M358K: November 5^{th}, 2021.

Statistical Inference for Two Proportions.

Our parameters of interest:

p_i, $i=1,2$ ... the (sub)population proportion for the (sub)population $i=1,2$

e.g., p_1 corresponds to the subpopulation who get the placebo;
p_2 corresponds to the subpopulation who get the treatment.

Sample sizes are denoted by n_1 and n_2.

Assume that the two samples are independent.

For large n_i, $i=1,2$, we know the approximate dist'n of:

- the count r.v.s:

$$X_i \approx Normal\left(mean = n_i \cdot p_i\,,\ sd = \sqrt{n_i \cdot p_i(1-p_i)}\right) \qquad i=1,2$$

and

- the sample proportion r.v.s:

$$\boxed{\hat{P}_i = \frac{X_i}{n_i} \approx Normal\left(mean = p_i\,,\ sd = \sqrt{\frac{p_i(1-p_i)}{n_i}}\right)} \qquad i=1,2$$

We base our statistical inference for $p_1 - p_2$ on $\hat{P}_1 - \hat{P}_2$.

$$\boxed{\hat{P}_1 - \hat{P}_2 \approx Normal\left(mean = p_1 - p_2\,,\ \boxed{sd = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}\right)}$$

Confidence Interval

C... confidence level

$$\text{pt. estimate} \pm \underline{\text{margin of error}}$$

$$z^* \cdot \boxed{\text{stderror}}$$

$$z^* = \Phi^{-1}\left(\frac{1+C}{2}\right) = \text{qnorm}\left(\frac{1+C}{2}\right)$$

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

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

w/ $\hat{p}_1, \hat{p}_2$ are the observed sample proportions

---

Hypothesis Testing.

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

Test statistic: $\hat{P}_1 - \hat{P}_2$

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

$$\boxed{\hat{P}_1 - \hat{P}_2 \approx \text{Normal}\left(\text{mean} = 0, \; sd = \sqrt{\frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}} = \sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\right)}$$

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

Let $\hat{p}_1$ and $\hat{p}_2$ be the observed sample proportions.

Q: What's the form of the observed z-statistic?

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

w/ the estimate $\hat{p}$ based on all the available data.

We pool the two samples to estimate p, i.e.,

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

w/ $x_i, i=1,2$ ... # of successes in sample i

Note:

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

To calculate the p-value, we find:

- $H_a: p_1 < p_2$: p-value : $\mathbb{P}[Z < z]$
- $H_a: p_1 \neq p_2$: p-value : $\mathbb{P}[Z < -|z|] + \mathbb{P}[Z > |z|]$
- $H_a: p_1 > p_2$: p-value : $\mathbb{P}[Z > z]$

If a significance level α is given, then ....

if p-value $\leq \alpha$, we REJECT THE NULL.

if p-value $> \alpha$, we FAIL TO REJECT THE NULL.