STA130 Notebook

Ian Huang
University of Toronto

The source code for this notebook as well as others can be found on GitHub

`https://github.com/iahuang/uoft-notebooks`

Contents

2 Module 2 - Visualizing and Describing Distributions

2.1 Types of Data

Data primarily falls into two categories, quantitative (or numerical) and categorical (or qualitative). Other types of data include text, image, video, sound, but even those can be represented by sets of numerical and categorical variables.

Numerical Data

Numerical data is divided into two categories: continuous variables and discrete variables.

Definition. (Discrete Variable)

Discrete variables are variables whose values are a countable set. Formally, a variable V is **discrete** if there exists a one-to-one correspondence between the possible values of V and $\mathbb{N}$.

Examples of discrete variables are generally representations of the *count* of something, such as the number of students in a classroom.

Definition. (Continuous Variables)

Continuous variables are variables whose values are an uncountable set.

Examples of discrete variables are generally representations of the *measure* of something, such as velocity.

Categorical Data

Definition. (Categorical Variables)

Categorical variables are variables that take a finite, fixed, set of values. If the set of values take a logical order, than the variable is called **ordinal**; otherwise, it is called **nominal**. If there are only two values it can take, then it is called **binary**.

Some examples of categorical variables include:

- Blood type: "A", "B", "AB", "O"
- Color: "Red", "Yellow", "Green", etc.
- The roll of a six-sided die: 1, 2, 3, 4, 5, 6
- Handedness: "Right", "Left"

Categorical variables can be represented with numbers, for instance,

$$\text{Handedness} = \begin{cases} 0 & \text{if "Right-handed"} \\ 1 & \text{if "Left-handed"} \end{cases}$$

However, handedness is still a categorical variable and not a numeric variable, as the difference between 0 and 1 represents no meaningful quantity or measure. Similarly, the outcome of a six-sided die roll is also considered categorical.

Data Types in R

Data Type	R Keyword	Description
Integer	`int`	A 32-bit signed representation of an element of $\mathbb{Z}$. Usually used to represent discrete variables
Double	`dbl`	A double-precision floating point representation of an element of $\mathbb{R}$. Usually used to represent continuous variables
Logical	`lgl`	Takes the value `TRUE` or `FALSE`. Usually used to represent binary categorical variables
Character	`chr`	Used to represent strings (text)
Factor	`fct`	A categorical variable. Similar to an `enum` type

2.2 Visualizing and Describing Numerical Data

In general, the *distribution* of a variable refers to the possible range of its values in a dataset and how often each value comes up. One consequential question that is sometimes asked is "is value a as likely as value b?".

Features of Numerical Distributions

Numerical distributions can be described using features such as *center* and *spread*.

- **Center**: describes a "typical" value of the variable.
- **Spread**: describes how concentrated the values of the variable are or how varied they are.
- **Shape**: describes the overall pattern of values for this variable.

The *mean* is a common way to measure the center of a variable distribution.

Definition. (Mean)

The **mean** of a distribution for some variable x, written $\overline{x}$, where $x_1, x_2, \ldots, x_n$ are observed values of x, is

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

The *median* is another way to measure the center of a variable distribution.

Definition. (Median)

The **median** of a distribution for some variable x, where $x_1, x_2, \ldots x_n$ are the *sorted* observations of x is

$$\begin{cases} x_{\frac{n}{2}} & n \text{ is even} \\ \frac{x_{\frac{n-1}{2}} + x_{\frac{n+1}{2}}}{2} & n \text{ is odd} \end{cases}$$

More intuitively, the median is some number m for which half of the observations for x are less than m, and half are larger than m.

The *variance* is one way to measure the "spread" of a variable distribution. The variance roughly measures the average squared difference between all values and the mean, in other words, an estimate for $(x_i - \overline{x})^2$ given some random observation x_i.

Definition. (Variance)

The **variance** of a distribution for some variable x, where $x_1, x_2, \ldots, x_n$ are the the observed values, then the variance, written s^2 or σ^2, is

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Furthermore, the *standard deviation* is another way to measure the spread of a numerical distribution. It is computed as the square root of the variance. It is often used over the variance, as it takes the same units as the variable from which it was derived.

Definition.

The **standard deviation**, written s or σ, of a variable distribution is the square root of its variance.

$$s = \sqrt{s^2}$$

Shape is more difficult to describe mathematically, and the description of the shape of a distribution is usually derived from a visualization of the distribution instead.

