Introduction to Statistical Methods

Fall 2021 | Exam 1

1. A researcher wants to determine the mean duration of poverty (measured in number of months) among families in poverty in an NYC neighborhood. In September 2016, the researcher randomly selected a cross-sectional sample of 100 families that are currently in poverty in this neighborhood. The average duration of poverty is 16 months in the sample and the sample standard deviation is 8. For now, let's assume that the random sample is representative of families in poverty in NYC.

   (a) [2pts] What is the standard error of the sample mean?

   $se = s/\sqrt{n} = 8/\sqrt{100} = 0.8$

   (b) [2pts] Construct a 95% confidence interval for the mean duration of poverty (assuming normal distribution of the test statistic, $z_{0.025} = 1.96$).

   $CI = [16 - 1.96 \cdot se, 16 + 1.96 \cdot se] = [14.43, 17.56]$

2. Continue with Problem 1. Now suppose that the researcher is interested in whether the mean duration of poverty is different between single-parent and two-parent families. In the sample, 40 families are single-parent families and 60 families are two-parent families. The sample mean duration of poverty is 20 months among single-parent families and 12 months among two-parent families. The sample standard deviations are 10 and 9 respectively.

   (a) [2pts] State H_0 and H_a.

   Let μ_1 and μ_2 denote the population mean duration of poverty among single- and two-parent families, respectively. Then the hypotheses are:

   H_0: $\mu_1 = \mu_2$

   H_a: $\mu_1 \neq \mu_2$

   (b) [2pts] What is the standard error for $\bar{y}_1 - \bar{y}_2$?

   $se = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{10^2}{40} + \frac{9^2}{60}} = 1.96$

   (c) [2pts] What is the test statistic for this significance test?

$t = \frac{20-12-0}{1.96} = 4.08$

(d) [2pts] Assuming normal distribution for the test statistic, $z_{0.025} = 1.96$ and $z_{0.05} = 1.645$, is your null hypothesis rejected at the significance level of 0.05?

Yes. $t > 1.96$ and thus $p < 0.05$.

3. (2pts) Consider a random variable X and a randomly selected sample of n observations: $x_1, x_2, x_3, ..., x_n$. Let $\bar{x}$ be the sample mean. Show that:

$$\Sigma_{i=1}^{n}(x_i - \bar{x}) = 0$$

.

$\Sigma_{i=1}^{n}(x_i - \bar{x}) = \Sigma_{i=1}^{n}x_i - \Sigma_{i=1}^{n}\bar{x} = \Sigma_{i=1}^{n}x_i - n \cdot \frac{\Sigma_{i=1}^{n}x_i}{n} = \Sigma_{i=1}^{n}x_i - \Sigma_{i=1}^{n}x_i = 0$

4. (Bonus 2pts) Actually, in Question 1, it can be shown that the way in which the researcher collects the sample will generate a sample mean that tends to *overestimates* (i.e. be greater than) the population mean duration among families that are in poverty. Why? Please be brief in your answer.

If we take a cross-sectional sample of *families in poverty*, we are more likely to include those that have spent a longer duration in poverty because those families are more likely to be living in poverty at a given time point. In other words, the selection of families into the sample is likely to bias towards those with a longer duration.

Collecting prospective data that follow the same group of individuals or families over time would be a better strategy to capture families with a shorter poverty duration.