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*

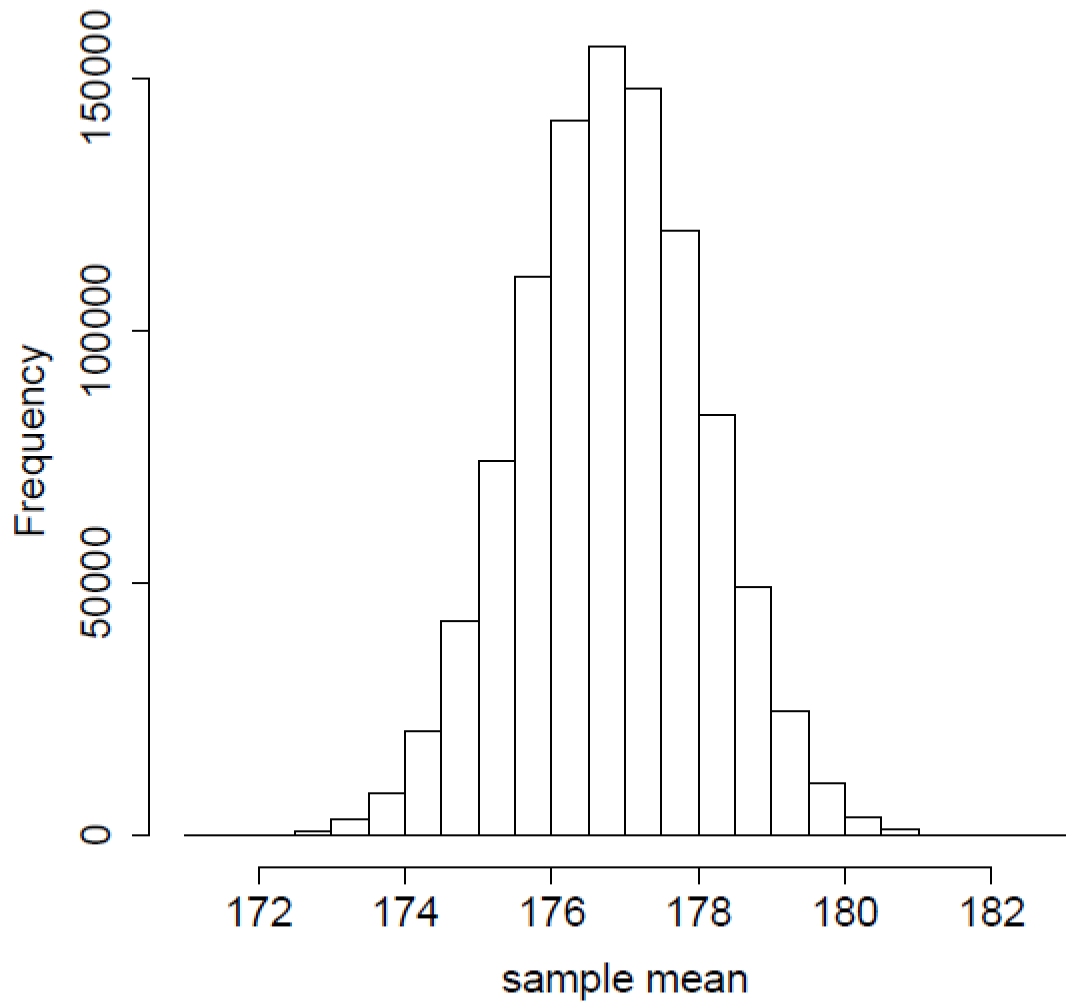
Recall that the sample, y, is random. Each element of y was selected randomly from the population. We could have selected a different sample of size $n = 20$. For example, in a parallel universe, we could have gotten y^* = {175.9, 175.3, 182.2, 178.6, 175.2, 180.3, 178.3, 183.7, 176.0, 167.4, 178.7, 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 $\bar{y}$.

