which represents the difference between these two means.

To determine if this main effect of Pollution is statistically significant and not due to chance, we can perform a paired samples t-test comparing the mean no-Pollution scores to the mean Pollution scores for each subject (tank). Alternatively, a one-sample t-test on the Pollution effect scores, testing against a mean difference of zero, would yield the same conclusion.

For the paired samples t-test, we compare the mean scores of the no-Pollution condition to the Pollution condition for each tank. We use the Mean_No_Pollution and Mean_Pollution columns from our data frame:

```
#> 
#>  Welch Two Sample t-test
#> 
#> data:  as.numeric(fake_data$Mean_No_Pollution[1:5]) and as.numeric(fake_data$Mean_Polluti
#> t = 8.6, df = 6.502, p-value = 8.726e-05
#> alternative hypothesis: true difference in means is not equal to 0
#> 95 percent confidence interval:
#>  3.099079 5.500921
#> sample estimates:
#> mean of x mean of y 
#>      11.1       6.8
```

For the one-sample t-test, we test whether the mean difference (Pollution effect) is significantly different from zero:

```
#> 
#>  One Sample t-test
#> 
#> data:  as.numeric(fake_data$Pollution_Effect[1:5])
#> t = 7.6615, df = 4, p-value = 0.00156
```

	All Conditions						
	Low Flow		High Flow		Flow Means		Flow Effect
	No Pollution	Pollution	No Pollution	Pollution	Low Flow	High Flow	Difference
subject	A	B	C	D	AB	CD	CD.minus.AB
1	10	5	12	9	7.5	10.5	3
2	8	4	13	8	6	10.5	4.5
3	11	3	14	10	7	12	5
4	9	4	11	11	6.5	11	4.5
5	10	2	13	12	6	12.5	6.5
Means					6.6	11.3	4.7

```
#> alternative hypothesis: true mean is not equal to 0
#> 95 percent confidence interval:
#>  2.741724 5.858276
#> sample estimates:
#> mean of x
#>       4.3
```

If we were to write-up our results for the main effect of Pollution we could say something like this:

The main effect of Pollution was significant, $t(4) = 7.66$, $p = 0.001$. The mean biodiversity was higher in the no-Pollution condition (M = 11.1) than the Pollution condition (M = 6.8).

10.5.5 Main effect of Flow

The main effect of Flow compares the overall means for all scores in the no-Flow and Flow conditions, collapsing over the Flow conditions.

The yellow columns show the no-Flow scores for each subject (tank). The blue columns show the Flow scores for each subject.

The overall means for for each subject, for the two Flow conditions are shown to the right. For example, subject 1 had a 10 and 5 in the no-Flow condition, so their mean is 7.5.

We are interested in the main effect of Flow. This is the difference between the AB column (average of subject scores in the no-Flow condition) and the CD column (average of the subject scores in the Flow condition). These differences for each subject are shown in the last green column. The overall means, averaging over subjects are in the bottom green row.

Just looking at the means, we can see there was a main effect of Flow. The mean number of species was 11.3 in the Flow condition, and 6.6 in the no-Flow condition. So, the size of the main effect of Flow was 4.7.

Is a difference of this size likely o unlikely due to chance? We could conduct a paired-samples *t*-test on the AB vs. CD means, or a one-sample *t*-test on the difference scores. They both give the same answer:

Here's the paired samples version:

```
#>
#>  Paired t-test
#>
#> data:  CD and AB
#> t = 8.3742, df = 4, p-value = 0.001112
#> alternative hypothesis: true mean difference is not equal to 0
#> 95 percent confidence interval:
#>  3.141724 6.258276
#> sample estimates:
#> mean difference
#>             4.7
```

Here's the one sample version:

```
#>
#>  One Sample t-test
#>
#> data:  CD - AB
#> t = 8.3742, df = 4, p-value = 0.001112
#> alternative hypothesis: true mean is not equal to 0
#> 95 percent confidence interval:
#>  3.141724 6.258276
#> sample estimates:
#> mean of x
#>       4.7
```

If we were to write-up our results for the main effect of Flow we could say something like this:

The main effect of Flow was significant, t(4) = 8.37, p = 0.001. The mean number of species was higher in the Flow condition (M = 11.3) than the no-Flow condition (M = 6.6).

	All Conditions						
	Low Flow		High Flow		Pollution Effects		Interaction Effect
	No Pollution	Pollution	No Pollution	Pollution	Low Flow	High Flow	Difference
subject	A	B	C	D	A-B	C-D	(A-B)-(C-D)
1	10	5	12	9	5	3	2
2	8	4	13	8	4	5	-1
3	11	3	14	10	8	4	4
4	9	4	11	11	5	0	5
5	10	2	13	12	8	1	7
Means					6	2.6	3.4

10.5.6 Interaction between Pollution and Flow

Now we are ready to look at the interaction. Remember, the whole point of this fake study was what? Can you remember?

Here's a reminder. We wanted to know if giving Flows versus not would change the size of the Pollution effect.

Notice, neither the main effect of Pollution, or the main effect of Flow, which we just went through the process of computing, answers this question.

In order to answer the question we need to do two things. First, compute Pollution effect for each subject when they were in the no-Flow condition. Second, compute the Pollution effect for each subject when they were in the Flow condition.

Then, we can compare the two Pollution effects and see if they are different. The comparison between the two Pollution effects is what we call the **interaction effect**. Remember, this is a difference between two difference scores. We first get the difference scores for the Pollution effects in the no-Flow and Flow conditions. Then we find the difference scores between the two Pollution effects. This difference of differences is the interaction effect (green column in the table)

The mean Pollution effects in the no-Flow (6) and Flow (2.6) conditions were different. This difference is the interaction effect. The size of the interaction effect was 3.4.

How can we test whether the interaction effect was likely or unlikely due to chance? We could run another paired-sample *t*-test between the two Pollution effect measures for each subject, or a one sample *t*-test on the green column (representing the difference between the differences). Both of these *t*-tests will give the same results:

Here's the paired samples version:

```
#>
#>  Paired t-test
```

```
#>
#> data:  A_B and C_D
#> t = 2.493, df = 4, p-value = 0.06727
#> alternative hypothesis: true mean difference is not equal to 0
#> 95 percent confidence interval:
#>  -0.3865663  7.1865663
#> sample estimates:
#> mean difference
#>             3.4
```

Here's the one sample version:

```
#>
#>  One Sample t-test
#>
#> data:  A_B - C_D
#> t = 2.493, df = 4, p-value = 0.06727
#> alternative hypothesis: true mean is not equal to 0
#> 95 percent confidence interval:
#>  -0.3865663  7.1865663
#> sample estimates:
#> mean of x
#>       3.4
```

Oh look, the interaction was not significant. At least, if we had set our alpha criterion to 0.05, it would not have met that criteria. We could write up the results like this. The two-way interaction between between Pollution and Flow was not significant, $t(4) = 2.493$, $p = 0.067$.

Often times when a result is "not significant" according to the alpha criteria, the pattern among the means is not described further. One reason for this practice is that the researcher is treating the means as if they are not different (because there was an above alpha probability that the observed differences were due to chance). If they are not different, then there is no pattern to report.

There are differences in opinion among reasonable and expert statisticians on what should or should not be reported. Let's say we wanted to report the observed mean differences, we would write something like this:

The two-way interaction between between Pollution and Flow was not significant, t(4) = 2.493, p = 0.067. The mean Pollution effect in the no-Flow condition was 6 and the mean Pollution effect in the Flow condition was 2.6.

10.5.7 Writing it all up

We have completed an analysis of a 2x2 repeated measures design using paired-samples t-tests. Here is what a full write-up of the results could look like.

The main effect of Pollution was significant, $t(4) = 7.66$, $p = 0.001$. The mean number of species was higher in the no-Pollution condition (M = 11.1) than the Pollution condition (M = 6.8).

The main effect of Flow was significant, $t(4) = 8.37$, $p = 0.001$. The mean number of species was higher in the Flow condition (M = 11.3) than the no-Flow condition (M = 6.6).

The two-way interaction between between Pollution and Flow was not significant, $t(4) = 2.493$, $p = 0.067$. The mean Pollution effect in the no-Flow condition was 6 and the mean Pollution effect in the Flow condition was 2.6.

Interim Summary. We went through this exercise to show you how to break up the data into individual comparisons of interest. Generally speaking, a 2x2 repeated measures design would not be analyzed with three paired-samples t-test. This is because it is more convenient to use the repeated measures ANOVA for this task. We will do this in a moment to show you that they give the same results. And, by the same results, what we will show is that the p-values for each main effect, and the interaction, are the same. The ANOVA will give us F-values rather than t values. It turns out that in this situation, the F-values are related to the t values. In fact, $t^2 = F$.

10.5.8 2x2 Repeated Measures ANOVA

We just showed how a 2x2 repeated measures design can be analyzed using paired-sampled t-tests. We broke up the analysis into three parts. The main effect for Pollution, the main effect for Flow, and the 2-way interaction between Pollution and Flow. We claimed the results of the paired-samples t-test analysis would mirror what we would find if we conducted the analysis using an ANOVA. Let's show that the results are the same. Here are the results from the 2x2 repeated-measures ANOVA, using the `aov` function in R.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Residuals	4	3.70	0.925	NA	NA
Pollution	1	92.45	92.450	58.698413	0.0015600
Residuals1	4	6.30	1.575	NA	NA
Flow	1	110.45	110.450	70.126984	0.0011122
Residuals2	4	6.30	1.575	NA	NA
Pollution:Flow	1	14.45	14.450	6.215054	0.0672681
Residuals	4	9.30	2.325	NA	NA

Let's compare these results with the paired-samples t-tests.

Main effect of Pollution: Using the paired samples t-test, we found $t(4)$ =7.6615, p=0.00156. Using the ANOVA we found, F(1,4) = 58.69, p=0.00156. See, the p-values are the same, and $t^2 = 7.6615^2 = 58.69 = F$.

Main effect of Flow: Using the paired samples t-test, we found $t(4)$ =8.3742, p=0.001112. Using the ANOVA we found, F(1,4) = 70.126, p=0.001112. See, the p-values are the same, and $t^2 = 8.3742^2 = 70.12 = F$.

Interaction effect: Using the paired samples t-test, we found $t(4)$ =2.493, p=0.06727. Using the ANOVA we found, F(1,4) = 6.215, p=0.06727. See, the p-values are the same, and $t^2 = 2.493^2 = 6.215 = F$.

There you have it. The results from a 2x2 repeated measures ANOVA are the same as you would get if you used paired-samples t-tests for the main effects and interactions.

10.6 2x2 Between-subjects ANOVA

You must be wondering how to calculate a 2x2 ANOVA. We haven't discussed this yet. We've only shown you that you don't have to do it when the design is a 2x2 repeated measures design (note this is a special case).

We are now going to work through some examples of calculating the ANOVA table for 2x2 designs. We will start with the between-subjects ANOVA for 2x2 designs. We do essentially the same thing that we did before (in the other ANOVAs), and the only new thing is to show how to compute the interaction effect.

Remember the logic of the ANOVA is to partition the variance into different parts. The SS formula for the between-subjects 2x2 ANOVA looks like this:

$SS_{\text{Total}} = SS_{\text{Effect IV1}} + SS_{\text{Effect IV2}} + SS_{\text{Effect IV1xIV2}} + SS_{\text{Error}}$

In the following sections we use tables to show the calculation of each SS. We use the same example as before with the exception that **we are turning this into a between-subjects design**. There are now 5 different subjects in each condition, for a total of 20 subjects. As a result, we remove the subjects column.

10.6.1 SS Total

We calculate the grand mean (mean of all of the score). Then, we calculate the differences between each score and the grand mean. We square the difference scores, and sum them up. That is SS_{Total}, reported in the bottom yellow row.

	All Conditions				Difference from Grand Mea		
	Low Flow		High Flow		Low Flow		Hig
	No Pollution	Pollution	No Pollution	Pollution	No Pollution	Pollution	No Pollutio
	A	B	C	D	A-GrandM	B-GrandM	C-GrandM
	10	5	12	9	1.05	-3.95	3.05
	8	4	13	8	-0.95	-4.95	4.05
	11	3	14	10	2.05	-5.95	5.05
	9	4	11	11	0.05	-4.95	2.05
	10	2	13	12	1.05	-6.95	4.05
Means	9.6	3.6	12.6	10			
Grand Mean	8.95						
sums							
SS Total							

	All Conditions				Pollution Mean - G		
	Low Flow		High Flow		Low Flow		
	No Pollution	Pollution	No Pollution	Pollution	No Pollution	Pollution	No Poll
	A	B	C	D	NDM-GM A	DM-GM B	NDM-G
	10	5	12	9	2.15	-2.15	2.15
	8	4	13	8	2.15	-2.15	2.15
	11	3	14	10	2.15	-2.15	2.15
	9	4	11	11	2.15	-2.15	2.15
	10	2	13	12	2.15	-2.15	2.15
Means	9.6	3.6	12.6	10			
Grand Mean	8.95	No Pollution	11.1	Pollution	6.8		
sums							
SS Pollution							

10.6.2 SS Pollution

We need to compute the SS for the main effect for Pollution. We calculate the grand mean (mean of all of the scores). Then, we calculate the means for the two Pollution conditions. Then we treat each score as if it was the mean for it's respective Pollution condition. We find the differences between each Pollution condition mean and the grand mean. Then we square the differences and sum them up. That is $SS_{\text{Pollution}}$, reported in the bottom yellow row.

These tables are a lot to look at! Notice here, that we first found the grand mean (8.95). Then we found the mean for all the scores in the no-Pollution condition (columns A and C), that was 11.1. All of the difference scores for the no-Pollution condition are 11.1-8.95 = 2.15. We also found the mean for the scores in the Pollution condition (columns B and D), that was 6.8. So, all of the difference scores are 6.8-8.95 = -2.15. Remember, means are the balancing point in the data, this is why the difference scores are +2.15 and -2.15. The grand mean 8.95

	All Conditions				Flow Mean - GM		
	Low Flow		High Flow		Low Flow		
	No Pollution	Pollution	No Pollution	Pollution	No Pollution	Pollution	No Pollu
	A	B	C	D	NRM-GM A	NRM-GM B	RM-GM
	10	5	12	9	-2.35	-2.35	2.35
	8	4	13	8	-2.35	-2.35	2.35
	11	3	14	10	-2.35	-2.35	2.35
	9	4	11	11	-2.35	-2.35	2.35
	10	2	13	12	-2.35	-2.35	2.35
Means	9.6	3.6	12.6	10			
Grand Mean	8.95	Low Flow	6.6	High Flow	11.3		
sums							
SS Flow							

is in between the two condition means (11.1 and 6.8), by a difference of 2.15.

10.6.3 SS Flow

We need to compute the SS for the main effect for Flow. We calculate the grand mean (mean of all of the scores). Then, we calculate the means for the two Flow conditions. Then we treat each score as if it was the mean for it's respective Flow condition. We find the differences between each Flow condition mean and the grand mean. Then we square the differences and sum them up. That is SS_{Flow}, reported in the bottom yellow row.

Now we treat each no-Flow score as the mean for the no-Flow condition (6.6), and subtract it from the grand mean (8.95), to get -2.35. Then, we treat each Flow score as the mean for the Flow condition (11.3), and subtract it from the grand mean (8.95), to get +2.35. Then we square the differences and sum them up.

10.6.4 SS Pollution by Flow

We need to compute the SS for the interaction effect between Pollution and Flow. This is the new thing that we do in an ANOVA with more than one IV. How do we calculate the variation explained by the interaction?

The heart of the question is something like this. Do the individual means for each of the four conditions do something a little bit different than the group means for both of the independent variables?

For example, consider the overall mean for all of the scores in the Low Flow group, we found that to be 6.6 Now, was the mean for each no-Flow group in the whole design a 6.6? For example, in the no-Pollution group, was the mean for column A (the no-Flow condition in that group) also 6.6? The answer is no, it was 9.6. How about the Pollution group? Was the mean for the Flow condition in the Pollution group (column B) 6.6? No, it was 3.6. The mean of 9.6 and 3.6 is 6.6. If there was no hint of an interaction, we would expect that the means for the Flow condition in both levels of the Pollution group would be the same, they would both be 6.6. However, when there is an interaction, the means for the Flow group will depend on the levels of the group from another IV. In this case, it looks like there is an interaction because the means are different from 6.6, they are 9.6 and 3.6 for the no-Pollution and Pollution conditions. This is extra-variance that is not explained by the mean for the Flow condition. We want to capture this extra variance and sum it up. Then we will have measure of the portion of the variance that is due to the interaction between the Flow and Pollution conditions.

What we will do is this. We will find the four condition means. Then we will see how much additional variation they explain beyond the group means for Flow and Pollution. To do this we treat each score as the condition mean for that score. Then we subtract the mean for the Pollution group, and the mean for the Flow group, and then we add the grand mean. This gives us the unique variation that is due to the interaction. We could also say that we are subtracting each condition mean from the grand mean, and then adding back in the Pollution mean and the Flow mean, that would amount to the same thing, and perhaps make more sense.

Here is a formula to describe the process for each score:

$\bar{X}_{\text{condition}} - \bar{X}_{\text{IV1}} - \bar{X}_{\text{IV2}} + \bar{X}_{\text{Grand Mean}}$

Or we could write it this way:

$\bar{X}_{\text{condition}} - \bar{X}_{\text{Grand Mean}} + \bar{X}_{\text{IV1}} + \bar{X}_{\text{IV2}}$

When you look at the following table, we apply this formula to the calculation of each of the differences scores. We then square the difference scores, and sum them up to get $SS_{\text{Interaction}}$, which is reported in the bottom yellow row.

10.6.5 SS Error

The last thing we need to find is the SS Error. We can solve for that because we found everything else in this formula:

$SS_{\text{Total}} = SS_{\text{Effect IV1}} + SS_{\text{Effect IV2}} + SS_{\text{Effect IV1xIV2}} + SS_{\text{Error}}$

Even though this textbook meant to explain things in a step by step way, we guess you are tired from watching us work out the 2x2 ANOVA by hand. You and me both, making these tables was a lot of work. We have already shown you how to compute the SS for error before, so we will not do the full example here. Instead, we solve for SS Error using the numbers we have already obtained.

	All Conditions				Interaction Differ		
	Low Flow		High Flow		Low Flow		
	No Pollution	Pollution	No Pollution	Pollution	No Pollution	Pollution	No
	A	B	C	D	A-NP-LF+GM	B-P-LF+GM	C-NP
	10	5	12	9	0.85	-0.85	-0.85
	8	4	13	8	0.85	-0.85	-0.85
	11	3	14	10	0.85	-0.85	-0.85
	9	4	11	11	0.85	-0.85	-0.85
	10	2	13	12	0.85	-0.85	-0.85
Means	9.6	3.6	12.6	10			
Grand Mean	8.95						
sums							
SS Interaction							

$SS_Error = SS_Total- SS_Effect IV1 - SS_Effect IV2 - SS_Effect IV1xIV2 $

$SS_Error = 242.95 - 92.45 - 110.45 - 14.45 = 25.6 $

10.6.6 Check your work

We are going to skip the part where we divide the SSes by their dfs to find the MSEs so that we can compute the three F-values. Instead, if we have done the calculations of the SSes correctly, they should be same as what we would get if we used R to calculate the SSes. Let's make R do the work, and then compare to check our work.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Pollution	1	92.45	92.45	57.78125	0.0000011
Flow	1	110.45	110.45	69.03125	0.0000003
Pollution:Flow	1	14.45	14.45	9.03125	0.0083879
Residuals	16	25.60	1.60	NA	NA

A quick look through the column `Sum Sq` shows that we did our work by hand correctly. Congratulations to us! Note, this is not the same results as we had before with the repeated measures ANOVA. We conducted a between-subjects design, so we did not get to further partition the SS error into a part due to subject variation and a left-over part. We also gained degrees of freedom in the error term. It turns out with this specific set of data, we find p-values of less than 0.05 for all effects (main effects and the interaction, which was not less than 0.05 using the same data, but treating it as a repeated-measures design)

10.7 Fireside chat

Sometimes it's good to get together around a fire and have a chat. Let's pretend we're sitting around a fire.

It's been a long day. A long couple of weeks and months since we started this course on statistics. We just went through the most complicated things we have done so far. This is a long chapter. What should we do next?

Here's a couple of options. We could work through, by hand, more and more ANOVAs. Do you want to do that? I don't, making these tables isn't too bad, but it takes a lot of time. It's really good to see everything that we do laid bare in the table form a few times. We've done that already. It's really good for you to attempt to calculate an ANOVA by hand at least once in your life. It builds character. It helps you know that you know what you are doing, and what the ANOVA is doing. We can't make you do this, we can only make the suggestion. If we keep doing these by hand, it is not good for us, and it is not you doing them by hand. So, what are the other options.

