BIOS 4110: Introduction to Biostatistics

Breheny

Lab #9

The Central Limit Theorem is very important in the realm of statistics, and today's lab will explore the application of it in both categorical and continuous data. We will review using it for hypothesis testing and for confidence interval construction. Recall these important formulas:

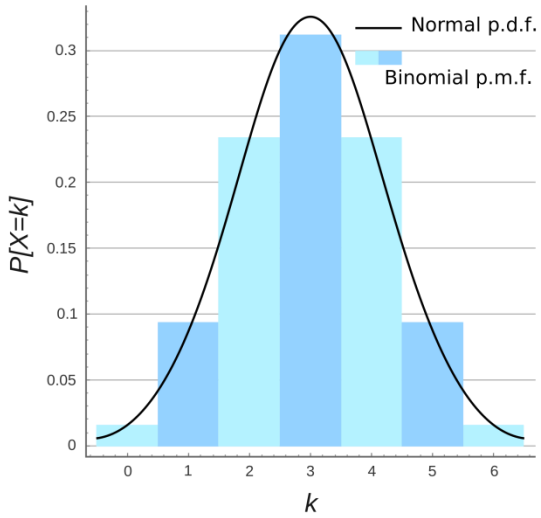
I. CATEGORICAL

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

- $\sqrt{\frac{p_0(1-p_0)}{n}}$ is the standard error of the sample proportion
- $\hat{p}$ is the sample proportion
- p_0 is the hypothesized value of the proportion

This formula is used for *hypothesis testing*. The standard error uses the hypothesized value of the true proportion in the population of interest. Recall in the large sample setting, the distribution of the sample proportion is approximately normal; therefore we can use the Z distribution to approximate the probability of observing a sample proportion as extreme, or more extreme, given that the true proportion is equal to p_0. The following figure shows the *exact* probability mass function as the bars, and the normal probability density function as the line. Notice how they are approximately the same.

NOTE: n = 6, p = 0.5



(1- α/2)x100% Confidence interval for the true proportion, **p**

$$\hat{p} - Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- $\hat{p}$ is the sample proportion
- $Z_{\alpha/2}$ is the critical value (depends on the value of α)
- Only works in large sample setting; $\hat{p}$ can't be close to 0 or 1 and n has to be relatively large
- BOUNDED BY 0 AND 1

If we are interested in providing a confidence interval for the true proportion in the population, we can use the formula above.

We are (1- α/2)x100% confident that the true proportion of [INSERT] in [POPULATION] is between [LOWER, UPPER].

II. CONTINUOUS

$$T = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$$

- $\bar{x}$ is the sample mean
- s is the sample standard deviation
- n is the sample size
- μ_0 is the hypothesized mean
- df is the degrees of freedom, which is equal to *n* - 1

NOTES: This statistic is based off the Student's curve, which is very similar to the normal curve. However, the tails are thicker and we have to account for the degrees of freedom. This accounts for the extra variability because we have to estimate the standard deviation.

When *n* is greater than about 50, the normal distribution and the Student's curve are about the same. In practice, Student's curve is used much more often. Note the degrees of freedom for *T*.

(1- α/2)x100% Confidence interval for the true mean, **μ**

$$\bar{x} - T_{\alpha/2,df}s/\sqrt{n}\ , \bar{x} + T_{\alpha/2,df}s/\sqrt{n}$$

We are (1- α/2)x100% confident that the true mean [INSERT] in [POPULATION] is between [LOWER, UPPER]

Confidence Intervals

Interpretation

- If you were to repeat this process an infinite number of times, 95% of interval estimates for μ created this way will contain the true parameter value μ.
- We treat the population mean μ as being fixed. Any particular interval may or may not contain the true population mean μ.
- We say we are '95% confident' the interval does contain μ because the procedure used to construct this interval produces a correct interval estimate 95% of the time.
- We do not say there is a 95% probability that μ lies between these values.

III. PRACTICE PROBLEMS

1.

A sample of Alzheimer's patients are tested to assess the amount of time in stage IV sleep. It has been hypothesized that individuals suffering from Alzheimer's Disease may spend less time per night in the deeper stages of sleep. Number of minutes spent is Stage IV sleep is recorded for sixty-one patients. The sample produced a mean of 45 minutes (S=14 minutes) of stage IV sleep over a 24 hour period of time.

a) Will we need to know the underlying distribution of Stage IV sleep of the population of individuals suffering from Alzheimer's Disease to make inferences on the sample mean? Explain.

No we do not. We have a sufficiently large sample size, therefore the sample mean will be approximately normal.

b) Suppose we want to test if the true mean amount of time in stage IV sleep for this population is 49 minutes, which is the mean of the general population. Set up the hypothesis test, find the critical value, report the p-value and conclusion.

$t = \frac{45-49}{14/\sqrt{61}}$. = -2.232…P(T < -2.32)*2 = 2*pt(-2.232, df = 60) = 0.029 = pvalue. We have sufficient evidence that it is in fact different from the general population (pvalue = 0.029).

c) Compute a 95 percent confidence interval for this data. What does this information tell you about a particular individual's (an Alzheimer's patient) stage IV sleep?

Critical value for T at significance level 0.05 and df = 60 is 2.

$45 \pm 2 * \frac{14}{\sqrt{61}}$ = (41.41, 48.59). The null hypothesis of 49 minutes is not in the 95% confidence interval. Therefore, we conclude the average amount of sleep spent in stage IV in Alzheimer's patients is less than the general population mean.

2. The distribution of weights for the population of males in the United States is approximately normal with mean μ = 172.2 and standard deviation σ = 29.8 pounds.

a) What is the probability that a randomly selected man weighs less than 130 pounds?

$z = \frac{130-172.2}{29.8}$ = -1.416. P(Z < -1.416) = 0.078

b) What is the probability that he weighs more than 210 pounds?

$z = \frac{210-172.2}{29.8}$ = 1.268. P(Z > 1.268) = 0.102

c) What is the probability that among five males selected at random from the population, at least one will have a weight outside the range 130 to 210 pounds?

1 – dbinom(0,5,0.18) = 0.629

3. Suppose a new flu vaccine is developed for the state of Iowa; however, there are adverse side effects associated with it. We wish to test the hypothesis that the proportion of people who experience an adverse effect is the same as it is with the standard flu vaccine, which is 30%. Suppose we have a random sample of size 1000, and 270 people had an adverse effect to the new flu vaccine.

a) Set up the hypothesis test and calculate the sample proportion.

H0: p = 0.3 H_1: p ≠ 0.3 $\hat{p}$ = 270/1000 = 0.27

b) Use binom.test(). Test the hypothesis stated in the problem at the 5% significance level and calculate a confidence interval. What do you conclude?

binom.test(270, 1000, 0.3) : pvalue = 0.038; 95% confidence interval is (0.243, 0.299). We have sufficient evidence to conclude the true proportion of people who experience an adverse effect for

this new flu vaccine is indeed lower than the true proportion of adverse effects from the standard flu vaccine (p = 0.038).

c) Use the normal approximation to find the p-value and confidence interval.

$z = \frac{0.27-0.3}{\sqrt{\frac{0.3(1-0.3)}{1000}}} =$ -2.07… P(Z < -2.07)*2 = 2*pnorm(-2.07) = 0.038

95% C.I. = $0.27 \pm 1.96\sqrt{\frac{0.27(1-0.27)}{1000}}$ = (0.24, 0.298).

We can see it is very similar to the exact confidence interval and we draw the same conclusion.