$$
\begin{aligned}
\mathrm{Var}[\bar{y}] &= \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} y_i\right] \\
&= \frac{1}{n^2}\mathrm{Var}\left[\sum_{i=1}^{n} y_i\right] \\
&= \frac{1}{n^2}\mathrm{Var}\left[y_1 + y_2 + \cdots + y_n\right] \\
&= \frac{1}{n^2}\left(\mathrm{Var}[y_1] + \mathrm{Var}[y_2] + \cdots + \mathrm{Var}[y_n]\right) \\
&= \frac{1}{n}\left(\sigma_y^2 + \sigma_y^2 + \cdots + \sigma_y^2\right) \\
&= \frac{n\sigma_y^2}{n^2} = \frac{\sigma_y^2}{n}
\end{aligned}
\tag{3.3}
$$

- Gauss-Markov theorem proves this is minimum variance
- We'll also need this to prove consistency, and for hyp. testing

3.2.5 Consistency

Suppose we had a lot of information. ($n \to \infty$)

What value should we get for our estimator?

How would state this mathematically?

Q) Prove that the sample mean is a consistent estimator for the population mean.

Q) Define the terms unbiasedness, efficiency, and consistency.

3.3 Hypothesis tests (known σ_y^2)

$$\begin{aligned} H_0 &: \mu_y = \mu_{y,0} \\ H_A &: \mu_y \neq \mu_{y,0} \end{aligned} \tag{3.4}$$

- Estimate μ_y (using $\bar{y}$ for example)
- See if $\bar{y}$ appears "close" to $\mu_{y,0}$
  - Remember, $\bar{y}$ is random! (and Normal)
- If it's close → fail to reject
- If it's far → reject

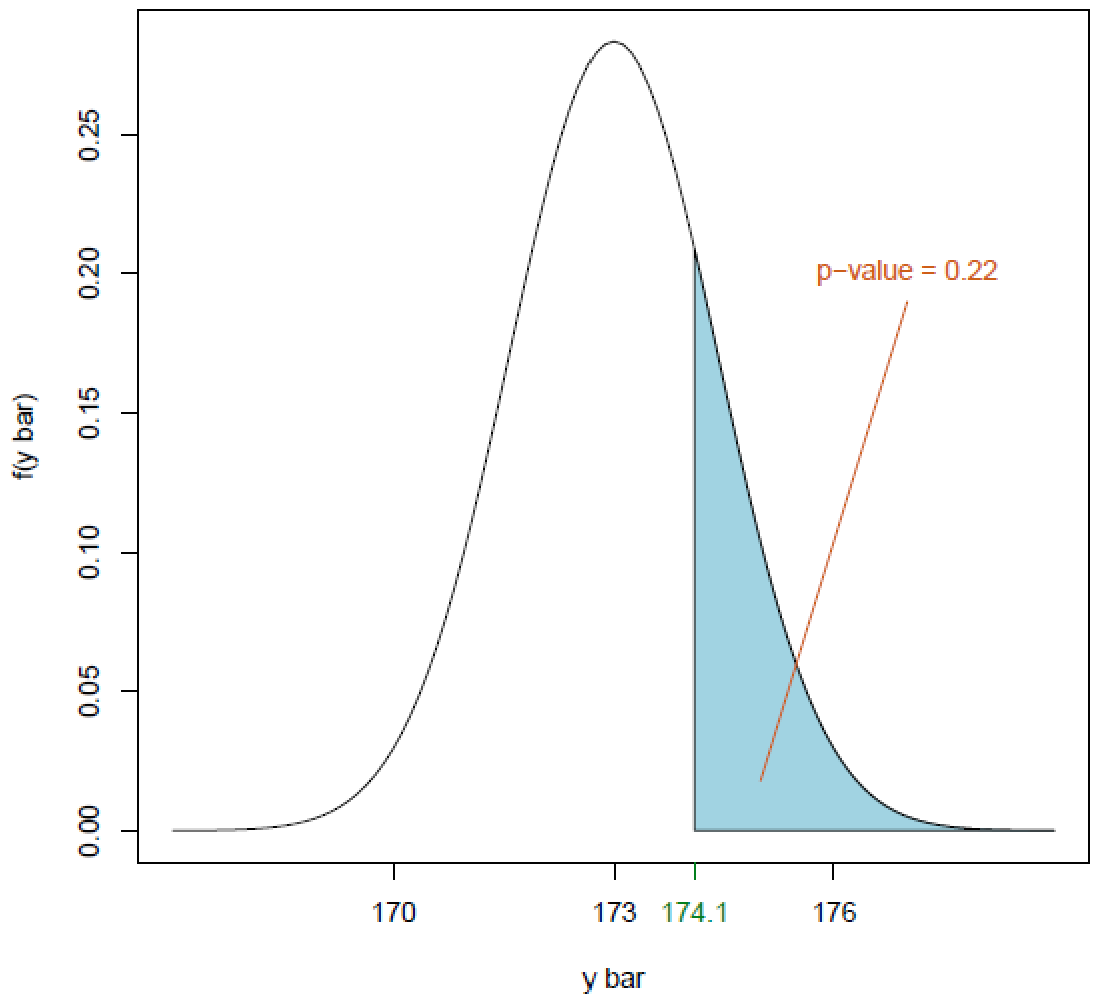
Example:

- Hypothesize that mean height of a U of M student is 173cm

$$H_0 : \mu_y = 173$$
$$H_A : \mu_y \neq 173 \qquad (3.5)$$

- Collect a sample: $y = \{173.9, 171.7, \ldots, 172.0\}$
- Calculate $\bar{y} = 174.1$
- Suppose (very unrealistically that we know that) $\sigma_y^2 = 39.7$
- What now?

Figure 3.2: Normal distribution with $\mu = 173$ and $\sigma^2 = {}^{39.7}/_{20}$. Shaded area is the probability that the normal variable is greater than 174.1.



The p-value for the above test is 0.44. How to interpret this?

3.3.1 Significance of a test

3.3.2 Type I error

3.3.3 Type II error (and power)

3.3.4 Test statistics

- Just a more convenient way of getting the p-value for the test
- Each hypothesis test would present us with a new normal curve that we would have to draw, and calculate a new area (see fig. 3.2)
- Instead: *standardize*
- This gives us *one curve for all testing problems* (the standard normal curve)
- Calculate a bunch of areas under the curve, and tabulate them
- Not an issue with modern computers, but this is still the way we do things
- How to get a z test statistic?
- Do a z test for our heights example.

Table 3.2: Area under the standard normal curve, to the right of z.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002

3.3.5 Critical values

3.3.6 Confidence intervals

What is the probability that our z statistic will be within a certain interval, if the null hypothesis is true? For example, what is the following probability?

$$\Pr\left(-1.96 \leq z \leq 1.96\right)? \tag{3.12}$$

$$\Pr\left(-1.96 \leq \frac{\bar{y} - \mu_{y,0}}{\sqrt{\sigma_y^2/n}} \leq 1.96\right) = 0.95 \tag{3.13}$$

Finally, we solve equation 3.13 so that the null hypothesis $\mu_{y,0}$ is in the middle of the probability statement:

$$\Pr\left(\bar{y} - 1.96 \times \sqrt{\frac{\sigma_y^2}{n}} \leq \ \mu_{y,0} \ \leq \bar{y} + 1.96 \times \sqrt{\frac{\sigma_y^2}{n}}\right) = 0.95 \tag{3.14}$$

3.4 Hypothesis Tests (unknown σ_y^2)

- Much more realistically, σ_y^2 (variance of *y*) will be unknown.
- Recall that: $Var[y] = \sigma_y^2/n$
- $z = \frac{\bar{y}-\mu_{y,0}}{s.e.(\bar{y})} = \frac{\bar{y}-\mu_{y,0}}{\sqrt{\sigma_y^2/n}}$
- So, we need to estimate σ_y^2 in order to perform hypothesis tests.

3.4.1 Estimating σ_y^2

- A "natural" estimator:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 \qquad (3.15)$$

- Is this a good estimator? Why or why not?
- A better estimator:

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \qquad (3.17)$$

- Degrees-of-freedom correction

So:

$$\text{Estimated variance of } \bar{y} = \frac{s_y^2}{n}$$

We can implement hypothesis testing by replacing the unknown σ_y^2 with its estimator s_y^2. The z test statistic now becomes:

$$\frac{\bar{y} - \mu_{y,0}}{\sqrt{s_y^2/n}} = t$$

Note: for large *n*, the *t* test is equivalent to the *z* test