ONLINE MASTERS IN **DATA SCIENCE**

DSC 215 - PROBABILITY AND STATISTICS FOR DATA SCIENCE

DIFFERENCE OF TWO PROPORTIONS

UC San Diego

COMPUTER SCIENCE & ENGINEERING

HALICIOĞLU DATA SCIENCE INSTITUTE

Difference of Two Proportions

Objective

- Extend the techniques we've learned for estimating confidence intervals and for hypothesis testing, so they apply to differences in (population) proportions
    - $p_1 - p_2$

Example

- In medical experiments, we often split patients into Control and Treatment groups. We are interested in understanding the difference in patient outcomes.

Difference of Two Proportions

Idea:

- Identify a point (sample) estimate $\hat{p}_1 - \hat{p}_2$
- Verify that $\hat{p}_1 - \hat{p}_2$ can be approximated by a normal distribution
- Compute the associated Standard Error
- Apply our (modified) inference framework to compute CI's or conduct hypothesis tests

Verifying that $\hat{p}_1 - \hat{p}_2$ Can Be Approximated by a Normal Distribution

- Like before, $\hat{p}_1 - \hat{p}_2$ can be modeled as being drawn from a normal distribution when two conditions hold:

1. **Independence:** Data are independent **within and between** the two groups.
2. **Success/Failure Condition:** The S/F condition hold for each of the two groups separately, i.e., for $i = 1, 2$

$$np_i \geq 10 \text{ and } n(1 - p_i) \geq 10$$

Confidence Intervals for $p_1 - p_2$

- The standard error in this case is given by:

$$SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

- **Can you guess why?** Hint: What is the variance of the difference of 2 variables?
- The confidence interval is then

$$I = \left(\hat{p}_1 - \hat{p}_2 - z^\star \times SE,\ \ \hat{p}_1 - \hat{p}_2 + z^\star \times SE\right)$$

- Which we can write as

$$(\hat{p}_1 - \hat{p}_2) \pm z^\star \times SE$$

Confidence Intervals for $p_1 - p_2$

- **Example:** In an experiment for patients who were given a particular medicine after a heart attack, patients were randomly divided into a treatment group (received a blood thinner) and a control group (did not receive a blood thinner). The variable of interest is survival after 24 hrs.

	Survived	Died	Total
Control	11	39	50
Treatment	14	26	40
Total	25	65	90

Example is from OpenIntro Statistics (Chapter 6)

Confidence Intervals for $p_1 - p_2$

- **Question:** Create and interpret a $90\,\%$ confidence interval of the difference between the survival rates in the study.
- **Solution:**
- We first check the conditions (Independence and S/F).
- Independence is satisfied because this is a randomized experiment, and S/F is satisfied because we have at least 10 successes and 10 failures in each of the control and treatment groups.
- Next, we calculate the survival rate for treatment group: $\hat{p}_t = 14/40 = 0.35$ and the control group $\hat{p}_c = 11/50 = 0.22$

$$\implies \hat{p}_t - \hat{p}_c = 0.13$$

Confidence Intervals for $p_1 - p_2$

- **Question:** Create and interpret a $90\ \%$ confidence interval of the difference between the survival rates in the study.
- **Solution (Continued):**

Now we calculate $SE \approx \sqrt{\dfrac{\hat{p}_t(1-\hat{p}_t)}{n_t} + \dfrac{\hat{p}_c(1-\hat{p}_c)}{n_c}}$

To get $SE \approx \sqrt{\dfrac{0.35(1-0.35)}{40} + \dfrac{0.22(1-0.22)}{50}} = 0.095$

Confidence Intervals for $p_1 - p_2$

- **Question:** Create and interpret a $90\ \%$ confidence interval of the difference between the survival rates in the study.
- **Solution (Continued):**

Finally, for a $90\ \%$ Confidence Interval, $z^\star = 1.65$, so our confidence interval is

$$I = \left(\hat{p}_t - \hat{p}_c - z^\star \times SE,\ \ \hat{p}_t - \hat{p}_c + z^\star \times SE\right)$$

$$\implies I = \left(0.13 - 1.65 \times 0.095,\ \ 0.13 + 1.65 \times 0.095\right)$$

$$\implies I = \left(-0.027,\ 0.287\right)$$

Confidence Intervals for $p_1 - p_2$

- **Question:** Create and interpret a $90\ \%$ confidence interval of the difference between the survival rates in the study.
- **Solution (Continued):**
- **Interpretation:**

We are 90% confident that blood thinners have a difference of -2.7% to +28.7% impact on survival rate for patients (like those in the study).

However, 0% is contained in the interval, so we cannot say at this confidence level, whether blood thinners help or harm in this context.