UNIVERSITY OF TEXAS AT AUSTIN

Problem Set # 17

Difference in two proportions.

Problem 17.1. 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?

Solution:

In our usual notation, we have

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

So, the standard error is

$$SE(\hat{P}_1 - \hat{P}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} = \sqrt{\frac{0.175 \times 0.825}{200} + \frac{0.35 \times 0.65}{80}} = 0.0597.$$

Problem 17.2. Suppose that, in our usual notation, $\hat{p}_1 = 0.5, \hat{p}_2 = 0.4, n_1 = 30$ and $n_2 = 40$. What is the $p-$value for testing

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

Solution: Under the null hypothesis, the two proportion parameters are equal. So, the estimate for the proportion parameter of the entire population is

$$\hat{p} = \frac{\hat{p}_1 \times n_1 + \hat{p}_2 \times n_2}{n_1 + n_2} = \frac{20 + 16}{70} = \frac{18}{35}.$$

Thus, the observed value of the $z-$statistic under the null hypothesis is

$$z_{obs} = \frac{0.5 - 0.4}{\sqrt{\left(\frac{18}{35}\right)\left(\frac{17}{35}\right)\left(\frac{1}{30} + \frac{1}{40}\right)}} \approx 0.8284169.$$

The associated $p-$value is

$$2(1 - \Phi(z_{obs})) = 0.4074345.$$

Problem 17.3. A simple random sample of 60 households in Whoville is taken. In the sample, there are 45 households that decorate their houses with lights for the holidays.

A simple random sample of 50 households is also taken from the neighboring Whoburgh. In the sample, there are 40 households that decorate their houses.

(i) What is a 95% confidence interval for the difference in population proportions of households that decorate their houses with lights for the holidays?

Solution:

The sample standard error is

$$\sqrt{\frac{(45/60)(15/60)}{60}+\frac{(40/50)(10/50)}{50}}=0.0795.$$

So, the required 95% confidence interval is

$$(0.75-0.80)\pm 1.96\times 0.0795=-0.05\pm 0.1558.$$

(ii) If you want to test the hypothesis whether one of the two cities has more festive inhabitants, i.e., whether one of the two cities has a higher proportion of decorated domiciles or not, what $p-$value would you obtain?

Solution:

Let p_1 be the population proportion of decorated houses in Whoville, and let p_2 be the population proportion of decorated houses in Whoburgh. The hypothesis test we want to conduct is

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

Under the null hypothesis of proportion equality, the estimate of the proportion of decorated houses is

$$\hat{p}=\frac{45+40}{60+50}=0.77.$$

So, the standard error under the null hypothesis equals

$$SE_0=\sqrt{0.77\times 0.23\times\left(\frac{1}{60}+\frac{1}{50}\right)}=0.0806.$$

The observed value of the $z-$score under the null is, hence,

$$\frac{0.75-0.80}{0.0806}=-0.62.$$

Finally, the $p-$value is

$$2(1-\Phi(0.62))=0.5352.$$

Of course, one would fail to reject the null hypothesis for any sensible choice of a significance level.

Problem 17.4. Caveat tester.

It is estimated that 780,000 surgical site infections (SSIs) occur each year. SSIs are the second most common type of healthcare-associated infections in U.S. hospitals and account for an extra $3.5 to $10 billion in healthcare costs per year. The national SSIs rate is 1.9%. A Georgetown medical office was interested in determining if their SSI rate were smaller than the national average. Out of a sample of 277 patients in their study, only one infection occurred.

(i) (1 point) What is the sample size n in this study?
Solution: 277

(ii) (1 point) What is the count of SSIs in this study?
Solution: 1

(iii) (1 point) What is the observed sample proportion?
Solution: 1/277

(iv) (2 points) What is the name of the approximate distribution of the the sample proportion (according to the CLT)?
Solution: Normal.

(v) (3 points) What is the standard error for $\hat{P}$?
Solution:

$$\sqrt{\frac{(1/277)(266/277)}{277}} = 0.0035.$$

(vi) (3 points) What is the 95% confidence interval for the population proportion parameter?
Solution:

This is not necessarily an admissible procedure since there was only one occurrence of the event of interest in our sample. Formally, we have

$$\frac{1}{277} \pm 1.96 \times 0.0035 = 0.0036 \pm 0.0068$$

Since the proportion parameter needs to be non-negative, the confidence interval is really $(0, 0.0104)$.

(vii) (5 points) Test the hypothesis that the Georgetown medical office SSIs rate is less than the national average SSIs rate at the significence level 0.01.
Solution: We are testing

$$H_0 : p = 0.019 \quad vs. \quad H_a : p < 0.019.$$

Even though $0.019 \times 277 > 5$, it is still simpler and more accurate to use the binomial distribution for the counts (and not the normal approximation). The probability of observing $\hat{p} = 0.0036$ or a smaller proportion under the null hypothesis, i.e., the $p-$value of the above test, is approximately

$$(0.019)^{277} + 277(0.019)^{276}(1 - 0.019) \approx 0.$$

What would our $p-$value be if we decided to use the normal approximation? The observed $z-$score is, under the null hypothesis,

$$\frac{0.0036 - 0.019}{\sqrt{\frac{0.019(1-0.019)}{277}}} = -1.88.$$

Therefore, the $p-$value would be about 0.0301.