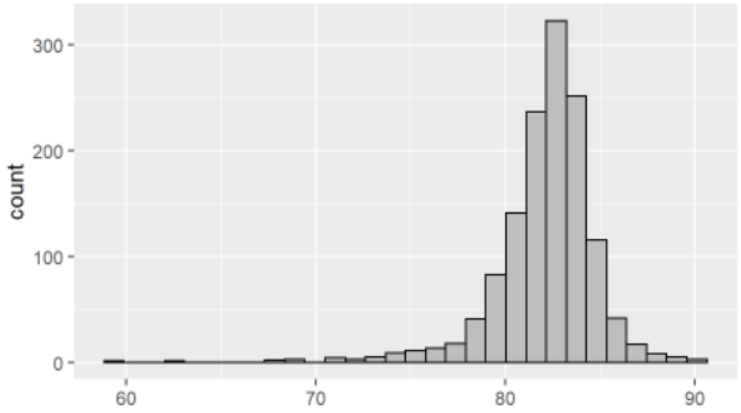
Histograms

A histogram is a representation of the distribution of numerical data. Consider a set $S = \{S_1, S_2, \ldots, n\}$ representing a dataset of values for some numerical variable. First, we take the range of the dataset, $\max S$ and $\min S$, and divide it into n (usually equal) "bins". For each ith bin, represented by the interval $[b_i, b_{i+1})$, count the number of values in S fall in this interval:

$$c_i = \sum_{i=1}^{n} \begin{cases} 0 & b_i \leq S_i < b_{i+1} \\ 1 & \text{otherwise} \end{cases}$$

Then, we represent each ith bin using a bar with a height proportional to c_i. The bars should be placed directly adjacent to each other and their width on the x-axis should be proportional to the size of the interval they represent.

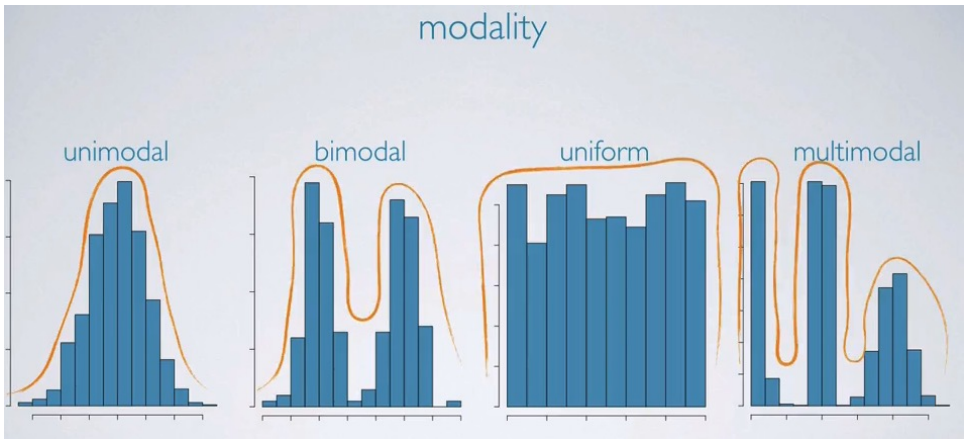
The choice of how many bins, n, is subjective. In general, the goal is to choose an n such that interesting features of the distribution are highlighted such as shape, center and spread, without inaccurately highlighting too many "spikes" or noise.



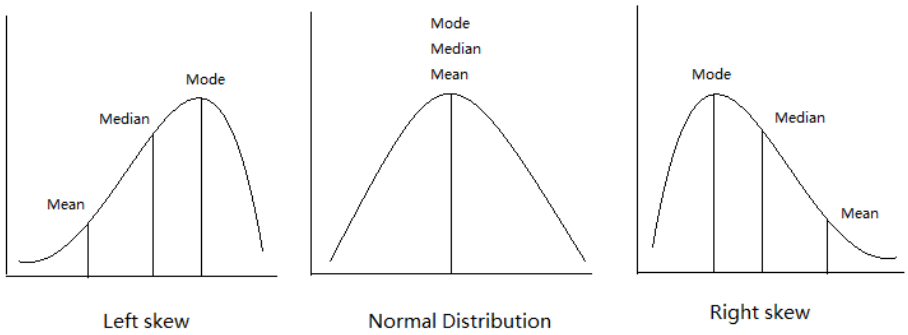
An example of a histogram with $n = 30$ bins

Histograms can be used to determine the shape of a numerical variable distribution. If most of the data is to the right, as is the case in the figure above, then we say that the distribution is *left-skewed*. Conversely, if most of the data is to the left, then we say that the distribution is *right-skewed*. If the data is centered, then we say it is *symmetric*.

Another way to characterize the data is by its *modality* or number of "peaks" in a distribution. The figure above represents a *unimodal* distribution, since it shows one clear "peak". The following diagram shows other descriptions of modality with a corresponding example histogram.



Here is a useful diagram showing the relative positions of various common statistical measures in a histogram:



Histograms in R

The code to produce a histogram with 30 bins is as follows:

```
ggplot(data=DATA, aes(x=VARIABLE)) +
    geom_histogram(
        color="black",
        fill="gray",
        bins=30
    ) +
    labs(x="X-AXIS LABEL")
```

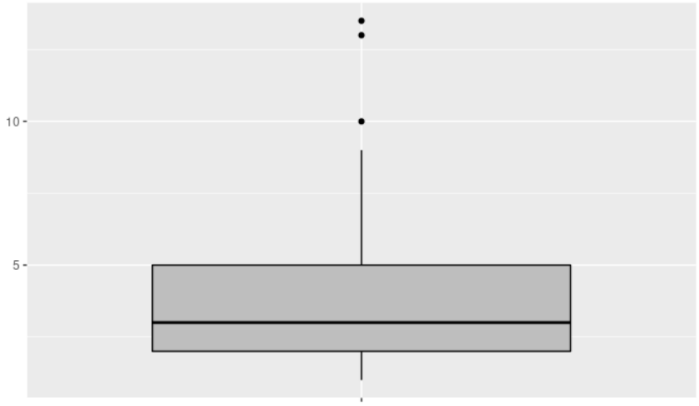
where `DATA` is some tibble, and `VARIABLE` is the variable being plotted.

