Math 247: The Central Limit Theorem for Sample Proportions (Section 7.3)
and Introduction to Hypothesis Testing (Section 8.1)

The goal here is to discuss the Sampling Distribution for $\hat{p}$, and turn the Sampling Distribution into a tool for finding statistically significant proportion (Section 8.1)

> Remember, since we rarely have census data, we have to INFER what's true about a population by using sample data!

We can use the Normal Distribution as an approximation for the Sampling Distribution of $\hat{p}$, as long as the following conditions are met in the problem:

Conditions:

1. **Random and Independent:** We must have a random sample from a population and the observations must be independent from one another.

2. **Large Sample:** The sample size is large enough to make the distribution approximately normal.

   How to check this: **Expected Count for each group must be at least 10**:
   Success group: $E = np$ (if p is unknown, use $E = n\hat{p}$)
   Failure group: $E = n(1-p)$ (if p is unknown, use $E = n(1-\hat{p})$)

3. **Large Population:** The population is at least 10 times greater than the sample. This is because we're *sampling without replacement* and the large sample from a small population would change the proportions!

If all these conditions are satisfied, we can use the Normal Distribution to approximate probabilities for $\hat{p}$.

Important values and formulas:

Assumed or Stated proportion

p = *population proportion* — Center!

$\hat{p}$ = *sample proportion* — given the real proportion is p, these are all the possible sample proportions we could get

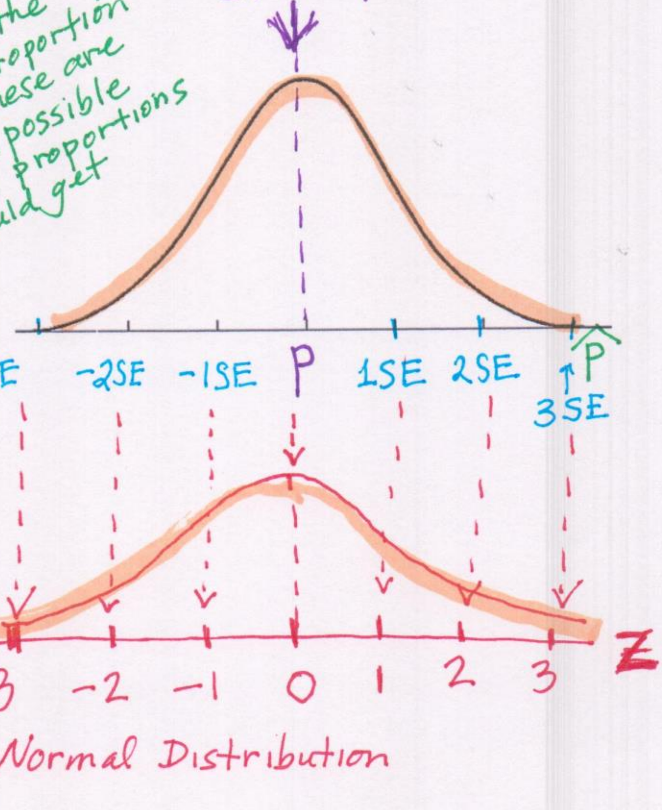
$$\hat{p} = \frac{x}{n} \quad \begin{matrix} \text{— successes} \\ \text{— sample size} \end{matrix}$$

n = *sample size*

also noted as

$$\sigma_{\hat{p}} = SE = \underbrace{\sqrt{\frac{p(1-p)}{n}}}_{\text{Larger samples will make SE smaller!}}$$

$$\boxed{z = \frac{\text{observed} - \text{center}}{\text{SD}} = \frac{\hat{p}-p}{SE}}$$



As always, Z measures how unusual (or likely) a particular value is in the Normal Distribution

Example: Is poverty increasing in SLO? The U.S. Constitution requires the government to conduct a census every 10 years. (This cost about \$15 billion in 2010!) In the last census (year 2010), the data showed 12.8% of the residents in SLO county were living below the poverty line. In 2016, the American Community Survey (also conducted by the U.S. Census Bureau) found that in a sample of 3000 residents in SLO county, 426 (14.2%) were below the poverty line. *(source: http://www.slohealthcounts.org/indicators/index/view?indicatorId=347&localeId=277)*

$n = 3000$

→ this is a "proportion" (percent)

another proportion → came from SAMPLE!

Since the proportion (14.2%) didn't come from census data, could this apparent increase just be the result of sampling variability (also known as "Sampling Error")? Or, was this a statistically significant increase in the proportion of people living in poverty? How can we tell?

Hypothesis Test!

To answer this, we can find the probability that 14.2% or more of the people in the sample would be living in poverty **if** the true population proportion was actually still 12.8%.

We can organize this analysis into a Hypothesis Test. → What are the 4 steps we learned before?

Before beginning, it's super helpful to set up a Parking Lot.

Parking Lot:

Population Proportion: $p = .128$ (assumed)

Sample size: $n = 3000$

Sample: $\hat{p} = .142$ { this is 426 out of 3000 $\hat{p} = \frac{426}{3000} = .142$

Step 1: Hypothesize

Assume, for argument's sake, that nothing has changed in SLO; i.e., that poverty levels are the same as before.

This assumption (the "**IF**" above) is called the "Null Hypothesis" and is denoted H_O.
Write the Null hypothesis in words and in symbols.

