M5_autograded_2

November 29, 2021

1 Module 5 Autograded Assignment

[417]:
```
# Load Necessary Libraries
library(testthat)
```

2 Problem 1

You are a proctor for a Data Science exam, and just gave a test to 15 students. You want to get an idea for the true standard deviation of the scores, using the scores you just recieved. Assume that the underlying score population is normally distributed.

[418]:
```
scores = c(53.02, 69.2, 81.96, 40.62, 76.24, 99.78, 94.49, 71.6, 76.95, 37.68,␣
↪37.59, 59.22, 92.44, 81.22, 63.74)
```

Part A) Using the data stored in the variable `scores`, calculate a 95% confidence interval for the standard deviation of the data. Your confidence interval should be two tailed, and cut off an equal proportion of area on each side. Save the lower value as `p1.lower` and the upper value as `p1.upper`. Round your answers to two decimal places.

[419]:
```
p1.upper = NA
p1.lower = NA
alpha = 0.05

# your code here
n1 = length(scores)
b1 = qchisq(1-(alpha/2), n1 - 1)
a1 = qchisq(alpha/2, n1 - 1)
var1 = var(scores)
p1.upper = round(sqrt(((n1-1) * var1) / a1), 2)
p1.lower = round(sqrt(((n1-1) * var1) / b1), 2)
```

[420]:
```
p1.upper
p1.lower
```

31.93

14.82

[421]:
```
# Hidden Test Cell
```

Part B) You consult with your coworkers, and determine that the historical standard deviation of scores for the test is 15. Based on your results from **Part A**, does your data agree with the historical value at an $\alpha = 0.05$ significance level? Save your answer into variable `p1.b`. Answer the boolean `TRUE` if the observed data does agree, and `FALSE` if it does not.

[422]:
```
p1.b = NA

# your code here
p1.b = TRUE
```

[423]:
```
# Hidden Test Cell
```

3 Problem 2

It's Halloween, and Ralphie has 6 large bags, each filled with a combination of 4 different kinds of candy. Each bag should have an approximately equal amounts of each flavor. But, being a buffalo, Ralphie is not sure that she spread the candy out evenly. She has asked your help to determine whether the proportions of flavors within each bag is the same.

Each bag contains hundreds of pieces of candy, so we can't count them all by hand. Instead, you take a sample of 50 pieces from each bag. The table below displays the counts of each type of candy within each sample of 50 pieces:

	Candy A	Candy B	Candy C	Candy D
Bag 1	14	18	11	6
Bag 2	10	20	12	9
Bag 3	13	14	15	8
Bag 4	15	15	10	10
Bag 5	11	17	13	9
Bag 6	11	14	14	11

Part A) Before you start testing anything about the candy counts, think about what kind of test you will need. You are comparing the underlying distribution of different categorical variables, to determine if they are the same. Which test will be most useful?

1. Z-test
2. t-test
3. Chi-Square Goodness of Fit test
4. Chi-Square Test of Independence
5. Some other test

Select the *integer* corresping to your answer, and save it into variable `p2.a`.

[ ]:

[424]:
```
p2.a = NA

# your code here
p2.a = 3
```

[425]:
```
# Hidden Test Cell
```

Part B) Determine whether the proportions of candy are equal by solving for the p-value of the appropriate test. Save your answer as `p2.b`. Round your answer to three decimal places.

[426]:
```
p2.b = NA

# your code here
a2 = 14 + 10 + 13 + 15 + 11 + 11
b2 = 18 + 20 + 14 + 15 + 17 + 14
c2 = 11 + 12 + 15 + 10 + 13 + 14
d2 = 6 + 9 + 8 + 10 + 9 + 11
candy = c(a2, b2 , c2, d2)
p2.b = round(chisq.test(candy, p= c(0.25, 0.25, 0.25, 0.25))$p.value, 3)
```

[427]:
```
# Hidden Test Cell
```

Part C) Based on the results from **Part B**, is there an equal proportion of candy in each bag? Assume a significance level of $\alpha = 0.05$. Save your answer into `p2.c`. Save the boolean value `TRUE` if there is an equal proportion within each bag, and `FALSE` if there is not an equal proportion.

[428]:
```
p2.c = NA

# your code here
p2.c = FALSE
```

[429]:
```
# Hidden Test Cell
```

Part D) Suppose that the results of your test from **Part B** indicate that the candies do have different proportions. From the test alone, are you able to determine which of the candies within the group have different proportions? Save your answer as `p2.d`. Save the boolean `TRUE` if you can determine which have different proportions, and save `FALSE` if you can not.

[430]:
```
p2.d = NA

# your code here
p2.d = FALSE
```

[431]:
```
# Hidden Test Cell
```

Part E) After some thinking, Ralphie has changed her mind and believes that the candies actually follow the below proportions:

$$p_A = 0.25$$
$$p_B = 0.35$$
$$p_C = 0.25$$
$$p_D = 0.25$$

At a significance level of $\alpha = 0.05$, determine if your observed samples agree with Ralphie's newly proposed proportions. Save the p-value of your calculations as `p2.e.pval`. Round this answer to three decimal places. Into variable `p2.e`, save `TRUE` if your data does agree with the new proportions, or `FALSE` if it does not.

[432]:
```
p2.e = NA
p2.e.pval = NA

# your code here
p2.e = TRUE
p2.e.pval = round(chisq.test(candy, p= c(0.25, 0.35, 0.25, 0.15))$p.value, 3)
```

[433]:
```
# Hidden Test Cell
```

4 Problem 3

A recent public opinion poll surveyed a simple random sample of 400 individuals. Respondents where classified by their age group (0-20, 20-40, 40-60, 60+) and by their preference of pet (Dog or Cat). Results are shown in the table below.

	Dog	Cat
0-20	41	28
20-40	76	54
40-60	80	56
60+	38	27

Does people's age group affect their pet preferences?

Part A) What kind of test should you use to solve this problem?

1. Z-test
2. t-test
3. Chi-Square Goodness of Fit test
4. Chi-Square Test of Independence
5. Some other test

Select the *integer* of the most appropriate test and save it into variable `p3.a`.

[434]:
```
p3.a = NA

# your code here
p3.a = 4
```

[435]:
```
# Hidden Test Cell
```

Part B) Determine whether people's age affect their preference in pet. Use an $\alpha = 0.05$ significance level. Using the test you selected in **Part A**, save the p-value of your test statistic as `p3.b.pval`. Round this value to three decimal places. If people's preference of pet is determined by their age, save `TRUE` into variable `p3.b`. Otherwise, save `FALSE` into `p3.b`.

[436]:
```
p3.b.pval = NA
p3.b = NA
alpha = 0.05

# your code here
data = matrix(c(41, 28, 76, 54, 80, 56, 38, 27), ncol=2, byrow=TRUE)
colnames(data) = c('Dog', 'Cat')
rownames(data) = c('0-20', '20-40', '40-60', '60+')
data = as.table(data)
p3.b.pval = round(chisq.test(data)$p.value, 3)
p3.b = TRUE
```

[437]:
```
p3.b.pval
p3.b
```

0.999

TRUE

[438]:
```
# Hidden Test Cell
```

Part C) Suppose that the results of your test from **Part B** indicate that you should reject the null hypothesis. Does this also mean that the different age groups *causes* the difference in pet preferences? Save the boolean `TRUE` or `FALSE` into variable `p3.c`.

[439]:
```
p3.c = NA

# your code here
p3.c = FALSE
```

[440]:
```
# Hidden Test Cell
```