Lecture 15:

- Statistics = Applications of the Central Limit Theorem
- Sampling Theory
- Point Estimates: warmup -- when the population parameters are known
- Confidence Intervals -- when the population parameters are known
Review: CLT and the Normal Distribution

Now suppose we consider the random variable \( \bar{X}_n \) representing the mean of the \( X_i \), i.e.,

\[
\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}
\]

The Central Limit Theorem

As \( n \) gets large, the random variable \( \bar{X}_n \) converges to the distribution \( N \left( \mu, \frac{\sigma^2}{n} \right) \).

There are several crucial things to remember about the CLT:

1. The mean \( \mu \) of \( \bar{X}_n \) is the same as the \( X_i \).
2. But the standard deviation \( \frac{\sigma}{\sqrt{n}} \) gets smaller as \( n \) gets larger, and approaches 0 as \( n \) approaches \( \infty \).
3. The distributions of the \( X_i \) do NOT MATTER at all, and as long as they have a common mean and standard deviation, they can be completely different distributions. Typically, however, these are separate "pokes" of the same random variable.
4. We can use the strong properties of the normal distribution, such as the "68-95-99 rule," to quantify the randomness inherent in the sampling process. This will be the fundamental fact we will use in developing the various statistical procedures in elementary statistics.
Review: CLT and the Normal Distribution

The 68 – 95 – 99 Rule

For the normal distribution, the values less than one standard deviation away from the mean account for 68.27% of the set; while two standard deviations from the mean account for 95.45%; and three standard deviations account for 99.73%.

Actually, we can be more precise...

$$+/- \ 1 \ \text{sigma} = 0.682689492137$$
$$+/- \ 2 \ \text{sigma} = 0.954499736104$$
$$+/- \ 3 \ \text{sigma} = 0.997300203937$$
$$+/- \ 4 \ \text{sigma} = 0.999936657516$$
$$+/- \ 5 \ \text{sigma} = 0.999999426697$$
$$+/- \ 6 \ \text{sigma} = 0.999999998027$$
$$+/- \ 7 \ \text{sigma} = 0.999999999997$$
Review: CLT and the Normal Distribution

Example: Let $X \sim N(66, 3^2)$. We calculated the mean for $n = 100$, so we should get a standard deviation smaller by a factor of 10:

\[
\frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{100}} = \frac{3}{10} = 0.3
\]

\[
\bar{X}_{100} = \frac{X_1 + X_2 + \cdots + X_{100}}{100}
\]

\[
\bar{X}_{100} \sim N(66, (0.3)^2)
\]

```
mu = 66
sigma = 3
n = 100 # try for 1, 2, 5, 10, 30, 100
num_trials = 10000
display_sample_mean_normal(mu, sigma, n, num_trials, 2)
```
Sampling Theory

Recall: Sampling is the process of randomly selecting outcomes from a population, which is really just a random variable; the terminology for samples is slightly different for characteristics of the sample and population:

**Population X**

Randomly sample n outcomes

A "trial" is one such selection of n samples.

**Sample of size n:**

Sampling is generally done with replacement, but if the population is very large (perhaps infinite) it does not matter!

**Sample Statistics**

- mean $\mu$
- variance $\sigma^2$
- standard deviation $\sigma$
- mean $\bar{x}$
- variance $s^2$
- standard deviation $s$
Sampling Theory

The sample statistics are estimators of the population parameters. They are also random variables (a function of the original random variable X). We will focus on the sample mean:

$$\bar{x} = \bar{x}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

In particular, we will use the CLT and focus on the sampling distribution of the sample mean, e.g.,

$$\bar{x} \sim N(66, (0.3)^2)$$
**Sampling Theory**

**Analogy:** You want to know the height of BU students. Every day you select 100 students and measure them and take the mean. This is one trial (one “poke” of the sample mean random variable) and produces one number (a sample statistic). This sampling distribution of the sample mean is what results when you do 10,000 trials, or 10,000 “pokes” of the sample mean random variable.

It’s random variables, functions of random variables, and distributions all over again!
Sampling Theory When Population Parameters are Known

This is a warm-up to the real situation.....

Suppose (humor me!) that you have the actual height data about all BU students, including the mean and standard deviation, but then you LOSE all the data, but somehow you remember that the standard deviation is \( \sigma = 3 \) inches.

