Answer Key: Problem Set 1

Applied Stats/Quant Methods 1

Jeffrey Ziegler

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.*

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:

```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.*

   First, let's calculate the t-score for the 90% confidence interval with degrees of freedom equal to 24 (remember that df $= n - 1$ for the t-distribution, and we're using the t-distribution because we have a small sample size). For the 90% confidence interval, the lower tail is equal to 0.05 ($(1-\alpha)/2 = (1-0.90)/2 = 0.05$).

```
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)
# capture the number of observations
n <- length(y)

# (a) Calculate the 90% confidence interval for the student IQ
# Step 1: get t-score
t <- qt(0.05, n-1, lower.tail = F)

# Step 2: Calculate lower and upper parts for the 90%
```

Second, let's calculate the mean ($\bar{y}$), the sample standard deviation S, and then $\hat{\sigma}_{\bar{y}} = \dfrac{S}{\sqrt{n}}$. This allows us to calculate our 90% confidence interval for the student IQ as $\bar{y} \pm T \times \hat{\sigma}_{\bar{y}}$, which equals $98.44 \pm 1.71 \times \frac{13.09}{5} = [93.96,\ 102.92]$.

```
upper_CI <- mean(y)+(t*(sd(y)/sqrt(n)))

# print CIs with mean
c(lower_CI, mean(y), upper_CI)

# double check our answer
t.test(y, conf.level = 0.9)$"conf.int"
```

We can interpret this result by saying that if we took 100 samples, the true population mean of the student IQ in the school should fall within the interval in 90 of those samples.

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 $\alpha = 0.05$.

   First, let's set up our null hypothesis: we want to know whether the mean of the sample ($\bar{y}$ or $\hat{\mu}$) is greater than the theoretical mean (μ_0). So, using proof by contradiction, $H_0 : \hat{\mu} \leq \mu_0$. Next, let's compute the standard error and our test statistic to get a p-value. Remember, since this is a one-sided test, we don't want both tails, so `lower.tail=F`.

```
# (b) Step 1: Calculate the standard error
SE <- sd(y)/sqrt(n)
# Step 2: Calculate the test statistic for this hypothesis testing of
    mean
t <- (mean(y) - 100)/SE
```

```
# Get the p-value from t-distribution
pvalue <- pt(t, n-1, lower.tail = F)
```

We can see that the p-value (≈ 0.72) is not equal to or below the $\alpha = 0.05$ threshold, so we would say that we do not find sufficient evidence to reject the null hypothesis that the average IQ of the students in this school is less than or equal to the population average IQ score ($\mu_0 \leq 100$). This makes sense; it's unlikely that we would have enough evidence to suggest that the average in the sample was larger than the population mean given that it is in fact lower.

We can also check our answer by using the `t.test` function in `R`. Note that if you only run this function and do not describe the steps of conducting a hypothesis test, that is not enough for full credit.

```
# Or another way to do this hypothesis testing is to use the function t.test directly
t.test(y, mu = 100, conf.level = 0.95, alternative = "greater")
```

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*

(a) *Explore the* ***`expenditure`*** *data set and import data into* ***`R`****.*

```
# read in expenditure data
expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/StatsI_Fall2023/main/datasets/expenditure.txt", header=T)
# inspect data through summary
summary(expenditure)
```

```
STATE          Y                   X1               X2                X3
AK     : 1   Min.   : 49.00   Min.   :1053   Min.   :334.0   Min.   :326.0
AL     : 1   1st Qu.: 68.25   1st Qu.:1698   1st Qu.:374.2   1st Qu.:426.2
AR     : 1   Median : 81.00   Median :1897   Median :395.0   Median :568.0
AZ     : 1   Mean   : 85.04   Mean   :1912   Mean   :404.7   Mean   :561.7
CA     : 1   3rd Qu.:102.00   3rd Qu.:2096   3rd Qu.:419.5   3rd Qu.:661.2
CO     : 1   Max.   :142.00   Max.   :2817   Max.   :637.0   Max.   :899.0
```

(b) *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)?*

```r
# create a matrix scatter plot to
# visualize the relationship among Y, X1, X2 and X3
# so not the first column of expenditure
pdf("plot_2a.pdf")
pairs(expenditure[,2:5], main = "")
dev.off()
```

The correlation (r) between Y and X_1 is 0.53, which indicates a moderate correlation and is consistent with the two subplots in the top-left of Figure 1. However, the correlation between Y and $X2$, and Y and $X3$, are weak (-0.21, 0.22).

We can also see that there is a positive relationship between X_1 and X_3, though there is not much of a relationship between X_1 and X_2, or X_2 and X_3.

(c) *Please plot the relationship between* Y *and* Region*? On average, which region has the highest per capita expenditure on public education?*

