Lecture 12: Sampling Distributions & Standard Errors

Chapter 4.1

Goals for Today

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

1. point estimate
2. sampling distribution
3. 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.

Ex: 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

Behavior of Point Estimates

Ex: Say we draw a random sample of size $n = 100$ from a large population that is normally distributed with $\mu = 5$ and $\sigma = 2$.

Two Important Questions:

1. Is $\overline{x}$ going to be exactly 5?
2. Say we get $\overline{x} = 5.025$. If we repeat this procedure: i.e. generate a new sample of size $n = 100$ and compute $\overline{x}$), will we get $\overline{x} = 5.025$?

We need to characterize this random error.

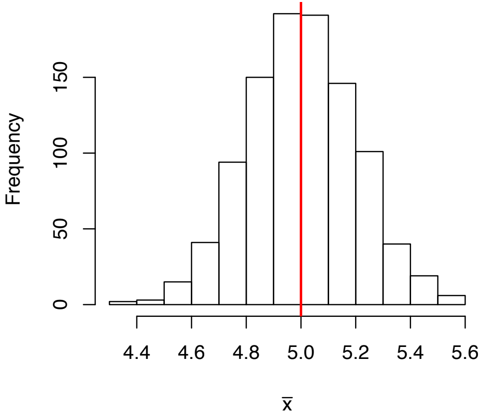
Behavior of Point Estimates

Let's repeat this procedure, say, 1000 times:

1st time	We get $\overline{x} = 4.831$
2nd time	We get $\overline{x} = 5.104$
3rd time	We get $\overline{x} = 4.965$
...	
1000th time	We get $\overline{x} = 4.957$

Sampling Distribution

This histogram is the 1000 instances of $\overline{x}$, where each $\overline{x}$ is based on a sample of $n = 100$. This is the sampling distribution of $\overline{x}$:



Sampling Distributions

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

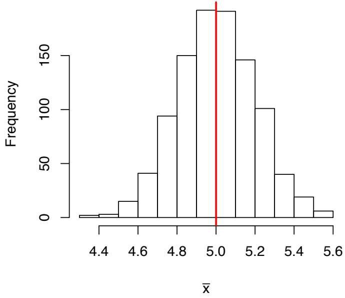
Every instance of a point estimate can be thought of as a draw from the sampling distribution.

If the sampling is representative (unbiased) then the sampling distribution will be centered around the true population parameter (in our case μ).

Sampling Distributions

Measure of Spread

What about spread? $[4.6, 5.4]$ contains roughly 95% of the data.



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

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. In other words, the "typical" error.

Confusing: the standard error is a specific kind of standard deviation.

Standard Error of $\overline{x}$

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

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

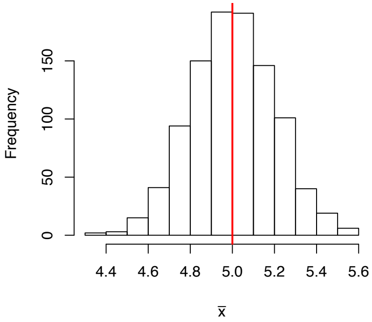
Rule of thumb for independence: You need a simple random sample consisting of less than 10% of the population.

Notice: $\sqrt{n}$ in the denominator: as n increases, SE decreases! This is why sample size matters.

Back to Histogram

Samples were of size $n = 100$ with $\sigma = 2$. We estimated that the SD of the sampling distribution was 0.2. Using the formula:

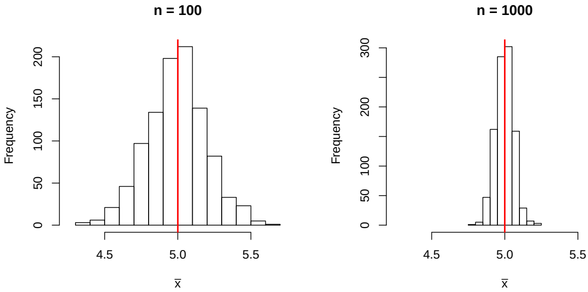
$$SE = \frac{\sigma}{\sqrt{n}} = \frac{2}{\sqrt{100}} = \frac{2}{10} = 0.2$$



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

Compare 1000 instances of $\overline{x}$ when

- $n = 100$. $SE = \frac{\sigma}{\sqrt{n}} = \frac{2}{\sqrt{100}} = 0.2$
- $n = 1000$. $SE = \frac{\sigma}{\sqrt{n}} = \frac{2}{\sqrt{1000}} = 0.0632$. Smaller!



Both are "accurate", but the estimates on the right are "more precise."

Repeated Sampling

Popular question: What's up with this "1000" instances? Why would you take 1000 different samples of size n?

Answer: No, in practice you would not sample repeatedly: you do this only once for the largest n possible.

Rather the 1000 instances of $\overline{x}$ is a theoretical exercise to illustrate that $\overline{x}$'s are random and we characterize its randomness by its sampling distribution and its standard error.

Standard Error of the Sample Mean

In this example we knew σ; typically we won't. However, when

- $n \geq 30$
- the distribution of the population is not strongly skewed

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

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

Example

Say in you take a simple random sample of 100 runners in a race and you are interested in their ages:

- $\overline{x} = 35.05$
- $s = 8.97$

Assuming that the 100 runners consist of less than 10% of the population, the standard error of $\overline{x}$ is

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

Population Distribution vs Sampling Distribution

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