The other options are to work at a slightly higher level. We will discuss some research designs, and the ANOVAs that are appropriate for their analysis. We will conduct the ANOVAs using R, and print out the ANOVA tables. This is what you do in the lab, and what most researchers do. They use software most of the time to make the computer do the work. Because of this, it is most important that you know what the software is doing. You can make mistakes when telling software what to do, so you need to be able to check the software's work so you know when the software is giving you wrong answers. All of these skills are built up over time through the process of analyzing different data sets. So, for the remainder of our discussion on ANOVAs we stick to that higher level. No more monster tables of SSes. You are welcome.

10.8 Factorial summary

We have introduced you to factorial designs, which are simply designs with more than one IV. The special property of factorial designs is that all of the levels of each IV need to be crossed with the other IVs.

We showed you how to analyse a repeated measures 2x2 design with paired samples-tests, and what an ANOVA table would look like if you did this in R. We also went through, by hand, the task of calculating an ANOVA table for a 2x2 between subjects design.

The main point we want you take away is that factorial designs are extremely useful for determining things that cause effects to change. Generally a researcher measures an effect of interest (their IV 1). Then, they want to know what makes that effect get bigger or smaller. They want to exert experimental control over their effect. For example, they might have a theory that says doing X should make the effect bigger, but doing Y should make it smaller.

They can test these theories using factorial designs, and manipulating X or Y as a second independent variable.

In a factorial design each IV will have it's own main effect. Sometimes the main effect themselves are what the researcher is interested in measures. But more often, it is the interaction effect that is most relevant. The interaction can test whether the effect of IV1 changes between the levels of IV2. When it does, researchers can infer that their second manipulation (IV2) causes change in their effect of interest. These changes are then documented and used to test underlying causal theories about the effects of interest.

11 More On Factorial Designs

Portions adapted from the Factorial ANOVA chapter, contributors Keryn Bain, Rachel Blakey, Stephanie Brodie, Corey Callaghan, Will Cornwell, Kingsley Griffin, Matt Holland, James Lavender, Andrew Letten, Shinichi Nakagawa, Shaun Nielsen, Alistair Poore, Gordana Popovic, Fiona Robinson and Jakub Stoklosa. "Environmental Computing" https://environmentalcomputing.net/

In this chapter, we're diving deeper into factorial designs, a cornerstone of understanding complex data in environmental science. You're already familiar with the idea of having more than one independent variable (IV) in your experiments. These IVs can be structured in various ways: all between-subjects, all within-subjects (like repeated measures), or a mix of both. ANOVA is our trusty tool to analyze these designs, giving us insights into each IV's main effect and their interactions.

11.1 Looking at main effects and interactions

Factorial designs are very common in environmental research. You'll often come across studies showing results from these designs. It's crucial to be comfortable interpreting these results. The key skill here is to recognize patterns of main effects and interactions in data graphs. This can get tricky with more than two IVs, each having multiple levels.

11.1.1 2x2 designs

Let's explore 2x2 designs. Here, you can expect two main effects and one interaction. You'll compare means for each main effect and interaction. There are eight possible outcomes in such a design:

1. no IV1 main effect, no IV2 main effect, no interaction
2. IV1 main effect, no IV2 main effect, no interaction
3. IV1 main effect, no IV2 main effect, interaction
4. IV1 main effect, IV2 main effect, no interaction
5. IV1 main effect, IV2 main effect, interaction

6. no IV1 main effect, IV2 main effect, no interaction
7. no IV1 main effect, IV2 main effect, interaction
8. no IV1 main effect, no IV2 main effect, interaction

OK, so if you run a 2x2, any of these 8 general patterns could occur in your data. That's a lot to keep track of isn't it? As you develop your skills in examining graphs that plot means, you should be able to look at the graph and visually guesstimate if there is, or is not, a main effect or interaction. You will need you inferential statistics to tell you for sure, but it is worth knowing how to know see the patterns.

Let's visualize these outcomes using R. We'll create bar and line graphs to illustrate these patterns. Bar graphs are great for seeing differences in means directly, while line graphs help us spot interactions – look for crossing lines as a hint of interaction. Figure 11.1 shows the possible patterns of main effects and interactions in bar graph form. Here is a legend for the labels in the panels.

- 1 = there was a main effect for IV1.
- ~1 = there was **not** a main effect for IV1
- 2 = there was a main effect for IV2
- ~2 = there was **not** a main effect of IV2
- 1x2 = there was an interaction
- ~1x2 = there was **not** an interaction

Figure 11.2 shows the same eight patterns in line graph form:

In line graphs, interactions are more apparent. Parallel lines suggest no interaction, while crossing lines indicate potential interactions. The position of points relative to each other helps identify main effects. Things get complicated fast. When designing experiments, aim for the minimum number of independent variables (IVs) and levels needed to answer your research question. This approach makes interpreting your data more straightforward and your conclusions clearer. Whenever you see that someone ran a 4x3x7x2 design, your head should spin. It's just too complicated.

11.2 Interpreting main effects and interactions

Understanding main effects and interactions is essential for accurately interpreting research data, especially in complex fields like environmental science.

A **main effect** refers to the consistent impact of an independent variable (IV) on a dependent variable (DV). For example, in environmental studies, consider the effect of a specific fertilizer (IV) on plant growth (DV). If using this fertilizer consistently results in increased growth compared to not using it, we observe a clear main effect. This effect remains true regardless of other variables such as soil type or weather conditions.

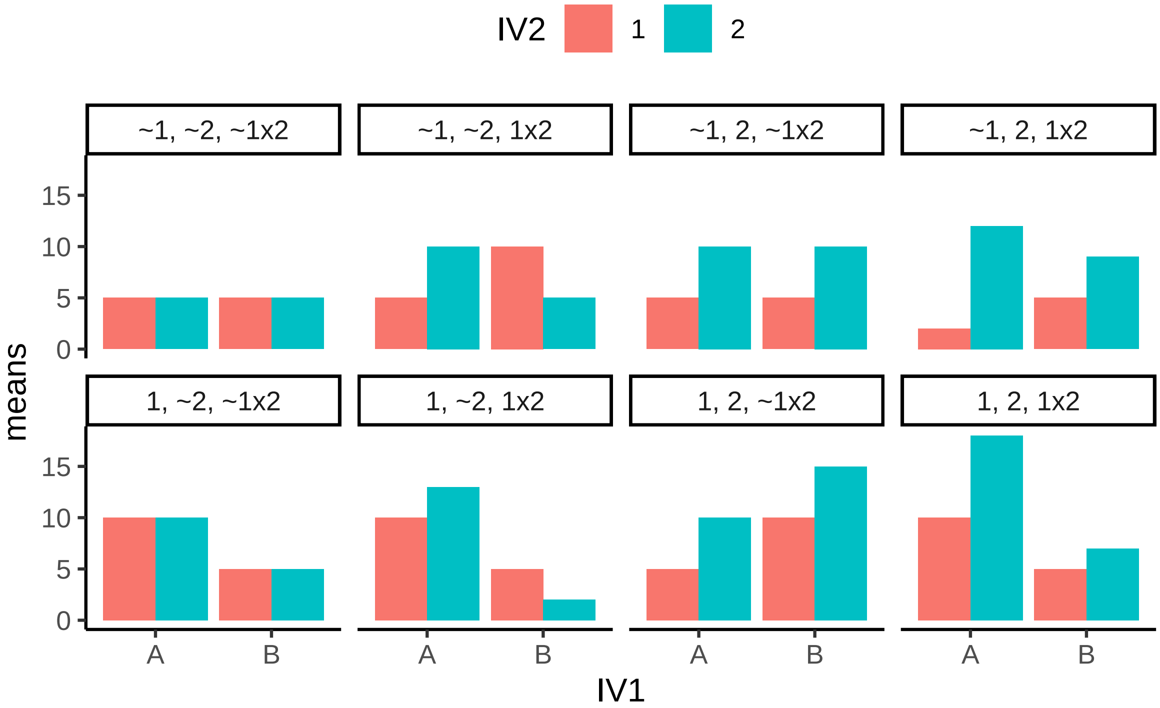


Figure 11.1: 8 Example patterns for means for each of the possible kinds of general outcomes in a 2x2 design.

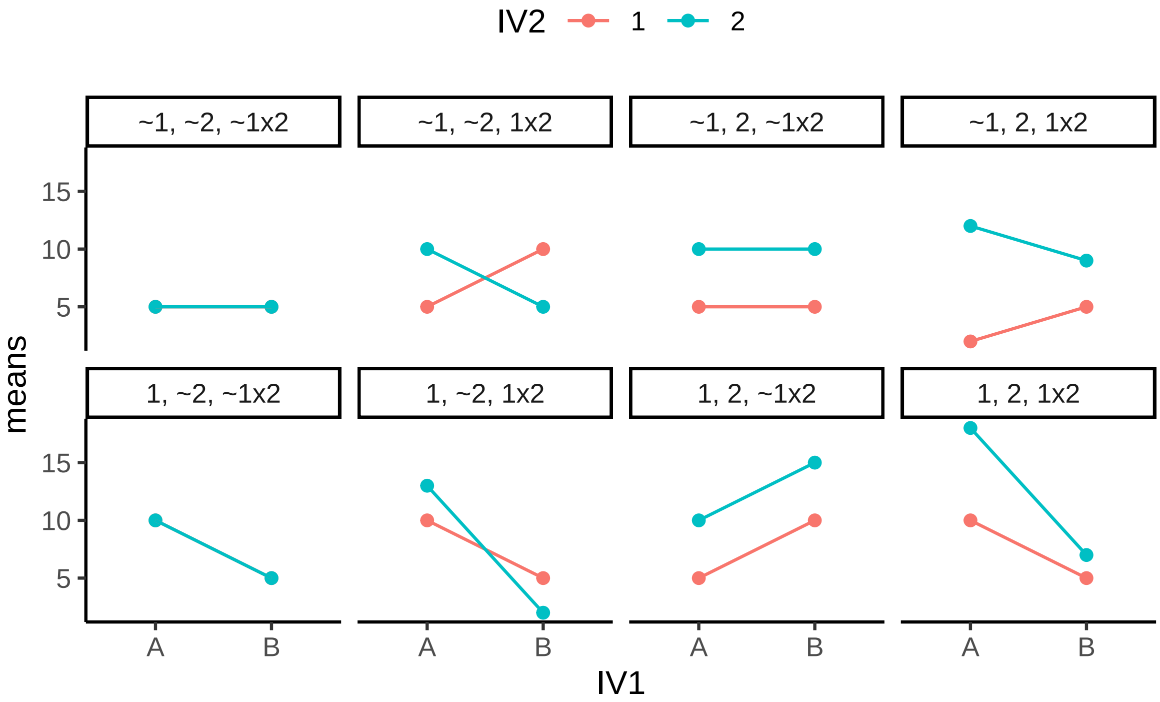


Figure 11.2: Line graphs showing 8 possible general outcomes for a 2x2 design.

Often, it is convenient to think of main effects as a consistent influence of one manipulation. However, the picture changes when we introduce an interaction. An interaction occurs when the effect of one IV depends on another IV. By definition, an interactino means that some main effect is **not** behaving consistently across different situations. For instance, the impact of our fertilizer might vary depending on the level of sunlight or the type of soil, indicating an interaction between these factors and the fertilizer. This interaction disrupts the consistency of the main effect, suggesting that the effect of the fertilizer is not uniform across all conditions.

Researchers often phrase their findings to highlight this complexity: "We found a main effect of X, **BUT**, this main effect was qualified by an interaction between X and Y." The use of "BUT" here is crucial. It signals that the main effect cannot be fully understood without considering the interaction. The interaction indicates that the influence of the IV changes under different conditions, making it essential to consider these variables together for a complete understanding.

In environmental science, this becomes particularly relevant when studying ecosystems or climate interactions, where multiple variables interplay in complex ways. The interpretation of main effects and interactions in such contexts is not just about identifying individual effects but understanding how these effects change in different environmental settings.

Here are two generalized examples to help you make sense of these issues:

11.2.1 A consistent main effect and an interaction

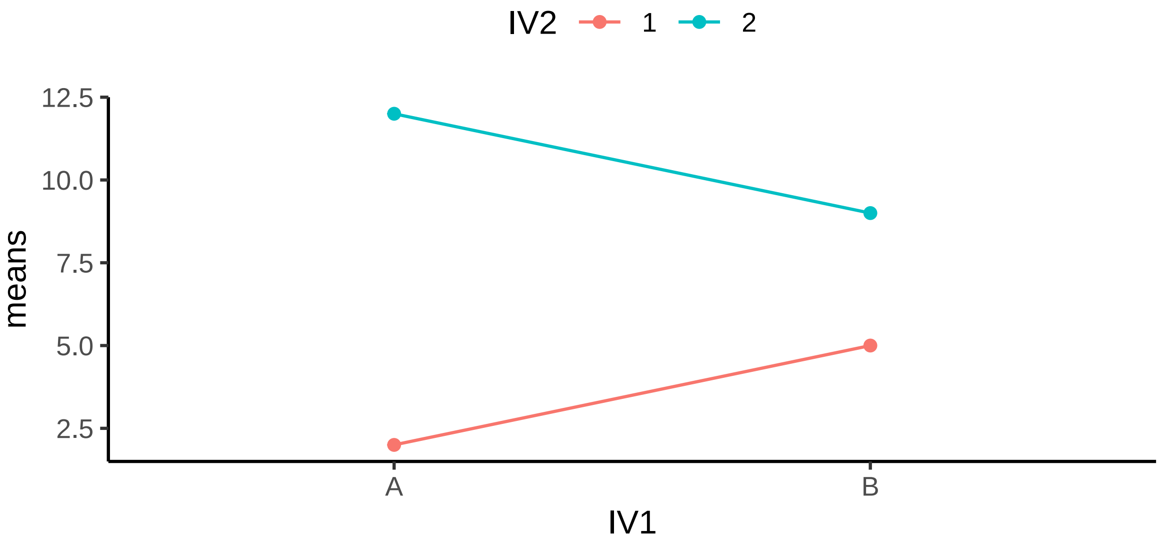


Figure 11.3: Example means showing a generally consistent main effect along with an interaction

Figure 11.3 shows a main effect and interaction. There is a main effect of IV2: the level 1 means (red points and line) are both lower than the level 2 means (aqua points and line). There is also an interaction. The size of the difference between the red and aqua points in the A condition (left) is bigger than the size of the difference in the B condition.

How would we interpret this? We could say there WAS a main effect of IV2, BUT it was qualified by an IV1 x IV2 interaction.

What's the qualification? The size of the IV2 effect changed as a function of the levels of IV1. It was big for level A, and small for level B of IV1.

What does the qualification mean for the main effect? Well, first it means the main effect can be changed by the other IV. That's important to know. Does it also mean that the main effect is not a real main effect because there was an interaction? Not really, there is a generally consistent effect of IV2. The green points are above the red points in all cases. Whatever IV2 is doing, it seems to work in at least a couple situations, even if the other IV also causes some change to the influence.

11.2.2 An inconsistent main effect and an interaction

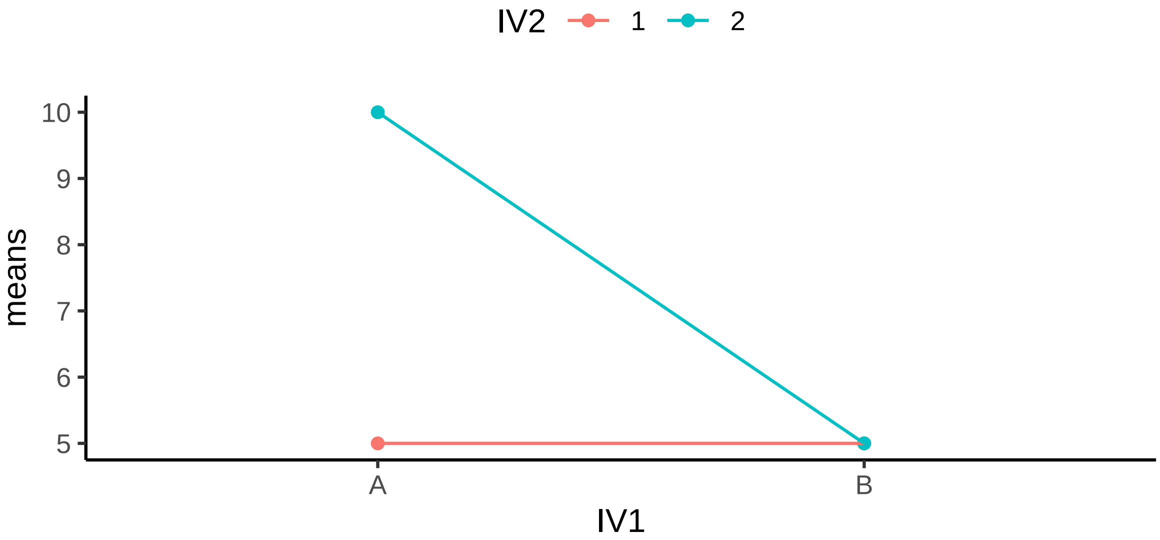


Figure 11.4: Example data showing how an interaction exists, and a main effect does not, even though the means for the main effect may show a difference

Figure 11.4 shows another 2x2 design. You should see an interaction here straight away. The difference between the aqua and red points in condition A (left two dots) is huge, and there is 0 difference between them in condition B. Is there an interaction? Yes!

Are there any main effects here? With data like this, sometimes an ANOVA will suggest that you do have significant main effects. For example, what is the mean difference between level 1 and 2 of IV2? That is the average of the green points ((10+5)/2 = 15/2= 7.5) compared to the average of the red points (5). There will be a difference of 2.5 for the main effect (7.5 vs. 5).

Starting to see the issue here? From the perspective of the main effect (which collapses over everything and ignores the interaction), there is an overall effect of 2.5. In other words, level 2 adds 2.5 in general compared to level 1. However, we can see from the graph that IV2 does not do anything in general. It does not add 2.5s everywhere. It adds 5 in condition A, and nothing in condition B. It only does one thing in one condition.

What is happening here is that a "main effect" is produced by the process of averaging over a clear interaction.

How would we interpret this? We might have to say there was a main effect of IV2, BUT we would definitely say it was qualified by an IV1 x IV2 interaction.

What's the qualification? The size of the IV2 effect completely changes as a function of the levels of IV1. It was big for level A, and nonexistent for level B of IV1.

What does the qualification mean for the main effect? In this case, we might doubt whether there is a main effect of IV2 at all. It could turn out that IV2 does not have a general influence over the DV all of the time, it may only do something in very specific circumstances, in combination with the presence of other factors.

11.3 Mixed Designs

In this book, we've explored various research designs, emphasizing that they can take different forms. These designs can be categorized as either between-subjects, where different subjects are in each group, or within-subjects, where the same subjects participate in all conditions. When you combine these approaches in a single study, you create what's known as a mixed design.

A mixed design occurs when one of your independent variables (IVs) is treated as a between-subjects factor, while another is treated as a within-subjects factor. This blend offers a unique approach to examining how different variables interact and affect the outcome.

In environmental science research, mixed designs are particularly useful for studying complex interactions between variables that vary both within and between subjects. For instance, consider a study examining the impact of a new agricultural technique (IV1) on crop yield (DV). This technique could be applied to different plots of land (between-subjects factor), while also measuring the impact across different seasons (within-subjects factor). Such a design allows researchers to understand not only the overall effectiveness of the technique but also how its impact varies seasonally.

The key to successfully navigating mixed designs lies in understanding how to calculate the appropriate statistical measures. Specifically, the F-values for each effect in an ANOVA (Analysis of Variance) are constructed using different error terms, depending on whether the IV is a within-subjects or between-subjects factor. While it's possible to run an ANOVA with any combination of between and within-subjects IVs, the complexity increases with the number of variables and their categorizations.

As this is an introductory text, we won't delve into the detailed formulas for constructing ANOVA tables with mixed designs. More advanced textbooks offer comprehensive discussions on this topic, and many resources are available online for those interested in deeper exploration.

11.4 More complicated designs

Up until now we have focused on the simplest case for factorial designs, the 2x2 design, with two IVs, each with 2 levels. It is worth spending some time looking at a few more complicated designs and how to interpret them.

11.4.1 3x2 design

In a 3x2 design there are two IVs. IV1 has three levels, and IV2 has two levels. Typically, there would be one DV. Let's apply this to an environmental science scenario.First, let's make the design concrete.

Imagine a study examining the impact of different irrigation methods (IV1: drip irrigation vs. sprinkler irrigation) on crop yield (DV) across three types of soil (IV2: sandy, loamy, clayey). The main effects would be the overall impact of irrigation method and soil type on crop yield, while the interaction would explore how these effects vary together.

