PS01 - My Answers (Quants 1)

Sarah Magdihs

09.10.2025

Question 1: 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:.

```
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.

   `Answer:`

   Let's start with the long version: As shown below, I calculated the sample size, the mean, the standard deviation and the standard error in order to then calculate the confidence interval "by hand". I use a t-value because n is smaller than 30.

   `R Code:`

```
n <- length(y)
mean_y <- mean(y)
sd_y <- sd(y)
se_y <- sd_y/sqrt(n)

t_value <- qt((1-0.9)/2, n-1, lower.tail = FALSE)
#Technically, I could also do the following:
#t_value <- qt(0.95, n-1) OR qt(0.05, n-1, lower.tail = FALSE)
#Each of these version will give me the positive value at the upper tail

#CI
upper90 <- mean_y + t_value*se_y
lower90 <- mean_y - t_value*se_y

CI90 <- c(lower90, upper90)
print(CI90)
```

```
[1]  93.95993 102.92007
```

Thus, the 90 per cent confidence intervall is [93.96;102.92]. This means that - with repeated sampling - the confidence interval contains the true parameter at least 90 per cent of the time. Hence, we can be 90 per cent confident that the interval [93.96;102.92] contains the population mean.

You can also cross-check this with a t-test:

R Code:

```
ttest_1 <- t.test(y, conf.level = .90)
```

```
data:  y
t = 37.593, df = 24, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
90 percent confidence interval:
93.95993 102.92007
sample estimates:
mean of x
98.44
```

As we can see, the t-test also shows that the CI = [93.96;102.92].

(Besides also doing other things).

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 $\alpha = 0.05$.

Based on the information, this is a one-sided t-test, as the counselor wants to test whether the average student IQ at her school is **higher** than the mean of the population.

Thus, the hypotheses are as following:

H0: mean is equal to or smaller than 100

Ha: mean bigger than 100

Since alpha = 0.05, the confidence level is 95 per cent.

Answer:

Let's start with the short version again: .

R Code:

```r
t.test(y, mu = 100, alternative = "greater", conf.level = 0.95)
```

```
data:  y
t = -0.59574, df = 24, p-value = 0.7215
alternative hypothesis: true mean is greater than 100
95 percent confidence interval:
93.95993      Inf
sample estimates:
mean of x
98.44
```

Based on this, the Null-Hypothesis cannot be rejected (p=0.7215, which is bigger than 0.05). Note that the p-value denotes the probability of seeing a value as extreme as this one (or higher) if the H0 was true.

Moreover, we can see that the mean of the counselor's students is actually lower than the average for all students in the country (mean= 98.44). This is not only indicated by the 'mean of x' but also by the negative t-value.

Furthermore, as discussed in the lecture, we can technically also do this step by step ourselves.

R Code:

```r
#I only create mu_0 here, since the other elements where already created for Task 1.
mu_0 <- 100

#create test statistic
TS <- (mean_y-mu_0)/se_y
#p-value: since it is a one-sided test (on the right side), it needs to be 1-Probability as pt would otherwise give us the probability that T =< TS (so basically the space under the curve to the left of the TS)
p_value <- 1-pt(q=TS, df = n-1)

#data frame
ttest_by_hand <- c(Mean = mean_y, StdError =se_y, t = TS, df = n-1, p_value = p_value)
print(ttest_by_hand)
```

```
      Mean   StdError          t         df    p_value
98.4400000  2.6185747 -0.5957439 24.0000000  0.7215383
```

Question 2: 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.

Okay, so let's load and inspect our data first:

R Code:

```
expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/StatsI_2025/main/datasets/expenditure.txt", header=T)
head(expenditure)
str(expenditure)
```

