Overview
Learning Objectives
Module Video
Random Variables and Sources of Variation
Types of Variables and Levels of Measurement
Major Types of Variables
Levels of Measurement
The Statistical Data Investigation Process
Major Types of Statistical Data Investigations
Surveys
Experiments
Observational or Correlational Studies
References

Module 1

Statistics - Dealing Confidently with Uncertainty

James Baglin

Last updated: 13 July, 2020

(https://www.flickr.com/photos/nicubunuphotos/5262645427/in/photolist-923r8Z-nyxedm-6YDv8R-n9TF7K-cJ1EZC-8Bxhpm-4GvDnN-8CHn7V-qK4Yi3-p7pAFK-rAVvmS-n4AFc-4qmbzr-oiMr1D-5XFXnm-qQLgfX-5BS2ug-7mMjSk-oDQQmk-5YiHnq-nVhd96-ctnn9d-kJDwuy-dL7Z1r-dLdtKy-sjZBe-dL7Xpn-avBiZM-DM9Lob-eHPCzg-rpowhA-ie9Nnf-4Ymw3A-7Z6agp-7wsKrw-9g66tY-8ctRb-pWSkJo-6x7aaT-ejLW8o-dLdrMW-nPa6st-aDFzyG-dL7YcX-3eSJiQ-g1bufo-dXqmf6-dZBApS-8Q4Rtb-3L8UiY)

Overview

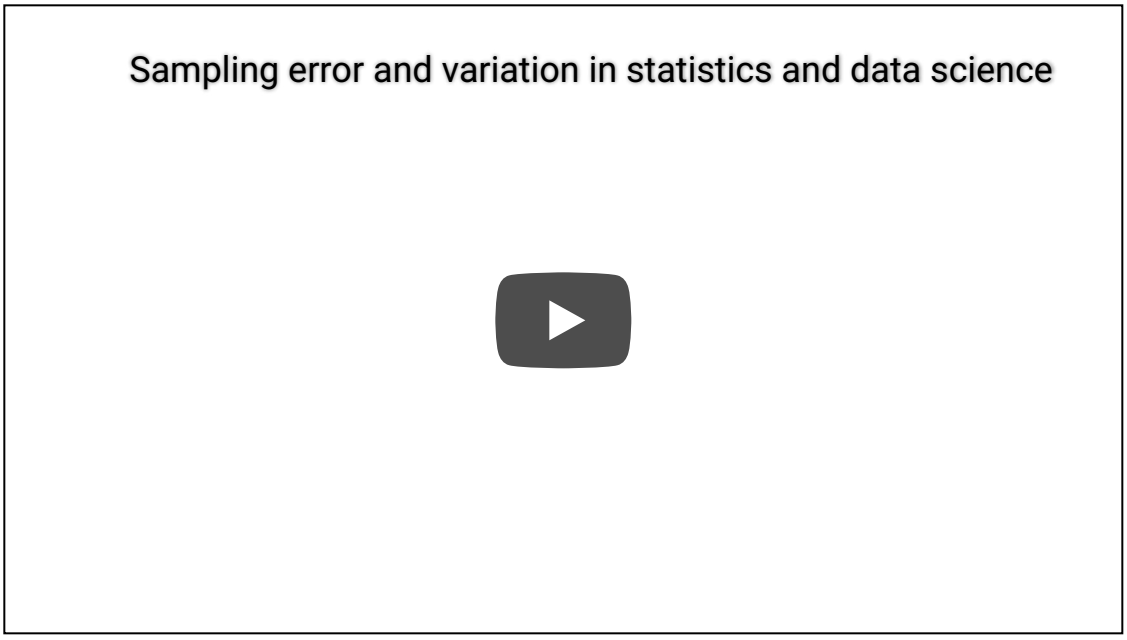
##Summary

Learning Objectives

The learning objectives associated with this module are:

- Define and understand the omnipresence of variability by stating major sources and types of variation.
- Define the discipline and science of statistics, and distinguish it from mathematics.
- Define a variable, their major types, and levels of measurement.
- Understand the basics stages of the statistical data investigation process.
- Understand the major types of statistical data investigations, namely, surveys, observational studies and experiments.
- Discuss the idea of statistical inference and how the use of samples give rise to uncertainty.
- Explore the central concepts of variability in a simple statistical data investigation.