Boxplots

A boxplot is a way of visualizing some specific features of a numerical distribution, including spread and center. A boxplot is usually constructed using five values:

- Q_2: the **median** or 50th percentile
- Q_1: the 25th percentile
- Q_3: the 7th percentile
- IQR: the *inter-quartile range*, computed as $Q_3 - Q_1$
- $1.5 \cdot \text{IQR}$: the standard threshold for determining *outliers*

Here you can see an example of a boxplot:



First consider the grey box. The top of the box is placed at Q_3, the line in the middle is the median, Q_2, and the bottom of the box is Q_1. The height of the box is therefore the IQR. Then, two whiskers are drawn above and below the boxes to a distance of 1.5 times the IQR. The bottom takes the lowest value in the distribution within this range, and similarly, the upper whisker takes the largest value in the distribution with this range. Values outside the range of the whiskers are called *outliers* and are represented with a dot. Specifically, the end of the bottom whisker is placed at

$$\min\{x \in X : x \geq Q_1 - 1.5(Q_3 - Q_1)\}$$

and the end of the top whisker at

$$\max\{x \in X : x \leq Q_3 + 1.5(Q_3 - Q_1)\}$$

where X is the set of variable observations used to construct the boxplot.

If the data is skewed, the position and size of the box in a boxplot will represent that. If the size of the box is fairly small relative to the range of the data, it may be reasonable to conclude that the data is also unimodal.

Boxplots in R

```
ggplot(data=DATA, aes(x=CATEGORY, y=VARIABLE)) +
    geom_histogram(
        color="black",
        fill="gray"
    ) +
    labs(
        x="X-AXIS LABEL",
        y="Y-AXIS LABEL"
    )
```

The code to render a boxplot is similar to that used for histograms. In this case, `CATEGORY` is the name of a categorical variable to split the plots by category. If only one plot is desired, use `x=""`, where an empty string is used in place of `CATEGORY`. For instance, consider a tibble with rows of numerical variable `x` and categorical variable `color` $\in \{$`RED`, `BLUE`$\}$. If we wanted to create a boxplot with all values of `x`, we would use `aes(x="", ...)`. If we wanted to create two separate boxplots, one with values of `x` with the category "RED", and another to the side with values of `x` with the category "BLUE", we would use `aes(x=color, ...)`

Barplots

Barplots are visualizations of categorical data. A bar plot presents a bar for each possible category, with the length of each bar proportional to the number of observations of that category.
TODO: Write this

3 Module 3 - Tidy Data

3.1 What is Tidy Data?

Definition

Tidy data is characterized by data that is "easy" to understand by computers. "Un-tidy" data can be recognized by its difficulty to use in code without prior manipulation. Here is a more precise definition:

Definition. (Tidy data)

Tidy data is characterized by three rules:

1. Each variable must have its own column.
2. Each observation must have its own row.
3. Each value must have its own cell.

In R, cells are allowed to be "empty" (i.e. having a value of `NA`). Even if data contains missing or `NA` values, it may still be considered tidy.

Examples

Consider the following data:

Student	HW1	HW2	HW3
Bob	75	93	88
Alice	55	87	78
Sarah	66	91	87

Although this data appears neat (i.e. it is easily understood and well organized), it is not considered tidy. The rule that is violated here is **2. Each observation must have its own row**. If we consider a student's grade on some assignment to be a variable, it becomes apparent that we have three observations of this variable per row. Instead, to make this data "tidy", we would want it formatted like so:

Student	Grade	Assignment Number
Bob	75	1
Bob	93	2
Bob	88	3
...	...	...

3.2 Data Wrangling

Data wrangling is the practice of transforming data to make it more useful. Some examples of data wrangling include:

- Making data *tidy*
- Adding or removing variables/columns
- Removing missing observations

Data Wrangling Functions - `filter()`

```
data %>% filter(CONDITION)
```

The above expression returns a copy of `data`, including only rows for which the boolean expression `CONDITION` is true. Remember that in R, `&` refers to logical AND rather than the conventional `&&`, and `|` refers to logical OR rather than the conventional `||`.

For instance, given the tibble `grades` (represented below)

student	grade	assn_no
Bob	75	1
Bob	93	2
Bob	88	3
...	...	...

the code

```
grades <- grades %>% filter(grade > 80)
```

would remove all instances of grades below 80%.

The `is.na()` function can also be used in conjunction with `filter` to remove or extract missing (`NA`) values.

Data Wrangling Functions - `select()`

```
data %>% select(VAR1, VAR2, VAR3, ...)
```

`select` takes the passed dataframe and a list of variables/columns, and returns a copy with only those variables/columns.

