M358K: November 4th, 2020.

Hypothesis testing for the population proportion p.

We test:

$$H_0: p = p_0 \quad \text{vs.} \quad H_a: \begin{cases} p < p_0 \\ p \neq p_0 \\ p > p_0 \end{cases}$$

Test statistic = ?

Sample proportion of "successes"; we look @ its sampling distribution under the null hypothesis

For a large enough sample, we know that the count r.v.:

$$X \text{ "}\sim\text{" } Normal(mean = n \cdot p_0,\ var = n \cdot p_0(1-p_0))$$

$$\Rightarrow \hat{P} = \frac{X}{n} \text{ "}\sim\text{" } Normal\left(mean = p_0,\ \boxed{var = \frac{p_0(1-p_0)}{n}}\right)$$

The observed sample proportion is denoted by $\hat{p}$. Then, the corresponding z-statistic is (under the null hypothesis)

$$\boxed{z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}}$$

p-value: the probability of observing what we observed or something more extreme under the null

- IF $H_a: p < p_0$, then p-value $= \mathbb{P}[Z < z]$
- IF $H_a: p \neq p_0$, then p-value $= \mathbb{P}[Z < -|z|] + \mathbb{P}[Z > |z|]$
- IF $H_a: p > p_0$, then p-value $= \mathbb{P}[Z > z]$

Let's say that a significance level α is given.

$$\begin{cases} \text{IF } \underline{p\text{-value} \leq \alpha}, \text{ } \underline{\text{then}} \text{ we REJECT THE NULL } H_0. \\ \text{IF } p\text{-value} > \alpha, \text{ } \underline{\text{then}} \text{ we FAIL TO REJECT THE NULL } H_0. \end{cases}$$

University of Texas at Austin

Problem Set # 13

Hypothesis testing: One-sample proportion.

Problem 13.1. *Source: Problem* **8.99** *from the Moore/McCabe/Craig.*
Castaneda v. Partida is an important court case in which statistical methods were used as part of a legal argument. When reviewing this case, the Supreme Court used the phrase "two or three standard deviations" as a criterion for statistical significance. This Supreme Court review has served as the basis for many subsequent applications of statistical methods in legal settings. (The two or three standard deviations referred to by the Court are values of the z statistic and correspond to p-values of approximately 0.05 and 0.0026.)

In Castaneda the plaintiffs alleged that the method for selecting juries in a county in Texas was biased against Mexican Americans. For the period of time at issue, there were $181,535$ persons eligible for jury duty, of whom $143,611$ were Mexican Americans. Of the 870 people selected for jury duty 339 were Mexican Americans.

(i) (1 point) What proportion of eligible jurors were Mexican Americans?

$$\boxed{p_0 = \frac{143{,}611}{181{,}535} = 0.7911}$$

(ii) (2 points) Let p denote the probability that a randomly selected juror is a Mexican American. Formulate the null and alternative hypotheses to be tested.

$$H_0: p = 0.7911 \quad \text{vs.} \quad H_a: p < p_0 = 0.7911$$

(iii) (1 point) What is the sample proportion of jurors who were Mexican American?

$$\boxed{\hat{p} = \frac{339}{870} = 0.3897}$$

(iv) (4 points) Compute the $z-$statistic, and find the $p-$value.

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.3897 - 0.7911}{\sqrt{\frac{0.7911\,(0.2089)}{870}}} = -29. \ldots$$

$\Rightarrow$ p-value is virtually zero.!

(v) (2 points) How would you summarize your conclusions? (A finding of statistical significance in this circumstance does not constitute proof of discrimination. It can be used, however, to establish a prima facie case. The burden of proof then shifts to the defense.)

There is evidence in favor of the prima facie case!

Statistical Inference for Two Proportions.

Our parameters of interest:

$p_i, i=1,2$ ... the population proportion for the subpopulation $i=1,2$.

e.g., p_1... corresponds to the subpopulation who get the sugar pill;

p_2... corresponds to the subpopulation who get the actual treatment.

Sample of size n_i, $i=1,2$ from the subpopulation $i=1,2$ is planned for.

Assume that the two samples are INDEPENDENT.

For large n_i, $i=1,2$, we know that for the count r.v.s.

$$X_i \text{ "}\sim\text{" } Normal(mean = n_i \cdot p_i, \ var = n_i \cdot p_i(1-p_i)) \quad i=1,2$$

$\Rightarrow$ for the sample proportion r.v.s

$$\boxed{\hat{P}_i = \frac{X_i}{n_i} \text{ "}\sim\text{" } Normal\left(mean = p_i, \ var = \frac{p_i(1-p_i)}{n_i}\right)} \quad i=1,2.$$

We want to do statistical inference on $\boxed{p_1 - p_2}$

It's sensible to focus on:

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