For instance, drip irrigation might consistently produce higher yields than sprinkler irrigation, showing a main effect of IV1. Soil type might also independently affect yield, with loamy soil perhaps leading to the highest yields, followed by clayey and sandy soils, indicating a main effect of IV2. An interaction would occur if, for example, the advantage of drip irrigation over sprinkler irrigation is more pronounced in sandy soil compared to clayey soil. Note that these examples are hypothetical to illustrate the concept.

The factorial ANOVA will test:

- whether there are any differences in crop yield among the three levels of soil type
- whether there are any differences in crop yield among the two levels of irrigation
- whether there is any interaction between irrigation type and soil type

You have three null hypotheses:

- There is no difference between the means for each level of soil type:

H$_0$: $\mu_{Clay} = \mu_{Loam} = \mu_{Sand}$

- There is no difference between the means for each level of irrigation:

H$_0$: $\mu_{Drip} = \mu_{Sprinkler}$

- There is no interaction between the factors.

Remember, this is far better than running two separate single factor ANOVAs that contrast irrigation effects for each level of soil type because you have more statistical power (higher degrees of freedom) for the tests of interest, and you get a formal test of the interaction between factors which is often scientifically interesting.

We might expect data like shown in Figure 11.5:

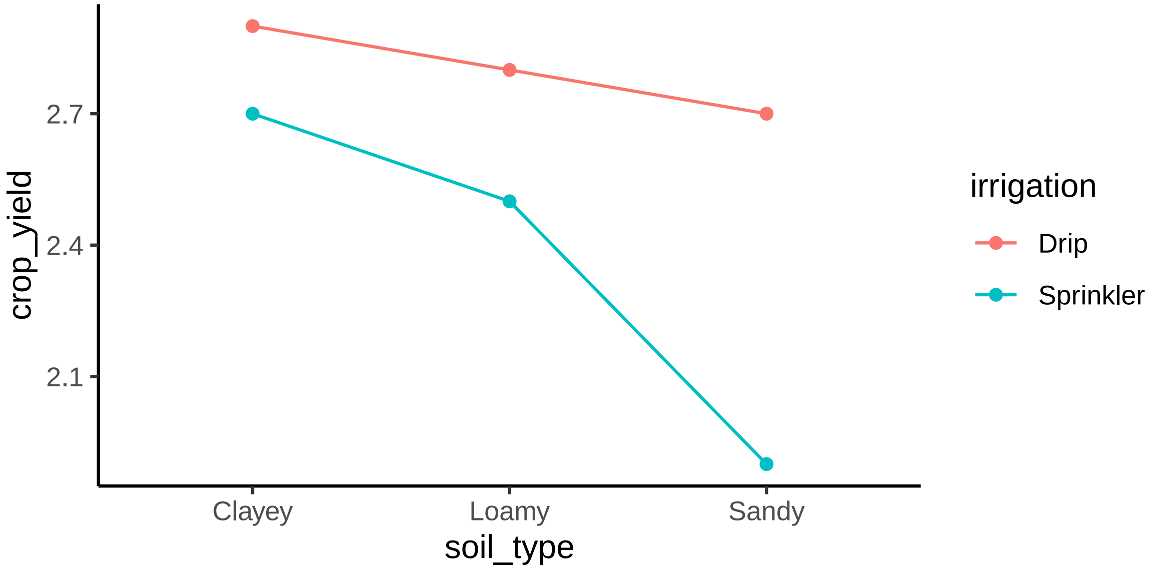


Figure 11.5: Example means for a 3x2 factorial design in environmental science

The figure shows some pretend means in all conditions. Let's talk about the main effects and interaction.

1. **Main Effect of Irrigation Method**: The main effect of the irrigation method is evident. Drip irrigation (represented by red line) generally leads to higher crop yields compared to sprinkler irrigation (represented by aqua line).

2. **Main Effect of Soil Type**: The main effect of soil type is clearly present. Clayey soils show the highest yield, followed by loamy soils, then sandy soils
3. **Interaction Between Irrigation Method and Soil Type**: Is there an interaction? Yes, there is. Remember, an interaction occurs when the effect of one IV depends on the levels of an another. The advantage of drip irrigation over sprinkler irrigation is more pronounced in sandy soil compared to clayey soil. So, the size of the irrigation effect (drip vs. sprinkler) changes with the type of soil. There is evidence in the means for an interaction. You would have to conduct an inferential test on the interaction term to see if these differences were likely or unlikely to be due to sampling error.

If there was no interaction and no main effect of soil type, we would see something like the pattern in Figure 11.6.

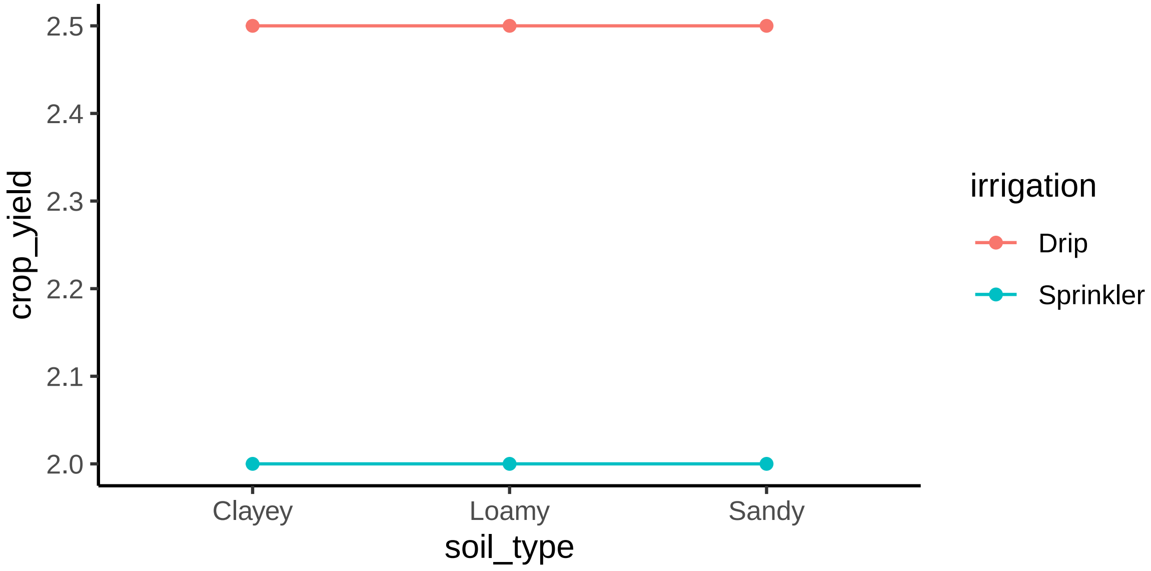


Figure 11.6: Example means for a 3x2 design in environmental science with only one main effect

What would you say about the interaction if you saw the pattern in Figure 11.7?

The correct answer is that there is evidence in the means for an interaction. Remember, we are measuring the irrigation effect (effect of drip vs. sprinkler) three times. The irrigation effect is the same for clayey and loamy soils, but it is much smaller for sandy soils. The size of the irrigation effect depends on the levels of the soil type IV, so here again there is an interaction.

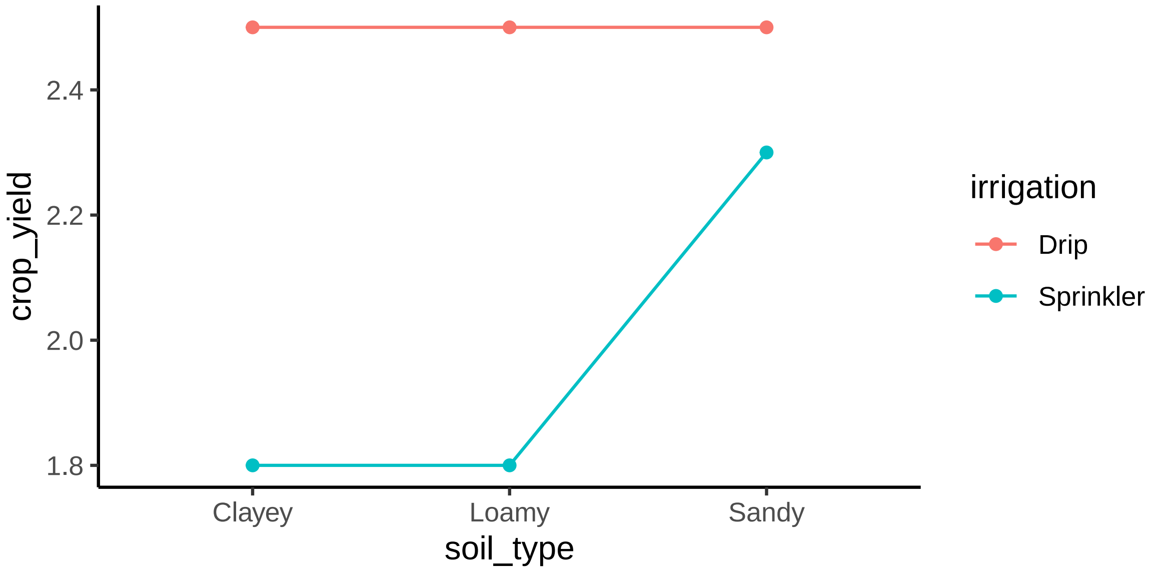


Figure 11.7: Example means for a 3x2 design in environmental science showing a different interaction pattern

11.4.2 2x2x2 designs

Let's take it up a notch and look at a 2x2x2 design. In a 2x2x2 design, there are three independent variables (IVs), each with two levels. This design allows for the examination of three main effects, three two-way interactions, and one three-way interaction.

We'll add another independent variable to our example from before: crop type (wheat vs. corn) as our third IV. So overall, in this 2x2x2 design, we'll consider three independent variables (IVs): irrigation method (IV1: drip vs. sprinkler), soil type (IV2: sandy vs. clayey), and crop type (IV3: wheat vs. corn). The dependent variable (DV) is still crop yield. This design helps us understand not just individual effects but also how these factors interact in various combinations.

n Figure 11.8, we have two panels: one for corn and one for wheat. You can think of the 2x2x2 as two 2x2 designs, one for each crop type. The key takeaway? Both wheat and corn show similar patterns, indicating a 2x2 interaction between irrigation method and soil type. We observe main effects for irrigation and soil type, but no main effect for crop type, and importantly, no three-way interaction.

But what exactly is a three-way interaction? It occurs when the pattern of a 2x2 interaction differs across the levels of the third variable. Let's visualize this with Figure 11.9.

We are looking at a 3-way interaction between irrigation type, crop type, and soil type. What is going on here?

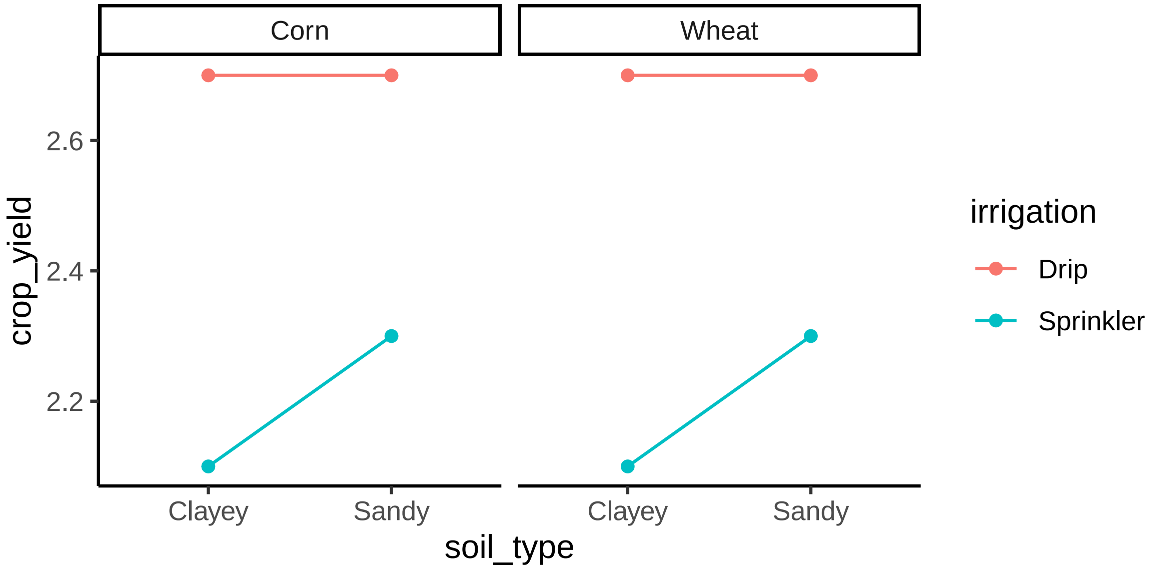


Figure 11.8: Example means from a 2x2x2 design in environmental science with no three-way interaction.

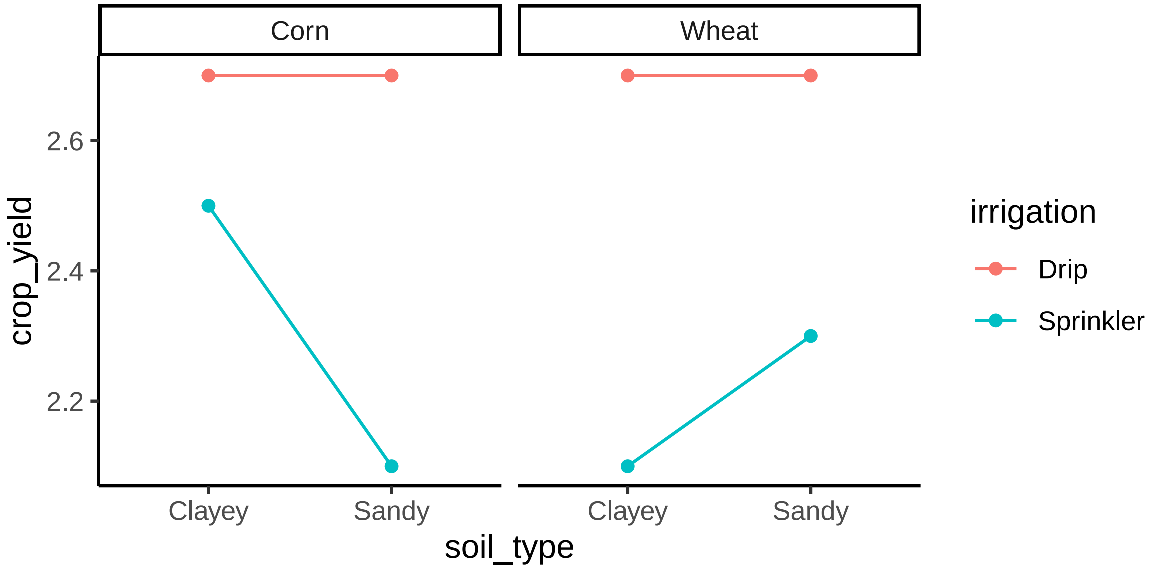


Figure 11.9: Example means from a 2x2x2 design in environmental science with a three-way interaction.

For corn crop yields, we see that there is a smaller irrigation effect in clayey soils, but the effect of irrigation gets bigger in sandy soils. A pattern like this might make sense, sandy soils don't retain much water so the irrigation method might matter more.

The wheat crop yields show a different pattern. Here, the irrigation effect is large in clayey soils and smaller in sandy soils. This difference in patterns between corn and wheat yields indicates a three-way interaction among irrigation type, soil type, and crop type.

In other words, the 2x2 interaction for the corn is **different** from the 2x2 interaction for the wheat. This can be conceptualized as an interaction between the two interactions, and as a result there is a three-way interaction, called a 2x2x2 interaction.

A general pattern here. Imagine you had a 2x2x2x2 design. That would have a 4-way interaction. What would that mean? It would mean that the pattern of the 2x2x2 interaction changes across the levels of the 4th IV. If two three-way interactions are different, then there is a four-way interaction.This becomes very complicated very quickly, another reminder of why simplicity in design is desirable.

11.4.3 Understanding and Interpreting Interactions in Environmental Science

So, you've got a handle on what interactions are and what they might look like. But there's still a key question hanging in the air: **Why do interactions matter?**

11.4.3.1 Interpreting the Results

Remember our example exploring the impact of different irrigation methods on crop yield across various soil types? The data in Figure 11.5 revealed something interesting: drip irrigation significantly boosts crop yield in sandy soil, but this effect diminishes in loamy and clayey soils. This is what we call an interaction effect: the impact of the irrigation method (drip vs. sprinkler) on crop yield varies depending on the soil type.

This interaction is important. It tells us that the effectiveness of an irrigation method is not uniform across all soil types. It suggests that environmental factors (like soil type) can influence how well an intervention (like irrigation method) works.

11.4.3.2 Practical Implications in Environmental Science

Understanding these interactions has real-world implications for environmental management and policy-making. Consider these examples:

1. **Resource Allocation**: By understanding how different soil types interact with various irrigation methods, farmers can tailor their agricultural practices more precisely. For instance, in areas with clayey soil, which retains water well, less frequent irrigation might be more suitable, reducing water usage and preserving natural resources.

2. **Climate Adaptation Strategies**: Understanding these interactions could play a role in developing climate adaptation strategies. For regions facing increased rainfall variability due to climate change, selecting the right combination of soil management and irrigation techniques can help in maintaining crop yields despite changing weather patterns.

3. **Policy Formulation**: Insights from these interactions can guide the creation of more nuanced agricultural policies. For example, providing subsidies or incentives for adopting certain irrigation methods in specific soil types could optimize crop yield and promote sustainability.

Think about it: are there other environmental factors where understanding interactions could be crucial for effective management and policy-making?

12 Analysis of Covariance (ANCOVA)

12.1 General Linear Models (GLM)

General Linear Models (GLMs) are a class of models that encompass various types of statistical analyses, including ANOVA, and regression. They're flexible enough to handle different types of data and relationships:

In our journey through statistical modeling, we've encountered three primary parametric models, each suited for different types of data scenarios:

- **No Groups and No Relationships (H0):**
  - This scenario often emerges when our ANOVA or regression analysis yields non-significant results.
  - **What We Report:** The grand mean and overall variance or standard deviation. Alternatively, we can use a confidence interval to encapsulate this information.
- **Two or More Categories with Significantly Different Means (t-test, ANOVA):**
  - Here, we delve into data where group means are distinct and significant.
  - **What We Report:** Group means. In classical ANOVA, we might report a common variance or standard deviation, or opt for group-specific measures (potentially through confidence intervals), ensuring to note that variances are not significantly different. With Welch's t-test, we focus on group-specific variance or standard deviation.
- **Two Continuous Variables with a Significant Linear or Monotonic Relationship (Regression):**
  - This model applies when there's a significant linear relationship between two continuous variables.
  - **What We Report:** The equation of the regression line, along with its confidence limits. We can also select specific x-values and provide confidence intervals for the predicted y-values at those points.

12.1.1 Model Statements:

- **No Groups/Relationships:** $y \sim N(\mu, \sigma^2)$
- **Categories with Different Means:** $y_{i,j} = \mu + \tau_i + \epsilon$ (t-test allows for unequal variance: $y_{i,j} = \mu + \tau_i + \epsilon_i$)
- **Linear Relationship:** $y = \beta_0 + \beta_1 x + \epsilon$

12.1.2 Flexibility and Applicability of GLMs

Ms enable us to model more complex relationships beyond the basic categorical X with continuous Y (as in t-tests and ANOVA) or continuous X with continuous Y (as in regression). They can handle diverse data types and model relationships that aren't strictly linear, making them crucial for studying environmental systems.

So far, in our exploration of parametric tests, we have primarily focused on two types of relationships:

- **A categorical X and a continuous Y:** This includes tests like the t-test and ANOVA.
- **A continuous X and a continuous Y:** This is typically analyzed using linear regression.

General Linear Models (GLMs) significantly expand our analytically capabilities. They are not limited to just these basic scenarios but offer a more versatile toolkit. GLMs are particularly adept at handling:

- **More Complex Relationships:** They can model scenarios where relationships between variables are not strictly linear.
- **Diverse Data Types:** GLMs are suitable for various data types, including count data, which is often encountered in environmental studies.

12.1.3 Understanding ANOVA as a linear model:

ANOVA, or Analysis of Variance, is traditionally viewed as a technique to compare means across multiple groups. However, at its core, ANOVA is a linear model. It's a special case where the predictors are categorical, not continuous. This distinction is important but subtle. Let's delve into an example that illustrates this concept clearly.

This example should be familiar from lecture. Imagine a researcher is exploring the effects of metal contamination on the species richness of sessile marine invertebrates. They're particularly interested in the impact of copper and the orientation of the substrate on which these organisms live. To investigate this, they conduct a factorial experiment, measuring species richness across different levels of copper enrichment (None, Low, High) and substrate orientation (Vertical, Horizontal). This setup allows us to not only consider each factor separately but also examine their potential interaction.