Data Wrangling Functions - `mutate()`

```
data %>% mutate(NEW_VARIABLE=VALUE)
data %>% mutate(NEW_VARIABLE=case_when(
    CONDITION1 ~ VALUE1,
    CONDITION2 ~ VALUE2,
    ...
))
```

TODO: Write this

Data Wrangling Functions - `arrange()`

```
data %>% arrange(VAR)       # sort by ascending
data %>% arrange(desc(VAR)) # sort by descending
```

`arrange` returns a sorted copy of a dataframe, sorted by a given variable.

Summary Tables in R

A summary table is an easy way to view certain statistics about a dataframe. Consider the tibble `data` representing the following data:

name	age
Jerry	24
Alex	31
Winston	40

The code

```
data %>% summarize(
    n=n(),
    mean_age=mean(age),
    median_age=median(age),
)
```

Produces the following tibble:

n	mean_age	median_age
3	31.667	31

Below are a list of functions that can be used in a summary table:

Function	Description
`n()`	Returns the number of observations
`mean()`	Returns the mean
`min()`	Returns the minimum
`max()`	Returns the maximum
`median()`	Returns the median
`var()`	Returns the variance (σ^2)
`sd()`	Returns the standard deviation (σ)
`sum()`	Returns the sum

Using the function `group_by` first will separate the statistics by the name of the passed categorical variable.

4 Module 4 - One Proportion Hypothesis Testing

4.1 Introduction to Statistical Inference

Definition.

Statistical inference is the process of coming to conclusions or decisions based on statistical information.

Conclusions made using statistical information are subject to uncertainty; in statistical inference, nothing should be stated as fact.

Statistical inference often uses techniques involving *sampling*; taking a small subset of the potentially available information and analyzing that. The following terms are often used in this context:

- **Population**: A complete set of items, events, or data that is of interest for statistical inference.
  - The population can be thought of as "the entire thing we are trying to study".
  - The population may be too large to feasibly observe, or may be hypothetical or infinite in size.
  - The goal of statistical inference is to devise some conclusion about the population without having to observe the entire population.
- **(Population) Parameter**: In statistical inference, the parameter is some "true" measure of the population that is trying to be inferred.
  - Sometimes written as p or π
- **Sample**: Refers to a manageable subset of the population that can be reasonably observed.
  - A good sample is one that is *representative* of the population; oftentimes, it is best to choose a random sample.
- **Test Statistic**: A quantity or measure derived from the sample that can be used to estimate the parameter.
  - Often denoted $\hat{p}$ or $\hat{\pi}$.
- **Sampling Distribution**: A probability distribution of test statistics for all possible samples of the same sample size.
  - Often represented in a histogram.

Here are some common symbols used for various statistical measures used in hypothesis testing:

Statistic Type	Population	Sample
Proportion	p	$\hat{p}$
Mean	μ	$\bar{x}$
Median	(no common symbol)	$\tilde{x}$

Hypothesis Testing

Hypothesis testing is a form of statistical inference, and involves coming up with a hypothesis and trying to prove that it is incorrect. Hypothesis testing generally takes the following steps:

1. Devise a *null hypothesis* and an *alternative hypothesis*.

   (a) The null hypothesis:

   > Definition.
   >
   > The **null hypothesis** is a conjecture that a hypothesis test will try to disprove. In general, the null hypothesis represents the "default" outcome in statistical inference, usually something along the lines of "no effect" or "no difference".
   >
   > In a court of law, the null hypothesis is "innocent" under the philosophy of "innocent until proven guilty".

   (b) The alternative hypothesis:

   > Definition.
   >
   > The **alternative hypothesis** is a statement that exists as the "alternative" option to the null hypothesis. The null hypothesis and the alternative hypothesis must be mutually exclusive, and in general, the alternative hypothesis is simply the opposite of the null hypothesis.

   (c) The null hypothesis is often written H_0, and the alternative hypothesis is written as H_1 or H_A. When stating the hypothesis, we write

   $$H_0 : P \quad \text{and} \quad H_A : Q$$

   where P and Q are some statements representing the null and alternative hypotheses respectively. Because the null hypothesis is a label, not a variable, we use a colon.

2. Create a sample and compute a test statistic based on the sample.

3. Estimate a sampling distribution, usually by simulation. We assume the null hypothesis is true.

4. Calculate a p-value to quantify evidence against the null hypothesis.

   > Definition.
   >
   > A **p-value** is defined as the probability of obtaining a result at least as extreme as the test statistic.

5. Draw a conclusion based on the p-value. For this, the following guideline is conventionally used:

p-value	**Evidence**
$p > 0.10$	No evidence against H_0
$0.05 < p < 0.10$	Weak evidence against H_0
$0.01 < p < 0.05$	Moderate evidence against H_0
$0.001 < p < 0.01$	Strong evidence against H_0
$p < 0.001$	Very strong evidence against H_0

Statistical Significance

In general, we have some value α, under which if $p < \alpha$, then the null hypothesis is false. Conventionally, we use $\alpha = 0.05$. If this is the case, that is, $p < \alpha$, then the findings are said to be *statistically significant*.

Example

One common example for statistical inference is measuring the "fairness" of a coin. Consider a coin that can land on either heads or tails, and we want to answer the question of "is this coin equally likely to land on heads as it is tails, that is, is this coin *fair*?

