Statistics Review

- A statistic is a *function* of a *sample of data*
- An *estimator* is a statistic
- Population parameter → unknown
- Estimator → used to estimate an unknown population parameter
- The sample, *y*, will be considered random
- Since *y* is random, estimators using *y* will be random

Since estimators are random, they have a _____, given a special name: sampling distribution.

We will obtain properties of the sampling distribution to see if the estimator is "good" or not.

3.1 Random Sampling from the Population

- Typically, we want to know something about a *population*
- The population is considered to be very large (infinite), and contains some unknown "truth"
- We likely won't observe the whole population, but a *sample* from the pop.
- We'll use the sample, *y*, to estimate that something

Example: suppose we want to know the mean height of a male U of M student

Let y = height of a male student

- Population: all male students
- Population parameter of interest: μ_Y

We can't afford to observe the whole pop.

We'll have to collect a *sample*, y.

[Picture]

We want the sample to reflect the population.

Question: How should the sample be selected from the population?

In particular we want the sample to be i.i.d.

- Identically
- Independently
- Distributed

So, the sample *y* is random!!

- Could have gotten a different *y*
- Parallel universe

Table 3.1: Entire population of heights (in cm). The true (unobservable) population mean and variance are $\mu_y = 176.8$ and $\sigma_y^2 = 39.7$.

177.3	170.2	187.2	178.3	170.3	179.4	181.2	180.0	**173.9**
178.7	**171.7**	160.5	183.9	175.7	175.9	**182.6**	181.7	180.2
181.5	176.5	**162.1**	180.3	175.6	**174.9**	**165.7**	172.7	178.9
175.3	178.7	175.6	166.4	173.1	173.2	175.6	183.7	181.3
174.2	180.9	179.9	171.2	171.0	178.6	181.4	175.2	**182.2**
171.7	178.4	**168.1**	186.0	**189.9**	173.4	168.7	180.0	175.1
175.7	180.8	176.2	170.8	177.3	**163.4**	**186.3**	177.1	191.2
171.0	180.3	**169.5**	167.2	178.0	172.9	176.0	176.5	**171.9**
175.1	184.2	165.3	180.2	178.3	183.4	**173.9**	178.6	177.9
184.5	184.1	180.9	187.1	179.9	167.1	**172.0**	167.4	**172.7**
171.6	186.6	182.4	185.5	174.8	178.8	192.8	179.3	**172.0**

How could i.i.d. be violated in the heights example?

Example: mean income of Canadians. How could i.i.d. be violated?

How should we estimate the mean height?

3.2 Estimators and Sampling Distributions

An estimator uses the sample *y* to "guess" something about the pop.

We collect our sample, $y = \{173.9, 171.7, 182.6, 181.5, 162.1, 174.9, 165.7, 182.2, 171.7, 168.1, 189.9, 175.7, 163.4, 186.3, 169.5, 171.9, 173.9, 172.0, 172.7, 172.0\}$. How should we use this sample to *estimate* the mean height?

3.2.1 Sample mean

A popular choice for estimating a population mean is by using a *sample mean* (or *sample average* or just *average*)

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad (3.1)$$

- From heights example: $\bar{y} = 174.1$, $\mu_y = 176.8$
- There are many ways to estimate μ_y. Examples?
- Why is (3.1) so popular?
- How good is $\bar{y}$ at estimating μ_y in general?
- To answer these questions: idea of a *sampling distribution*

