Problem Set 1

Applied Stats/Quant Methods 1

Due: October 1, 2023

Instructions

- Please show your work! You may lose points by simply writing in the answer. If the problem requires you to execute commands in `R`, please include the code you used to get your answers. Please also include the `.R` file that contains your code. If you are not sure if work needs to be shown for a particular problem, please ask.
- Your homework should be submitted electronically on GitHub.
- This problem set is due before 23:59 on Sunday October 1, 2023. No late assignments will be accepted.
- Total available points for this homework is 80.

Question 1 (40 points): Education

A school counselor was curious about the average of IQ of the students in her school and took a random sample of 25 students' IQ scores. The following is the data set:

```r
y <- c(105, 69, 86, 100, 82, 111, 104, 110, 87, 108, 87, 90, 94, 113, 112, 98,
    80, 97, 95, 111, 114, 89, 95, 126, 98)
```

1. Find a 90% confidence interval for the average student IQ in the school.

2. Next, the school counselor was curious whether the average student IQ in her school is higher than the average IQ score (100) among all the schools in the country.

   Using the same sample, conduct the appropriate hypothesis test with $\alpha = 0.05$.

```r
# Q1: Find 90% confidence interval for the average student IQ in the school

# As stated in Logan (n.d.), a confidence interval is "the interval that contains
# the population parameter with probability 1 −  ." In other words, the CI is the
# range of values within which the relevant population statistic is expected to sit
# within − to the stated level of certainty. This is calculated using values from
# a randomly−selected sample.

# If calculating manually for the t−test method, this requires finding the sample
# mean, it's standard deviation and then standard error for this mean, the
# corresponding t−score for the chosen confidence level, and the margin of error.
# Link: Logan (n.d.) https://bookdown.org/logan_kelly/r_practice/p09.html

# There are two ways of finding the 90% confidence interval − firstly, through
# using the t−distribution and manually calculating each component value before
# inputting these into the qt() function:

# Find the sample mean ie the central tendency
mean_y <- mean(y)

# Find the standard error for the population statistic
n_y <- length(y) # find the n−size of the sample
standard_deviation_y <- sd(y) # also possible to calculate this manually by finding
# the sample variability using var() and finding its squareroot ie sqrt(var(y))
standard_error_for_y <- standard_deviation_y/sqrt(n_y)

# Find the corresponding t−score for 90% confidence interval − we are not assuming
# direction, and so we are using a two−tail test
confidence_interval <- 0.1
degrees_of_freedom <- n_y - 1
t_score <- qt(p = confidence_interval/2, df = degrees_of_freedom, lower.tail = FALSE) # where
# p is the vector of probabilities for each tail, df is one less the N−size as we
# exclude the mean_y value, and we indicate we are using two−tail test

# Find the margin of error
margin_of_error_y <- t_score * standard_error_for_y

# Find the confidence interval
lower_bound <- mean_y - margin_of_error_y
```

```r
upper_bound <- mean_y + margin_of_error_y

# Print the 90% confidence intervals and sample y mean, using round() to give
# each value to two decimal places.
CI_values <- c(round(lower_bound, 2), round(mean_y, 2), round(upper_bound, 2))
names(CI_values) <- c("lower-bound", "sample mean", "upper-bound")
print(CI_values)

# The same can be approximated using Z-scores instead. The corresponding Z-score
# to the 90% confidence interval is 1.64.

lower_bound_z <- mean_y - (1.64 * standard_deviation_y / sqrt(n_y))
upper_bound_z <- mean_y + (1.64 * standard_deviation_y / sqrt(n_y))
# And now printing these values
CI_values_z <- c(round(lower_bound_z, 2), round(mean_y, 2), round(upper_bound_z, 2))
names(CI_values_z) <- c("lower-bound", "sample mean", "upper-bound")
print(CI_values_z)

# Note the values are similar, though the approximated confidence interval is
# slightly reduced (due to the Z-score slightly under-shooting the true 90% threshold)
```

```r
# Q2:  Next, the school counselor was curious whether the average student IQ in
# her school is higher than the average IQ score (100) among all the schools in
# the country. Using the same sample, conduct the appropriate hypothesis test
# with    = 0.05.

# Now we are conducting a one-tail significance test for the difference between two
# means in order to reject the null hypothesis that mean_y is equal to or lower than
# mean_population

mean_population <- 100
t.test(x = y, mu = mean_population, alternative = "greater")

# The result indicates that p > 0.05, failing to reject the null hypothesis ie
# we fail to reject that the class's average IQ score is equal to or lower than
# the average IG score among all schools in the country.
```

Question 2 (40 points): Political Economy

Researchers are curious about what affects the amount of money communities spend on addressing homelessness. The following variables constitute our data set about social welfare expenditures in the USA.

`State`	*50 states in US*
`Y`	*per capita expenditure on shelters/housing assistance in state*
`X1`	*per capita personal income in state*
`X2`	*Number of residents per 100,000 that are "financially insecure" in state*
`X3`	*Number of people per thousand residing in urban areas in state*
`Region`	*1=Northeast, 2= North Central, 3= South, 4=West*

Explore the `expenditure` data set and import data into `R`.

```
# inputting these into the qt() function:
```

- Please plot the relationships among *Y*, *X1*, *X2*, and *X3*? What are the correlations among them (you just need to describe the graph and the relationships among them)?

- Please plot the relationship between *Y* and *Region*? On average, which region has the highest per capita expenditure on housing assistance?

- Please plot the relationship between *Y* and *X1*? Describe this graph and the relationship. Reproduce the above graph including one more variable *Region* and display different regions with different types of symbols and colors.