For this, we say that the *population* is the outcomes of the set of all possible coin flips. Then, our parameter p is the proportion of those coin flips that land on heads. As we can see, our population is theoretically infinite, and therefore infeasible to measure directly; however, we can loosely define our parameter using the following expression[1]

$$p = \lim_{n\to\infty} \left(\frac{1}{n} \sum_{i=1}^{n} F_i \right)$$

Where, if the coin were to be flipped an arbitrarily large number of times, F_n would be the result of the nth coin flip, 1 if heads, and 0 if tails.

Let us continue by defining our hypotheses. As stated earlier, the null hypothesis is often similar to the phrase "no difference". Therefore, we will assume that there is no difference between the probability of the coin landing on heads and the coin landing on tails. Therefore, we define our hypotheses as follows:

$$H_0 : p = 0.5 \quad \text{and} \quad H_A : p \neq 0.5$$

Given that our population is infinite, we need to define a finite sample of the population on which to perform inference. Let n be our sample size, which may be chosen arbitrarily. For this example, we will use $n = 50$. Then, we can compute the test statistic as

$$\hat{p} = \frac{1}{50} \sum_{i=1}^{50} F_i$$

Where in this case F_n is the observed nth result of actually flipping the coin. For this example, we will pretend that $\hat{p} = 0.56$, that is, 28 out of 50 coin flips were heads.

To estimate a sampling distribution to find out where our test statistic falls, we will assume that the null hypothesis is true, that is, $p = 0.5$, and simulate a distribution of test statistics for samples also of size $n = 50$. This can be done using the *binomial distribution function*, but for simplicity, we can also use a computer program to flip an imaginary coin many times. Remember that the p-value is defined as the probability that we observe a result in our sampling distribution as or more extreme than our test statistic. We define "as or more extreme" in this case as being the same distance or farther away from 0.5 than the test statistic. Therefore, our p-value is the proportion of values in our sampling distribution that are as or more extreme than $\hat{p}$. Let $s_1, s_2, \ldots, s_N$ be the results of our our simulated sampling distribution where s_n is the proportion of a simulated fair coin landing on heads, where the coin is flipped 50 times, just like in

[1] We use limit notation in this case, as using ∞ directly in our expression for p would look strange.

our sample. To make our sampling distribution as accurate as possible, we will make N very large, say, $N = 100\,000$.

$$p\text{-value} = \frac{1}{N}\sum_{i=1}^{N}\begin{cases}1 & |s_i - 0.5| \geq |\hat{p} - 0.5| \\ 0 & \text{otherwise}\end{cases}$$

Using a computer program to fill in values for s_i, we find a p-value of about 0.40. By the table in the previous section, we see that we have no evidence against our null hypothesis that our coin is fair, and therefore our *conclusion* is that the coin is, in fact fair.

5 Module 5 - Randomization Testing

5.1 Motivating Example - Two Group Hypothesis Testing

Consider an experimental drug that is designed to lower blood pressure. We have two groups, a *treatment* group, which receives this drug, and a *control* group which does not (or receives a placebo, etc.). The population in this case, would be a hypothetical infinite set of patients in each group, and we want to find whether the *mean* blood pressure is lower in the group that received the treatment. Hence, we call our parameters $\mu_{treatment}$ and $\mu_{control}$ respectively.

Our null hypothesis is that the drug has *no effect*, that is, $\mu_{treatment} = \mu_{control}$. We then label our hypotheses

$$H_0 : \mu_{treatment} = \mu_{control} \quad \text{and} \quad H_A : \mu_{treatment} \neq \mu_{control}$$

or alternatively,

$$H_0 : \mu_{treatment} - \mu_{control} = 0 \quad \text{and} \quad H_A : \mu_{treatment} - \mu_{control} \neq 0$$

Say we then derive a *sample* of 50 patients for both the treatment and control group, running the experiment accordingly. We then have values for our *test statistic*, $\bar{x}$, the mean blood pressures of each sample group. Say we find that[2]

$$\bar{x}_{treatment} = 92.1 \quad \text{and} \quad \bar{x}_{control} = 103.4$$

Hence,

$$\bar{x} = \bar{x}_{treatment} - \bar{x}_{control} = -11.3$$

How do we derive a p-value to measure the likelihood that our null hypothesis is false, that is, how do we know whether the drug was effective?

5.2 Randomization Testing

Continuing with the earlier example involving a hypothetical drug trial, we need a method to simulate a sampling distribution under the null hypothesis to see whether or not our test statistic was "likely" or not. Under the null hypothesis, we assume that there is no difference between the treatment and control groups. Therefore, given some observation for a blood pressure measure, we could say that under the null hypothesis, it is equally likely that measure belonged to the control group or the treatment group.

Hence, one way to simulate a sampling distribution is to take the all the blood pressure measures of the treatment group and all blood pressure measures into the control group and combine them into a series of observations $x_1, x_2, \ldots x_n$. We then take the "labels" of each observation, either "Control" or "Treatment" (in our example there is an equal number of both), and combine them into a series of labels $L_1, L_2, \ldots, L_n$ as well. We then shuffle the labels and observations around independently, and construct a set of tuples (x_i, L_i) that effectively rematches each label to a random observation.