```
> head(expenditure)
STATE  Y   X1  X2  X3 Region
1    ME 61 1704 388 399      1
2    NH 68 1885 272 598      1
3    VT 72 1745 397 370      1
4    MA 72 2394 458 868      1
5    RI 62 1966 157 899      1
6    CT 91 2817 162 690      1
> str(expenditure)
'data.frame': 50 obs. of  6 variables:
$ STATE : chr  "ME" "NH" "VT" "MA" ...
$ Y     : int  61 68 72 72 62 91 120 99 70 82 ...
$ X1    : int  1704 1885 1745 2394 1966 2817 2685 2521 2127 2184 ...
$ X2    : int  388 272 397 458 157 162 494 153 152 187 ...
$ X3    : int  399 598 370 868 899 690 728 826 656 674 ...
$ Region: int  1 1 1 1 1 1 1 1 1 2 ...
```

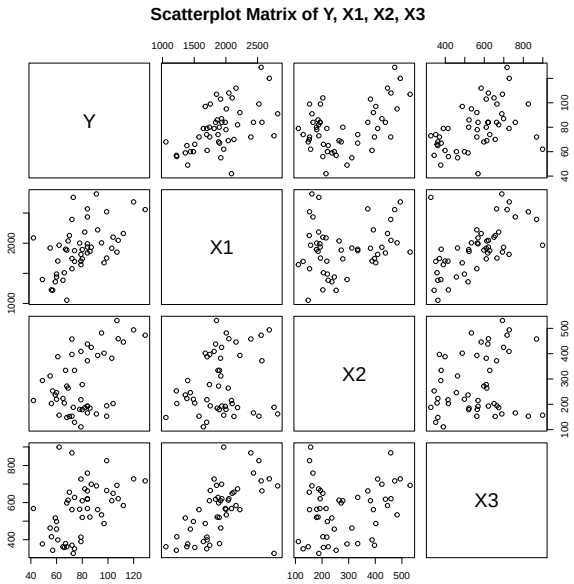
1. Please plot the relationships among Y, X1, X2, and X3 ? What are the correlations among them (you just need to describe the graph and the relationships among them)?

Answer:

In order to plot the relationships between these variables, I will use a scatterplot. However, I don't want to plot each combination individually, as that is really tedious. Instead, let's use the pairs command.

R Code:

```
pdf("Plot_Matrix.pdf")
pairs(expenditure[,c(2,3,4,5)],
      main = "Scatterplot Matrix of Y, X1, X2, X3")
dev.off()
```



So, that covers the plotting. To assess the relationships, looking at data is important, but sometimes correlations (or assosciations) can be hard to gauge visually. So let's use a correlation matrix before we move to the interpretation.

R Code:

```
cor_var <- expenditure[,c(2,3,4,5)]
correlation_matrix <- cor(cor_var)
print(correlation_matrix)
round(correlation_matrix, 2)
```

```
       Y   X1   X2   X3
Y  1.00 0.53 0.45 0.46
X1 0.53 1.00 0.21 0.60
X2 0.45 0.21 1.00 0.22
X3 0.46 0.60 0.22 1.00
```

Now for the description/interpretation:

Generally, the created plot includes all possible correlation plot between the four variables. The diagonal only shows the variable names, since each variable obviously perfectly correlates with itself. The correlations above and below the diagonal are mirrored (since it shows, for example, the correlation of Y and X1, as well as X1 and Y). So, I will focus on the plots below the diagonal.

There seems to be a moderate positive correlation between Y and X1; Y and X2; Y and X3. Thus, based on the plots and the correlation coefficients, states that have a higher per capita personal income/more financially insecure residence/a higher urban population density, tend to spend more money on shelters/housing assistance.

Moreover, X1 and X3 have an r = 0.6, which means that there could be collinearity issues when a regression model uses both variables as predictors.

Lastly, X1 and X2 are only weakly correlated (but seem somewhat linear).X2 and X3 have a low correlation coefficient (r=0.22) which indicates that there is at best a weak linear correlation. This also makes sense when looking at the graphs: The scatterplot of X2 and X3 looks like their assosciation may be better described by a quadratic function.

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

Answer:

To look at the relationship between Expenditure and Region, we can use boxplots. However, when I inspected the data, it showed that Region is an integer - which will be an issue for a boxplot. So let's make it a factor.

R Code:

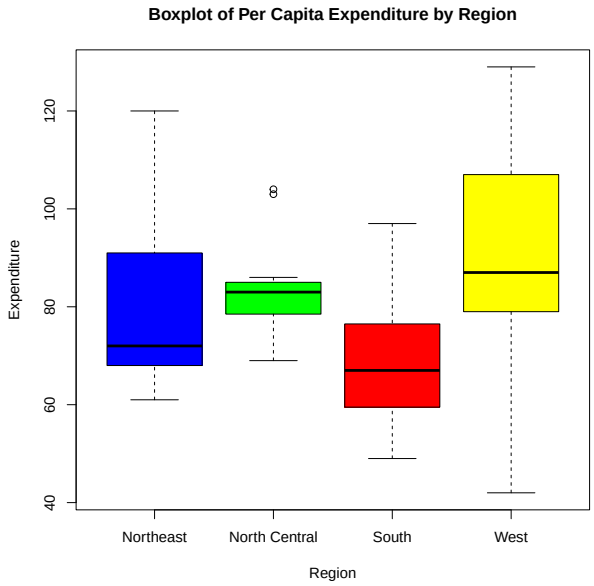
```r
str(expenditure)

#Issue: Region is an integer
expenditure$Region <- factor(expenditure$Region,
                             levels = c(1, 2, 3, 4),
                             labels = c("Northeast", "North Central", "
    South", "West"))
str(expenditure)
```

Now, we can make boxplots:

R Code:

