Statistical Inference

Adapted from:

- Significant Statistics, Chapter 6 - Foundations of Inference. John Morgan Russell (2020).
- Statistical Thinking, Chapter 9 - Hypothesis Testing. Russell A. Poldrack (2019).

It is often necessary to "guess", infer, or generalize about the outcome of an event in order to make a decision. Politicians study polls to guess their likelihood of winning an election. Teachers choose a particular course of study based on what they think students can comprehend. Doctors choose the treatments needed for various diseases based on their assessment of likely results. You may have visited a casino where people play games chosen because of the belief that the likelihood of winning is good. You may have chosen your course of study based on the probable availability of jobs.

Statistical Inference

The goal of statistical inference is to **generalise** - to make statements about a population based on a sample.

Statistical inference uses **what we know about probability** to make our **best "guesses"** from *samples* about what we **don't know** about the *population*.

Statistical Inference

Main forms of statistical inference

1. Point estimation
   - Using sample data to **calculate** a single statistic as an *estimate* of an unknown population parameter
   - Example: What is the average height of undergraduates at this university? What is the average construction cost of an office building in London? What was it in 2019?
2. Confidence intervals
   - An interval built **around a point estimate** for an unknown population parameter.
3. Hypothesis testing
   - A decision making procedure for determining **whether sample evidence supports a hypothesis**.

These three examples make up the main forms of statistical inference. However, there are many other forms of statistical inference, such as regression analysis - e.g. How much does building energy use change as occupancy increases?

Point Estimation

Suppose you were trying to determine the mean rent of a two-bedroom apartment in your town. You might look in the classified section of the newspaper, write down several rents listed, and average them together. You would have obtained a point estimate of the true mean. If you are trying to determine the percentage of times you make a basket when shooting a basketball, you might count the number of shots you make and divide that by the number of shots you attempted. In this case, you would have obtained a point estimate for the true proportion.

The most natural way to estimate features of the population (parameters) is to use the **corresponding summary statistic** calculated from the sample. Some common point estimates and their corresponding parameters are found in the following table:

Parameter	Measure	Statistic
μ	Mean of a single population	$\bar{x}$
p	Proportion of a single population	$\hat{p}$
μ_D	Mean difference of two dependent populations	$\bar{x}_D$
$\mu_1 - \mu_2$	Difference in means of two independent populations	$\bar{x}_1 - \bar{x}_2$
$p_1 - p_2$	Difference in proportions of two population	$\hat{p}_1 - \hat{p}_2$
σ^2	Variance of a single population	S^2
σ	Standard deviation of a single population	S

Point Estimation

Parameter	Statistic
μ	$\bar{x}$
p	$\hat{p}$
$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$
$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$
σ^2	S^2
σ	S

Suppose the mean weight of a sample of 60 adults is 173.3 lbs; this sample mean is a point estimate of the population mean weight, μ.

Remember: this is one of many samples that we could have taken from the population.

If a different random sample of 60 individuals were taken from the same population, the new sample mean would likely be different as a result of sampling variability. While estimates generally vary from one sample to another, the population mean is a fixed value.

Suppose a poll suggested the US President's approval rating is 45%. We would consider 45% to be a point estimate of the approval rating we might see if we collected responses from the entire population. This entire-population response proportion is generally referred to as the parameter of interest. When the parameter is a proportion, it is often denoted by p, and we often refer to the sample proportion as $\hat{p}$ (pronounced "p-hat"). Unless we collect responses from every individual in the population, p remains unknown, and we use $ as our estimate of p.

How would one estimate the difference in average weight between men and women? Suppose a sample of men yields a mean of 185.1 lbs and a sample of women men yields a mean of 162.3 lbs. What is a good point estimate for the difference in these two population means? We will expand on this in following chapters.

Unbiased Estimation
Sampling variability

We have established that different samples yield different statistics due to sampling variability. These statistics have their own distributions, called sampling distributions, that reflect this as a random variable. The sampling distribution of a sample statistic is the distribution of the point estimates based on samples of a fixed size, n, from a certain population. It is useful to think of a particular point estimate as being drawn from a sampling distribution.

Recall the sample mean weight calculated from a previous sample of 173.3 lbs. Suppose another random sample of 60 participants might produce a different value of x, such as 169.5 lbs. Repeated random sampling could result in additional different values, perhaps 172.1 lbs, 168.5 lbs, and so on. Each sample mean can be thought of as a single observation from a random variable X. The distribution of X is called the sampling distribution of the sample mean, and has its own mean and standard deviation like the random variables discussed previously. We will simulate the concept of a sampling distribution using technology to repeatedly sample, calculate statistics, and graph them. However, the actual sampling distribution would only be attainable if we could theoretically take an infinite amount of samples.