The ANOVA framework provides three null hypotheses to test:

1. There are no differences in species richness across the copper levels.
2. There are no differences in species richness between substrate orientations.
3. There is no interaction effect between copper levels and substrate orientation.

These hypotheses can be represented in a linear model as follows:

$y_{ijk} = \beta_0 + \beta_{Copper_i} + \beta_{Orientation_j} + \beta_{Copper \times Orientation_{ij}} + \varepsilon_{ijk}$

Here, β_0 is the grand mean (intercept), β_{Copper_i} and $\beta_{Orientation_j}$ are the main effects of the copper levels and orientation, respectively, and $\beta_{Copper \times Orientation_{ij}}$ represents the interaction between these factors. The ε_{ijk} term captures the residual variance, or the random deviation of each observation from the model prediction. The significance of each beta coefficient is tested to determine if it makes a meaningful contribution to the model. A non-significant beta suggests that the corresponding factor or interaction does not have a distinct effect on the outcome, and therefore, might be omitted from the model for parsimony.

When conducting an ANOVA, we're essentially fitting a linear model with categorical predictors. These predictors are represented by beta coefficients in our model, which show the unique contribution of each factor (like copper levels or substrate orientation) to the outcome variable (such as species richness). The critical question we ask is whether each beta coefficient is significant—does it make a meaningful difference to our model?

In statistical terms, we assess this by examining the F-statistic. The F-statistic is a ratio that compares the amount of variance explained by a particular factor to the variance not explained by the model (within-group variance or error). A higher F-statistic indicates that the factor explains a significant portion of the variability in the outcome variable, while a lower F-statistic suggests that the factor does not contribute much to our understanding of the outcome variable.

$F = \frac{MS_{Treatment}}{MS_{Error}}$.

In an ANOVA with two factors, we calculate the F-statistic for each factor and their interaction. We don't need to show the actual calculation here—what matters is that the F-statistic tells us if the factor makes a significant contribution. If it does, we keep the beta coefficient in our model. If not, we may consider omitting it for a simpler model.

It turns out that ANOVAs are just a type of linear model in which the predictor variable is categorical. In practice, we can perform ANOVA using the `lm()` function in R, treating the categorical predictors as factors. This allows us to use the familiar beta notation and interpret ANOVA as a linear regression model with categorical predictors.

12.2 Moving Beyond Regression and ANOVA to ANCOVA

As we've explored statistical modeling, we've understood the power of both regression and ANOVA. Now, let's merge these concepts to broaden our analytical horizon. ANCOVA, or Analysis of Covariance, allows us to examine relationships across different categories, enhancing our ability to compare and understand varying trends within our data.

Consider these questions that may arise in environmental research:

- Are dbh (Diameter at Breast Height) and height related similarly for tulip poplars and oaks?
- Are biomass and BTUs (British Thermal Units) related similarly for corn stover and Miscanthus?
- Does the exposure to PFAS correlate with the lifetime incidence of cancer uniformily across low- and high-income American populations?

This specific subsection of general linear models is known as **Analysis of Covariance – ANCOVA**. Each of these questions challenges us to compare relationships across categories, which is precisely where ANCOVA shines.

12.2.1 Understanding ANCOVA

With ANCOVA, if we identify a significant difference between two linear relationships, our model will represent two distinct lines. The challenge then becomes integrating these two lines into a single equation.

12.2.1.1 Conceptualizing the Transformation

Consider the following equations:

- `y = 2x + 3`
- `y = 3x - 4`

The question we pose is: What modifications are required to transform the first equation into the second? This involves determining the adjustments needed in terms of x (the slope) and the constant term. Understanding this transformation is key to grasping how ANCOVA allows us to compare different linear relationships within a single model framework.

12.3 Introducing the Indicator or Dummy Variable

In General Linear Models (GLMs), indicator or dummy variables allow us to include categorical variables in models traditionally designed for continuous variables.

12.3.1 Understanding Indicator Variables

Indicator variables are used to encode categories. For `n` categories, you need `n-1` indicator variables.

Consider a scenario with 5 species of wombat. We can code these species using 4 indicator variables:

Ind1	Ind2	Ind3	Ind4	Species
1	0	0	0	Species 1
0	1	0	0	Species 2
0	0	1	0	Species 3
0	0	0	1	Species 4
0	0	0	0	Species 5

In the simplest case of two categories, only one indicator variable is needed:

- 0 = Wombat Species 1
- 1 = Wombat Species 2

12.3.1.1 Flipping the Switch: Indicator Variables in Action

In the context of GLMs, "turning on" an indicator variable means assigning it a value of 1. This action activates certain terms in the equation that are multiplied by the indicator variable, thereby altering the model's output.

When an indicator variable (`xc`) is set to 1, it effectively activates any terms in the equation that are multiplied by `xc`. This can change the slope and/or intercept of the regression line, depending on how `xc` is used in the equation.

For example, in the equation `y = 2xl + 3 + 1xcxl - 7xc`:

- When `xc` = 0 (turned off), the equation simplifies to `y = 2xl + 3`. Here, the terms `1xcxl` and `-7xc` are deactivated because they are multiplied by `xc`, which is 0.
- When `xc` = 1 (turned on), the equation becomes `y = (2xl + 3) + (1xl - 7)`. In this case, the terms `1xcxl` and `-7xc` are activated, altering the slope and intercept of the line.

12.3.1.2 Visualizing the Effect

The R plots below demonstrate the change in the regression line when the indicator variable is toggled between being active (`xc = 1`) and inactive (`xc = 0`).

When `xc` = 0:

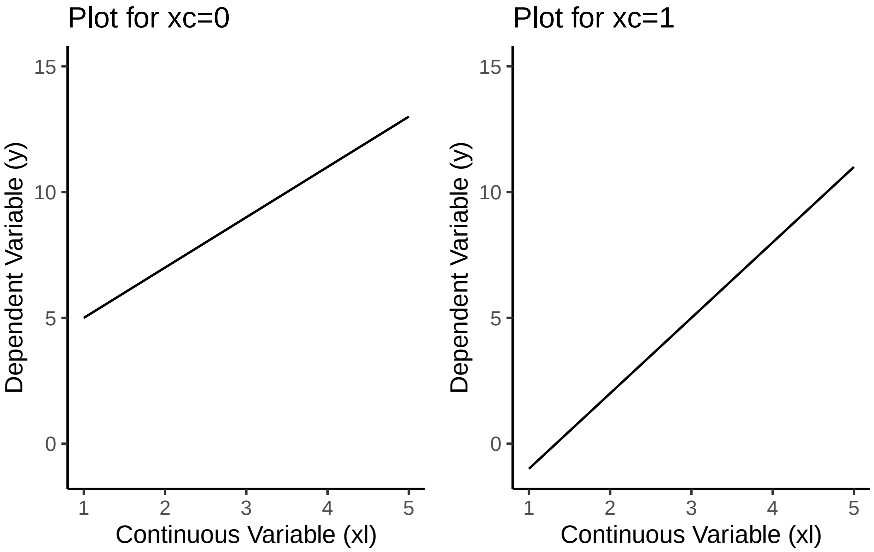
Equation Simplified: `y = 2xl + 3`

When `xc` = 1:

Equation Modified: `y = (2xl + 3) + (1xl - 7)`

The plot on the left shows the regression line when the indicator variable is inactive. The equation simplifies, reflecting a scenario where the categorical variable does not influence the outcome.

The plot on the right illustrates the regression line when the indicator variable is active. The equation now includes additional terms, showcasing how the presence of the categorical variable changes the relationship between `xl` and `y`.



Using indicator variables, we can create models in which regression lines change completely depending on which category we're modeling. If we need to tweak both the intercept and the slope, then we will need two new regression parameters – β_2 and β_3.

12.3.2 Interaction Term in GLMs

Just like in ANOVA, interaction terms in GLMs tell us whether one variable's effect on the outcome changes when another variable comes into play.

- **Simple Definition:** In statistics, an interaction term helps us understand if the effect of one factor (like temperature) on an outcome (like plant growth) changes when another factor (like rainfall) is also considered. It's like asking, "Does the relationship between temperature and plant growth change when we also consider how much it rains?"

12.3.2.1 The Math Behind The Interactions

- **Model Equation Explained:**

Let's break down a typical equation:

$$y = \beta_0 + \beta_1 x_l + \beta_2 x_c + \beta_3 x_l x_c + \epsilon$$

Here, $x_l x_c$ is the interaction term, and β_3 is its coefficient.

- y is what we're trying to predict (like plant growth).
- β_0 is the starting point of our prediction when all other factors are zero.
- $\beta_1 x_l$ shows how our prediction changes with changes in a continuous variable (like temperature).
- $\beta_2 x_c$ shows the change with a categorical variable (like type of plant).
- $\beta_3 x_l x_c$ is the key player here. It shows how the effect of our continuous variable (temperature) changes across different categories (types of plants).
- ϵ is the error term, accounting for variations we can't explain with our model.
- $\beta_1 x_l$ and $\beta_2 x_c$ represent the main effects of the continuous and categorical variables, respectively.

12.3.2.2 When is the Interaction Term Significant?

- **Significance of $\beta_3 x_l x_c$:**

If $\beta_3 x_l x_c$ (our interaction term) is significant, it means the relationship between our continuous variable (like temperature) and our outcome (plant growth) is different for different categories (like types of plants). If $\beta_3 x_l x_c$ is not significant, it suggests that the effect of our continuous variable is consistent across categories, and we might not need this term in our model.

12.3.2.3 What are the options for our model?

1. **Full Model with Interaction** ($\beta_0, \beta_1, \beta_2, \beta_3$):
   - When both the categorical ($x_c$) and continuous (x_l) variables are significant, and there is a significant interaction (β_3), the model unfolds into two distinct linear equations for each category of$x_c$. This indicates that the relationship between the continuous variable and the outcome differs depending on the category of the categorical variable.
   - Equations:
     - For $x_c = 0$: $y = \beta_0 + \beta_1 x_l$
     - For $x_c = 1$: $y = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_l$
2. **Model without Interaction ((_0, _1, _2)):**
   - If both the categorical and continuous variables are significant, but there is no interaction, the model simplifies to parallel lines for each category of $x_c $„ indicating that the slope (effect of x_l, is consistent across categories.
   - Equations:
     - For $x_c = 0$: $y = \beta_0 + \beta_1 x_l$
     - For $x_c = 1$: $y = (\beta_0 + \beta_2) + \beta_1 x_l$
3. **Simple Linear Regression** (β_0, β_1):
   - If only the continuous variable (x_l) is significant, the model reduces to a simple linear regression, indicating a linear relationship betweenx_l and y, regardless of the category of$x_c$.
   - Equation: $y = \beta_0 + \beta_1 x_l$
4. **Two Different Means (T-test Equivalent)** (β_0, β_2):
   - If the continuous variable is not significant but the categorical variable is, the model effectively becomes a comparison of means between two groups (similar to a t-test).
   - Equations:
     - For $x_c = 0$: $y = \beta_0$
     - For $x_c = 1$: $y = \beta_0 + \beta_2$
5. **Single Mean (One Sample Data)** (β_0):
   - If neither variable is significant, the model reduces to a single mean, indicating no effect of either the continuous or categorical variable.
   - Equation: $y = \beta_0$

12.3.3 Terminology in GLMs

Understanding the terminology used in General Linear Models (GLMs) is important for grasping the concepts and effectively communicating your findings. Here are some key terms:

- **Fixed Factor:** This refers to a categorical variable with specific, predefined categories. For example, if you're studying the effect of different seasons (spring, summer, autumn, winter) on plant growth, 'season' is a fixed factor because it represents specific, distinct categories of interest.
- **Random Factor:** This is a categorical variable where the categories represent a random sample from a larger population. For instance, if you're examining the output quality from different machines in a factory, and these machines are randomly selected from a larger set, then 'machine' is a random factor.
- **Covariate:** This is your continuous variable that varies along with the dependent variable (Y). In environmental studies, this could be something like temperature, rainfall, or pollution levels, which you suspect might influence your outcome of interest (like species distribution or plant growth).

12.4 So, What Exactly is ANCOVA?

Analysis of Covariance (ANCOVA) is a statistical method that combines the principles of ANOVA and regression. It's designed to compare categorical independent variables—such as different treatments or groups—while controlling for the influence of continuous variables. These continuous variables, known as covariates, are typically not the primary focus but could affect the outcome.

For ANCOVA, you need:

1) A continuous dependent variable (the main effect you're studying)
2) At least one continuous independent variable (covariate)
3) At least one categorical independent variable (which can be either a fixed or random factor)

12.4.1 The Null Hypothesis in ANCOVA

Demystifying the Concept: The null hypothesis in ANCOVA suggests that once we adjust for the covariates, the categorical independent variables do not significantly affect the dependent variable. Essentially, we're testing whether the apparent differences between groups are genuine or just statistical noise.

Mathematically Speaking: In a simple ANCOVA model, the null hypothesis posits that the adjusted means for each level of the categorical variable are equivalent, once we account for the covariate:

$H_0 : \mu_1 = \mu_2 = ... = \mu_t$

$H_A : \mu_i \neq \mu_j^*$ for some $i \neq j$

This translates to: "After controlling for the covariate, the different levels of the categorical independent variable do not lead to different outcomes."

12.4.2 Testing the Null Hypothesis

Conducting the Test: When we perform ANCOVA, we examine the null hypothesis through an omnibus F-test. If we find sufficient evidence to reject the null hypothesis, we infer that significant differences exist among the adjusted means of the dependent variable across the categorical groups, even after considering the covariate.

Interpreting the Findings: Rejecting the null hypothesis indicates that the categorical independent variable has a significant impact on the dependent variable, aside from the covariate's effect. It's crucial to remember, though, that statistical significance doesn't automatically translate to practical significance. We must always consider the results within the specific context and aims of the study.

12.5 Illustrating ANCOVA with an Example

12.5.1 Setting the Stage

Imagine we're environmental scientists studying the impact of air pollution—a continuous variable—on bird species distribution, our dependent variable. We want to see if this relationship differs between urban and rural areas—our categorical variable. For this purpose, we've created a synthetic dataset to model the scenario.

12.5.2 Crafting the Model

With our data in hand, we'd construct an ANCOVA model to discern whether location type influences bird distribution, beyond what can be explained by pollution levels alone. This would involve coding our categorical variable (urban vs. rural) into a dummy variable and including air pollution as a covariate in the model. I've done this for you already by creating the synthetic dataset.

- **Dataset Overview:** Our dataset comprises 200 observations, capturing bird species density in various urban and rural locations, along with corresponding levels of air pollution.

12.5.3 Why We're Running the ANCOVA

- **Research Question:** Does the effect of air pollution on bird species density differ between urban and rural areas?
- **Expectation:** We predict that air pollution will have a more pronounced negative effect on bird species density in urban areas than in rural areas.

12.5.4 Null and Alternative Hypotheses in our example ANCOVA

1. **Omnibus Null Hypothesis** (H_0): The null hypothesis posits that once we adjust for air pollution, the mean bird species density does not differ between urban and rural areas.

$H_0 : \mu^*_{\text{Urban}} = \mu^*_{\text{Rural}}$

$H_A : \mu^*_{\text{Urban}} \neq \mu^*_{\text{Rural}}$

Where μ^*_{Urban} and μ^*_{Rural} represent the adjusted means of bird species density for urban and rural areas, respectively..

2. **Null Hypothesis for Slope** (H_{01}): This hypothesis examines if the effect of air pollution on bird species density is consistent between urban and rural areas.

   - H_{01}: There is no interaction between air pollution and area type; the slopes are parallel.
   - H_{A1}: There is an interaction; the effect of air pollution on bird species density differs between urban and rural areas.

$H_{01} : \beta_{\text{interaction}} = 0$

$H_{A1} : \beta_{\text{interaction}} \neq 0$

3. **Null Hypothesis for Intercept** (H_{02}): If we find no significant interaction, we then consider the intercepts. This hypothesis tests whether the baseline bird species density, at zero air pollution, differs between urban and rural areas.

   - H_{02}: The intercepts are the same, indicating no difference in bird species density between urban and rural areas at zero air pollution.
   - H_{A2}: The intercepts differ, suggesting an inherent difference in bird species density between urban and rural areas, independent of air pollution.

$H_{02} : \beta_{0,\text{Urban}} = \beta_{0,\text{Rural}}$

$H_{A2} : \beta_{0,\text{Urban}} \neq \beta_{0,\text{Rural}}$

Rejecting H_{02} would imply that there is an inherent difference in bird species density due to area type, even before considering the air pollution levels.

These hypotheses guide our ANCOVA model and analysis, aiming to isolate the effects of urbanization from the influence of air pollution on avian populations.

12.5.5 Analysis

12.5.5.1 Explore the data

```
#>   AreaType AirPollution BirdDensity
#> 1    Urban        148.9          25
#> 2    Urban         39.7          39
#> 3    Urban         37.6          24
#> 4    Rural         58.8          30
#> 5    Urban         54.9          18
#> 6    Rural         38.2          35
```

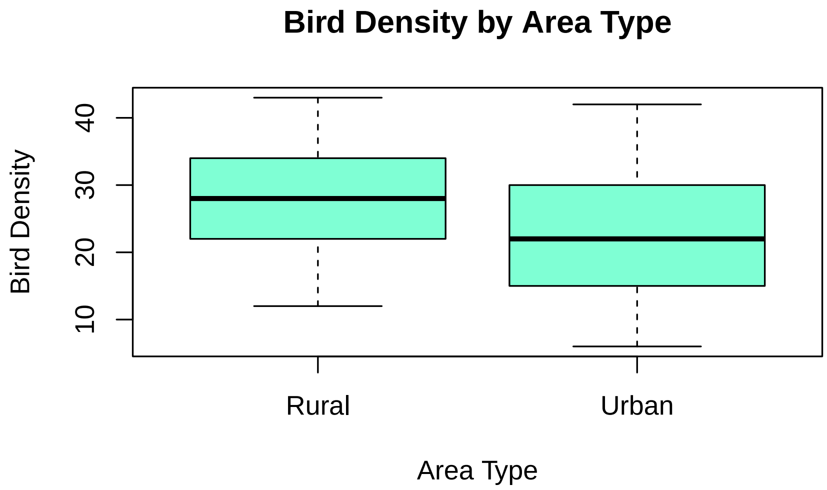
Let's create a summary of our current dataset

```
summary(data)
#>    AreaType          AirPollution     BirdDensity
#>  Length:200         Min.   : 25.80   Min.   : 6.00
#>  Class :character   1st Qu.: 54.45   1st Qu.:18.00
#>  Mode  :character   Median : 83.60   Median :25.00
#>                     Mean   : 86.14   Mean   :24.83
#>                     3rd Qu.:117.72   3rd Qu.:32.00
#>                     Max.   :149.90   Max.   :43.00
```

This will give us a quick statistical summary of our variables.