```r
# generate boxplot with comparisons for different values of Region
pdf("plot_2b.pdf")
boxplot(expenditure$Y~expenditure$Region, xlab="Region", ylab="Y", main="
    ")
dev.off()
```

Figure 2 displays the box-plot of Y by *Region* side-by-side, which is appropriate because *Region* is a categorical variable and Y is a quantitative variable. The above code generates a side-by-side box-plot for the variables Y and *Region*. From Figure 2, we can see that Region 4 has the highest per capita expenditure on public education.

Figure 1: Scatterplot of relationship between Y, X_1, X_2, and X_3.

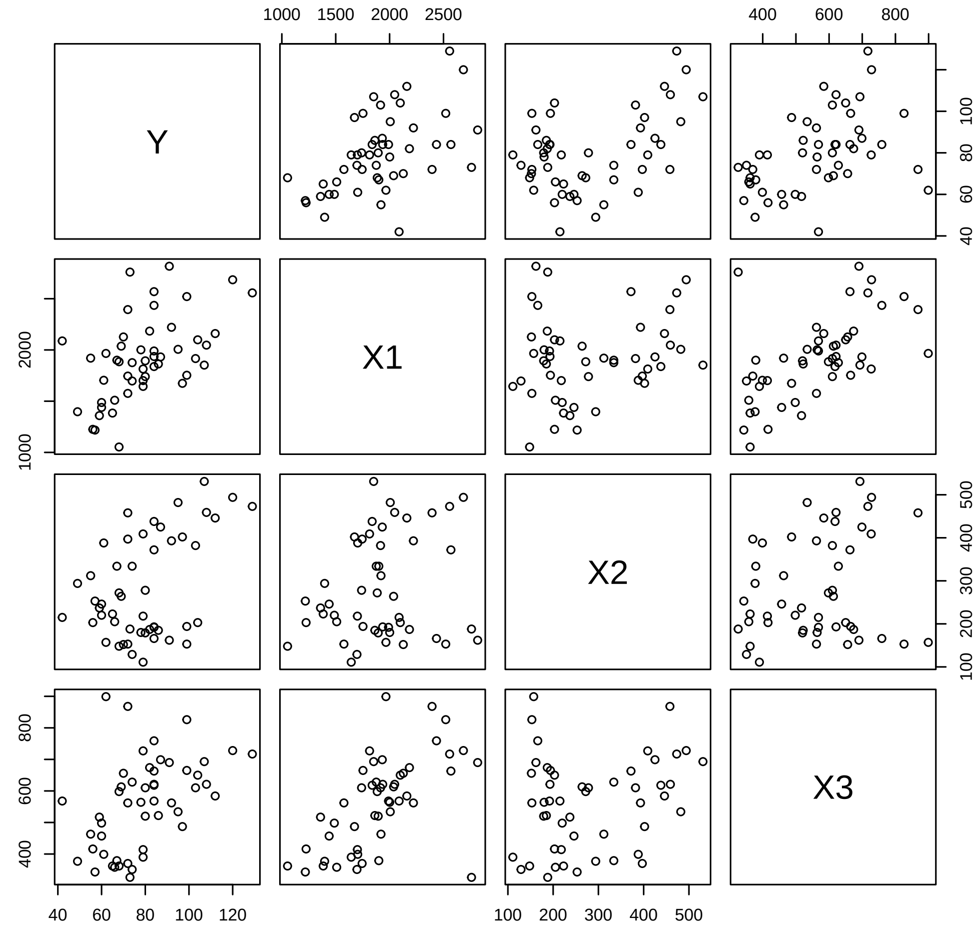
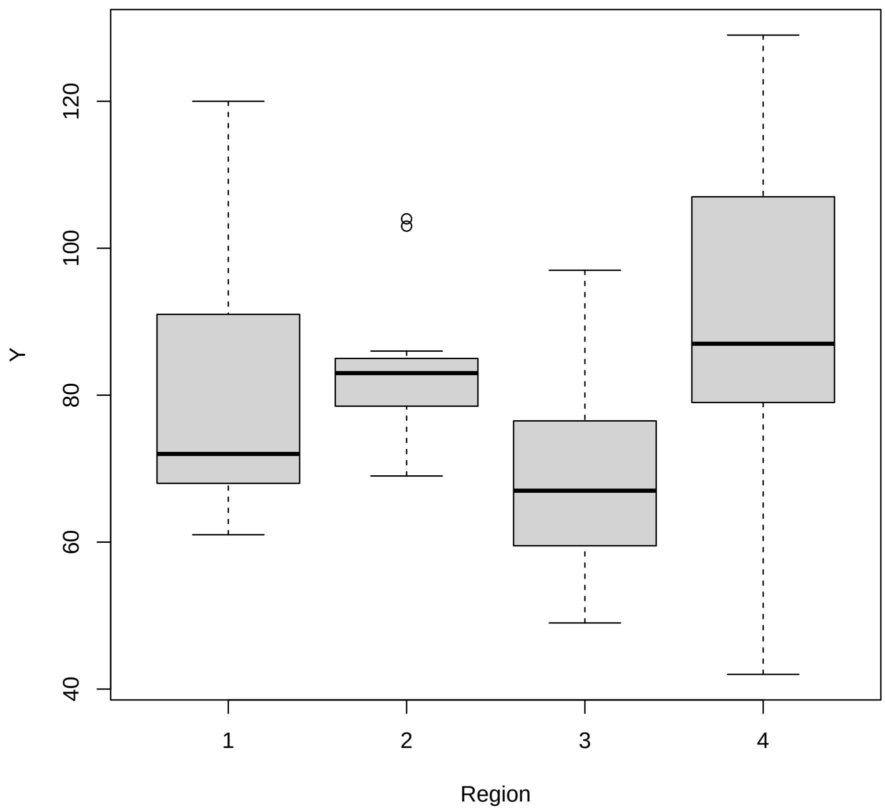


Figure 2: Boxplot of Y by *Region*.



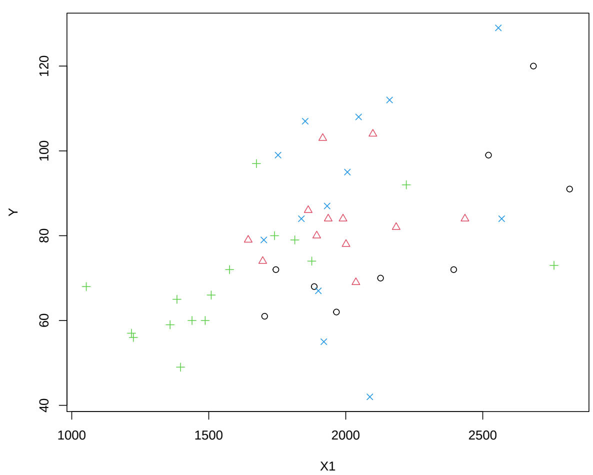
(d) *Please plot the relationship between* Y *and* X1*? Describe this graph and the relationship. Reproduce the above graph including one more variable* Region *and display different regions with different types of symbols and colors.*

