AP Statistics

When we want to compare **TWO PROPORTIONS** (groups) we need to combine the variances of each group. We always add variances, never subtract even if we want to compare differences because when we combine or take the difference, the variations of each sample come together to give us more variation rather than less.

- In a two-sample problem, the groups we want to compare are **Population 1** and **Population 2**. Comparing two populations:

Population	Population proportion	Sample size	Sample proportion
1	p_1	n_1	$\hat{p}_1$
2	p_2	n_2	$\hat{p}_2$

- We compare populations by doing **inference about the difference**, $p_1 - p_2$, between the population proportions. The statistic that estimates this difference is the **difference between the two sample proportions**, $\hat{p}_1 - \hat{p}_2$.

THE STANDARD DEVIATION OF THE DIFFERENCES BETWEEN TWO PROPORTIONS

The standard deviations of the sample proportions are $SD_{\hat{p}_1} = \sqrt{\frac{p_1 q_1}{n_1}}$ and $SD_{\hat{p}_2} = \sqrt{\frac{p_2 q_2}{n_2}}$, so the variance of the difference in proportions is $Var_{(\hat{p}_1 - \hat{p}_2)} = \left(\sqrt{\frac{p_1 q_1}{n_1}}\right)^2 + \left(\sqrt{\frac{p_2 q_2}{n_2}}\right)^2 = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}$.

The standard deviation of the difference in proportions is $SD_{(\hat{p}_1 - \hat{p}_2)} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$.

We usually don't know the true values of p_1 and p_2. When we have the sample proportions in hand from the data, we use them to estimate the variances. So the **standard error** is $SE_{(\hat{p}_1 - \hat{p}_2)} = \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$.

SAMPLING DISTRIBUTIONS OF TWO PROPORTIONS $\hat{p}_1 - \hat{p}_2$

- Assumptions: $\hat{p}_1 - \hat{p}_2$ is an unbiased estimator of $p_1 - p_2$
- The **difference** in the proportion of $\hat{p}_1 - \hat{p}_2$ approximates $p_1 - p_2$, that is, the difference of sample proportions is an unbiased estimator of the difference of population proportions.
- The **variance** of $\hat{p}_1 - \hat{p}_2$ is $Var_{(\hat{p}_1 - \hat{p}_2)} = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}$, provided that the sample proportions are independent.
- For large samples, the distribution of $\hat{p}_1 - \hat{p}_2$ is approximately normal.

Assumptions and Conditions:

- **The Independence Assumption:** Within each group, the data should be based on results for independent individuals. We can't check that for certain, but we can check the following:
  - **Randomization Condition:** The data in each group should be drawn independently and at random from a homogeneous population or generated by a randomized comparative experiment.
  - **10% Condition:** If the data are sampled without replacement, the sample should not exceed 10% of the population.

- **Independent Groups Assumption:** The two groups we're comparing must also be independent of each other. Usually, the independence of the groups from each other is evident from the way the data were collected.

■ **The Sample Size Assumption:** Each of the two groups must be big enough. As with individual proportions, we need larger groups to estimate proportions that are near 0% or 100%. We usually check the Success/Failure Condition for each group.

- **Success/Failure Condition:** Both groups are big enough that at least 10 successes and at least 10 failures have been observed in each. $np_1 \geq 10,\ nq_1 \geq 10,\ np_2 \geq 10,\ nq_2 \geq 10$

Example, p. 506: Finding the standard error of a difference in proportions

A recent survey of 886 randomly selected teenagers (12-17) found that more than half of them had online profiles. Some researchers are concerned about the possible access to personal information about teens in public places on the internet. There appear to be differences between boys and girls in their online behavior. Among teens ages 15 to 17, 57% of the 248 boys had posted profiles compared to 70% of the 256 girls. What is the standard error of the difference in sample proportions?

$$SE_{(\hat{p}_{boys}-\hat{p}girls)} = \sqrt{\frac{\hat{p}_{boys}\hat{q}_{boys}}{n_{boys}} + \frac{\hat{p}_{girls}\hat{q}_{girls}}{n_{girls}}} = \sqrt{\frac{(0.57)(0.43)}{248} + \frac{(0.70)(0.30)}{256}} = 0.0425$$

CONFIDENCE INTERVALS FOR TWO PROPORTIONS $p_1 - p_2$

- SRS of size n_1, from a population with proportion p_1 of success, and SRS of size n_2, from a population with proportion p_2 of success.
- When n_1 and n_2 are large, level **C** confidence interval for $p_1 - p_2$ is $(\hat{p}_1 - \hat{p}_2) \pm z^* SE$.
- In this formula, the **standard error SE** of $\hat{p}_1 - \hat{p}_2$ is $SE_{(\hat{p}_1-\hat{p}_2)} = \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$, and z^* is the upper $\frac{1-C}{2}$ standard normal critical value.
- Use this CI when the populations $\geq 10n_1,\ 10n_2$; however, sometimes we use $n_1\hat{p}_1,\ n_1\hat{q}_1,\ n_2\hat{p}_2,\ n_2\hat{q}_2 \geq 5$.

"POOLING" is the name given to a technique used to obtain a more precise estimate of the standard deviation of a sample statistic by combining the estimates given by two (or more) independent samples. When calculating confidence intervals for a difference of two means, we do not pool. In other statistical situations we may or may not pool, depending on the situation and the populations being compared. For example, the theory behind analysis of variance and the inferences for simple regression are based on pooled estimates of variance. The rules for inference about two proportions firmly go both(!) ways. We always use a pooled estimate of the standard deviation (based on a pooled estimate of the proportion) when carrying out a hypothesis test whose null hypothesis is $p_1 = p_2$ --but not when constructing a confidence interval for the difference in proportions.

SIGNIFICANCE TESTS FOR COMPARING TWO PROPORTIONS $p_1 - p_2$

- To test the hypothesis

  H_0: $p_1 - p_2 = p_0$, most times it will be H_0: $p_1 - p_2 = 0$

  Calculate the z-statistic, first find the pooled sample proportion,

  <u>We always pool when we use SE for significance tests of two proportions!!!</u> Why? Since we assume that $p_1 = p_2$, there should be only one $\hat{p}$, well which one do we use? The **pooled proportion!**

- $$\hat{p}_{\text{pooled}} = \frac{\text{count of successes in both samples combined}}{\text{count of observations in both samples combined}} = \frac{x_1 + x_2}{n_1 + n_2}$$

  $$z\text{-statistic},\quad z = \frac{(\hat{p}_1 - \hat{p}_2) - p_0}{\sqrt{\frac{\hat{p}_{\text{pooled}}\hat{q}_{\text{pooled}}}{n_1} + \frac{\hat{p}_{\text{pooled}}\hat{q}_{\text{pooled}}}{n_2}}}$$