```
pdf("boxplot_Y_Region.pdf")
boxplot(expenditure$Y ~ expenditure$Region,
        main="Boxplot of Per Capita Expenditure by Region",
        ylab="Expenditure",
        xlab="Region",
        col = c("blue","green","red","yellow"))
dev.off()
```

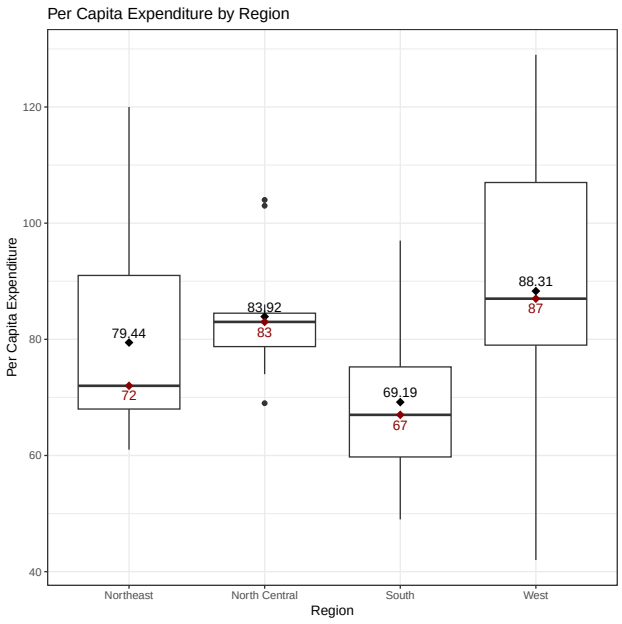


However, the second question is about the average, but boxplots give us the median. In this case, I think the difference should not be drastic. But since averages our outlier sensitive, let's make sure it actually doesn't make a difference.

R Code:

```
pdf("boxplot_with_average.pdf")
ggplot(data = expenditure, aes(x =Region, y = Y)) +
  geom_boxplot() +
  stat_summary(fun = mean, geom = "point", shape = 18, size = 3, color =
    "black") +
  stat_summary(fun = mean, geom = "text", aes(label = round(after_stat(y)
    ,2)),
                vjust = -0.5, color = "black") +
  stat_summary(fun = median, geom = "point", shape = 18, size = 3, color
    = "darkred") +
  stat_summary(fun = median, geom = "text", aes(label = round(after_stat(
    y),2)),
```

```
                  vjust = 1.5, color = "darkred") +
  labs(title = "Per Capita Expenditure by Region",
       x = "Region", y = "Per Capita Expenditure") +
  theme_bw()
dev.off()
```



As expected, in this case it doesn't really make a difference regarding the interpretation (despite the mean and median slightly diverging).

Here, we see that on average states in the Region "West" have the highest per capita expenditure on shelters/housing assistance.

3. 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..

Answer:

Let's start with the simple plot:

R Code:

```
pdf("Task2_3_basic.pdf")
plot(expenditure$X1, expenditure$Y,
     main="Relationship Between Per Capita Expenditure and Per Capita
    Personal Income",
     ylab="Expenditure",
     xlab="Personal Income",
     cex.main = 0.95,
```

```
7
8       abline(lm(Y ~ X1, data = expenditure), col = "red", lwd = 2) )
9 dev.off()
```

As already mentioned during the discussion of the scatterplots and potential correlations, we can see here that there is some kind of positive linear assosciation between per capita personal income and per capita expenditure on shelters/housing assistance. More concretely, states that have higher per capita personal income tend to spend more money on shelters/housing assistance.

The added regression line helps visualise this relationship more effectively.

Lastly, let's reproduce the above graph including one more variable Region and display different regions with different types of symbols and colors.

R Code:

```
1 #define colors per region
2 colors_regions <- c("Northeast" = "lightblue",
3                     "North Central" = "lightpink",
4                     "South" = "lightgreen",
5                     "West" = "purple")
6
7 #define symbols per region
8 symbols_regions <- c("Northeast" = 15,  # square
9                     "North Central" = 17,  # triangle
10                    "South" = 18,  # diamond
11                    "West" = 19)  # circle
```

```
12
13 #now the graph
14 pdf("Task2_3_with_Colours_and_Symbols.pdf")
15 plot(expenditure$X1, expenditure$Y,
16      main="Relationship Between Per Capita Expenditure and Per Capita
       Personal Income",
17      ylab="Expenditure",
18      xlab="Personal Income",
19      cex.main = 0.95,
20      col = colors_regions[expenditure$Region],
21      pch = symbols_regions[expenditure$Region])
22
23 legend("topleft", legend = levels(expenditure$Region), col = colors_
       regions, pch = symbols_regions, title = "Region")
24 dev.off()
```