ONLINE MASTERS IN **DATA SCIENCE**

DSC 215 - PROBABILITY AND STATISTICS FOR DATA SCIENCE

DIFFERENCE OF TWO PROPORTIONS:

HYPOTHESIS TESTS

UC San Diego

COMPUTER SCIENCE & ENGINEERING

HALICIOĞLU DATA SCIENCE INSTITUTE

Difference of Two Proportions

Recall our framework:

- 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

S/F Condition for Using a Normal Distribution

- To verify that $\hat{p}_1 - \hat{p}_2$ can be approximated by a normal distribution, we use the ***pooled proportion*** to check the success/failure condition.

$$\hat{p}_{pooled} := \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

- **Example:**
  - In a randomized study on (say) survival rates associated with some treatment, a total of $n_1 = 500$ control, and $n_2 = 500$ treatment samples are collected.
  - 25 patients in the control group, and 35 patients in the treatment group die. So,

$$\hat{p}_{pooled} = \frac{25 + 35}{1000} = 0.06$$

Conditions for Using a Normal Distribution

- **Example (continued):** Since $\hat{p}_{pooled} = \dfrac{25 + 35}{1000} = 0.06$, we can calculate
  - $n_1 \times \hat{p}_{pooled} = 500 \times 0.06 = 30$
  - $n_1 \times (1 - \hat{p}_{pooled}) = 470$
  - $n_2 \times \hat{p}_{pooled} = 500 \times 0.06 = 30$
  - $n_1 \times (1 - \hat{p}_{pooled}) = 470$
- All are ≥ 10, so we can safely model the difference between the proportions as a normal distribution.

Computing the Standard Error

- **Remark:** If our null hypothesis is that there is no difference between the treatment and control groups, then $\hat{p}_{pooled}$ is our best estimate of the proportions p_1 and p_2.
- **We also use $\hat{p}_{pooled}$ in computing the standard error:**

$$SE = \sqrt{\frac{\hat{p}_{pooled}(1 - \hat{p}_{pooled})}{n_1} + \frac{\hat{p}_{pooled}(1 - \hat{p}_{pooled})}{n_2}}$$

- *(This is because we want to know how likely it is to have data at least as extreme as our sample, assuming the null – that $\hat{p}_1 = \hat{p}_2 = \hat{p}_{pooled}$)*

Hypothesis Tests for $p_1 - p_2$

- **Example:** In an experiment, patients were randomly divided into a treatment group and a control group. The variable of interest is survival during a 30 yr followup period.

	Survived	Died
Control	500	44425
Treatment	505	44405

Example is from OpenIntro Statistics (Chapter 6)

Hypothesis Tests for $p_1 - p_2$

- **Question:** Set up your hypotheses to test whether there was a difference in deaths between the two groups.
- **Solution:**

H_0 : **Death rate is the same for treatment and control groups**

$$p_t - p_c = 0$$

H_A : **Death rate is different for patients in the control group and treatment group**

$$p_t - p_c \neq 0$$

Hypothesis Tests for $p_1 - p_2$

- **Question:** Evaluate the hypotheses with a significance level of $5\,\%$.
- **Solution:**

First, we compute our point estimate $\hat{p}_t - \hat{p}_c$. Notice that

$$n_t = 44925,\ n_c = 44910, \text{ so}$$

$$\hat{p}_t = \frac{500}{44925} = 0.01113 \quad \text{and} \quad \hat{p}_c = \frac{505}{44910} = 0.01125$$

$$\implies \hat{p}_t - \hat{p}_c = -0.00012$$

Hypothesis Tests for $p_1 - p_2$

- **Question:** Evaluate the hypotheses with a significance level of $5\,\%$.
- **Solution (continued):**
- **Second,** we check the conditions (Independence and S/F).
- Independence is satisfied because this is a randomized experiment, and S/F is satisfied because:

$$\hat{p}_{pooled} = \frac{500 + 505}{44925 + 44910} = 0.0112$$

$$\implies n_t \times \hat{p}_{pooled},\ n_t(1 - \hat{p}_{pooled}), n_c \times \hat{p}_{pooled},\ n_c(1 - \hat{p}_{pooled})$$

are all greater than 10.

Hypothesis Tests for $p_1 - p_2$

- **Question:** Evaluate the hypotheses with a significance level of $5\,\%$.
- **Solution (continued):**
- **Third,** we calculate the standard error

$$SE = \sqrt{\frac{\hat{p}_{pooled}(1 - \hat{p}_{pooled})}{n_t} + \frac{\hat{p}_{pooled}(1 - \hat{p}_{pooled})}{n_c}} = 0.00070$$

Hypothesis Tests for $p_1 - p_2$

- **Question:** Evaluate the hypotheses with a significance level of $5\ \%$.
- **Solution (Continued):**
- **Fourth,** we calculate the p-value
- *Recall: "What is the probability that $\hat{p}$ is at least as far in the tails as $\hat{p}_t - \hat{p}_c$ under the null distribution?"*
- *We can do it by integration (using software) or using Z-scores:*

$$Z = \frac{\textit{point estimate} \quad - \textit{null value}}{SE} = \frac{-0.00012 - 0}{0.00070} = -\ 0.17$$

$$\implies \ p\text{-}\textit{value} = 0.865 \ \ (\textit{why?})$$

Hypothesis Tests for $p_1 - p_2$

- **Question:** Evaluate the hypotheses with a significance level of $5\,\%$.
- **Solution (Continued):**

- So, we have that p-value ≥ 0.05 and we **do not reject the null hypothesis.**

- **Interpretation:** The difference in deaths between the control and treatment can be reasonably explained by chance.