Module Video



Random Variables and Sources of Variation

(https://www.flickr.com/photos/jamescridland/61344581O/in/photolist-Wd54U-he6YZR-gbKUXy-eT6YzG-JeuWPk-rnPWD6-shiEjn-pzFDUM-pJMr9Q-s7WH5k-qEo1W8-pEc8w8-oZ18Ec-nGNZPi-byz3v5-fnZeah-dDxSDz-onw5WB-fNhxD-2JUtd-a2XHk4-dGzHeJ-pJ6YbZ-aoAAvE-oVLQYD-9es325-dBvZA8-itN5fE-7Dwdq7-e9yGzU-ib2qiC-Dv28A-dyW8N8-6vLdsH-ct7RT3-o2QzWW-knkntv-d34hh3-H9R99w-doMAuS-e9dPqs-nnCdf8H-abkLWW-pJbHrQ-4CCfd9-i3vALF-H9BWgS-HvqoZc-e9F5R6-hZF26c)

Statistics, as defined by MacGillivray, Utts and Heckard (2014), is the “discipline and science of obtaining, understanding, modelling, interpreting, and using data in all real and complex systems and processes that involve uncertainty and variation” (p.15). Data, variation, and uncertainty are at the core of statistics. Statistical data refer to **variables**, which are defined as any characteristic or value of a unit that can change or vary. The idea of a **unit** is very broad and can refer to a person, a time point, a device, or system. Variation is all around us. This idea is referred to as the “omnipresence of variability” and is the reason why the field of statistics emerged. There are many forms of variation that you need understand. These can be summarised into four main categories.

- **Natural or Real Variation**: This refers to inherent, natural variability that is left over when all the other sources of variability are accounted for. Take, for example, a person’s height. Height varies greatly in the population, but there a many other variables that can explain why one person is taller than another. Males tend to be taller than females. Adults are taller than children. However, even if we compared males of a similar age, height will still vary. This is the natural or “real” variability that statistics seeks to measure and understand. Natural variability is what makes the world an interesting place.
- **Explainable Variation**: This is the variation in one variable that we can explain by considering another variable. The statistical tests and models that you will learn in this course seek to test relationships and effects that different variables have on each other. You already know heaps of examples of variables that “explain” other variables. For example, you know that height can help explain variation in weight, smoking can help us understand why some people are at a greater risk of lung cancer, gender can explain variation in the risk of heart disease, and the amount of hours spent studying can help explain exam scores.
- **Sampling Error**: Take a sample from a population, make a measurement and record the result. Now, repeat the process many times. The sample results will all

differ to a certain degree. This type of variability is known as sampling variability. Statistical inference and hypothesis testing, to be introduced in later modules, deals with this specific form of variability and the implications it has on drawing conclusions from our studies. We will also consider this important source of variation in an interesting demonstration at the end of this module.

- **Non-sampling Variation**: This refers to any artificial variability induced by the research process. As researchers, you try to understand real variability, while acknowledging, accounting or controlling for induced variability. Induced variability can come from many factors. The following are some common examples:
    - **Measurement**: Sometimes referred to as observational error, this is the variability associated with how we have measured our variables. All measures of a variable are imperfect. We need to understand the reliability and validity of our measurements to account for measurement variability. Measurement variability can come from two main sources:
        - *Measures*: There are usually many different ways to measure a variable. Different measures have different levels of reliability and validity. You should always use the most reliable and valid measure available. However, practical constraints may prevent this (e.g. time and money). All good measures will report their reliability and validity. If you create your own measure during the course of your research, you will need to test the reliability and validity yourself.
        - *Devices*: You may use the same type of measure, but use different devices to record your results. For example, using two different weighing scales to measure your samples. The problem is that different devices may introduce variability due to calibration, or natural variability between devices.
    - **Accident**: This is exactly what it sounds like, just silly mistakes that can invalidate your data. As researcher, we must do everything we can to reduce such mistakes. Accidents can happen at different levels.
        - *During collection*: You might write down the wrong measurement, make a typo, miss a question on a questionnaire or lose participant records.
        - *Processing*: Errors can be made when entering and saving data, when manipulating data, and when cleaning up the data. This is very annoying and is why you should include checks when processing your data.

Types of Variables and Levels of Measurement

A **measurement** occurs when a variable is actually measured or recorded. For example, you measure a person's heart rate (measurement). Heart rate is measured on a **scale**, or the unit of measurement, for example, beats per minute (BPM). A set of measurement or records from a variable is called **data** and can come from either a **sample**, a sub group of a population, or the **population** itself. When an entire population is measured, this is referred to as a **census**.