Furthermore, you need to know the mean height, but you don’t have a lot of time, and in any case you only need an approximation (an estimate) of the true mean \( \mu \).

What to do? Sample 100 randomly-selected students (one trial) and use the sample mean as your estimate! (Think, polling: you ask 100 random people who they voted for.)

When you report your result, you have an estimate, and you can use the CLT to give precise information about how accurate your estimate is. This is called a Confidence Interval...
Confidence Intervals When Population Parameters are Known

So you know that the actual standard deviation is $\sigma = 3$ inches and you want to estimate the unknown actual mean height $\mu$ by using one trial, one “poke” of the sample mean estimator $\bar{x}$, and you know by the CLT what the sampling distribution looks like. You just don’t know where the centerpoint $\mu$ is:

$$\bar{x} \sim N(\mu, (0.3)^2)$$

![Diagram showing confidence intervals]

- 68.27% confidence interval: $\mu - 0.9$ to $\mu + 0.9$
- 95.45% confidence interval: $\mu - 0.6$ to $\mu + 0.6$
- 99.73% confidence interval: $\mu - 0.3$ to $\mu + 0.3$

$$\sigma_{\bar{x}_{100}} = \frac{3}{\sqrt{100}} = \frac{3}{10} = 0.3$$
Confidence Intervals When Population Parameters are Known

But what you DO know is that whatever number you get for $\bar{x}$ from one trial of measuring 100 students, you have 68.27% chance of being within 0.3 inches of the true mean, 95.45% chance of being within 0.6 inches, and 99.73% of being within 0.9 inches:
Confidence Intervals When Population Parameters are Known

But what you DO know is that whatever number you get for $\bar{x}$ from one trial of measuring 100 students, you have 68.27% chance of being within 0.3 inches of the true mean, 95.45% chance of being within 0.6 inches, and 99.73% of being within 0.9 inches:

$\bar{x} \sim N(\mu, (0.3)^2)$
Confidence Intervals When Population Parameters are Known

But what you DO know is that whatever number you get for $\bar{x}$ from one trial of measuring 100 students, you have 68.27% chance of being within 0.3 inches of the true mean, 95.45% chance of being within 0.6 inches, and 99.73% of being within 0.9 inches:

$$\bar{x} \sim N(\mu, (0.3)^2)$$
Confidence Intervals When Population Parameters are Known

But what you DO know is that whatever number you get for $\bar{x}$ from one trial of measuring 100 students, you have 68.27% chance of being within 0.3 inches of the true mean, 95.45% chance of being within 0.6 inches, and 99.73% of being within 0.9 inches:

$\bar{x} \sim N(\mu, (0.3)^2)$
Confidence Intervals When Population Parameters are Known

But what you DO know is that whatever number you get for $\bar{x}$ from one trial of measuring 100 students, you have 68.27% chance of being within 0.3 inches of the true mean, 95.45% chance of being within 0.6 inches, and 99.73% of being within 0.9 inches:
Confidence Intervals When Population Parameters are Known

But notice that what we are really talking about is the probability of the distance $|\bar{x} - \mu|$ being within bounds guaranteed by the CLT:

- $P(|\bar{x} - \mu| \leq \sigma) = 0.6827$
- $P(|\bar{x} - \mu| \leq 2\sigma) = 0.9545$
- $P(|\bar{x} - \mu| \leq 3\sigma) = 0.9973$

$x \sim N(\mu, (0.3)^2)$
Confidence Intervals When Population Parameters are Known

But notice that what we are really talking about is the probability of the distance $|\bar{x} - \mu|$ being within bounds guaranteed by the CLT:

$\bar{x} \sim N(\mu, (0.3)^2)$

$P(|\bar{x} - \mu| \leq \sigma) = 0.6827$

$P(|\bar{x} - \mu| \leq 2\sigma) = 0.9545$

$P(|\bar{x} - \mu| \leq 3\sigma) = 0.9973$
Confidence Intervals When Population Parameters are Known

But then because the normal is symmetric, it does not matter if we change our perspective to use a sampling distribution centered on $\mu$ or on $\bar{x}$:

$$P(|\bar{x} - \mu| \leq \sigma) = 0.6827$$
$$P(|\bar{x} - \mu| \leq 2\sigma) = 0.9545$$
$$P(|\bar{x} - \mu| \leq 3\sigma) = 0.9973$$
Confidence Intervals When Population Parameters are Known