We may then compute a test statistic $\bar{x}$ for our sampling distribution based on our simulated sample as so:

$$\begin{aligned}\bar{x} &= \text{mean treatment} - \text{mean control} \\ &= \frac{\sum \text{treatment observations}}{\#\,\text{treatment}} - \frac{\sum \text{control observations}}{\#\,\text{control}}\end{aligned}$$

[2] Using arbitrary units for blood pressure

Specifically,

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i \cdot \left(\begin{cases} 1 & L_i = \text{“Treatment”} \\ 0 & \text{otherwise} \end{cases} \right)}{\sum_{i=1}^{n} \begin{cases} 1 & L_i = \text{“Treatment”} \\ 0 & \text{otherwise} \end{cases}} - \frac{\sum_{i=1}^{n} x_i \cdot \left(\begin{cases} 1 & L_i = \text{“Control”} \\ 0 & \text{otherwise} \end{cases} \right)}{\sum_{i=1}^{n} \begin{cases} 1 & L_i = \text{“Control”} \\ 0 & \text{otherwise} \end{cases}}$$

Repeating the shuffling and recomputation of $\bar{x}$ an arbitrary number of times produces a sampling distribution, from which a p-value can be calculated as the proportion of simulated $\bar{x}$ that are farther away from 0 than our test statistic.

5.3 Type I and Type II Errors

Hypothesis testing is imperfect by nature. Sometimes, it is possible to arrive at the wrong conclusions.

Type I Error

A type 1 error, sometimes referred to as a *false positive*, refers to the incorrect rejection of a null hypothesis, that is, we conclude that the null hypothesis is false, when it is actually true. Some examples include:

- A COVID-19 test returning positive, when the user was not infected.
- An innocent person being convicted.

In a hypothesis test, this is often because the observed test statistic was just very unusual.

Type II Error

A type 2 error, sometimes referred to as a *false negative*, refers to the mistaken acceptance of a null hypothesis, that is, we conclude that the null hypothesis is true, when it is actually false. Some examples include:

- A COVID-19 test returning negative, when the user is actually infected.
- A medical screening fails to refer someone to a physician for further testing.
- An guilty person being deemed innocent.

6 Module 6 - Bootstrap and Confidence Intervals

6.1 Introduction to Estimation

Motivation

There are a few problems with p-value based hypothesis testing:

1. Scientific studies are often too reliant on p-values; it is possible to intentionally misuse p-values to make misleading arguments.[3]
2. The choice of $\alpha = 0.05$ as the threshold for statistical significance, while conventional, is still arbitrary.
3. Hypothesis testing can only tell us if something is true or not; it doesn't give us any more information than that.

Types of Sampling

Moving forward, it is worth differentiating the notions of *sampling with replacement* and *sampling without replacement*. The typical idea of "taking a sample of size n from a population" does not usually involve replacement. Specifically, when we sample without replacement, the same member of the population cannot be sampled twice.

Sampling *with replacement*, on the other hand, can be thought of using an analogy of a bag of colored marbles. When we take a sample of size n with replacement, we take a marble from the bag, record its color, put the marble back in the bag (*replace* it), and repeat n times. If we were to sample *without replacement*, we would not put each marble back after taking it out.

What is Estimation?

Confidence interval estimation involves using a sample to make a reasonable guess about the value of some population parameter, and placing an interval around that guess to represent a conclusion about the rest of the population.

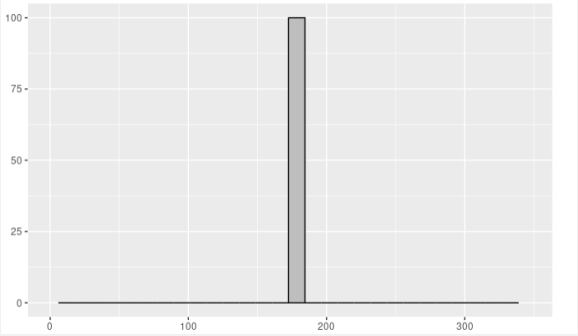
Estimation, like hypothesis testing, uses notions of populations, samples, parameters, test statistics, and sampling distributions. Imagine a population such as the heights of every male adult on Earth, and we want to find the population mean, that is, the average male height. We take a random sample of 100 men, and say that we find the average height of this sample to be about 5'8" (173 cm). Unlike hypothesis testing, there is no "null hypothesis" under which to sample, so what would a sampling distribution for this statistic look like, and how would we find one? The most theoretically straightforward answer would be to simply take more random samples from the population of the same size (100 men).

Sampling Distributions and Sample Size

How would the sample size affect the shape of a sampling distribution of means? Consider two extremes of sample size n, one where n is the size of the population itself, and another where $n = 1$. If we were to sample without replacement samples the size of the entire population, each sample mean

[3] This is sometimes called *p-hacking*.

would simply be the population mean. A histogram for that distribution would have a single value, a distinct mode, and be perfectly symmetrical.