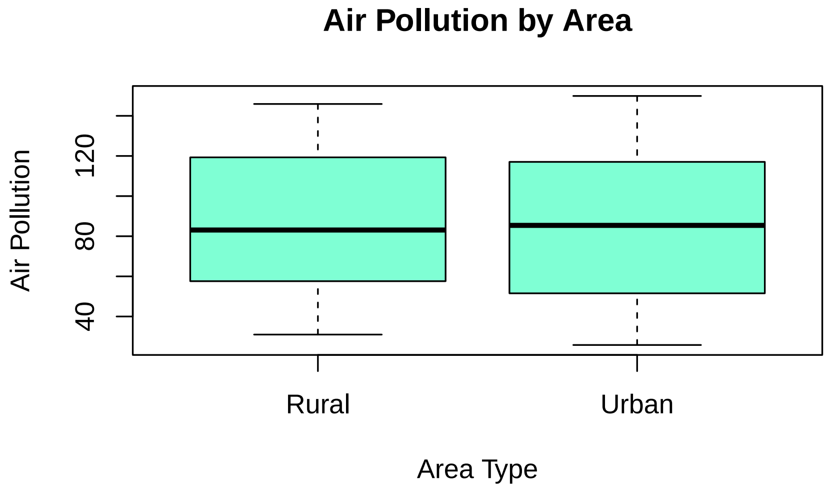
Now, let's visualize the distribution of bird density in different areas.

```
# Boxplot for bird density by area type
boxplot(BirdDensity ~ AreaType, data = data,
        main = "Bird Density by Area Type",
        xlab = "Area Type", ylab = "Bird Density",
        col = "aquamarine", border = "black")
```



And the same for air pollution:

```
boxplot(AirPollution ~area_type, data=data, main="Air Pollution by Area", xlab="Area Type", y
```



12.5.6 Results of the ANCOVA Model

The anova results

```
ancova <- aov(BirdDensity ~ AirPollution*AreaType, data=data)
summary(ancova)
#>                        Df Sum Sq Mean Sq F value  Pr(>F)
#> AirPollution            1      6     5.9   0.086  0.7690
#> AreaType                1   1688  1688.0  24.788 1.4e-06 ***
#> AirPollution:AreaType   1    291   291.3   4.277  0.0399 *
#> Residuals             196  13347    68.1
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The linear model results

```
ancova_lm <- lm(BirdDensity ~ AirPollution*AreaType, data=data)
summary(ancova_lm)
#>
#> Call:
#> lm(formula = BirdDensity ~ AirPollution * AreaType, data = data)
#>
#> Residuals:
#>      Min       1Q   Median       3Q      Max
#> -17.9260  -6.2694   0.4038   5.9023  19.2857
#>
#> Coefficients:
#>                             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)                 31.54648    2.28553  13.803  < 2e-16 ***
#> AirPollution                -0.04333    0.02476  -1.750 0.081641 .
#> AreaTypeUrban              -11.57676    3.02164  -3.831 0.000172 ***
#> AirPollution:AreaTypeUrban   0.06695    0.03237   2.068 0.039943 *
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 8.252 on 196 degrees of freedom
#> Multiple R-squared:  0.1295, Adjusted R-squared:  0.1162
#> F-statistic: 9.717 on 3 and 196 DF,  p-value: 5.23e-06
```

12.5.6.1 Main Effects:

- **Air Pollution**: The ANCOVA model does not find a significant main effect of air pollution on bird density ($p = 0.7690$ from ANOVA), indicating no consistent impact across the areas. However, the linear model's marginal p-value ($p = 0.081641$) suggests a trend that might have been significant with a larger sample size or reduced variability.

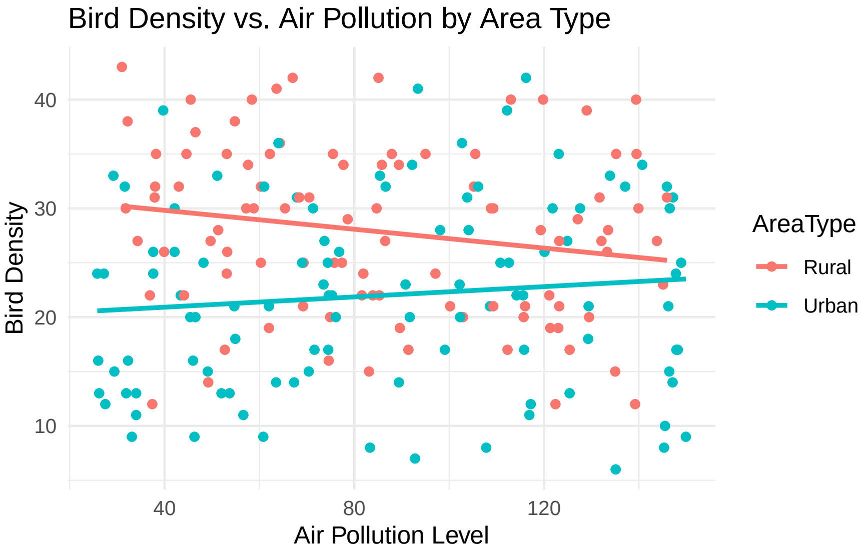
- **Area Type**: There is a significant main effect of AreaType on bird density (p < 0.001), indicating that there are considerable differences in bird density between rural and urban areas, with urban areas showing lower bird density.

12.5.6.2 Interpreting the Beta Coefficients:

- **Intercept**: Reflecting the expected bird density in rural areas at zero air pollution, the intercept is quite high, indicating a relatively healthy bird population in the absence of pollution.
- **Area Type**: The beta coefficient for our area type factor is significantly different from zero (p < 0.001), suggesting a distinct difference in bird density between urban and rural areas. However, this effect is captured more precisely in the interaction term.
- **Interaction Term**: The significant beta for the interaction term confirms our hypothesis that air pollution's effect on bird density is not uniform across urban and rural areas.

12.5.6.3 Visual Representation:

The scatter plot with regression lines will demonstrate these relationships visually, showing the differing trends for urban and rural areas suggested by the significant interaction term.



12.5.6.4 Interpretation:

The significant interaction term indicates that the simple main effect of air pollution is not adequate to describe its impact on bird density. In urban areas, there seems to be a slightly positive or less negative relationship compared to rural areas, which could reflect a range of urban-specific factors affecting bird populations differently than in rural areas.

12.5.7 Contextualizing the Findings:

The lack of a significant main effect for air pollution in the presence of a significant interaction suggests complex underlying ecological dynamics. These findings call for nuanced environmental management strategies that address the specific challenges and conditions of urban and rural habitats to support bird conservation effectively. By considering the interaction term's significance, we gain a more accurate understanding of the ecological effects of air pollution, which is essential for developing targeted conservation policies.

12.6 Chapter Summary

In this chapter, we explored the fundamental principles and applications of Analysis of Covariance (ANCOVA), a powerful statistical tool that extends beyond the capabilities of ANOVA by incorporating covariates. We began by understanding the conceptual framework of ANOVA as a linear model, setting the stage for the more complex ANCOVA analysis. Through our discussions, we emphasized the importance of the null and alternative hypotheses in ANCOVA, using practical examples like the study of bird species density in different environments to illustrate these concepts. The chapter highlighted how ANCOVA adjusts for the effects of additional variables, allowing us to more accurately isolate and understand the impact of our factors of interest. By integrating real-world scenarios and focusing on clear, practical applications, we aimed to demystify the process of hypothesis testing in ANCOVA, making it accessible and relevant to environmental science.

13 Thinking about answering questions with data

You might be happy that this is the last chapter (so far) of this textbook. At this point we are in the last weeks of our introductory statistics course. It's called "introductory" for a reason. There's just too much out there to cover in one short semester. In this chapter we acknowledge some of the things we haven't yet covered, and treat them as things that you should think about. If there is one take home message that we want to get across to you, it's that when you ask questions with data, you should be able to **justify** how you answer those questions.

13.1 Effect-size and power

If you already know something about statistics while you were reading this book, you might have noticed that we neglected to discuss the topic of effect-size, and we barely talked about statistical power. We will talk a little bit about these things here.

First, it is worth pointing out that over the years, at least in Psychology, many societies and journals have made recommendations about how researchers should report their statistical analyses. Among the recommendations is that measures of "effect size" should be reported. Similarly, many journals now require that researchers report an "a priori" power-analysis (the recommendation is this should be done before the data is collected). Because these recommendations are so prevalent, it is worth discussing what these ideas refer to. At the same time, the meaning of effect-size and power somewhat depend on your "philosophical" bent, and these two ideas can become completely meaningless depending on how you think of statistics. For these complicating reasons we have suspended our discussion of the topic until now.

The question or practice of using measures of effect size and conducting power-analyses are also good examples of the more general need to think about about what you are doing. If you are going to report effect size, and conduct power analyses, these activities should not be done blindly because someone else recommends that you do them, these activities and other suitable ones should be done as a part of justifying what you are doing. It is a part of thinking about how to make your data answer questions for you.

13.1.1 Chance vs. real effects

Let's rehash something we've said over and over again. First, researchers are interested in whether their manipulation causes a change in their measurement. If it does, they can become confident that they have uncovered a causal force (the manipulation). However, we know that differences in the measure between experimental conditions can arise by chance alone, just by sampling error. In fact, we can create pictures that show us the window of chance for a given statistic, these tells us roughly the range and likelihoods of getting various differences just by chance. With these windows in hand, we can then determine whether the differences we found in some data that we collected were likely or unlikely to be due to chance. We also learned that sample-size plays a big role in the shape of the chance window. Small samples give chance a large opportunity make big differences. Large samples give chance a small opportunity to make big differences. The general lesson up to this point has been, design an experiment with a large enough sample to detect the effect of interest. If your design isn't well formed, you could easily be measuring noise, and your differences could be caused by sampling error. Generally speaking, this is still a very good lesson: better designs produce better data; and you can't fix a broken design with statistics.

There is clearly another thing that can determine whether or not your differences are due to chance. That is the effect itself. If the manipulation does cause a change, then there is an effect, and that effect is a real one. Effects refer to differences in the measurement between experimental conditions. The thing about effects is that they can be big or small, they have a size.

For example, you can think of a manipulation in terms of the size of its hammer. A strong manipulation is like a jack-hammer: it is loud, it produces a big effect, it creates huge differences. A medium manipulation is like regular hammer: it works, you can hear it, it drives a nail into wood, but it doesn't destroy concrete like a jack-hammer, it produces a reliable effect. A small manipulation is like tapping something with a pencil: it does something, you can barely hear it, and only in a quiet room, it doesn't do a good job of driving a nail into wood, and it does nothing to concrete, it produces tiny, unreliable effects. Finally, a really small effect would be hammering something with a feather, it leaves almost no mark and does nothing that is obviously perceptiple to nails or pavement. The lesson is, if you want to break up concrete, use a jack-hammer; or, if you want to measure your effect, make your manipulation stronger (like a jack-hammer) so it produces a bigger difference.

13.1.2 Effect size: concrete vs. abstract notions

Generally speaking, the big concept of effect size, is simply how big the differences are, that's it. However, the biggness or smallness of effects quickly becomes a little bit complicated. On the one hand, the raw difference in the means can be very meaningful. Let's saw we are measuring performance on a final exam, and we are testing whether or not a miracle drug can make you do better on the test. Let's say taking the drug makes you do 5% better on the test, compared to not taking the drug. You know what 5% means, that's basically a whole letter grade. Pretty good. An effect-size of 25% would be even better right! Lot's of measures have a concrete quality to them, and we often want to the size of the effect expressed in terms of the original measure.

Let's talk about concrete measures some more. How about learning a musical instrument. Let's say it takes 10,000 hours to become an expert piano, violin, or guitar player. And, let's say you found something online that says that using their method, you will learn the instrument in less time than normal. That is a claim about the effect size of their method. You would want to know how big the effect is right? For example, the effect-size could be 10 hours. That would mean it would take you 9,980 hours to become an expert (that's a whole 10 hours less). If I knew the effect-size was so tiny, I wouldn't bother with their new method. But, if the effect size was say 1,000 hours, that's a pretty big deal, that's 10% less (still doesn't seem like much, but saving 1,000 hours seems like a lot).

Just as often as we have concrete measures that are readily interpretable, Psychology often produces measures that are extremely difficult to interpret. For example, questionnaire measures often have no concrete meaning, and only an abstract statistical meaning. If you wanted to know whether a manipulation caused people to more or less happy, and you used to questionnaire to measure happiness, you might find that people were 50 happy in condition 1, and 60 happy in condition 2, that's a difference of 10 happy units. But how much is 10? Is that a big or small difference? It's not immediately obvious. What is the solution here? A common solution is to provide a standardized measure of the difference, like a z-score. For example, if a difference of 10 reflected a shift of one standard deviation that would be useful to know, and that would be a sizeable shift. If the difference was only a .1 shift in terms of standard deviation, then the difference of 10 wouldn't be very large. We elaborate on this idea next in describing cohen's d.

13.1.3 Cohen's d

Let's look a few distributions to firm up some ideas about effect-size. Figure 13.1 has four panels. The first panel (0) represents the null distribution of no differences. This is the idea that your manipulation (A vs. B) doesn't do anything at all, as a result when you measure scores in conditions A and B, you are effectively sampling scores from the very same overall distribution. The panel shows the distribution as green for condition B, but the red one for condition A is identical and drawn underneath (it's invisible). There is 0 difference between these distributions, so it represent a null effect.

The remaining panels are hypothetical examples of what a true effect could look like, when your manipulation actually causes a difference. For example, if condition A is a control group, and condition B is a treatment group, we are looking at three cases where the treatment manipulation causes a positive shift in the mean of distribution. We are using normal curves with mean =0 and sd =1 for this demonstration, so a shift of .5 is a shift of half of a standard

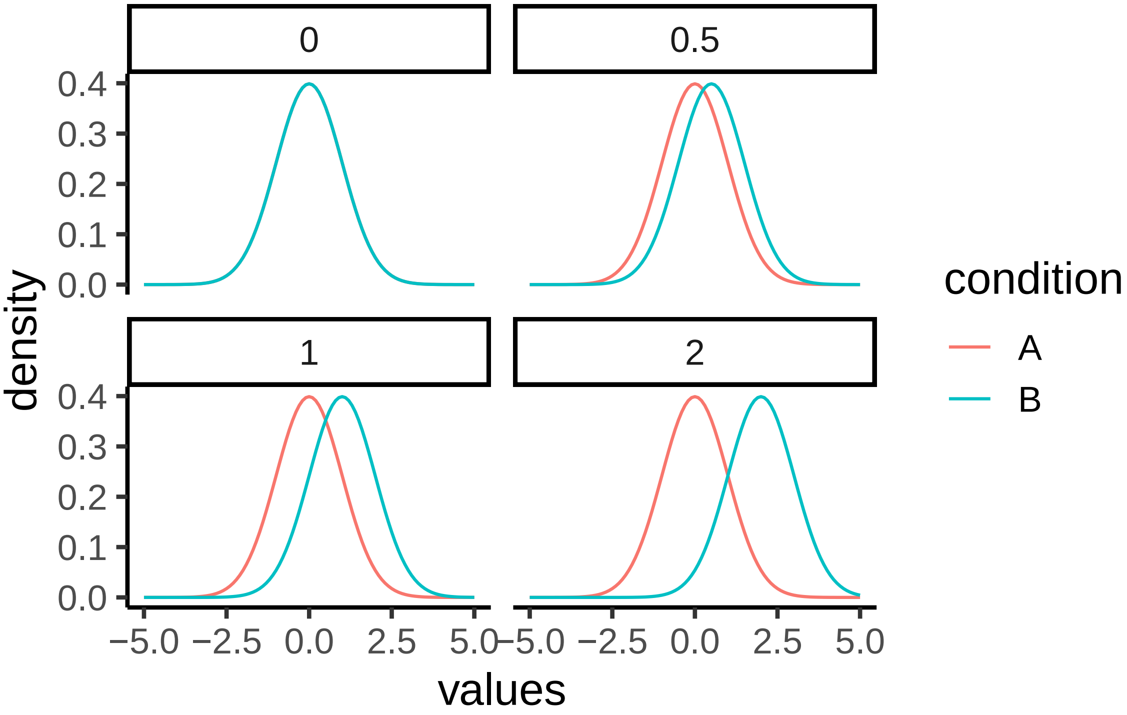


Figure 13.1: Each panel shows hypothetical distributions for two conditions. As the effect-size increases, the difference between the distributions become larger.

deviation. A shift of 1 is a shift of 1 standard deviation, and a shift of 2 is a shift of 2 standard deviations. We could draw many more examples showing even bigger shifts, or shifts that go in the other direction.

Let's look at another example, but this time we'll use some concrete measurements. Let's say we are looking at final exam performance, so our numbers are grade percentages. Let's also say that we know the mean on the test is 65%, with a standard deviation of 5%. Group A could be a control that just takes the test, Group B could receive some "educational" manipulation designed to improve the test score. These graphs then show us some hypotheses about what the manipulation may or may not be doing.

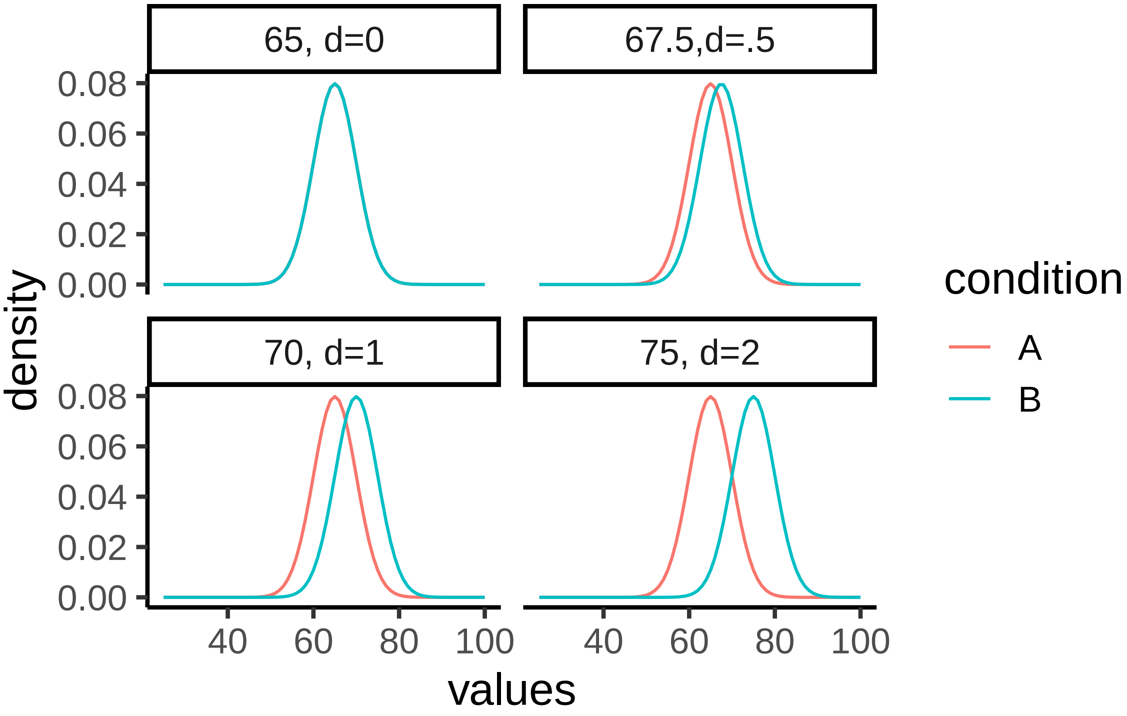


Figure 13.2: Each panel shows hypothetical distributions for two conditions. As the effect-size increases, the difference between the distributions become larger.

The first panel shows that both condition A and B will sample test scores from the same distribution (mean =65, with 0 effect). The other panels show shifted mean for condition B (the treatment that is supposed to increase test performance). So, the treatment could increase the test performance by 2.5% (mean 67.5, .5 sd shift), or by 5% (mean 70, 1 sd shift), or by 10% (mean 75%, 2 sd shift), or by any other amount. In terms of our previous metaphor, a shift of 2 standard deviations is more like jack-hammer in terms of size, and a shift of .5 standard deviations is more like using a pencil. The thing about research, is we often have no clue about whether our manipulation will produce a big or small effect, that's why we are conducting the research.

You might have noticed that the letter d appears in the above figure. Why is that? Jacob Cohen (**cohen1988?**) used the letter d in defining the effect-size for this situation, and now everyone calls it Cohen's d. The formula for Cohen's d is:

$d = \frac{\text{mean for condition 1} - \text{mean for condition 2}}{\text{population standard deviation}}$

If you notice, this is just a kind of z-score. It is a way to standardize the mean difference in terms of the population standard deviation.

It is also worth noting again that this measure of effect-size is entirely hypothetical for most purposes. In general, researchers do not know the population standard deviation, they can only guess at it, or estimate it from the sample. The same goes for means, in the formula these are hypothetical mean differences in two population distributions. In practice, researchers do not know these values, they guess at them from their samples.

Before discussing why the concept of effect-size can be useful, we note that Cohen's d is useful for understanding abstract measures. For example, when you don't know what a difference of 10 or 20 means as a raw score, you can standardize the difference by the sample standard deviation, then you know roughly how big the effect is in terms of standard units. If you thought a 20 was big, but it turned out to be only 1/10th of a standard deviation, then you would know the effect is actually quite small with respect to the overall variability in the data.

13.2 Power

When there is a true effect out there to measure, you want to make sure your design is sensitive enough to detect the effect, otherwise what's the point. We've already talked about the idea that an effect can have different sizes. The next idea is that your design can be more less sensitive in its ability to reliabily measure the effect. We have discussed this general idea many times already in the textbook, for example we know that we will be more likely to detect "significant" effects (when there are real differences) when we increase our sample-size. Here, we will talk about the idea of design sensitivity in terms of the concept of power. Interestingly, the concept of power is a somewhat limited concept, in that it only exists as a concept within some philosophies of statistics.

13.2.1 A digresssion about hypothesis testing

In particular, the concept of power falls out of the Neyman-Pearson concept of null vs. alternative hypothesis testing. Up to this point, we have largely avoided this terminology. This is perhaps a disservice in that the Neyman-Pearson ideas are by now the most common and widespread, and in the opinion of some of us, they are also the most widely misunderstood and abused idea, which is why we have avoided these ideas until now.

What we have been mainly doing is talking about hypothesis testing from the Fisherian (Sir Ronald Fisher, the ANOVA guy) perspective. This is a basic perspective that we think can't be easily ignored. It is also quite limited. The basic idea is this:

1. We know that chance can cause some differences when we measure something between experimental conditions.
2. We want to rule out the possibility that the difference that we observed can not be due to chance
3. We construct large N designs that permit us to do this when a real effect is observed, such that we can confidently say that big differences that we find are so big (well outside the chance window) that it is highly implausible that chance alone could have produced.
4. The final conclusion is that chance was extremely unlikely to have produced the differences. We then infer that something else, like the manipulation, must have caused the difference.
5. We don't say anything else about the something else.
6. We either reject the null distribution as an explanation (that chance couldn't have done it), or retain the null (admit that chance could have done it, and if it did we couldn't tell the difference between what we found and what chance could do)

Neyman and Pearson introduced one more idea to this mix, the idea of an alternative hypothesis. The alternative hypothesis is the idea that if there is a true effect, then the data sampled into each condition of the experiment must have come from two different distributions. Remember, when there is no effect we assume all of the data cam from the same distribution (which by definition can't produce true differences in the long run, because all of the numbers are coming from the same distribution). The graphs of effect-sizes from before show examples of these alternative distributions, with samples for condition A coming from one distribution, and samples from condition B coming from a shifted distribution with a different mean.

So, under the Neyman-Pearson tradition, when a researcher find a signifcant effect they do more than one things. First, they reject the null-hypothesis of no differences, and they accept the alternative hypothesis that there was differences. This seems like a sensible thing to do. And, because the researcher is actually interested in the properties of the real effect, they might be interested in learning more about the actual alternative hypothesis, that is they might want to know if their data come from two different distributions that were separated by some amount…in other words, they would want to know the size of the effect that they were measuring.

13.2.2 Back to power

We have now discussed enough ideas to formalize the concept of statistical power. For this concept to exist we need to do a couple things.

1. Agree to set an alpha criterion. When the p-value for our test-statistic is below this value we will call our finding statistically significant, and agree to reject the null hypothesis and accept the "alternative" hypothesis (sidenote, usually it isn't very clear which specific alternative hypothesis was accepted)
2. In advance of conducting the study, figure out what kinds of effect-sizes our design is capable of detecting with particular probabilites.

The power of a study is determined by the relationship between

1. The sample-size of the study
2. The effect-size of the manipulation
3. The alpha value set by the researcher.

To see this in practice let's do a simulation. We will do a t-test on a between-groups design 10 subjects in each group. Group A will be a control group with scores sampled from a normal distribution with mean of 10, and standard deviation of 5. Group B will be a treatment group, we will say the treatment has an effect-size of Cohen's $d = .5$, that's a standard deviation shift of .5, so the scores with come from a normal distribution with mean =12.5 and standard deivation of 5. Remember 1 standard deviation here is 5, so half of a standard deviation is 2.5.

The following R script runs this simulated experiment 1000 times. We set the alpha criterion to .05, this means we will reject the null whenever the *p*-value is less than .05. With this specific design, how many times out of of 1000 do we reject the null, and accept the alternative hypothesis?