But then because the normal is symmetric, it does not matter if we change our perspective to use a sampling distribution centered on $\mu$ or on $\bar{x}$:

$$\bar{x} - 0.9 \quad \bar{x} - 0.6 \quad \bar{x} - 0.3 \quad \bar{x} \quad \bar{x} + 0.3 \quad \bar{x} + 0.6 \quad \bar{x} + 0.9$$

$\mu \sim N(\bar{x}, (0.3)^2)$

- $P(|\bar{x} - \mu| \leq \sigma) = 0.6827$
- $P(|\bar{x} - \mu| \leq 2\sigma) = 0.9545$
- $P(|\bar{x} - \mu| \leq 3\sigma) = 0.9973$
Confidence Intervals When Population Parameters are Known

So we can **pretend** that the population mean is normally distributed around the sample mean (not true in general, but for one sample, it is effectively the same thing).
Confidence Intervals When Population Parameters are Known

So we can **pretend** that the population mean is normally distributed around the sample mean (not true in general, but for one sample, it is effectively the same thing).

\[
\mu \sim N\left(\bar{x}, \frac{\sigma^2}{n}\right)
\]
Confidence Intervals When Population Parameters are Known

So we can pretend that the population mean is normally distributed around the sample mean (not true in general, but for one sample, it is effectively the same thing).

\[ \mu \sim N\left( \bar{x}, \frac{\sigma^2}{n} \right) \]
Confidence Intervals When Population Parameters are Known

**Conclusion:** How to report results of sampling experiment to determine the mean of a population (when $\sigma$ of population is known):

Given a population $X$ with known standard deviation $\sigma$:

1. Choose a sample size $n$;
2. Choose a confidence level $CL$ (e.g., 95.45%);
3. Calculate the multiplier $k$ for $\sigma$ corresponding to $CL = P(\mu - k\sigma \leq \bar{x} \leq \mu + k\sigma)$;
4. Perform random sampling and calculate $\bar{x}$;
5. Report your results using the confidence interval corresponding to $CL$:
   
   “The mean of the population is $\bar{x} \pm k\sigma$ with a confidence of $CL$.”

```r
#c. Find P(-k<X<k) for standard normal
k = 2
CL = norm.cdf(x=k,loc=0,scale=1) - norm.cdf(x=-k,loc=0,scale=1)
print(CL)
0.954499736104
```
Confidence Intervals When Population Parameters are Known

Example -- Height of BU Students:

Suppose we know that the height of BU students has standard deviation \( \sigma = 3 \) inches.

1. Choose a sample size \( n = 100 \);
2. Choose a confidence level \( CL = 95.45\% \);
3. Calculate the multiplier \( k = 2 \);
4. Perform random sampling of 100 students and calculate \( \bar{x} = 66.134 \) inches;
5. Report your results using the confidence interval corresponding to \( CL \):

   “The mean height of BU students is 66.134 +/- 0.6 inches with a confidence of 95.45%.”

or change the confidence level if you wish:

   “The mean height of BU students is 66.134 +/- 0.9 inches with a confidence of 99.73%.”
Confidence Intervals When Population Parameters are Known

Caveat: There is a one-to-one correspondence between confidence levels and \( k \), but unfortunately these do not correspond to nice, round numbers on each side. So just be aware of whether you want, for example, “two standard deviations” or “95%” (which are different). Also realize that “95.45%” is an approximation of “two standard deviations”:

```python
#c. Find \( P(-k<X<k) \) for standard normal
CL = norm.cdf(x=2, loc=0, scale=1) - norm.cdf(x=-2, loc=0, scale=1)
print("CL for k = 2: " + str(CL))
CL = norm.cdf(x=3, loc=0, scale=1) - norm.cdf(x=-3, loc=0, scale=1)
print("CL for k = 3: " + str(CL))

#f give the endpoints of the range for the central alpha percent # of the distribution
print("\n90%: " + str(norm.interval(alpha=0.90, loc=0, scale=1)))
print("95%: " + str(norm.interval(alpha=0.95, loc=0, scale=1)))
print("99%: " + str(norm.interval(alpha=0.99, loc=0, scale=1)))

CL for k = 2: 0.954499736104
CL for k = 3: 0.997300203937

90%: (-1.6448536269514729, 1.6448536269514722)
95%: (-1.959963984540054, 1.959963984540054)
99%: (-2.5758293035489004, 2.5758293035489004)
```