Conversely, a sampling distribution of $n = 1$ would look very similar to the distribution of the data itself. It may be skewed, multimodal, etc. In general, as the sample size n increases:

1. The mean of the sampling distribution stays roughly the same and approximates the population mean.
2. The shape of the sampling distribution becomes less skewed.
3. The spread of the sampling distribution decreases.

If we were to repeatedly take more random samples from the population, we would get a very good representation of a sampling distribution. However, in the example where our population is the heights of every man on Earth, it is feasible to keep taking random samples from the population, otherwise we would survey the entire population itself.

The only way to create a sampling distribution, then, is to estimate one from the one sample itself.

6.2 Bootstrap Sampling

Bootstrap sampling is the process of estimating a sampling distribution from a single sample. If we assume that the sample taken is representative of the population (which should be the case if it was randomly sampled), then we can create a sampling distribution from the sample, by repeatedly *sampling the sample with replacement*. If we sampled without replacement, then the sampling distribution would be perfectly homogeneous; however, sampling with replacement allows some variability in our sampling distribution that emulates the variability in the population sampling distribution.

Although this method seems like it would not work–after all it seems to create new data out of thin air–it proves to be a useful method in estimating a sampling distribution, so long as the original sample is representative of the population.

6.3 Confidence Intervals

So far, we have been able to use bootstrap sampling to create a sampling distribution, but the question of how to use these methods to create an estimation of a population parameter still remains. For this, we use a measure called a confidence interval.

Definition

A 95% **confidence interval** for a population parameter is defined as some interval such that, for 95% of possible random samples from the population, that interval will contain the true parameter. Specifically, say we compute a bootstrap sampling distribution, represented using the variable x with observations $x_1, x_2, \ldots x_n$, where each x_i is the mean of some bootstrapped sample from our original "true" sample.

> In this context, 95% is called the **confidence level**, and is by far the most commonly used value, but other values such as 90% and 99% are also sometimes used. Specifically, for some a confidence level $c \in [0, 1]$, if we were to take many random samples from the population, computing a bootstrap confidence interval (with confidence level c) for each one, c represents the proportion of those confidence intervals that contain the actual parameter.

Recall that for some $p \in [0, 100]$, the pth percentile of some variable is the smallest value such that p% of those of the observations of that variable are less than p. Thus, a 95% confidence interval $[a, b]$ is defined where a is the 2.5th percentile, and b is the 97.5th percentile.

> Notice that the range of values between a and b covers 95% of the observed values. If we were to more intuitively pick a and b to be the 5th and 95th percentiles respectively, we would end up with a 90% confidence interval instead.

Interpreting Confidence Intervals

Say we concluded a 95% confidence interval of $[165.3, 182.1]$ (measured in centimeters) for our earlier sample of 100 men's heights. What we say from this information is that we are "95% confident that the mean adult male height globally is between 165.3 and 182.1 centimeters". Note that we don't say that there is a "95% chance", only that we are "95% confident".

7 Module 7 - Simple Linear Regression

7.1 Association of Numerical Variables

Association between two numerical variables refers to variables that can be related mathematically. A common way that variables can be associated is using a scatterplot, a diagram used to show correlation between observations of two variables x and y, by representing them as a set of points (x_i, y_i) on the Cartesian plane for each ith observation.

Features of Association

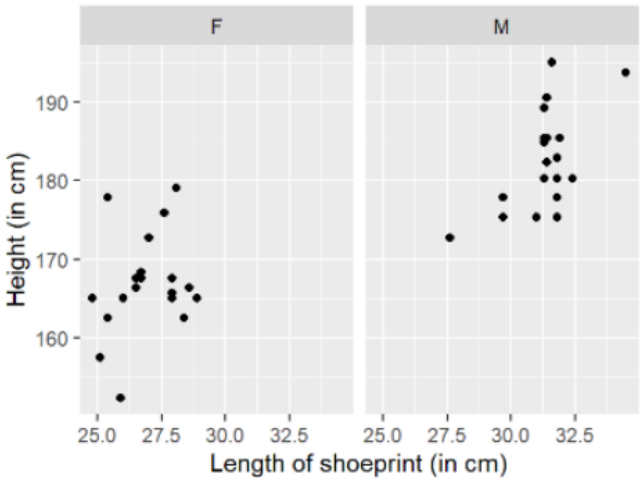
- **Form**: describes the pattern that two variables follow together (e.g. linear, non-linear, quadratic, etc.)
  - These labels are not mutually exclusive; if a pattern is quadratic, it is also non-linear by definition.
- **Direction**: describes an overall overall trend.
  - Two variables are said to be **positively associated** if values of one variable increase as the other increases.
  - Two variables are said to be **negatively associated** if values of one variable decrease as the other decreases.
- **Strength**: describes how concentrated the values of the variable are around the pattern.

Scatterplots in R

```
heights %>% ggplot(aes(x=X_VARIABLE, y=Y_VARIABLE, color=CATEGORY))
                + geom_point()
                + labs(x="X AXIS LABEL", y="Y AXIS LABEL"))
                + theme_minimal() # optional
```

The `color` argument is optional, and can be used to color-code the points based on the value of some categorical variable. Alternatively, adding `facet_wrap(~CATEGORY)` will split the data into two scatterplots. As an example, here is the output produced by the following code, where `heights` is a data frame where each row contains information about a person's height, sex, and shoe print length.

```
heights %>% ggplot(aes(x=shoePrint, y=height))
                + geom_point()
                + labs(x="Length of shoeprint (in cm)", y="Height (in cm)"))
                + theme_minimal()
                + facet_wrap(~sex)
```



Quantifying Linear Association - Correlation

Correlation is a value, written r, between -1 and 1, that summarizes the strength and direction of the linear relationship between two numerical variables.