```
#> [1] 179
```

The answer is that we reject the null, and accept the alternative 179 times out of 1000. In other words our experiment succesfully accepts the alternative hypothesis 17.9 percent of the time, this is known as the power of the study. Power is the probability that a design will succesfully detect an effect of a specific size.

Importantly, power is completely abstract idea that is completely determined by many assumptions including N, effect-size, and alpha. As a result, it is best not to think of power as a single number, but instead as a family of numbers.

For example, power is different when we change N. If we increase N, our samples will more precisely estimate the true distributions that they came from. Increasing N reduces sampling error, and shrinks the range of differences that can be produced by chance. Lets' increase our N in this simulation from 10 to 20 in each group and see what happens.

```
#> [1] 336
```

Now the number of significant experiments i 336 out of 1000, or a power of 33.6 percent. That's roughly doubled from before. We have made the design more sensitive to the effect by increasing N.

We can change the power of the design by changing the alpha-value, which tells us how much evidence we need to reject the null. For example, if we set the alpha criterion to 0.01, then we will be more conservative, only rejecting the null when chance can produce the observed difference 1% of the time. In our example, this will have the effect of reducing power. Let's keep N at 20, but reduce the alpha to 0.01 and see what happens:

```
#> [1] 148
```

Now only 148 out of 1000 experiments are significant, that's 14.8 power.

Finally, the power of the design depends on the actual size of the effect caused by the manipulation. In our example, we hypothesized that the effect caused a shift of .5 standard deviations. What if the effect causes a bigger shift? Say, a shift of 2 standard deviations. Let's keep N= 20, and alpha < .01, but change the effect-size to two standard deviations. When the effect in the real-world is bigger, it should be easier to measure, so our power will increase.

```
#> [1] 1000
```

Neat, if the effect-size is actually huge (2 standard deviation shift), then we have power 100 percent to detect the true effect.

13.2.3 Power curves

We mentioned that it is best to think of power as a family of numbers, rather than as a single number. To elaborate on this consider the power curve below. This is the power curve for a specific design: a between groups experiments with two levels, that uses an independent samples t-test to test whether an observed difference is due to chance. Critically, N is set to 10 in each group, and alpha is set to .05

In Figure 13.3 power (as a proportion, not a percentage) is plotted on the y-axis, and effect-size (Cohen's d) in standard deviation units is plotted on the x-axis.

A power curve like this one is very helpful to understand the sensitivity of a particular design. For example, we can see that a between subjects design with N=10 in both groups, will detect an effect of d=.5 (half a standard deviation shift) about 20% of the time, will detect an effect of d=.8 about 50% of the time, and will detect an effect of d=2 about 100% of the time. All of the percentages reflect the power of the design, which is the percentage of times the design would be expected to find a $p < 0.05$.

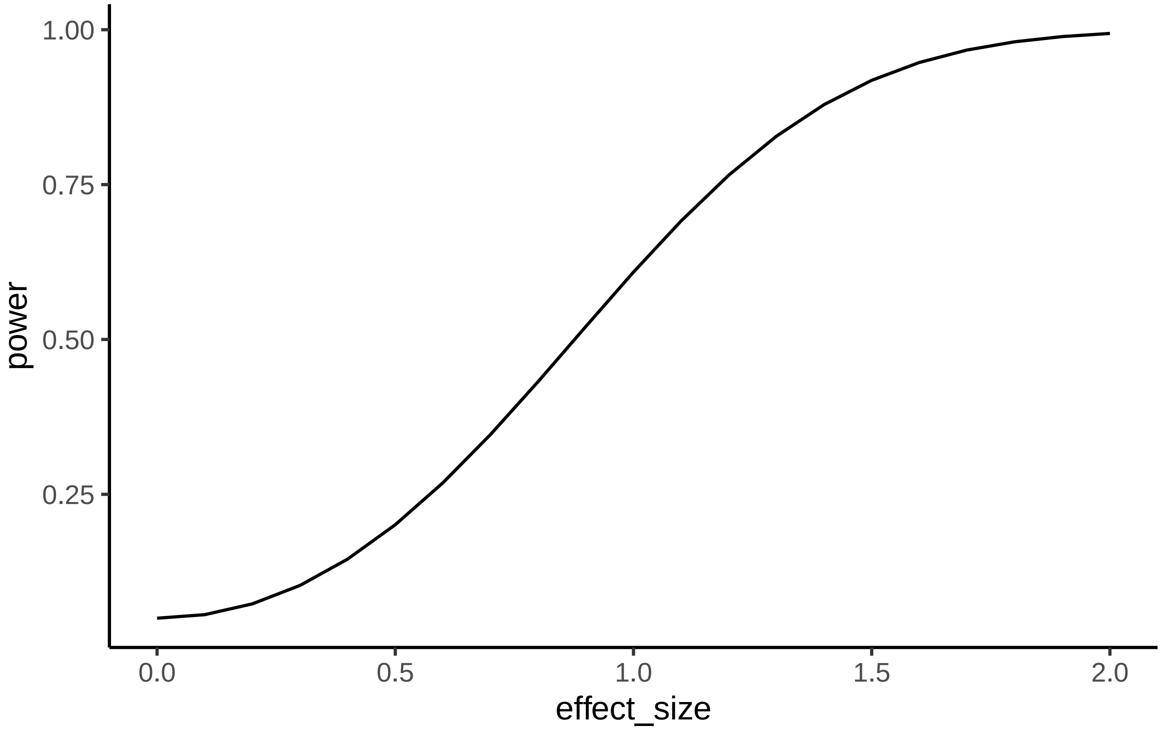


Figure 13.3: This figure shows power as a function of effect-size (Cohen's d) for a between-subjects independent samples t-test, with N=10, and alpha criterion 0.05.

Let's imagine that based on prior research, the effect you are interested in measuring is fairly small, d=0.2. If you want to run an experiment that will detect an effect of this size a large percentage of the time, how many subjects do you need to have in each group? We know from the above graph that with N=10, power is very low to detect an effect of d=0.2. Let's make Figure 13.4 and vary the number of subjects rather than the size of the effect.

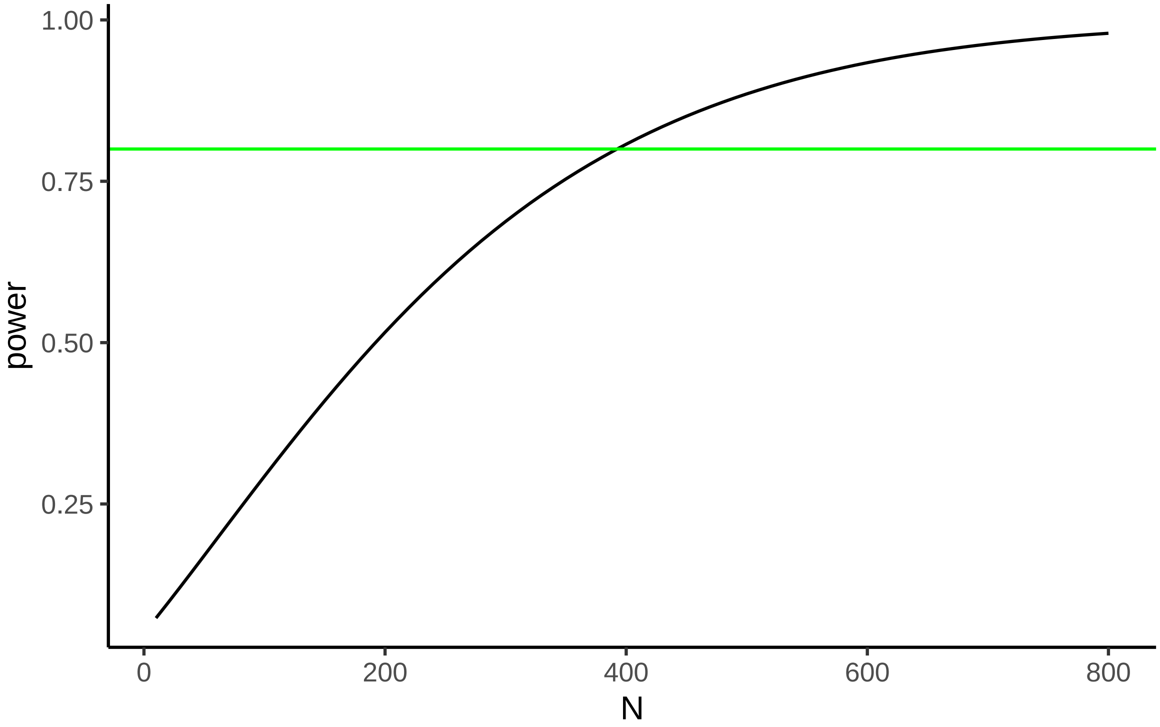


Figure 13.4: This figure shows power as a function of N for a between-subjects independent samples t-test, with d=0.2, and alpha criterion 0.05.

The figure plots power to detect an effect of d=0.2, as a function of N. The green line shows where power = .8, or 80%. It looks like we would nee about 380 subjects in each group to measure an effect of d=0.2, with power = .8. This means that 80% of our experiments would succesfully show $p < 0.05$. Often times power of 80% is recommended as a reasonable level of power, however even when your design has power = 80%, your experiment will still fail to find an effect (associated with that level of power) 20% of the time!

13.3 Planning your design

Our discussion of effect size and power highlight the importance of the understanding the statistical limitations of an experimental design. In particular, we have seen the relationship between:

1. Sample-size
2. Effect-size
3. Alpha criterion
4. Power

As a general rule of thumb, small N designs can only reliably detect very large effects, whereas large N designs can reliably detect much smaller effects. As a researcher, it is your responsibility to plan your design accordingly so that it is capable of reliably detecting the kinds of effects it is intended to measure.

13.4 Some considerations

13.4.1 Low powered studies

Consider the following case. A researcher runs a study to detect an effect of interest. There is good reason, from prior research, to believe the effect-size is d=0.5. The researcher uses a design that has 30% power to detect the effect. They run the experiment and find a significant p-value, (p<.05). They conclude their manipulation worked, because it was unlikely that their result could have been caused by chance. How would you interpret the results of a study like this? Would you agree with thte researchers that the manipulation likely caused the difference? Would you be skeptical of the result?

The situation above requires thinking about two kinds of probabilities. On the one hand we know that the result observed by the researchers does not occur often by chance (p is less than 0.05). At the same time, we know that the design was underpowered, it only detects results of the expected size 30% of the time. We are face with wondering what kind of luck was driving the difference. The researchers could have gotten unlucky, and the difference really could be due to chance. In this case, they would be making a type I error (saying the result is real when it isn't). If the result was not due to chance, then they would also be lucky, as their design only detects this effect 30% of the time.

Perhaps another way to look at this situation is in terms of the replicability of the result. Replicability refers to whether or not the findings of the study would be the same if the experiment was repeated. Because we know that power is low here (only 30%), we would expect that most replications of this experiment would not find a significant effect. Instead, the experiment would be expected to replicate only 30% of the time.

13.4.2 Large N and small effects

Perhaps you have noticed that there is an intriguiing relationship between N (sample-size) and power and effect-size. As N increases, so does power to detect an effect of a particular size. Additionally, as N increases, a design is capable of detecting smaller and smaller effects with greater and greater power. For example, if N was large enough, we would have high power to detect very small effects, say d= 0.01, or even d=0.001. Let's think about what this means.

Imagine a drug company told you that they ran an experiment with 1 billion people to test whether their drug causes a significant change in headache pain. Let's say they found a significant effect (with power =100%), but the effect was very small, it turns out the drug reduces headache pain by less than 1%, let's say 0.01%. For our imaginary study we will also assume that this effect is very real, and not caused by chance.

Clearly the design had enough power to detect the effect, and the effect was there, so the design did detect the effect. However, the issue is that there is little practical value to this effect. Nobody is going to by a drug to reduce their headache pain by 0.01%, even if it was "scientifcally proven" to work. This example brings up two issues. First, increasing N to very large levels will allow designs to detect almost any effect (even very tiny ones) with very high power. Second, sometimes effects are meaningless when they are very small, especially in applied research such as drug studies.

These two issues can lead to interesting suggestions. For example, someone might claim that large N studies aren't very useful, because they can always detect really tiny effects that are practically meaningless. On the other hand, large N studies will also detect larger effects too, and they will give a better estimate of the "true" effect in the population (because we know that larger samples do a better job of estimating population parameters). Additionally, although really small effects are often not interesting in the context of applied research, they can be very important in theoretical research. For example, one theory might predict that manipulating X should have no effect, but another theory might predict that X does have an effect, even if it is a small one. So, detecting a small effect can have theoretical implication that can help rule out false theories. Generally speaking, researchers asking both theoretical and applied questions should think about and establish guidelines for "meaningful" effect-sizes so that they can run designs of appropriate size to detect effects of "meaningful size".

13.4.3 Small N and Large effects

All other things being equal would you trust the results from a study with small N or large N? This isn't a trick question, but sometimes people tie themselves into a knot trying to answer it. We already know that large sample-sizes provide better estimates of the distributions the samples come from. As a result, we can safely conclude that we should trust the data from large N studies more than small N studies.

At the same time, you might try to convince yourself otherwise. For example, you know that large N studies can detect very small effects that are practically and possibly even theoretically meaningless. You also know that that small N studies are only capable of reliably detecting very large effects. So, you might reason that a small N study is better than a large N study because if a small N study detects an effect, that effect must be big and meaningful; whereas, a large N study could easily detect an effect that is tiny and meaningless.

This line of thinking needs some improvement. First, just because a large N study can detect small effects, doesn't mean that it only detects small effects. If the effect is large, a large N study will easily detect it. Large N studies have the power to detect a much wider range of effects, from small to large. Second, just because a small N study detected an effect, does not mean that the effect is real, or that the effect is large. For example, small N studies have more variability, so the estimate of the effect size will have more error. Also, there is 5% (or alpha rate) chance that the effect was spurious. Interestingly, there is a pernicious relationship between effect-size and type I error rate

13.4.4 Type I errors are convincing when N is small

So what is this pernicious relationship between Type I errors and effect-size? Mainly, this relationship is pernicious for small N studies. For example, the following figure illustrates the results of 1000s of simulated experiments, all assuming the null distribution. In other words, for all of these simulations there is no true effect, as the numbers are all sampled from an identical distribution (normal distribution with mean =0, and standard deviation =1). The true effect-size is 0 in all cases.

We know that under the null, researchers will find p values that are less 5% about 5% of the time, remember that is the definition. So, if a researcher happened to be in this situation (where there manipulation did absolutely nothing), they would make a type I error 5% of the time, or if they conducted 100 experiments, they would expect to find a significant result for 5 of them.

Figure 13.5 reports the findings from only the type I errors, where the simulated study did produce $p < 0.05$. For each type I error, we calculated the exact p-value, as well as the effect-size (cohen's D) (mean difference divided by standard deviation). We already know that the true effect-size is zero, however take a look at this graph, and pay close attention to the smaller sample-sizes.

For example, look at the red dots, when sample size is 10. Here we see that the effect-sizes are quite large. When p is near 0.05 the effect-size is around .8, and it goes up and up as when p gets smaller and smaller. What does this mean? It means that when you get unlucky with a small N design, and your manipulation does not work, but you by chance find a "significant" effect, the effect-size measurement will show you a "big effect". This is the pernicious aspect. When you make a type I error for small N, your data will make you think there is no way it could be a type I error because the effect is just so big!. Notice that when N is very large, like

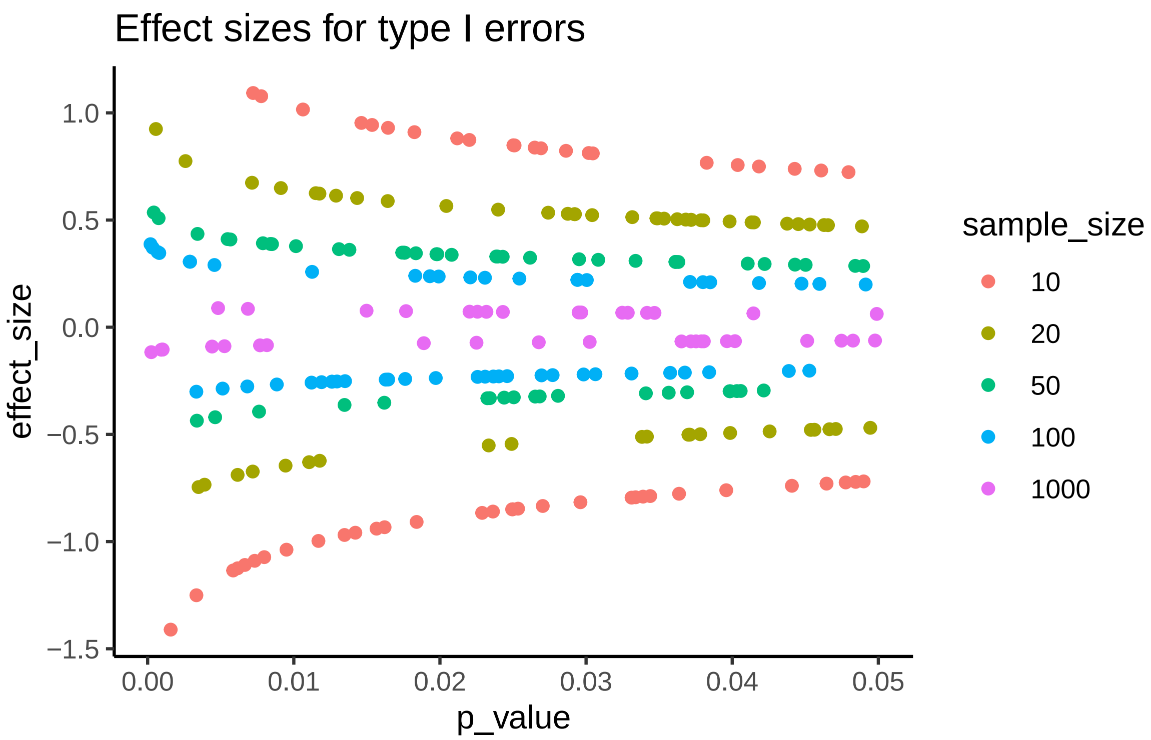


Figure 13.5: Effect size as a function of p-values for type 1 Errors under the null, for a paired samples t-test.

1000, the measure of effect-size approaches 0 (which is the true effect-size in the simulation shown in Figure 13.6).

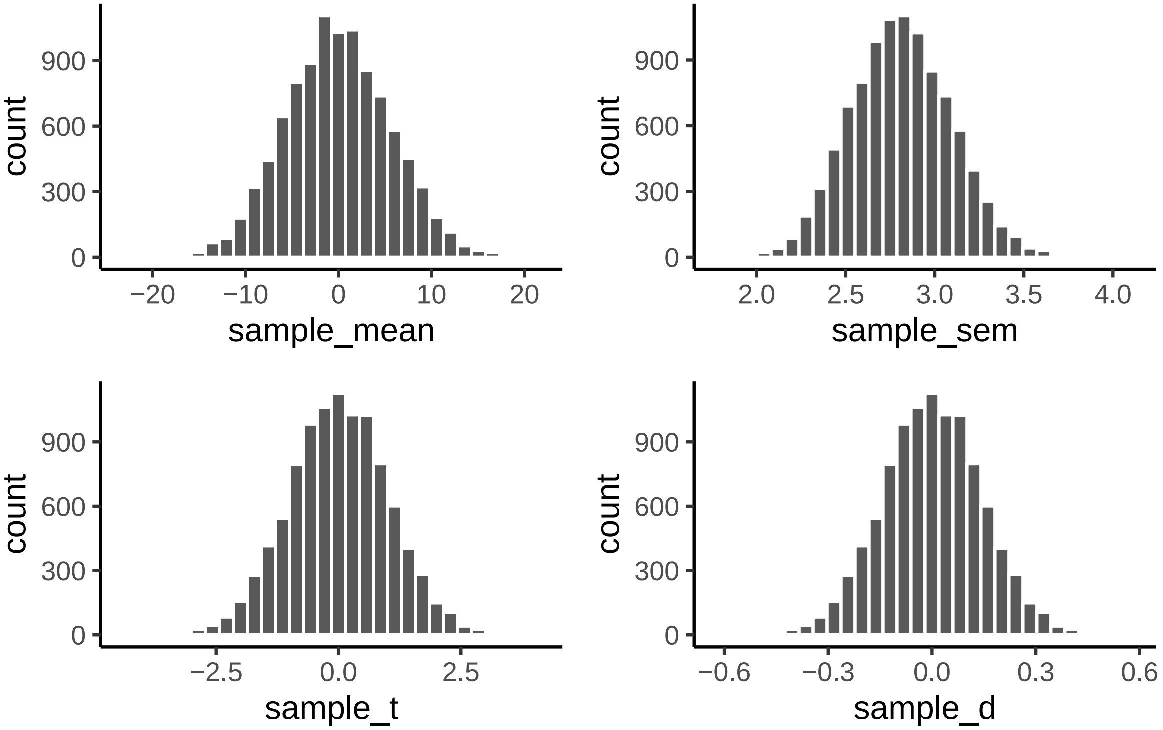


Figure 13.6: Each panel shows a histogram of a different sampling statistic.

14 GIFs

This is the place where I put the stats gifs as I make them. The gifs can downloaded from this page, or they can be downloaded from this folder on the github repo for this book https://github.com/CrumpLab/statistics/tree/master/gifs. Please feel free to use them however you wish. The source code for compiling the gifs in R is shown alongside each gif. The animations are made possible by the **gganimate** package.

This is a work in progress, subject to change and addition

14.1 Correlation GIFs

Note regression lines and confidence bands can be added using `geom_smooth(method=lm, se=T)`

14.1.1 N=10, both variables drawn from a uniform distribution