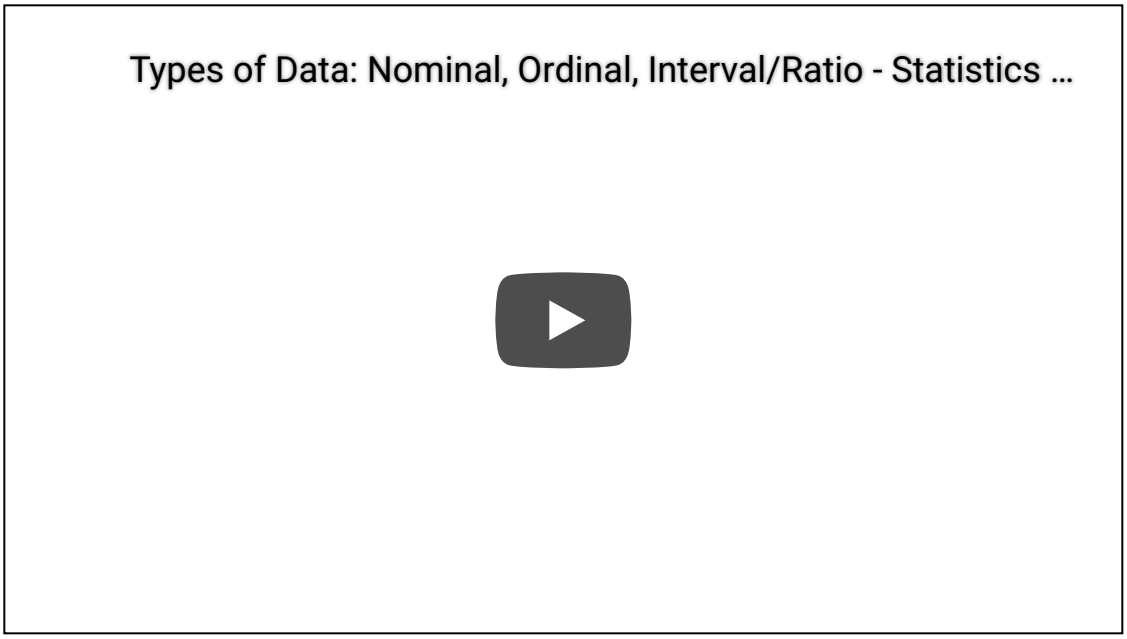
Major Types of Variables

The two major types of variables are **qualitative** and **quantitative**. The type of variables you collect and analyse have a direct bearing on the type of statistical summaries and analyses you can perform.

- **Qualitative** - Qualitative variables have different qualities, characteristics or categories, e.g. hair colour (black, brown, blonde,...), disease (cancer, heart disease,...), gender (male, female), country of birth (New Zealand, Japan,...). Qualitative variables are used to categorise quantitative data into groups or to tally frequencies of categories that can be converted to proportions and percentages.
- **Quantitative** - Quantitative variables measure a numerical quantity on each unit. Quantitative variables can be either **discrete** - can only assume a finite or countable number of values, e.g. marks on a test, birthday, number of people in a lecture, or **continuous** - the value can assume any value corresponding to the points on a number line, e.g. time (seconds), height (cm), weight (kg), age etc.

Levels of Measurement

When you measure a variable, qualitative and quantitative variables can take on different scales or levels of measurement. Levels of measurement have a direct bearing on the quantitative data analysis techniques you will need to use. We need to understand the language used to describe different scales. The following short video by Nicola Petty provides a great overview.



- **Categorical or Nominal (Qualitative)**. Categorical variables are group variables or categories if you will. There are no meaningful measurement differences such as rankings or intervals between the different categories. Categorical or nominal

variables include binary variables (e.g. yes/no, male/female) and multinomial variables (e.g. religious affiliation, hair colour, ethnicity, suburb).