H_O : $p = .128$

We could write $p_{2016} = .128$

The POPULATION of SLO residents in 2016 still has the same PROPORTION of residents living in poverty as before (in 2010)

As an alternative, we could hypothesize that something HAS changed. Before looking at the ACS data, we might have thought poverty was increasing or decreasing or weren't sure, so just wanted to test to see whether it could have changed. We would then have an "Alternative Hypothesis" and is denoted H_a.
Write the Alternative Hypothesis in words and symbols.

H_a : $p \neq .128$

(again $p_{2016} \neq .128$

The POPULATION has ~~more or~~ a larger or smaller PROPORTION of people living in poverty.

$\neq$ { means, there's been a change, things are different in 2016.

(We don't have 2016 CENSUS data, so we can't know for sure, BUT if we get SAMPLE data, we can see if it's suggesting there's been a change!)

Step 2: Prepare (Check conditions, make assumptions)

This is where the **Central Limit Theorem** comes in. We have to check that the conditions to use the Central Limit Theorem are satisfied by our sample. → This pertains to the SAMPLE info.

Check that the Conditions are satisfied

✓ a) **Random sample and Independent observations?**

Random sample from SLO Pop? Assume

Independent observations? Assume

✓ b) **Large Sample?**

Note: We have defined "success" and "failure" by how we've written our hypotheses.

Success group: People in poverty "Success" is NOT a value judgement!

✓ Check: $E = np \geq 10$? $n = 3000$, $p = .127$

$E = 3000 \cdot (.127) = 381 \geq 10$? Yes!

✓ Failure group: People not in poverty

Check: $E = n(1-p) \geq 10$?

$E = 3000(1 - .127) = 2619 \geq 10$? Yes!

OR $E = 3000 - 381 = 2619$

The Sample is large enough for us to use the Normal Distribution to find probabilities & to model this situation

✓ c) **Large Population?** $Pop \geq 10 \cdot n$

Since we're sampling without replacement, we have to make sure removing people by sampling isn't changing the overall proportion much. (see Chapter 5)

Pop of ALL SLO residents $\geq 10 \cdot n = 10(3000) = 30{,}000$?

Yes, SLO Pop is more than 30K

All conditions are satisfied, so it's safe to use the Normal model. "Safe" means our results will be reliable!

P-value!

Step 3: Compute to Compare:

Now we're going to find the probability that we would get an observation as or more extreme as what we did (the higher level of poverty) IF the null hypothesis was actually true.

If nothing has changed in SLO, how likely is to get a SAMPLE with 14.2% in poverty?

$p = .127$

- Draw a normal curve that represents the Sampling Distribution of $\hat{p}$. Use the Standard Error formula to find the standard deviation of the sampling distribution.

$$SE = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{.127(1-.127)}{3000}} = \boxed{.00608}$$

use p from null

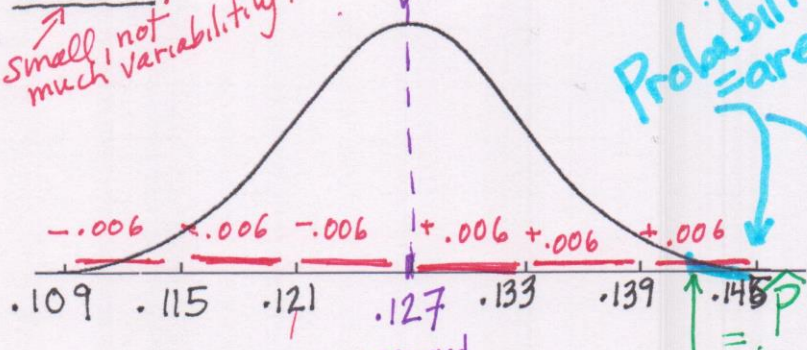
small, not much variability!

Null! $p = .127$

Probability = area

- Plot the Sample Proportion and shade in the area that represents the probability you're trying to find. First estimate the z-score then find it.

$$z = \frac{\text{observed} - \text{center}}{\text{SD}} = \frac{\hat{p} - p}{SE}$$



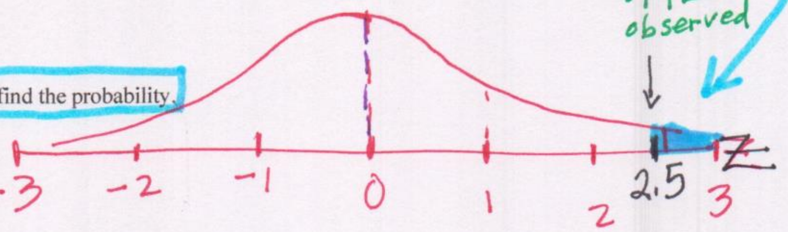
$$Z = \frac{.142 - .127}{.006}$$

$$\boxed{Z = 2.5}$$

- Use the Normal Distribution Calculator to find the probability.

$P(\hat{p} \geq .142)$

$P(Z \geq 2.5) = .00629$



Step 4: Interpret:

What does this probability tell us?

Both the Z-score and the probability tell us it is REALLY unlikely (only a .6% chance!) that we would get a SAMPLE with 14.2% of people in poverty IF the true value in the POPULATION was still 12.7% of people in poverty.

Would it be more reasonable to conclude that poverty levels have stayed the same or that poverty levels have changed, based on this result?

It is much more reasonable to conclude that poverty levels have actually increased – they have NOT stayed the same, based on the evidence in the sample!

Note: The actual P-value for this hypothesis test would be $2 \cdot (.00629) = .01258$ – We'll discuss this in the next section!

Note: the 7.3 homework will just have you checking conditions and computing probabilities (Steps 2 and 3).