```
all_df<-data.frame()
for(sim in 1:10){
  North_pole <- runif(10,1,10)
  South_pole <- runif(10,1,10)
  t_df<-data.frame(simulation=rep(sim,10),
                                  North_pole,
                                  South_pole)
  all_df<-rbind(all_df,t_df)
}


ggplot(all_df,aes(x=North_pole,y=South_pole))+
  geom_point()+
  geom_smooth(method=lm, se=FALSE)+
  theme_classic()+
  transition_states(
    simulation,
    transition_length = 2,
    state_length = 1
  )+enter_fade() +
  exit_shrink() +
  ease_aes('sine-in-out')
```

14.1.2 Correlation between random deviates from uniform distribution across four sample sizes

N= 10,50,100,1000 All values sampled from a uniform distribution

```
all_df<-data.frame()
for(sim in 1:10){
  for(n in c(10,50,100,1000)){
  North_pole <- runif(n,1,10)
  South_pole <- runif(n,1,10)
  t_df<-data.frame(nsize=rep(n,n),
                   simulation=rep(sim,n),
                                  North_pole,
                                  South_pole)
  all_df<-rbind(all_df,t_df)
  }
}


ggplot(all_df,aes(x=North_pole,y=South_pole))+
  geom_point()+
  geom_smooth(method=lm, se=FALSE)+
  theme_classic()+
  facet_wrap(~nsize)+
  transition_states(
    simulation,
    transition_length = 2,
    state_length = 1
  )+enter_fade() +
  exit_shrink() +
  ease_aes('sine-in-out')
```

14.1.3 Correlation between random deviates from normal distribution across four sample sizes

N= 10,50,100,1000 All values sampled from the same normal distribution (mean=0, sd=1)

```
all_df<-data.frame()
for(sim in 1:10){
  for(n in c(10,50,100,1000)){
  North_pole <- rnorm(n,0,1)
  South_pole <- rnorm(n,0,1)
  t_df<-data.frame(nsize=rep(n,n),
                   simulation=rep(sim,n),
                                  North_pole,
                                  South_pole)
  all_df<-rbind(all_df,t_df)
  }
}


ggplot(all_df,aes(x=North_pole,y=South_pole))+
  geom_point()+
  geom_smooth(method=lm, se=FALSE)+
  theme_classic()+
  facet_wrap(~nsize)+
  transition_states(
    simulation,
    transition_length = 2,
    state_length = 1
  )+enter_fade() +
  exit_shrink() +
  ease_aes('sine-in-out')
```

14.1.4 Correlation between X and Y variables that have a true correlation as a function of sample-size

```
library(MASS)
r<-.7

proportional_permute<-function(x,prop){
  indices<-seq(1:length(x))
  s_indices<-sample(indices)
  n_shuffle<-round(length(x)*prop)
  switch<-sample(indices)
  x[s_indices[1:n_shuffle]]<-x[switch[1:n_shuffle]]
  return(x)
}

all_df<-data.frame()
for(sim in 1:10){
  for(samples in c(10,50,100,1000)){
    #data <- mvrnorm(n=samples, mu=c(0, 0), Sigma=matrix(c(1, r, r, 1), nrow=2), empirical=T
    #North_pole <- data[, 1]  # standard normal (mu=0, sd=1)
    #South_pole <- data[, 2]

    North_pole <- runif(samples,1,10)
    South_pole <- proportional_permute(North_pole,.5)+runif(samples,-5,5)

    t_df<-data.frame(nsize=rep(samples,samples),
                   simulation=rep(sim,samples),
                                  North_pole,
                                  South_pole)
  all_df<-rbind(all_df,t_df)
  }
}

ggplot(all_df,aes(x=North_pole,y=South_pole))+
  geom_point()+
  geom_smooth(method=lm, se=FALSE)+
  theme_classic()+
  facet_wrap(~nsize)+
  transition_states(
    simulation,
    transition_length = 2,
    state_length = 1
  )+enter_fade() +
  exit_shrink() +
  ease_aes('sine-in-out')
```

14.1.5 Type I errors, sampling random deviates from normal distribution with regression lines

These scatter plots only show what would be type I errors (assuming alpha=.05). The X and Y values were both sampled from the same normal distribution (mean = 0, sd=1). 1000 simulations were conducted for each sample size (10,50,100,1000). For each, the animation shows 10 scatter plots where the observed "correlation" would have passed a significance test. According to definition, these correlations only arise from random normal deviates 5% of the time, but when they do arise for small sample sizes, they look fairly convincing.

```r
all_df<-data.frame()
for(n in c(10,50,100,1000)){
  count_sims<-0
  for(sim in 1:1000){
    North_pole <- rnorm(n,0,1)
    South_pole <- rnorm(n,0,1)
    if(cor.test(North_pole,South_pole)$p.value<.05){
      count_sims<-count_sims+1
    t_df<-data.frame(nsize=rep(n,n),
                     simulation=rep(count_sims,n),
                     North_pole,
                     South_pole)
    all_df<-rbind(all_df,t_df)

    if(count_sims==10){
      break
    }
    }
  }
}


ggplot(all_df,aes(x=North_pole,y=South_pole))+
  geom_point()+
  geom_smooth(method=lm, se=TRUE)+
  theme_classic()+
  facet_wrap(~nsize)+
  transition_states(
    simulation,
    transition_length = 2,
    state_length = 1
  )+enter_fade() +
  exit_shrink() +
  ease_aes('sine-in-out')
```

14.1.6 Cell-size and correlation

This simulation illustrates how the behavior of correlating two random normal samples as a function of cell-size. The sample-size is always set at N=10. For each panel, the simulation uses an increasing cell-size to estimate the mean for X and Y. When cell-size is 1, 10 X and Y values are drawn from the same normal (u=0, sd=1). When cell-size is 5, for each X,Y score in the plot, 5 samples were drawn from the same normal, and then the mean of the samples is plotted. The effect of cell-size shrinks the dot cloud, as both X and Y scores provide better estimates of the population mean = 0. Cell-size has no effect on the behavior of r, which swings around because sample-size N is small. These are all random, so there is always a 5% type I error rate (alpha =.05).

```r
get_sampling_means<-function(m,sd,cell_size,s_size){
  save_means<-length(s_size)
  for(i in 1:s_size){
    save_means[i]<-mean(rnorm(cell_size,m,sd))
  }
  return(save_means)
}

all_df<-data.frame()
for(n in c(1,5,10,100)){
  count_sims<-0
  for(sim in 1:10){
    North_pole <- get_sampling_means(0,1,n,10)
    South_pole <- get_sampling_means(0,1,n,10)
      count_sims<-count_sims+1
      t_df<-data.frame(nsize=rep(n,10),
                       simulation=rep(count_sims,10),
                       North_pole,
                       South_pole)
      all_df<-rbind(all_df,t_df)
  }
}


ggplot(all_df,aes(x=North_pole,y=South_pole))+
  geom_point()+
  geom_smooth(method=lm, se=TRUE)+
  theme_classic()+
  facet_wrap(~nsize)+
  ggtitle("Random scatterplots, N=10, Cell-size = 1,5,10,100")+
  transition_states(
    simulation,
    transition_length = 2,
    state_length = 1
  )+enter_fade() +
  exit_shrink() +
  ease_aes('sine-in-out')
```

14.1.7 Regression

We look at how the residuals (error from points to line) behave as the regression lines moves above and below it's true value. The total error associated with all of the red lines is represents by the grey area. This total error is smallest (minimized) when the black line overlaps with the blue regression line (the best fit line). The total error expands as the black line moves away from the regression. That's why the regression line is the least wrong (best fit) line to skewer the data (according to least squares definition)

```
d <- mtcars
fit <- lm(mpg ~ hp, data = d)
d$predicted <- predict(fit)   # Save the predicted values
d$residuals <- residuals(fit) # Save the residual values

coefs<-coef(lm(mpg ~ hp, data = mtcars))
coefs[1]
coefs[2]

x<-d$hp
move_line<-c(seq(-6,6,.5),seq(6,-6,-.5))
total_error<-length(length(move_line))
cnt<-0
for(i in move_line){
  cnt<-cnt+1
  predicted_y <- coefs[2]*x + coefs[1]+i
  error_y <- (predicted_y-d$mpg)^2
  total_error[cnt]<-sqrt(sum(error_y)/32)
}
```

```
move_line_sims<-rep(move_line,each=32)
total_error_sims<-rep(total_error,each=32)
sims<-rep(1:50,each=32)

d<-d %>% slice(rep(row_number(), 50))

d<-cbind(d,sims,move_line_sims,total_error_sims)


anim<-ggplot(d, aes(x = hp, y = mpg, frame=sims)) +
  geom_smooth(method = "lm", se = FALSE, color = "lightblue") +
  geom_abline(intercept = 30.09886+move_line_sims, slope = -0.06822828)+
  lims(x = c(0,400), y = c(-10,40))+
  geom_segment(aes(xend = hp, yend = predicted+move_line_sims, color="red"), alpha = .5) +
  geom_point() +
  geom_ribbon(aes(ymin = predicted+move_line_sims - total_error_sims, ymax = predicted+move_
  theme_classic()+
  theme(legend.position="none")+
  xlab("X")+ylab("Y")+
  transition_manual(frames=sims)+
  enter_fade() +
  exit_fade()+
  ease_aes('sine-in-out')

animate(anim,fps=5)
```

14.2 Sampling distributions

14.2.1 Sampling from a uniform distribution

Animation shows histograms for N=20, sampled from a uniform distribution, along with mean (red line). Uniform distribution in this case is integer values from 1 to 10.

```
a<-round(runif(20*10,1,10))
df<-data.frame(a,sample=rep(1:10,each=20))
df2<-aggregate(a~sample,df,mean)
df<-cbind(df,mean_loc=rep(df2$a,each=20))

library(gganimate)
```

```r
ggplot(df,aes(x=a, group=sample,frame=sample)) +
  geom_histogram() +
  geom_vline(aes(xintercept=mean_loc,frame = sample),color="red")+
  scale_x_continuous(breaks=seq(1,10,1))+
  theme_classic()+
  transition_states(
    sample,
    transition_length = 2,
    state_length = 1
  )+enter_fade() +
  exit_shrink() +
  ease_aes('sine-in-out')
```

14.2.2 Sampling from uniform with line showing expected value for each number

```r
a<-round(runif(20*10,1,10))
df<-data.frame(a,sample=rep(1:10,each=20))


library(gganimate)
ggplot(df,aes(x=a))+
  geom_histogram(bins=10, color="white")+
  theme_classic()+
  scale_x_continuous(breaks=seq(1,10,1))+
  geom_hline(yintercept=2)+
  ggtitle("Small N=20 samples from a uniform distribution")+
  transition_states(
    sample,
    transition_length = 2,
    state_length = 1
  )+enter_fade() +
  exit_shrink() +
  ease_aes('sine-in-out')
```

14.2.3 Sampling distribution of the mean, Normal population distribution and sample histograms

This animation illustrates the relationship between a distribution (population), samples from the distribution, and the sampling distribution of the sample means, all as a function of n

Normal distribution in red. Individual sample histograms in grey. Vertical red line is mean of individual sample. Histograms for sampling distribution of the sample mean in blue. Vertical blue line is mean of the sampling distribution of the sample mean.

Note: for purposes of the animation (and because it was easier to do this way), the histograms for the sampling distribution of the sample means have different sizes. When sample-size = 10, the histogram shows 10 sample means. When sample size=100, the histogram shows 100 sample means. I could have simulated many more sample means (say 10000) for each, but then the histograms for the sample means would be static.

The y-axis is very rough. The heights of the histograms and distributions were scaled to be in the same range for the animation.

```
get_sampling_means<-function(m,sd,s_size){
  save_means<-length(s_size)
  for(i in 1:s_size){
    save_means[i]<-mean(rnorm(s_size,m,sd))
  }
  return(save_means)
}

all_df<-data.frame()
for(sims in 1:10){
  for(n in c(10,50,100,1000)){
    sample<-rnorm(n,0,1)
    sample_means<-get_sampling_means(0,1,n)
    t_df<-data.frame(sims=rep(sims,n),
                     sample,
                     sample_means,
                     sample_size=rep(n,n),
                     sample_mean=rep(mean(sample),n),
                     sampling_mean=rep(mean(sample_means),n)
                     )
    all_df<-rbind(all_df,t_df)
  }
}


ggplot(all_df, aes(x=sample))+
  geom_histogram(aes(y=(..density..)/max(..density..)^.8),color="white",fill="grey")+
  geom_histogram(aes(x=sample_means,y=(..density..)/max(..density..)),fill="blue",color="whi
  stat_function(fun = dnorm,
                args = list(mean = 0, sd = 1),
                lwd = .75,
                col = 'red')+
  geom_vline(aes(xintercept=sample_mean,frame=sims),color="red")+
  geom_vline(aes(xintercept=sampling_mean,frame=sims),color="blue")+
  facet_wrap(~sample_size)+xlim(-3,3)+
  theme_classic()+ggtitle("Population (red), Samples (grey), \n and Sampling distribution of
  xlab("value")+
  transition_states(
    sims,
    transition_length = 2,
    state_length = 1
  )+enter_fade() +
  exit_shrink() +
  ease_aes('sine-in-out')
```

14.2.4 Null and True effect samples and sampling means

The null dots show 50 different samples, with the red dot as the mean for each sample. Null dots are all sampled from normal (u=0, sd=1). The true dots show 50 more samples, with red dots for their means. However, the mean of the true shifts between -1.5 and +1.5 standard deviations of 0. This illustrates how a true effect moves in and out of the null range.

```
all_df<-data.frame()
all_df_means<-data.frame()
dif_sim<-seq(-1.5,1.5,.25)
for(sim in 1:13){
  values<-c(rnorm(25*25,0,1),rnorm(25*25,dif_sim[sim],1))
  samples<-c(rep(seq(1:25),each=25),rep(seq(1:25),each=25))
  df<-data.frame(samples,values,sims=rep(sim,50*25),type=rep(c("null","true"),each=625))
  df_means<-aggregate(values~samples*type,df,mean, sims=rep(sim,50))
  all_df<-rbind(all_df,df)
  all_df_means<-rbind(all_df_means,df_means)
}

all_df<-cbind(all_df,means=rep(all_df_means$values,each=25))

ggplot(all_df,aes(y=values,x=samples))+
  geom_point(aes(color=abs(values)), alpha=.25)+
  geom_point(aes(y=means,x=samples),color="red")+
  theme_classic()+
  geom_vline(xintercept=25.5)+
  facet_wrap(~type)+
  geom_hline(yintercept=0)+
  theme(legend.position="none") +
  ggtitle("null=0, True effect moves from -1.5 sd to 1.5 sd")+
  transition_states(
    sims,
    transition_length = 2,
    state_length = 1
  )+enter_fade() +
  exit_shrink() +
  ease_aes('sine-in-out')
```

14.3 Statistical Inference

14.3.1 Randomization Test

This is an attempt at visualizing a randomization test. Samples are taken under two conditions of the IV (A and B). At the beginning of the animation, the original scores in the first condition are shown as green dots on the left, and the original scores in the second condition are the red dots on the right. The means for each group are the purple dots. During the randomization, the original scores are shuffled randomly between the two conditions. After each shuffle, two new means are computed and displayed as the yellow dots. This occurs either for all permutations, or for a large random sample of them. The animation shows the original scores being shuffled around across the randomizations (the colored dots switch their original condition, appearing from side to side).

For intuitive inference, one might look at the range of motion of the yellow dots. This is how the mean difference between group 1 and group 2 behaves under randomization. It's what chance can do. If the difference between the purple dots is well outside the range of motion of the yellow dots, then the mean difference observed in the beginning is not likely produced by chance.

```
study<-round(runif(10,80,100))
no_study<-round(runif(10,40,90))

study_df<-data.frame(student=seq(1:10),study,no_study)
mean_original<-data.frame(IV=c("studied","didnt_study"),
                          means=c(mean(study),mean(no_study)))
t_df<-data.frame(sims=rep(1,20),
                 IV=rep(c("studied","didnt_study"),each=10),
                 values=c(study,no_study),
                 rand_order=rep(c(0,1),each=10))

raw_df<-t_df
for(i in 2:10){
  new_index<-sample(1:20)
  t_df$values<-t_df$values[new_index]
  t_df$rand_order<-t_df$rand_order[new_index]
  t_df$sims<-rep(i,20)
  raw_df<-rbind(raw_df,t_df)
}

raw_df$rand_order<-as.factor(raw_df$rand_order)
rand_df<-aggregate(values~sims*IV,raw_df,mean)
names(rand_df)<-c("sims","IV","means")


a<-ggplot(raw_df,aes(x=IV,y=values,color=rand_order,size=3))+
  geom_point(stat="identity",alpha=.5)+
  geom_point(data=mean_original,aes(x=IV,y=means),stat="identity",shape=21,size=6,color="blac
  geom_point(data=rand_df,aes(x=IV,y=means),stat="identity",shape=21,size=6,color="black",fil
  theme_classic(base_size = 15)+
  coord_cartesian(ylim=c(40, 100))+
  theme(legend.position="none") +
  ggtitle("Randomization test: Original Means (purple),
          \n Randomized means (yellow)
          \n Original scores (red,greenish)")+
  transition_states(
    sims,
    transition_length = 1,
    state_length = 2
  )+enter_fade() +
  exit_shrink() +
  ease_aes('sine-in-out')

animate(a,nframes=100,fps=5)
```

14.3.2 Independent t-test Null

This is a simulation of the null distribution for an independent samples t-test, two groups, 10 observations per group.

This animation has two panels. The left panel shows means for group A and B, sampled from the same normal distribution (mu=50, sd =10). The dots represent individual scores for each of 10 observations per group.

The right panel shows a t-distribution (df=18) along with the observed t-statistic for each simulation.

`gganimate` does not yet directly support multiple panels as shown in this gif. I hacked together these two gifs using the `magick` package. Apologies for the hackiness.

