Lecture 12: Sampling Distributions

Chapter 4.1

Goals for Today

Start Chapter 4: Arguably the most important chapter of the book, as it goes to the heart of what statistical inference is. Three important definitions today:

1. Define what a point estimate is
2. Define the sampling distribution
3. Define the standard error

Point Estimates

Definition 1: Point estimates are functions of a random sample of n observations $x_1, \ldots, x_n$.

They estimate the value of some unknown population parameter.

Most common example: the sample mean

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{x_1 + \ldots + x_n}{n}$$

is a point estimate of the true population mean μ

Thought Experiment: Behavior of Point Estimates

Thought experiment: Say we draw a random sample of size $n = 100$ from a large population, where we know the true population mean $\mu = 5$ and $\sigma = 2$ (in real life, we won't know these values).

Let's use the point estimate $\overline{x}$ (the sample mean) to estimate μ.

Two Important Conceptual Questions:

1. If we compute $\overline{x}$ of these points, are we going to get exactly 5?
2. Say we do this once and $\overline{x} = 5.025$. If we repeat this procedure (i.e. generate a new sample of 100 points and compute $\overline{x}$) are we going to get $\overline{x} = 5.025$ exactly?

Thought Experiment: Behavior of Point Estimates

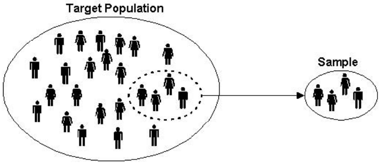
Let's repeat this procedure 1000 times (arbitrarily chosen):

Do this for the 1st time	We get, say, $\overline{x} = 4.831$
Do this for the 2nd time	We get, say, $\overline{x} = 5.104$
Do this for the 3rd time	We get, say, $\overline{x} = 4.965$
...	
Do this for the 1000th time	We get, say, $\overline{x} = 4.957$

Sampling Distributions

In other words, you are repeating the following procedure 1000 times:

- Draw a random sample of size $n = 100$ from the population:



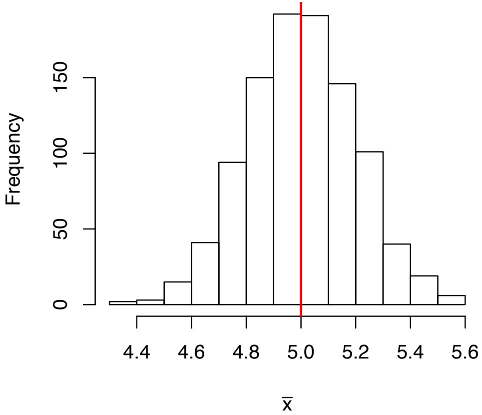
- Compute the sample mean $\overline{x}$ from the sample

The sampling distribution of $\overline{x}$

- describes how these different instances of $\overline{x}$ behave
- has its name because its values are based on samples

Sampling Distribution

Each element in this histogram is one of 1000 instances of $\overline{x}$ from the previous slide, where each $\overline{x}$ is computed from a sample of $n = 100$ values. This is the sampling distribution of $\overline{x}$:



Behavior of Point Estimates

Notice in the histogram:

- the 1000 instances of $\overline{x}$ are centered around the true population mean $\mu = 5$
- there is a spread of the values about the center.

The interval $[4.6, 5.4]$ contains roughly 95% of the data. Since

$$\begin{aligned}
\text{the length of the interval } [\mu - 2SD, \mu + 2SD] &= 4SD \\
\text{the length of the interval } [4.6, 5.4] &= 4SD \\
5.4 - 4.6 = 0.8 &= 4SD \\
SD &= 0.2
\end{aligned}$$

Sampling Distributions

Definition 2: the sampling distribution is the distribution of point estimates based on samples of fixed size n.

i.e. every instance of a point estimate can be thought of as having been drawn from the sampling distribution.

Standard Errors

Definition 3: The standard error is the standard deviation of the sampling distribution of a point estimate. It describes the uncertainty/variability associated with the point estimate.

Very confusing for people: the standard error is a specific kind of standard deviation.

i.e. it describes the typical error in our point estimate $\overline{x}$.