```r
# create scatterplot of Y and X1
# basic and then differentiate color by region
pdf("plot_2c1.pdf", width=8)
ggplot(expenditure, aes(x=X1, y=Y)) +
  geom_point() + labs(y="Y\n", x="\nX1") +
  theme_bw() +
  theme(axis.title = element_text(size=20),
        axis.text = element_text(size=15))
dev.off()

# make sure that Region is categorical
expenditure$Region <- as.factor(expenditure$Region)

pdf("plot_2c2.pdf", width=8)
plot(expenditure$X1, expenditure$Y,
     pch = as.integer(expenditure$Region),
     xlab="X1", ylab="Y", main="",
     col = expenditure$Region)
dev.off()

pdf("plot_2c3.pdf", width=8)
ggplot(expenditure, aes(x=X1, y=Y, colour=Region, shape=Region)) +
  geom_point()  +
labs(y="Y\n", x="\nX1") +
  theme_bw() +
  theme(axis.title = element_text(size=20),
        axis.text = element_text(size=15),
        legend.title = element_text(size=17),
        legend.text = element_text(size=15))
dev.off()
```

We're using a scatter plot because both of these variables are quantitative, and we can see from Figure 3 that there is a moderate positive linear correlation between X_1 and Y (which we noted in the above question). However, we can see in the right plot of Figure 3 that certain regions have much steeper (higher) or flatter (lower) correlations between Y and X_1, which suggests that the effect of X_1 of Y differs by region.

Figure 3: Differentiating between region by color and shape.

(a) Using base R.



(b) Using ggplot.