- **Ordinal (Qualitative)**. Ordinal data has a rank order by which it can be sorted but the differences between the ranks are not relative or measurable. Therefore, ordinal data is not strictly quantitative. For example, consider the 1st, 2nd and 3rd place in a race. We know who was faster or slower, but we have no idea by how much. We need to look at the race times.
- **Interval (Quantitative)**: An interval variable is similar to an ordinal variable except that the intervals between the values of the interval scale are equally spaced. Interval variables have an arbitrary zero point and therefore no meaningful ratios. An example is our calendar year. 1000 AD is not half of 2000 AD, and 20 degrees Celsius is not twice as "hot" as 10 degrees Celsius. This is because our calendar and Celsius scale have an arbitrary value for zero. Zero AD and zero degrees Celsius do not imply the presence of zero time or zero heat energy.
- **Ratio (Quantitative)**: A ratio variable is similar to an interval variable; however there is an absolute zero point and ratios are meaningful. An example is time given in seconds, length in centimeters, or heart beats per minute. A value of 0 implies the absence of a variable. We can also make statements like 30 seconds is twice the time of 15 seconds, 10 cm is half the height of 20 cm, and during exercise a person's resting heart beat almost doubles. Zero heart rate, call 000!

#Statistical Inference - The Big Idea of Statistics

The idea of the following sections is to give you a glimpse into the big picture of this course, that is, statistical inference. What is statistical inference?

> *"Statistical inference moves beyond the data in hand to draw conclusions about some wider universe, taking into account that variation is everywhere and the conclusions are uncertain"* (Moore 2007, xxviii)

As such, statistics has been referred to as the discipline involved with dealing confidently with uncertainty. Wild, Pfannkuch and Horton (2011) provided the analogy to help explain the big idea behind statistical inference.

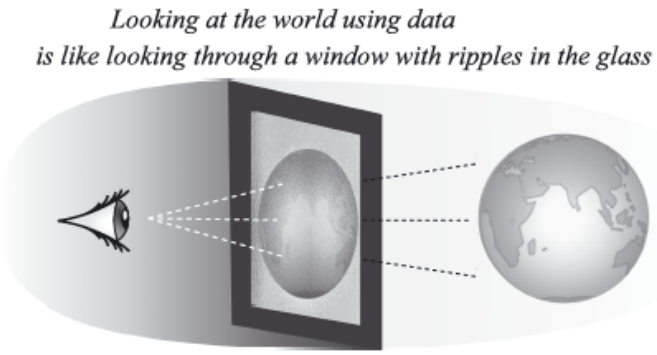


Fig. 2. 'What I see is not quite the way it really is'

Fig. 3. Distortions due to sampling

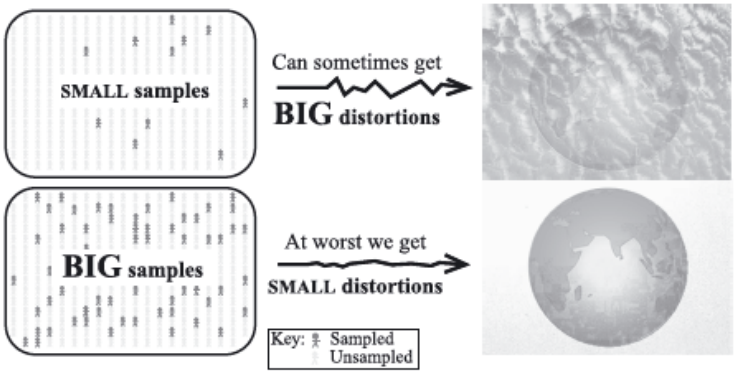


Fig. 4. Distortions related to sample size

Throughout this course you will develop a deeper understanding of statistical inference and the uncertainty associated with the use of samples. You will learn how samples impact data and the conclusions you draw. You will also learn how to measure and express your statistical uncertainty and confidently draw appropriate conclusions.

The Statistical Data Investigation Process

Statistical practice is much broader than analysing data. The statistical data investigation process describes how real problems are tackled from a statistical problem solving approach. It is how statistics is applied to investigate questions in science, medicine, agriculture, business, engineering, psychology, or anywhere data are needed and the data exhibit variation.

As discussed previously, variation is omnipresent. At almost all levels of government, industry, and science, data are measured, quantified, and interpreted in order to understand variation. By asking statistical questions of the data, taking observations and performing experiments, data can be used by investigators to seek patterns amongst great variation.

The entire process of a statistical data investigation involves everything from initial thoughts, through to planning, collecting and exploring data, and reporting findings. The process is depicted and summarised in the following slideshow, along with a brief description and key considerations at each stage. Click on each stage to read more.

