

STAT 446 HOMEWORK 4

Leslie Gains-Germain

1. Yes, since the 95% confidence interval for the true percentage of entries from the South does not contain the percentage of persons living in the South (30.9%), there is evidence that the percentage of entries from the South differs from 30.9%.

2. (a) $\hat{p} = 0.275$

(b) The 95% confidence interval for p is $0.275 \pm 1.96 \sqrt{\frac{5107-640}{5107} \frac{0.275*0.725}{639}}$ which is (0.243, 0.307).

(c) *The CEO is 95% confident that the proportion of **fulltime** employees who worked for the past 12 months that did not miss any workdays due to illness is between L and U .*

This is a misinterpretation of the confidence interval because of the word fulltime. The SRS of 640 employees was taken from all 5107 employees of the company, and I assume this includes both full-time, part-time and seasonal employees. Therefore, the interpretation should refer to “the proportion of all employees of this company who worked for the past 12 months.”

- (d) Let’s suppose that the CEO is not confused and he truly is only interested in full-time employees of his company. In the sample, however, they could have sampled full-time employees as well as part-time and seasonal employees. With respect to the CEO’s target population, this sample suffers from **overcoverage**.
- (e) There is a good chance that employees will lie about how many days of work they missed due to illness because they don’t want to look like a bad employee. If the employees are worried about what the CEO will think of them, they may understate the amount of days they truly missed due to illness (resulting in observational error).

3. With a standard deviation estimate of 1.92, to get a margin of error less than or equal to 0.5 at an α level of 0.05, the sample size must be at least 59. I added 2 to account for the fact that the sample mean follows a t-distribution (not a z-distribution) for small sample sizes.

$$n = 1.96^2 * 1.924839^2 / 0.5^2 = 57 + 2 = 59$$

4. (a) We want the margin of error to be less than or equal to 0.04. We solve, $1.96\sqrt{\frac{N-n}{N-1}\frac{\hat{p}(1-\hat{p})}{n}} \leq 0.04$. Solving, we find that a sample size of at least 580 is needed to be 95% confident that the sample proportion will be within 0.04 of the true proportion.

(b) First, we could have obtained a strange sample with a \hat{p} that was more than 1.96 standard deviations from the true p . Second, our estimate from \hat{p} from the preliminary study could have been farther from 0.5 than the true p , so that the standard error we used for the sample size determination underestimated the true standard deviation of the sampling distribution of \hat{p} . Third, our estimate of the population size, N , could have been too small so that our estimate of the finite population correction was too small. This would also cause us to underestimate the true standard deviation of the sampling distribution of \hat{p} which in turn would cause our determined sample size to be smaller than it should be.

5. I solve equation 2.25 for e :

$$\begin{aligned} n &= \frac{n_0}{1 + n_0/N} = \frac{z_{\alpha/2}^2 s^2}{e^2 + z_{\alpha/2}^2 s^2 / N} \\ e^2 &= z_{\alpha/2}^2 \frac{s^2}{n} - n z_{\alpha/2}^2 \frac{s^2}{nN} \\ e &= z_{\alpha/2} \sqrt{1 - \frac{n}{N}} \frac{s}{\sqrt{n}} \end{aligned}$$

6. If $n_0 > N$, then this would suggest that we should just sample the entire population and there would be no need to use equation 2.25 to solve for n .

7. I find the total cost function, take the derivative and set to 0 to find the value of n that minimizes the cost function.

$$\begin{aligned} L(n) + C(n) &= k\left(1 - \frac{n}{N}\right) \frac{s^2}{n} + c_0 + c_1 n \\ \frac{d}{dn}(L(n) + C(n)) &= \frac{-ks^2}{n^2} + c_1 = 0 \\ n &= \sqrt{\frac{ks^2}{c_1}} \end{aligned}$$

I then take the second derivative to prove that this value is a minimum.

$$\frac{d}{dn} \frac{-ks^2}{n^2} + c_1 = \frac{2ks^2}{n^3} > 0$$