```r
library(dplyr)
library(ggplot2)
library(magick)
library(gganimate)

A<-rnorm(100,50,10)
B<-rnorm(100,50,10)
DV <- c(A,B)
IV <- rep(c("A","B"),each=100)
sims <- rep(rep(1:10,each=10),2)
df<-data.frame(sims,IV,DV)

means_df <- df %>%
               group_by(sims,IV) %>%
               summarize(means=mean(DV),
                         sem = sd(DV)/sqrt(length(DV)))

stats_df <- df %>%
              group_by(sims) %>%
              summarize(ts = t.test(DV~IV,var.equal=TRUE)$statistic)

a<-ggplot(means_df, aes(x=IV,y=means, fill=IV))+
  geom_bar(stat="identity")+
  geom_point(data=df,aes(x=IV, y=DV), alpha=.25)+
  geom_errorbar(aes(ymin=means-sem, ymax=means+sem),width=.2)+
  theme_classic()+
  transition_states(
    states=sims,
    transition_length = 2,
    state_length = 1
  )+enter_fade() +
  exit_shrink() +
  ease_aes('sine-in-out')
```

```
a_gif<-animate(a, width = 240, height = 240)

b<-ggplot(stats_df,aes(x=ts))+
  geom_vline(aes(xintercept=ts, frame=sims))+
  geom_line(data=data.frame(x=seq(-5,5,.1),
                            y=dt(seq(-5,5,.1),df=18)),
            aes(x=x,y=y))+
  theme_classic()+
  ylab("density")+
  xlab("t value")+
  transition_states(
    states=sims,
    transition_length = 2,
    state_length = 1
  )+enter_fade() +
  exit_shrink() +
  ease_aes('sine-in-out')

b_gif<-animate(b, width = 240, height = 240)


d<-image_blank(240*2,240)

the_frame<-d
for(i in 2:100){
  the_frame<-c(the_frame,d)
}

a_mgif<-image_read(a_gif)
b_mgif<-image_read(b_gif)

new_gif<-image_append(c(a_mgif[1], b_mgif[1]))
for(i in 2:100){
  combined <- image_append(c(a_mgif[i], b_mgif[i]))
  new_gif<-c(new_gif,combined)
}

new_gif
```

14.3.3 Independent t-test True

This is a simulation of an independent samples t-test, two groups, 10 observations per group, assuming a true difference of 2 standard deviations between groups

This animation has two panels. The left panel shows means for group A (normal, mu=50, sd=10) and B (normal, mu=70, sd=10). The dots represent individual scores for each of 10 observations per group.

The right panel shows a t-distribution (df=18) along with the observed t-statistic for each simulation.

```r
library(dplyr)
library(ggplot2)
library(magick)
library(gganimate)

A<-rnorm(100,70,10)
B<-rnorm(100,50,10)
DV <- c(A,B)
IV <- rep(c("A","B"),each=100)
sims <- rep(rep(1:10,each=10),2)
df<-data.frame(sims,IV,DV)

means_df <- df %>%
               group_by(sims,IV) %>%
               summarize(means=mean(DV),
                         sem = sd(DV)/sqrt(length(DV)))

stats_df <- df %>%
              group_by(sims) %>%
              summarize(ts = t.test(DV~IV,var.equal=TRUE)$statistic)

a<-ggplot(means_df, aes(x=IV,y=means, fill=IV))+
  geom_bar(stat="identity")+
  geom_point(data=df,aes(x=IV, y=DV), alpha=.25)+
  geom_errorbar(aes(ymin=means-sem, ymax=means+sem),width=.2)+
  theme_classic()+
  transition_states(
    states=sims,
    transition_length = 2,
    state_length = 1
  )+enter_fade() +
  exit_shrink() +
  ease_aes('sine-in-out')

a_gif<-animate(a, width = 240, height = 240)

b<-ggplot(stats_df,aes(x=ts))+
  geom_vline(aes(xintercept=ts, frame=sims))+
  geom_vline(xintercept=qt(c(.025, .975), df=18),color="green")+
  geom_line(data=data.frame(x=seq(-5,5,.1),
                            y=dt(seq(-5,5,.1),df=18)),
            aes(x=x,y=y))+
  theme_classic()+
  ylab("density")+
  xlab("t value")+
  transition_states(
    states=sims,
    transition_length = 2,
    state_length = 1
  )+enter_fade() +
  exit_shrink() +
  ease_aes('sine-in-out')

b_gif<-animate(b, width = 240, height = 240)


d<-image_blank(240*2,240)

the_frame<-d
for(i in 2:100){
  the_frame<-c(the_frame,d)
}

a_mgif<-image_read(a_gif)
b_mgif<-image_read(b_gif)

new_gif<-image_append(c(a_mgif[1], b_mgif[1]))
for(i in 2:100){
  combined <- image_append(c(a_mgif[i], b_mgif[i]))
  new_gif<-c(new_gif,combined)
}

new_gif
```

14.3.4 T-test True sample-size

The top row shows 10 simulations of an independent sample t-test, with N=10, and true difference of 1 sd.

The bottom row shows 10 simulations with N=50.

The observed t-value occurs past the critical value (green) line much more reliably and often when sample size is larger than smaller.

```r
library(dplyr)
library(ggplot2)
library(magick)
library(gganimate)

A<-rnorm(100,60,10)
B<-rnorm(100,50,10)
DV <- c(A,B)
IV <- rep(c("A","B"),each=100)
sims <- rep(rep(1:10,each=10),2)
df<-data.frame(sims,IV,DV)

means_df <- df %>%
               group_by(sims,IV) %>%
               summarize(means=mean(DV),
                         sem = sd(DV)/sqrt(length(DV)))

stats_df <- df %>%
              group_by(sims) %>%
              summarize(ts = t.test(DV~IV,var.equal=TRUE)$statistic)

a<-ggplot(means_df, aes(x=IV,y=means, fill=IV))+
  geom_bar(stat="identity")+
  geom_point(data=df,aes(x=IV, y=DV), alpha=.25)+
  geom_errorbar(aes(ymin=means-sem, ymax=means+sem),width=.2)+
  theme_classic()+
  transition_states(
    states=sims,
    transition_length = 2,
    state_length = 1
  )+enter_fade() +
  exit_shrink() +
  ease_aes('sine-in-out')
```

```
a_gif<-animate(a, width = 240, height = 240)

b<-ggplot(stats_df,aes(x=ts))+
  geom_vline(aes(xintercept=ts, frame=sims))+
  geom_vline(xintercept=qt(c(.025, .975), df=18),color="green")+
  geom_line(data=data.frame(x=seq(-5,5,.1),
                            y=dt(seq(-5,5,.1),df=18)),
            aes(x=x,y=y))+
  theme_classic()+
  ylab("density")+
  xlab("t value")+
  transition_states(
    states=sims,
    transition_length = 2,
    state_length = 1
  )+enter_fade() +
  exit_shrink() +
  ease_aes('sine-in-out')

b_gif<-animate(b, width = 240, height = 240)


d<-image_blank(240*2,240)

the_frame<-d
for(i in 2:100){
  the_frame<-c(the_frame,d)
}

a_mgif<-image_read(a_gif)
b_mgif<-image_read(b_gif)

new_gif<-image_append(c(a_mgif[1], b_mgif[1]))
for(i in 2:100){
  combined <- image_append(c(a_mgif[i], b_mgif[i]))
  new_gif<-c(new_gif,combined)
}

new_gif

## increase sample-size
```

```
A<-rnorm(50*10,60,10)
B<-rnorm(50*10,50,10)
DV <- c(A,B)
IV <- rep(c("A","B"),each=50*10)
sims <- rep(rep(1:10,each=50),2)
df<-data.frame(sims,IV,DV)

means_df <- df %>%
               group_by(sims,IV) %>%
               summarize(means=mean(DV),
                         sem = sd(DV)/sqrt(length(DV)))

stats_df <- df %>%
              group_by(sims) %>%
              summarize(ts = t.test(DV~IV,var.equal=TRUE)$statistic)

a<-ggplot(means_df, aes(x=IV,y=means, fill=IV))+
  geom_bar(stat="identity")+
  geom_point(data=df,aes(x=IV, y=DV), alpha=.25)+
  geom_errorbar(aes(ymin=means-sem, ymax=means+sem),width=.2)+
  theme_classic()+
  transition_states(
    states=sims,
    transition_length = 2,
    state_length = 1
  )+enter_fade() +
  exit_shrink() +
  ease_aes('sine-in-out')

a_gif<-animate(a, width = 240, height = 240)

b<-ggplot(stats_df,aes(x=ts))+
  geom_vline(aes(xintercept=ts, frame=sims))+
  geom_vline(xintercept=qt(c(.025, .975), df=98),color="green")+
  geom_line(data=data.frame(x=seq(-5,5,.1),
                            y=dt(seq(-5,5,.1),df=98)),
            aes(x=x,y=y))+
  theme_classic()+
  ylab("density")+
  xlab("t value")+
  transition_states(
    states=sims,
    transition_length = 2,
    state_length = 1
  )+enter_fade() +
  exit_shrink() +
  ease_aes('sine-in-out')

b_gif<-animate(b, width = 240, height = 240)


d<-image_blank(240*2,240)

the_frame<-d
for(i in 2:100){
  the_frame<-c(the_frame,d)
}

a_mgif<-image_read(a_gif)
b_mgif<-image_read(b_gif)

new_gif2<-image_append(c(a_mgif[1], b_mgif[1]))
for(i in 2:100){
  combined <- image_append(c(a_mgif[i], b_mgif[i]))
  new_gif2<-c(new_gif2,combined)
}

## add new row

final_gif <- image_append(c(new_gif[1], new_gif2[1]),stack=TRUE)
for(i in 2:100){
  combined <- image_append(c(new_gif[i], new_gif2[i]),stack=TRUE)
  final_gif<-c(final_gif,combined)
}

final_gif
```

14.3.5 one-factor ANOVA Null

Three groups, N=10, all observations sampled from same normal distribution (mu=50, sd = 10)

```r
library(dplyr)
library(ggplot2)
library(magick)
library(gganimate)


A<-rnorm(100,50,10)
B<-rnorm(100,50,10)
C<-rnorm(100,50,10)
DV <- c(A,B,C)
IV <- rep(rep(c("A","B","C"),each=10),10)
sims <- rep(1:10,each=30)
df<-data.frame(sims,IV,DV)

means_df <- df %>%
  group_by(sims,IV) %>%
  summarize(means=mean(DV),
            sem = sd(DV)/sqrt(length(DV)))

stats_df <- df %>%
  group_by(sims) %>%
  summarize(Fs = summary(aov(DV~IV))[[1]][[4]][1])

a<-ggplot(means_df, aes(x=IV,y=means, fill=IV))+
  geom_bar(stat="identity")+
  geom_point(data=df,aes(x=IV, y=DV), alpha=.25)+
  geom_errorbar(aes(ymin=means-sem, ymax=means+sem),width=.2)+
  theme_classic(base_size = 20)+
  transition_states(
    states=sims,
    transition_length = 2,
    state_length = 1
  )+enter_fade() +
  exit_shrink() +
  ease_aes('sine-in-out')

b<-ggplot(stats_df,aes(x=Fs))+
  geom_vline(aes(xintercept=Fs))+
  geom_vline(xintercept=qf(.95, df1=2,df2=27),color="green")+
  geom_line(data=data.frame(x=seq(0,6,.1),
                            y=df(seq(0,6,.1),df1=2,df2=27)),
            aes(x=x,y=y))+
  theme_classic(base_size = 20)+
  ylab("density")+
  xlab("F value")+
  transition_states(
    states=sims,
    transition_length = 2,
    state_length = 1
  )+enter_fade() +
  exit_shrink() +
  ease_aes('sine-in-out')

a_gif<-animate(a,width=480,height=480)
b_gif<-animate(b,width=480,height=480)

a_mgif<-image_read(a_gif)
b_mgif<-image_read(b_gif)

new_gif<-image_append(c(a_mgif[1], b_mgif[1]))
for(i in 2:100){
  combined <- image_append(c(a_mgif[i], b_mgif[i]))
  new_gif<-c(new_gif,combined)
}

new_gif
```

14.3.6 Factorial Null

10 simulations, N=10 in each of 4 conditions in a 2x2 (between-subjects). All observations taken from the same normal distribution (mu=50, sd =10).

```
A<-rnorm(100,50,10)
B<-rnorm(100,50,10)
C<-rnorm(100,50,10)
D<-rnorm(100,50,10)
DV <- c(A,B,C,D)
IV1 <- rep(c("A","B"),each=200)
IV2<-rep(rep(c("1","2"),each=100),2)
sims <- rep(1:10,40)
df<-data.frame(sims,IV1,IV2,DV)

means_df <- df %>%
  group_by(sims,IV1,IV2) %>%
  summarize(means=mean(DV),
            sem = sd(DV)/sqrt(length(DV)))

stats_df <- df %>%
  group_by(sims) %>%
  summarize(FIV1 = summary(aov(DV~IV1*IV2))[[1]][[4]][1],
            FIV2 = summary(aov(DV~IV1*IV2))[[1]][[4]][2],
            F1x2 = summary(aov(DV~IV1*IV2))[[1]][[4]][3]
            )

a<-ggplot(means_df, aes(x=IV1,y=means,
                                           group=IV2,
                                           color=IV2))+
  geom_point(data=df,aes(x=IV1, y=DV,group=IV2),
             position=position_dodge(width=.2),
             size=2,
             alpha=.25)+
  geom_point(size=4)+
  geom_line(size=1.3)+
  geom_errorbar(aes(ymin=means-sem, ymax=means+sem),width=.2,
                color="black")+
  theme_classic(base_size = 20)+
  transition_states(
    states=sims,
    transition_length = 2,
    state_length = 1
  )+enter_fade() +
  exit_shrink() +
  ease_aes('sine-in-out')

b<-ggplot(stats_df,aes(x=FIV1))+
  geom_vline(aes(xintercept=FIV1),color="red",size=1.2)+
  geom_vline(aes(xintercept=FIV2),color="blue",size=1.2)+
  geom_vline(aes(xintercept=F1x2),color="purple",size=1.2)+
  geom_vline(xintercept=qf(.95, df1=1,df2=36),color="green",size=1.2)+
  geom_line(data=data.frame(x=seq(0,20,.1),
                            y=df(seq(0,20,.1),df1=1,df2=36)),
            aes(x=x,y=y))+
  theme_classic(base_size = 20)+
  ylab("density")+
  xlab("F value")+
  ggtitle(label="",subtitle="red=IV1, blue=IV2, \n purple=Interaction")+
  transition_states(
    states=sims,
    transition_length = 2,
    state_length = 1
  )

a_gif<-animate(a,width=480,height=480)
b_gif<-animate(b,width=480,height=480)

a_mgif<-image_read(a_gif)
b_mgif<-image_read(b_gif)

new_gif<-image_append(c(a_mgif[1], b_mgif[1]))
for(i in 2:100){
  combined <- image_append(c(a_mgif[i], b_mgif[i]))
  new_gif<-c(new_gif,combined)
}

image_animate(new_gif, fps = 10,dispose="none")
```

14.4 Distributions

14.4.1 Normal changing mean

```
some_means<-c(0,1,2,3,4,5,4,3,2,1)
all_df<-data.frame()
for(i in 1:10){
  dnorm_vec <- dnorm(seq(-10,10,.1),mean=some_means[i],sd=1)
  x_range   <- seq(-10,10,.1)
  means <- rep(some_means[i], length(x_range))
  sims <- rep(i, length(x_range))
  t_df<-data.frame(sims,means,x_range,dnorm_vec)
  all_df<-rbind(all_df,t_df)
}

ggplot(all_df, aes(x=x_range,y=dnorm_vec))+
  geom_line()+
  theme_classic()+
  ylab("probability density")+
  xlab("value")+
  ggtitle("Normal Distribution with changing Mean")+
   transition_states(
    sims,
    transition_length = 1,
    state_length = 1
  )
  #enter_fade() +
  #exit_shrink() +
  #ease_aes('sine-in-out')
```

14.4.2 Normal changing sd

```
some_sds<-seq(0.5,5,.5)
all_df<-data.frame()
for(i in 1:10){
  dnorm_vec <- dnorm(seq(-10,10,.1),mean=0,sd=some_sds[i])
  x_range   <- seq(-10,10,.1)
  sds <- rep(some_sds[i], length(x_range))
  sims <- rep(i, length(x_range))
  t_df<-data.frame(sims,sds,x_range,dnorm_vec)
  all_df<-rbind(all_df,t_df)
}

labs_df<-data.frame(sims=1:10,
                    sds=as.character(seq(0.5,5,.5)))

ggplot(all_df, aes(x=x_range,y=dnorm_vec, frame=sims))+
  geom_line()+
  theme_classic()+
  ylab("probability density")+
  xlab("value")+
  ggtitle("Normal Distribution with changing sd")+
  geom_label(data = labs_df, aes(x = 5, y = .5, label = sds))+
   transition_states(
    sims,
    transition_length = 2,
    state_length = 1
  )+
  enter_fade() +
  exit_shrink() +
  ease_aes('sine-in-out')
```

References

Anscombe, F. J. 1973. "Graphs in Statistical Analysis." *American Statistician* 27: 17–21.

Barnes, Mallory L. 2023. *Statistics for Environmental Science*.

Behmer, Lawrence P, and Matthew JC Crump. 2017. "Spatial Knowledge During Skilled Action Sequencing: Hierarchical Versus Nonhierarchical Representations." *Attention, Perception, & Psychophysics* 79 (8): 2435–48. https://doi.org/10.3758/s13414-017-1389-3.

Campbell, D. T., and J. C. Stanley. 1963. *Experimental and Quasi-Experimental Designs for Research*. Boston, MA: Houghton Mifflin.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Second. Lawrence Erlbaum.

Fisher, R. A. 1922. "On the Mathematical Foundation of Theoretical Statistics." *Philosophical Transactions of the Royal Society A* 222: 309–68.

Hothersall, D. 2004. *History of Psychology*. McGraw-Hill.

James, Ella L, Michael B Bonsall, Laura Hoppitt, Elizabeth M Tunbridge, John R Geddes, Amy L Milton, and Emily A Holmes. 2015. "Computer Game Play Reduces Intrusive Memories of Experimental Trauma via Reconsolidation-Update Mechanisms." *Psychological Science* 26 (8): 1201–15. https://doi.org/10.1177/0956797615583071.

Keynes, John Maynard. 1923. *A Tract on Monetary Reform*. London: Macmillan and Company.

Matejka, Justin, and George Fitzmaurice. 2017. "Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics Through Simulated Annealing." In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 1290–94. ACM. https://doi.org/10.1145/3025453.3025912.

Maul, Andrew. 2017. "Rethinking Traditional Methods of Survey Validation." *Measurement: Interdisciplinary Research and Perspectives* 15 (2): 51–69. https://doi.org/10.1080/15366367.2017.1348108.

Meehl, P. H. 1967. "Theory Testing in Psychology and Physics: A Methodological Paradox." *Philosophy of Science* 34: 103–15. https://doi.org/10.1086/288135.

Mehr, Samuel A, Lee Ann Song, and Elizabeth S Spelke. 2016. "For 5-Month-Old Infants, Melodies Are Social." *Psychological Science* 27 (4): 486–501. https://doi.org/10.1177/0956797615626691.

Pfungst, O. 1911. *Clever Hans (The Horse of Mr. Von Osten): A Contribution to Experimental Animal and Human Psychology*. Translated by C. L. Rahn. New York: Henry Holt.

Salsburg, David. 2001. *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*. Macmillan.

Skilleter, G. A. 1996. “An Experimental Test of Artifacts from Repeated Sampling in Soft-Sediments.” *Journal of Experimental Marine Biology and Ecology* 205 (1): 137–48. https://doi.org/10.1016/S0022-0981(96)02617-2.

Student, A. 1908. “The Probable Error of a Mean.” *Biometrika* 6: 1–2.

von Kügelgen, Julius, Luigi Gresele, and Bernhard Schölkopf. 2021. “Simpson’s Paradox in COVID-19 Case Fatality Rates: A Mediation Analysis of Age-Related Causal Effects.” *IEEE Transactions on Artificial Intelligence* 2 (1): 18–27. https://doi.org/10.1109/TAI.2021.3073088.