M358K: November 6th, 2020.

Statistical Inference for Two Proportions.

The counts from the two samples are independent:

$i=1,2$: X_i "$\sim$" Normal(mean $= n_i \cdot p_i$, var $= n_i \cdot p_i(1-p_i)$)

$\Rightarrow$ The proportions from the two samples:

$i=1,2$: $\hat{P}_i = \frac{X_i}{n_i}$ "$\sim$" Normal(mean $= p_i$, var $= \frac{p_i(1-p_i)}{n_i}$)

We're interested in: $\boxed{p_1 - p_2}$

So, we focus on:

$$\hat{P}_1 - \hat{P}_2 \text{ "}\sim\text{" Normal}\left(\text{mean} = p_1 - p_2, \ \boxed{\text{var} = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$$

Confidence Intervals

$$\underline{\text{point estimate}} \pm \underbrace{\text{margin of error}}_{z^* \cdot \text{std error}}$$

$$\hat{p}_1 - \hat{p}_2 \pm z^* \cdot \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

w/ $\hat{p}_i$ ... the observed sample proportion for sample i, $i=1,2$

University of Texas at Austin

Problem Set # 14

Difference in two proportions.

#1

#2

Problem 14.1. A simple random sample of 200 students is selected from a large university. In this sample, there are 35 minority students. A simple random sample of 80 students is selected from the community college in the same town. In this sample, there are 28 minority students. What is the standard error of the difference in sample proportions of minority students?

$$\hat{p}_1 = \frac{35}{200} = 0.175 \; ; \; \hat{p}_2 = \frac{28}{80} = 0.35$$

$$\text{std error}: \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} = \cdots = 0.05971$$

$\boxed{\text{For laughs:}}$ Pick a confidence level : $C = 95\%$

$$\Rightarrow z^* = 1.96$$

$$p_1 - p_2 = (\hat{p}_1 - \hat{p}_2) \pm z^*(\text{std error})$$

$$p_1 - p_2 = -0.175 \pm 1.96(0.05971) = -0.175 \pm 0.117$$

Problem 14.2. Suppose that, in our usual notation, $\hat{p}_1 = 0.5, \hat{p}_2 = 0.2, n_1 = 20$ and $n_2 = 30$. What is the $p-$value for testing

$$H_0 : p_1 = p_2 \quad vs. \quad H_a : p_1 \neq p_2.$$

Hypothesis Testing.

The test is of this form:

$$H_0: p_1 = p_2 \quad \text{vs.} \quad H_a: \begin{cases} p_1 > p_2 \\ \boxed{p_1 \neq p_2} \\ p_1 < p_2 \end{cases}$$

Our test statistic is:

$$\hat{P}_1 - \hat{P}_2 \text{ "}\sim\text{" Normal}\left(\text{mean} = p_1 - p_2,\ \text{var} = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)$$

Under the null hypothesis: $\boxed{p_1 = p_2 = p}$

$$\hat{P}_1 - \hat{P}_2 \text{ "}\sim\text{" Normal}\Big(\text{mean} = 0,$$

$$\text{var} = \frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2} = $$

$$= p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\Big)$$

Of course, we don't have the exact value of p

$\Rightarrow$ We use an estimate:

$$\hat{p} = \frac{n_1}{n_1+n_2}\cdot\hat{p}_1 + \frac{n_2}{n_1+n_2}\cdot\hat{p}_2$$

$$\hat{p} = \frac{\boxed{n_1\cdot\hat{p}_1} + \boxed{n_2\cdot\hat{p}_2}}{n_1+n_2}$$

$$\boxed{\hat{p} = \frac{x_1 + x_2}{n_1+n_2}}$$

POOLING

w/ x_i, $i = 1, 2 \ldots$ observed number of successes in sample $i = 1, 2$.

$$\boxed{\frac{\hat{P}_1 - \hat{P}_2}{\sqrt{p(1-p)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}} \text{ “}\sim\text{” } N(0,1)$$ under the null hypothesis.

Say, you observed $\hat{p}_1$ and $\hat{p}_2$; <u>then</u> the z-statistic is:

$$\boxed{z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}}$$

e.g., for a two-sided alternative hypothesis:

$H_a: p_1 \neq p_2$

the p-value is:

$$\mathbb{P}[Z > |z|] + \mathbb{P}[Z < -|z|]$$