LO 1. Define population proportion p (parameter) and sample proportion $\hat{p}$.

LO 2. Calculate the sampling variability of the proportion, the standard error, as $SE = \sqrt{\frac{p(1-p)}{n}}$, where p is the population proportion.

- Note that when the population proportion p is not known (almost always), this can be estimated using the sample proportion, $SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

LO 3. Recognize that the Central Limit Theorem (CLT) is about the distribution of point estimates, and that given certain conditions, this distribution will be nearly normal. In the case of the proportion the CLT tells us that if

- the observations in the sample are independent,
- the sample size is sufficiently large (checked using the success/failure condition: $np \geq 10$ and $n(1-p) \geq 10$,

then the distribution of the sample proportion will be nearly normal, centered at the true population proportion and with a standard error of $SE = \sqrt{\frac{p(1-p)}{n}}$.

$$\hat{p} \sim N\left(mean = p, SE = \sqrt{\frac{p(1-p)}{n}} \right)$$

LO 4. Note that if the CLT doesn't apply and the sample proportion is low (close to 0) the sampling distribution will likely be right skewed, if the sample proportion is high (close to 1) the sampling distribution will likely be left skewed.

LO 5. Remember that confidence intervals are calculated as

$$\text{point estimate} \pm \text{margin of error}$$

and test statistics are calculated as

$$\text{test statistic} = \frac{\text{point estimate - null value}}{\text{standard error}}$$

LO 6. Note that the standard error calculation for the confidence interval and the hypothesis test are different when dealing with proportions, since in the hypothesis test we need to assume that the null hypothesis is true – remember: p-value = P(observed or more extreme test statistic | H0 true).

- For confidence intervals use $\hat{p}$ (observed sample proportion) when calculating the standard error and checking the success/failure condition:

$$SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- For hypothesis tests use p0 (null value) when calculating the standard error and checking the success/failure condition:

$$SE_{\hat{p}} = \sqrt{\frac{p_0(1-p_0)}{n}}$$

- Such a discrepancy doesn't exist when conducting inference for means, since the mean doesn't factor into the calculation of the standard error, while the proportion does.

LO 7. Explain why when calculating the required minimum sample size for a given margin of error at a given confidence level, we use p^=0.5 if there are no previous studies suggesting a more accurate estimate.

- Conceptually: When there is no additional information, 50% chance of success is a good guess for events with only two outcomes (success or failure).

- Mathematically: Using $\hat{p} = 0.5$ yields the most conservative (highest) estimate for the required sample size.

LO 8. Note that the calculation of the standard error of the distribution of the difference in two independent sample proportions is different for a confidence interval and a hypothesis test.

- confidence interval and hypothesis test when $H_0 : p_1 - p_2$ = some value other than 0:

$$SE_{(\hat{p}_1-\hat{p}_2)} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

- hypothesis test when $H_0 : p_1 - p_2 = 0$:

$$SE_{(\hat{p}_1-\hat{p}_2)} = \sqrt{\frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{n_1} + \frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{n_2}},$$

,

where $\hat{p}_{pool}$ is the overall rate of success:

$\hat{p}_{pool}$ =number of successes in group 1 + number of successes in group 2n1+n2

LO 9. Note that the reason for the difference in calculations of standard error is the same as in the case of the single proportion: when the null hypothesis claims that the two population proportions are equal, we need to take that into consideration when calculating the standard error for the hypothesis test, and use a common proportion for both samples.