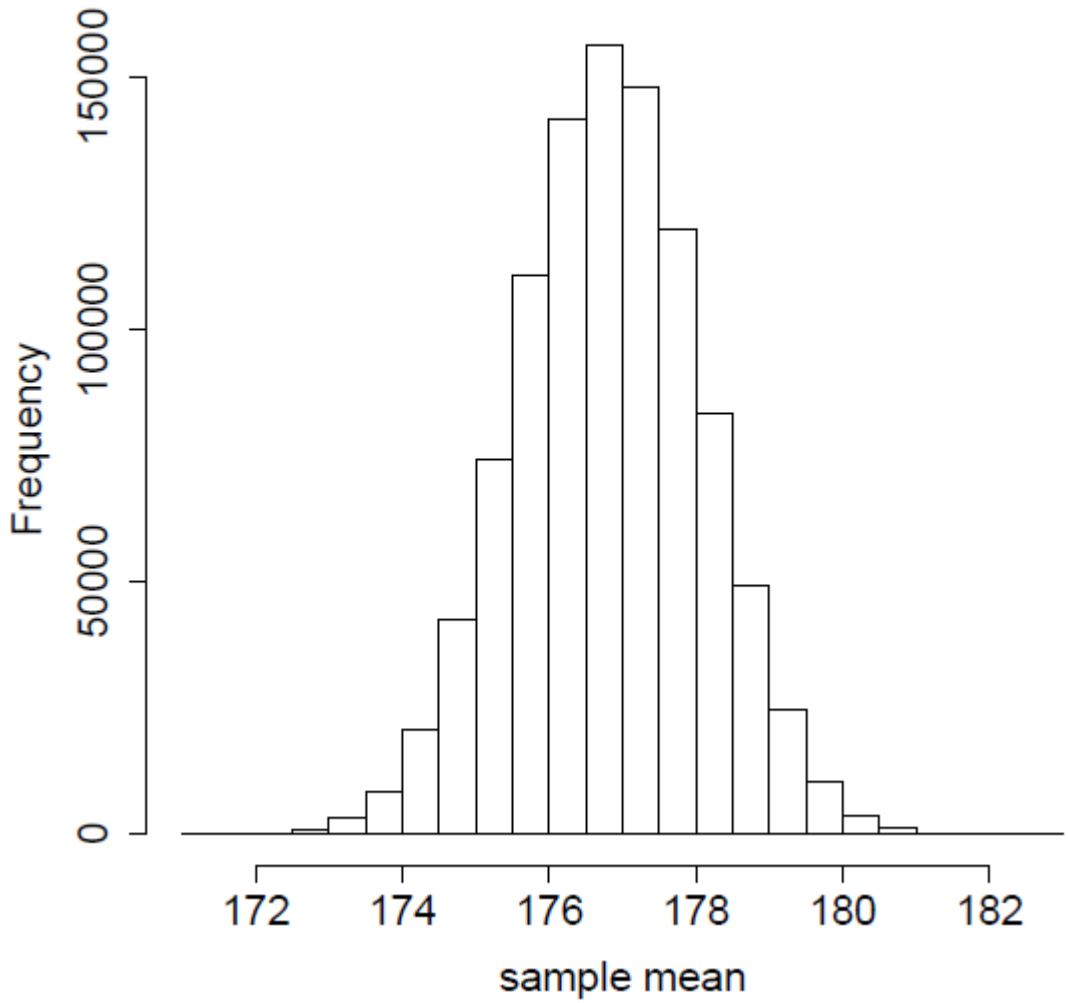
Recall that the sample, y, is random. Each element of y was selected randomly from the population. We could have selected a different sample of size $n = 20$. For example, in a parallel universe, we could have gotten $y^* = \{175.9, 175.3, 182.2, 178.6, 175.2, 180.3, 178.3, 183.7, 176.0, 167.4, 178.7, 178.7, 186.0, 175.6, 180.0, 168.7, 178.6, 173.1, 173.2, 187.1\}$, where the * in y^* denotes that we are in the parallel universe. In this parallel universe, we got $\bar{y}^* = 177.6$. But in every universe, the population (table 3.1), is the same.

- Randomly sample from the population → get y
  - y is random
- Use y to calculate $\bar{y}$
  - $\bar{y}$ is random
  - could have gotten a different sample → could have gotten a different $\bar{y}$
  - population is always the same (μ_y)

3.2.2 Sampling distribution of the sample mean

- $\bar{y}$ is random variable (it's an estimator, all estimators are random)
- random variables usually have probability functions
- $\bar{y}$ has a *sampling distribution* (probability function for an estimator)
- *sampling distribution* – imagine all possible values for $\bar{y}$ that you could get – plot a histogram
- Using a computer, I drew 1 mil. different random samples of n=20 from table 3.1. Calculate $\bar{y}$ each time. Plot histogram:

Figure 3.1: Histogram for 1 million $\bar{y}$s



Which probability function is right for $\bar{y}$? Why?

- Look at figure 3.1
- Notice the summation operator in equation 3.1
- Answer: _____ Reason: _____

$\bar{y}$ is random. We'll derive its:

- mean
- variance

Use these to determine if it's a "good" estimator via three statistical properties:

- Bias
- Efficiency
- Consistency

3.2.3 Bias

An estimator is unbiased if its expected value is equal to the population parameter it's estimating.

That is, $\bar{y}$ is unbiased if $E[\bar{y}] = \mu_y$

Unbiased if it gives "the right answer on average".

Biased if it gives the wrong answer on average.

$$
\begin{aligned}
\mathrm{E}\left[\bar{y}\right] &= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} y_i\right] \\
&= \frac{1}{n}\mathrm{E}\left[\sum_{i=1}^{n} y_i\right] \\
&= \frac{1}{n}\mathrm{E}\left[y_1 + y_2 + \cdots + y_n\right] \\
&= \frac{1}{n}\left(\mathrm{E}\left[y_1\right] + \mathrm{E}\left[y_2\right] + \cdots + \mathrm{E}\left[y_n\right]\right) \\
&= \frac{1}{n}\left(\mu_y + \mu_y + \cdots + \mu_y\right) \\
&= \frac{n\mu_y}{n} = \mu_y
\end{aligned}
\tag{3.2}
$$

3.2.4 Efficiency

An estimator is efficient if it has the smallest variance among all other potential estimators (for us, potential = linear, unbiased)

Need to get the variance of $\bar{y}$.