As you work through the statistical data investigation process, its useful to apply statistical habits of mind. Some useful examples proposed by Hollylynne Lee and Dung Tran (https://place.fi.ncsu.edu/local/catalog/course.php?id=4) include:

- Always consider the context of data
- Ensure the best measure of an attribute of interest is used
- Anticipate, look for, and describe variation
- Attend to sampling issues
- Embrace uncertainty, but build confidence in interpretations
- Use several visual and numerical representations to make sense of data
- Be a skeptic throughout an investigation

Sometimes a data investigation starts with a question, sometimes a hypothesis, sometimes a problem, and sometimes just a general situation to be explored. Statistical questions and problems are not the same as mathematical questions and problems. Don't confuse the two. Tukey (1953), a world famous statistician, best explained this when he wrote:

"Statistics is a science in my opinion, and it is no more a branch of mathematics than are physics, chemistry, and economics; for if its methods fail the test of experience – not the test of logic – they are discarded."

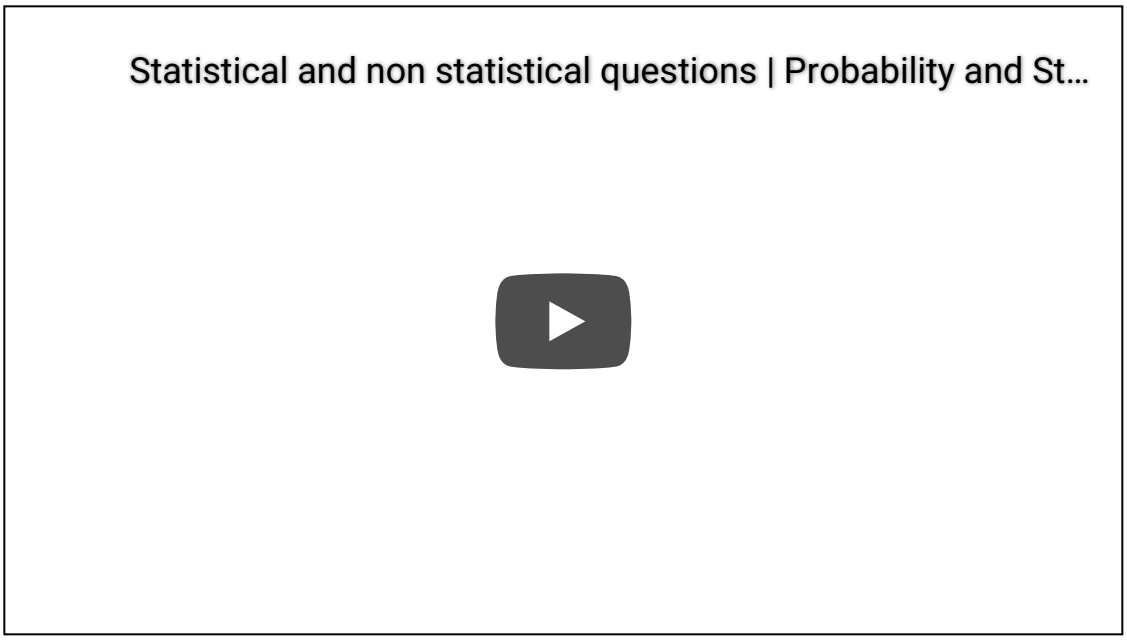
Furthermore, statistical questions can be differentiated based on the following:

- The use of context and data collection
- Measurement decisions
- Omnipresence of variability
- Dealing with uncertainty

In contrast, mathematics questions are characterised by the following:

- Problems can exist without context
- Measurements are assumed to be exact.
- No variability
- Deterministic answers

The following video from the Khan Academy explores how statistical questions are fundamentally different to maths questions.



