Lab 5: Hypothesis Testing Bonanza!

Statistics 12

Submission Instructions

1. Save your final answers in PDF format.
2. Upload your PDF to the CCLE submission link.
3. Submit the assignment as final.

Objectives

1. Learn how to conduct hypothesis testing for one proportion
2. Learn how to conduct hypothesis testing for two proportions
3. Learn how to conduct hypothesis testing for one mean
4. Interpreting hypothesis testing for linear regression

Collaboration Policy

In Lab you are encouraged to work in pairs or small groups to discuss the concepts on the assignments. However, DO NOT copy each other's work as this constitutes cheating. The work you submit must be entirely your own. If you have a question in lab, feel free to reach out to other groups or talk to me if you get stuck.

Overview: the pnorm() function

The pnorm() function computes probabilities from a normal distribution with specified mean and standard deviation. The function inputs a value in the q argument and computes the probability that a value drawn from a normal distribution will be **less than or equal** to q. The exact normal distribution to compare with is specified using the mean and sd arguments. The default mean is mean = 0 and the default standard deviation is sd = 1.

The optional argument lower.tail inputs a logical value (TRUE or FALSE) and changes the direction of the probability. The default is lower.tail = TRUE, so the pnorm() function will, as noted above, compute the probability that a value drawn from the normal distribution will be less than or equal to q. If we set lower.tail = FALSE, then the pnorm() function will compute the probability that a value drawn from the normal distribution will be greater than or equal to q.

For example, if we want the probability of observing a person with a height of 70 inches or less from a population that follows a normal distribution with a mean height of 68 inches and a standard deviation of 1.9 inches:

pnorm(70, mean = 68, sd = 1.9)

If we want the probability of observing a person with a height of 70 inches or more from a population that follows a normal distribution with a mean height of 68 inches and a standard deviation of 1.9 inches:

pnorm(70, mean = 68, sd = 1.9, lower.tail = FALSE)

Overview: The One-Sample z-Test for proportions

To do a (theory-based) hypothesis test (test of significance) for the proportions, we follow several steps:

1. State the null and alternative hypotheses about the population proportion, p_0.

2. From a sample of data, compute the sample proportion $\hat{p}$ and its standardized z-statistic $z = \frac{\hat{p} - p_0}{SE}$, where SE is assumed to be based on the null hypothesis and is calculated as $\sqrt{\frac{p_0 (1 - p_0)}{n}}$ where n is the sample size.

3. Check the validity conditions to apply the Central Limit Theorem.

4. If the validity conditions hold, compute the p-value by comparing the observed z-statistic to a standard normal distribution (using normal tables or pnorm()).

Exercise 1 – Hypothesis testing with one proportion.

We will be working with a modified Flint dataset which can be found on CCLE. Please download this file and read it into R. You may recall in Lab 2 that lead levels were considered dangerous if the result was greater than or equal to 15PPB. We are interested in determining if the proportion of dangerous lead levels in Flint is greater than 10%. Assume the Flint data **is a random sample** used to address this research question.

    a. We will conduct a hypothesis test for this research question. Using symbols from lecture, what are the null and alternative hypotheses? Is this a one-sided or a two-sided test?
    b. Calculate the sample proportion and sample standard deviation of the sample proportion of dangerous lead levels.
    c. Now, calculate the SE of sample proportions, and the z-value for this test. Consult the above instructions and/or the lecture slides for guidance.
    d. Using the z-statistic in (c), calculate the p-value associated with this test. You may use R's pnorm() function or a normal table, but please show all work.
    e. Using a significance level of 0.05, do you reject the null hypothesis?
    f. If greater than 10% of households in Flint contain dangerous lead levels, the EPA requires remediation action to be taken. Based on your results, what should you tell the EPA?

g. Another way to run this test is to use the prop.test() function using the mosaic package. You will need to know your sample size, and the number of "successes" in the sample. Use this function to conduct the same hypothesis test in (a)-(d) and obtain a p-value from the test. Using the same significance level of 0.05, do your results change? An example of the prop.test() function is shown in the two lines below:

We flipped 100 coins and 60 were heads. Is the long-run proportion of heads greater than 0.5?

prop.test(x = 60, n = 100, p = 0.5, alt = "greater")

h. Notice that the prop.test() output produced a confidence interval. Try using the help screen under the mosaic package for prop.test() to find the argument for the confidence interval. Modify the argument and re-run the code in (f) to produce a 99% confidence interval instead of a 95% interval.