Given a set of data points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ for a set of variable observations, the mean value for x, $\bar{x}$, and the mean value for y, $\bar{y}$, r is given as

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2 \cdot \sum_{i=1}^{n}(y_i - \overline{y})^2}$$

The correlation value has the following properties that describe the strength and direction of the linear association:

- $r > 0$ denotes a positive linear association.
- $r < 0$ denotes a negative linear association
- $r = 0$ denotes a perfectly flat linear association or no correlation at all.
- The higher the magnitude of r, $|r|$, the stronger the linear association.

The following code can be used to compute r in R.

```
corr(
        x=heights$shoePrint,
        y=heights$height
)
```

> Remember that r only quantifies the *linear* association between two variables. Even if they are associated perfectly in another way, say quadratically, they may have a very low r value.

7.2 Simple Linear Regression

The **simple linear regression model** assumes a "best" straight line that summarizes the real relationship between two variables x and y, and that the values only deviate from this line. The model can be summarized using the following equation:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Where

- Y_i: The response variable (or dependent variable, target variable, etc.) for observation i.
- x_i: The independent variable (or predictor, covariate, feature, input, etc.) for observation i.
- β_0: The intercept parameter.
- β_1: The slope parameter.
- ε_i: A random error term for observation i that accounts for deviation from the line defined by $y = \beta_1 x + \beta_0$.

7.3 Fitting a Regression Line

Recall that the simple linear regression model assumes the existence of some line described by the equation

$$Y_i = \beta_0 + \beta_1 x_i$$

(and an extra error term ε_i to account for deviations). But this does not tell us what the values of β_0 and β_1 are. There is no way to determine the "true" values of these parameters. But we can use various mathematical models to find best estimates for these values, denoted $\hat{\beta}_0$. and $\hat{\beta}_1$.

> In the context of linear regression, and often other statistical models, a carat or "hat" above a variable often denotes that this variable is an *estimate* or a *prediction*.

Least Squares Regression Line

The least squares regression method finds values for $\hat{\beta}_0$ and $\hat{\beta}_1$ that *minimize* the squared error between $\hat{y}$, the predicted values for y, and the true values of y, where

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

This method produces the following closed-form mathematical expressions for the estimated values of β_0 and β_1:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

and

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

where $\bar{x}$ and $\bar{y}$ are the sample means of the x and y observations respectively.

Interpretation of Regression Coefficients

The estimated values found above produce a way to predict new values for y, given some value for x, using the following equation:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

$\hat{y}$ is called the **fitted value** or **predicted value** and is the estimated value for y when the predictor takes a value of x. The equation above is called the **fitted regression line** and its slope is given by $\hat{\beta}_1$, and its y-intercept by $\hat{\beta}_0$. The difference between the observed and predicted value of y for the ith observation is called the **residual** $e_i = y_i - \hat{y}_i$.

Coefficient of Determination

The **coefficient of determination**, written R^2, is a number between 0 and 1 that measures the proportion of variability in y that is explained by a fitted regression line.

$$R^2 = 1 - \frac{\sum_{i=0}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{n}(y_i - \overline{y})^2} = r^2$$

- A value of R^2 close to 1 indicates that most of the variability in y is explained by the regression model
- A value of R^2 close to 0 indicates that very little of the variability in y is explained by the regression model
- R^2 is also equal to r^2, where r is the correlation value.

7.4 Linear Regression with Categorical Variables

If we wanted to have the independent variable x be a binary categorical variable instead of a numerical one, we use an **indicator variable**. For this, we assign a **baseline value** to one category, the value corresponding to $x = 0$, and assign $x = 1$ to the other.

For instance, we may have

$$x_i = \begin{cases} 1 & \text{individual } i \text{ is male} \\ 0 & \text{individual } i \text{ is female} \end{cases}$$

where "female" is the baseline value.

Interpreting $\hat{\beta}_0$ and $\hat{\beta}_1$

Recall that

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

As an example, let x be a categorical variable for sex as defined above, and let $\hat{y}$ be a prediction for height.

Thus,

- When $x = 0$, (e.g. when our predictor is that some person is female), we have

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot 0 = \hat{\beta}_0$$

- When $x = 1$, (e.g. when our predictor is that some person is male), we have

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot 0 = \hat{\beta}_0 + \hat{\beta}_1$$

Therefore, $\hat{\beta}_0$ is the predicted height for women (it is also the sample mean height for women), $\hat{\beta}_0 + \hat{\beta}_1$ is the predicted height for men (it is also the sample mean height for men), and $\hat{\beta}_1$ is the average difference in height between men and women.

8 Module 8 - Multiple Linear Regression

9 Module 9 - Classification

10 Module 10 - Ethics