Major Types of Statistical Data Investigations

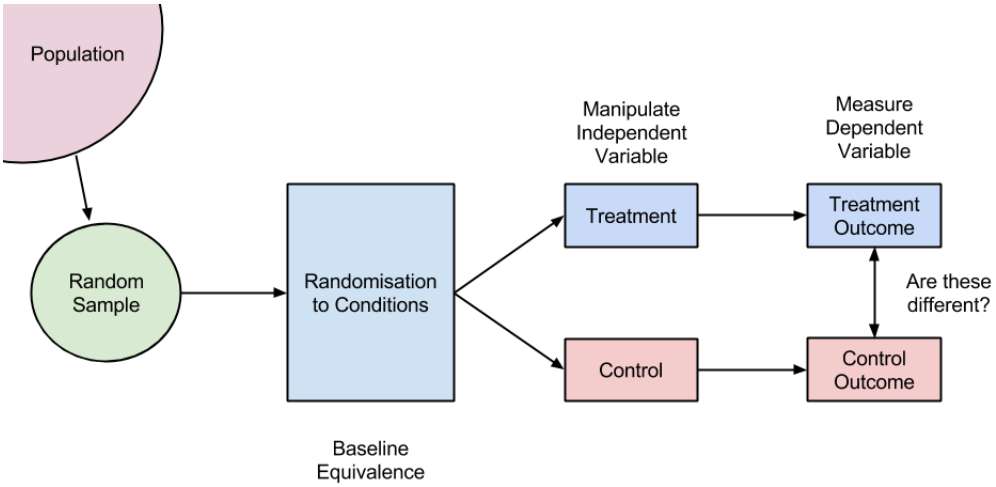
There are three major types of statistical data investigations that can be distinguished based on their aims and methods of data collection. We must understand the strengths and weaknesses of each types as these have an impact on the conclusions you draw. We will learn more about different research designs throughout the course when we look at different statistical methods.

Surveys

Surveys aim to gather data from the population, in the form of responses, in order to investigate or understand some type of phenomenon. For example, you might have had experience filling out course evaluation surveys, product surveys, customer satisfaction surveys, the Australian Census, and job evaluation surveys. Surveys are typically done using samples, but may also be completed by an entire population. When this happens, this is known as a census. The Australian Census seeks to count and describe the characteristics of the Australian population in order to help Australia plan for the future. Surveys are typically administered using paper-based or online questionnaires completed by the participant, but may also involve face-to-face or telephone interviews. Surveys may seem like the easiest and most cost effective way to gather large amounts of data from a population, but that is far from the truth. Surveys have many challenges, including selecting a good sample, poor response rates, response bias, and designing good questions.

Experiments

In the simplest **experimental** design, participants or some other unit are randomised to a **control** group or a **treatment** group. The investigator's manipulation of exposure to the treatment is what defines a true experiment. The treatment is referred to as an **independent variable**. **Randomisation** is used to maximise the chance of the groups being considered equivalent , in terms of their characteristics, at the start of the experiment, thus minimising systematic bias between groups. The control group does not receive the actual treatment, and may instead, receive an inert form of the treatment called a **placebo**. **Blinding** is used to prevent the results being influenced by participant expectations, by ensuring the participants are unaware of their allocated research group or the true nature behind the experiment. The investigator seeks to keep all other factors and variables constant throughout the experiment. At the end of the experiment, the investigator will measure the two groups on a **dependent** or **outcome variable**. Because of randomisation of participants and tight control, it is assumed that by the end of the experiment, any significant difference between groups on the dependent variable could ONLY be due to the treatment. Therefore, if a difference between groups is evident at the end of the experiment, its assumed to be the effect of the treatment. Thus, experiments seek to test cause and effect hypotheses. However, experiments are also the most difficult and time consuming of investigation types.



Observational or Correlational Studies

Observational or correlational research designs look for a relationship between at least two variables. Observational or correlational research do not attempt to manipulate or control an independent variable, which distinguishes it from an experiment. Therefore, these types of studies cannot test cause and effect. Instead, they are used to establish evidence of relationships, associations or correlations between variables that may suggest evidence of causal effects. Conclusions from observational and correlational investigations are always interpreted with this limitation in mind. On the plus side, these types of investigations allow researchers to study relationships between variables that cannot be manipulated in experiments. For example, ethically, you cannot randomise people to smoke cigarettes to test if it increases risk of cancer. However, you can observe and compare the incidence of cancer in people who voluntarily smoke to those who don't.

References

MacGillivray, H., J. M. Utts, and R. F Heckard. 2014. *Mind on statistics*. 2nd ed. South Melbourne, VIC: Cengage Learning Australia.

Moore, D. S. 2007. *The basic practice of statistics*. 4th ed. New York, NY: W. H. Freeman; Company.

Tukey, J. W. 1953. “The growth of experimental design in a research laboratory.” In *Research Operations in Industry*, 303–13. New York: King’s Crown Press.

Wild, C., M. Pfannkuch, M. Regan, and N. J. Horton. 2011. “Towards more accessible conceptions of statistical inference.” *Journal of the Royal Statistical Society* 174: 247–95. https://doi.org/10.1111/j.1467-985X.2010.00678.x (https://doi.org/10.1111/j.1467-985X.2010.00678.x).