Overview: the Two-Sample z-Test for proportions

The hypothesis test for two proportions follows a similar framework as the test for one proportion, but uses a different statistic. The z-statistic is calculated as $z = \frac{\hat{p}_1 - \hat{p}_2 - 0}{SE}$ and SE =

$\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$ where:

n_1 is the sample size in sample 1

n_2 is the sample size in sample 2

x_1 is the number of successes in sample 1

x_2 is the number of successes in sample 2

$\hat{p} = \frac{x1+x2}{n1+n2}$ is the combined (pooled) proportion of successes

$\hat{p}_1 = \frac{x1}{n1}$ is the proportion of successes in sample 1

$\hat{p}_2 = \frac{x2}{n2}$ is the proportion of successes in sample 2

OPTIONAL Exercise 2 – Hypothesis testing with two proportions.

Notice that the Flint data contains a column called Region which is either "North" or "South". We are interested in determining if the proportion of dangerous lead levels in the North region differs from the proportion of dangerous lead levels in the South region. Again assume the Flint data from CCLE is a random sample used to address this research question.

a. We will conduct a hypothesis test for this research question. What are the null and alternative hypotheses? Is this a one-sided or a two-sided test?
b. Using guidance from lecture, calculate every value you will need to produce a z-statistic for this test. Then, calculate the z-statistic. Please show all work.
c. Using the z-statistic in (b), calculate the p-value associated with this test. You may use R's pnorm() function or a normal table, but please show all work.

d. Using a significance level of 0.05, do you reject the null hypothesis? Interpret this result in the context of our research question. *Hint: is this a two-sided test?*
e. Another way to run this test is to use the prop.test() function with some slight modifications from exercise 1. Use the function to conduct the same hypothesis test in (a)-(d) and obtain a p-value from the test, again using a significance level of 0.05. Do your results change? A sample of the prop.test() function for two proportions is shown in the two lines below:

We flipped 100 pennies and 52 were heads. Then we flipped 80 dimes, and 47 were heads. Is
the long-run proportion of heads different for pennies and dimes?

prop.test(x = c(52,47), n = c(100,80), alt = "two.sided")

Overview: the One-Sample t-Test for means

To do a (theory-based) hypothesis test (test of significance) for the mean of a quantitative variable, we follow several steps:

1. State the null and alternative hypotheses about the mean parameter μ.

2. From a sample of data, compute the sample mean $\bar{x}$ and its standardized t-statistic $t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$, where s is the sample standard deviation and n is the sample size.

3. Check the validity conditions to apply the Central Limit Theorem.

4. If the validity conditions hold, compute the p-value by comparing the observed t-statistic to a t-distribution with $df = n - 1$.

Overview: the pt() Function

The pt() function computes probabilities from a t-distribution with specified degrees of freedom. The syntax for pt() is very similar to the pnorm() function. The pt() function inputs a value in the q argument and computes the probability that a value drawn from a t-distribution will be less than or equal to q. The exact t-distribution to compare with is specified using the df argument (df stands for degrees of freedom). There is no default value for the df, so you must input the df argument for the command to work.

The optional argument lower.tail inputs a logical value (TRUE or FALSE) and changes the direction of the probability. The default is lower.tail = TRUE, so the pt() function will, as noted above, compute the probability that a value drawn from the t-distribution will be less than or equal to q. If we set lower.tail = FALSE, then the pt() function will compute the probability that a value drawn from the t-distribution will be greater than or equal to q.

For example, if we want the probability of observing a value of -1.5 or less from statistics that follow a t-distribution with 29 degrees of freedom:

pt(-1.5,df=29)

Exercise 3 – Hypothesis testing with means

Copper is another metal which can be dangerous in high quantities. We believe the average copper levels in the state of Michigan's drinking water is 40 PPM. We are interested in finding if the copper level in Flint's water differs from the Michigan average. Again, assume the Flint data constitutes a random sample.

a. We will conduct a hypothesis test for this research question. What are the null and alternative hypotheses? Is this a one-sided or a two-sided test?

b. Calculate the sample mean and sample standard deviation of the sample copper levels in Flint.

c. Calculate the standard error of the sample mean for copper, $\bar{x}$.

d. Using the values in (b) and (c), calculate the t-test statistic and p-value associated with this test. Use pt() to obtain a test statistic. *Hint: is this a two-sided test?*

e. Using a significance level of 0.01, do you reject the null hypothesis? Interpret this result in the context of our research question.

f. Another way to run this test is to use the t.test() function using the mosaic package. Use this function to conduct the same hypothesis test in (a)-(d) and obtain a p-value from the test, again using a significance level of 0.05. Do your results change? An example of the prop.test() function is shown in the two lines below:

Using sample Flint lead data, do we believe the long run Flint lead average is not equal to 3?

t.test(flint$Pb, mu = 3, alt = "two.sided")