Each of the point estimates in the table above have their own unique sampling distributions which we will look at in the future

- What makes a statistical estimate of this parameter of interest a "Good" one? It must be both accurate and precise.
- Although variability in samples is present, there remains a fixed value for any population parameter.
- According to the law of large numbers, probabilities converge to what we expect over time.
- Point estimates follow this rule, becoming more accurate with increasing sample size.

Example Dataset - NHANES
National Health and Nutrition Examination Survey (NHANES) from the US Centers for Disease Control (CDC)

```
NHANES_adult |>
  select(c("SurveyYr", "Gender", "Age", "Race1", "Education", "Weight", "Height",
"Pulse", "Diabetes")) |>
  gt_preview() |>
  tab_header(title = "NHANES Dataset")
```

NHANES Dataset

	SurveyYr	Gender	Age	Race1	Education	Weight	Height	Pulse	Diabetes
1	2009_10	male	34	White	High School	87.4	164.7	70	No
2	2009_10	female	49	White	Some College	86.7	168.4	86	No
3	2009_10	female	45	White	College Grad	75.7	166.7	62	No
4	2009_10	male	66	White	Some College	68.0	169.5	60	No
5	2009_10	male	58	White	College Grad	78.4	181.9	62	No
6..4785									
4786	2011_12	male	60	White	College Grad	78.4	168.8	76	Yes

Unbiased Estimation
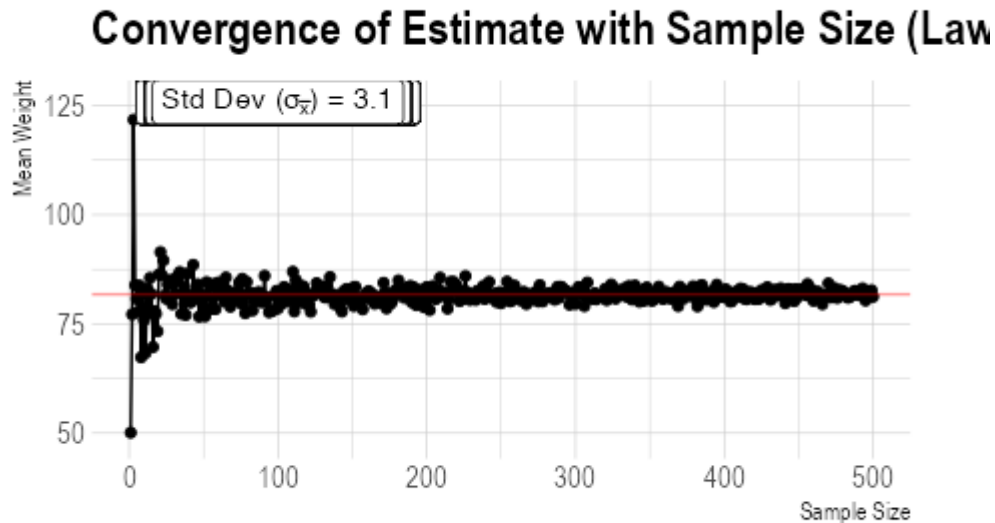
The accuracy of an estimate refers to how well it estimates the actual value of that parameter. Mathematically, this is true when that the expected value your statistic is equal to the value of that parameter. This can be visualized as the center of the sampling distribution appearing to be situated at the value of that parameter.

According to the law of large numbers, probabilities converge to what we expect over time. Point estimates follow this rule, becoming more accurate with increasing sample size. The figure above shows the sample mean weight calculated for random samples drawn, where sample size increases by 1 for each draw until sample size equals 500. The maroon dashed horizontal line is drawn at the average weight of all adults 169.7 lbs, which represents the population mean weight according to the CDC.

The figure above shows the sample mean weight calculated for random samples drawn, where sample size increases by 1 for each draw until sample size equals 500. The maroon dashed horizontal line is drawn at the average weight of all adults 169.7 lbs, which represents the population mean weight according to the CDC.

Note how a sample size around 50 may produce a sample mean that is as much as 10 lbs higher or lower than the population mean. As sample size increases, the fluctuations around the population mean decrease; in other words, as sample size increases, the sample mean becomes less variable and provides a more reliable estimate of the population mean.

In addition to accuracy, a precise estimate is also more useful. This means when repeatedly sampling, the values of the statistics seem pretty close together. The precision of an estimate can be visualized as the spread of the sampling distribution, usually quantified by the standard deviation.

The phrase "the standard deviation of a sampling distribution" is often shortened to the standard error. A smaller standard error means a more precise estimate and is also effected by sample size.

Bibliography