Standard Error of the Sample Mean $\bar{x}$

Given n independent observations from a population with standard deviation σ, the standard error of the sample mean is

$$SE = \frac{\sigma}{\sqrt{n}}$$

A good way to ensure independence of sample observations is to conduct a simple random sample consisting of less than 10% of the population.

Standard Error of the Sample Mean $\overline{x}$

Notice the $\sqrt{n}$ in the denominator: n increases, SE decreases!

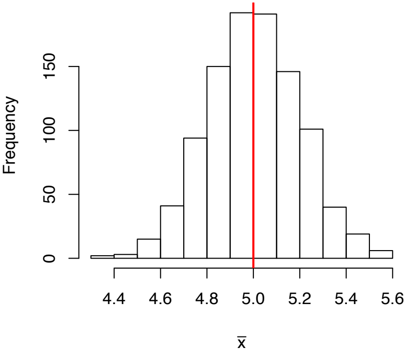
i.e. as the sample size gets bigger, the variability of our point estimate $\overline{x}$ decreases. This is why sample size matters!

Going back to the histogram. We drew samples of size $n = 100$ of data with $\sigma = 2$. We estimated earlier that the standard deviation of the sampling distribution was 0.2. Using the formula

$$SE = \frac{\sigma}{\sqrt{n}} = \frac{2}{\sqrt{100}} = \frac{2}{10} = 0.2$$

Standard Error of the Sample Mean $\overline{x}$

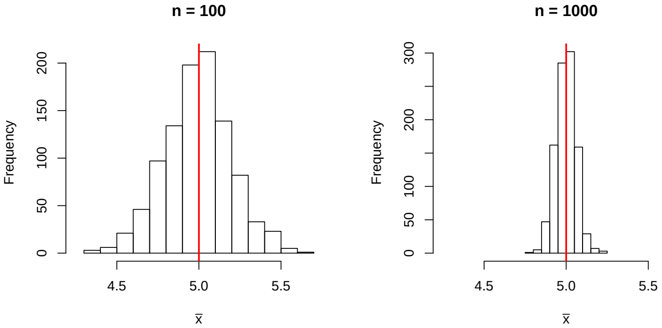
$$SE = \frac{\sigma}{\sqrt{n}} = \frac{2}{\sqrt{100}} = \frac{2}{10} = 0.2$$



Standard Error of the Sample Mean $\overline{x}$

Now compare the sampling distributions based on

- 1000 instances of $\overline{x}$ where each $\overline{x}$ is based on $n = 100$.
  So $SE = \frac{\sigma}{\sqrt{n}} = \frac{2}{\sqrt{100}} = 0.2$
- 1000 instances of $\overline{x}$ where each $\overline{x}$ is based on $n = 1000$.
  So $SE = \frac{\sigma}{\sqrt{n}} = \frac{2}{\sqrt{1000}} = 0.0632$. Smaller!



The estimates on the right are "more precise."

Standard Error of the Sample Mean

In this example we knew σ. However, in real life we won't know σ!

However, when

- the sample size is at least 30
- the population distribution is not strongly skewed

we can use the point estimate of the standard deviation from the sample. i.e. plug in s in place of σ:

$$SE = \frac{s}{\sqrt{n}}$$

Exercise 4.5 on Page 164

Say in you take a simple random sample of 100 runners and you find:

- the sample mean $\overline{x}$ of ages is 35.05
- the sample standard deviation of the runners ages is $s = 8.97$

Assuming that the 100 runners consist of less than 10% of the population (i.e. there are at least 1000 runners in the population), the standard error of the sample mean is

$$SE = \frac{s}{\sqrt{100}} = \frac{8.97}{10} = 0.897$$

Sampling Distributions

We can define the sampling distributions for any point estimate, not just $\overline{x}$:

- s
- the sample median
- the sample minimum/maximum

We will only focus on sample means, including the sample proportion $\widehat{p}$..

Recap

- Point estimates are based on a sample $x_1, \ldots, x_n$ and are used to estimate population parameters.
- The sampling distribution characterizes the (random) behavior of point estimates.
- The standard deviation of a sampling distribution is the standard error: it quantifies the uncertainty/variability of point estimates.

Next Time

- Confidence Intervals
- When quoting survey results, what does: "the results of this survey are estimated to be accurate within 3.1 percentage points, 19 times out of 20" mean?
- Big One: Central Limit Theorem