PS01 Response

Applied Stats/Quant Methods 1

Caitlín Cooney

Question 1 (50 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:

```
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.**

   The formula for finding a Confidence Interval is:

   Confidence Interval = (sample mean)+/-(confidence level. value)(standard error).*

   So I start by finding the sample mean (xbar), standard deviation (sd), and sample size (n).

   ```
   ybar <- mean(y)
   sd <- sd(y)
   n <- length(y)
   ```

   Find the confidence level value (z)

   ```
   conf.level <- 0.9
   z <- qt((1+conf.level)/2, df= length(y)-1) # this gets me to the 95th
      probability
   ```

This gets me to the 95th probability in order to establish a 90 percent confidence interval.

Next, find the standard error

```
se <- sd(y)/sqrt(n)
se
```

Multiply the critical value by the standard error

```
CI <- z*se
CI
```

We can now determine the lower and upper confidence interval boundaries:

```
lowerinterval <- ybar - CI
lowerinterval

upperinterval <- ybar + CI
upperinterval
```

Our Lower Interval = 93.95993, and Upper Interval = 102.9201.

With repeated sampling, we would anticipate that the mean IQ of students in the school falls between 90 out of 100 re-samplings from our population

We can check our work using the t.test function:

```
t.test(y, conf.level = 0.9, alternative = "two.sided")
```

The results are as follows:

```
 One Sample t-test

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
```

The t.test output shows us a Lower Confidence Interval of 93.95993 and an Upper Confidence Level of 102.92007, which matches our results obtained by hand, and allows us to be certain of our results.

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, state the null hypothesis and alternative hypothesis:

```
# H0: mu <= 100
# H1: mu > 100
```

```
mu <- 100
```

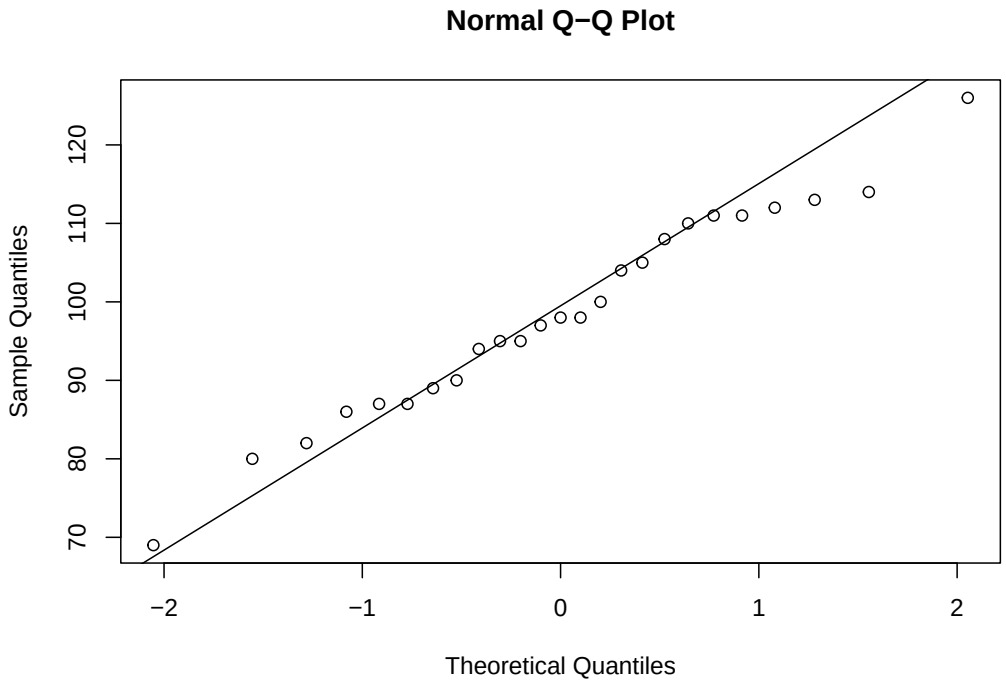
   Set significance level:

```
a = 0.05
```

   We use a QQ plot to determine if our IQ variable is normally distributed

```
qqnorm(y)
qqline(y,
       distribution = qnorm)
```

Figure 1: IQ Q-Q Plot



The scatter plot shows a strong positive correlation so we can conclude that our IQ variable is indeed normally distributed.

We will use a right-tailed t-test, because population size is unknown, our sample size is less than 30 and we are specifying the direction i.e. the alternative hypothesis states that the parameter is bigger than the value specified in the null hypothesis

Start by getting the mean:

```
ybar <- mean(y)
```

Then get the standard deviation:

```
sd <- sd(y)
```

Then the standard error

```
se <- sd(y)/sqrt(length(y)) # Create an object with our standard error
```

Calculate the degrees of freedom:

```
df <- n-1
```

Calculate our test statistic:

```
t <- (ybar-mu)/(sd/sqrt(n))
```

Create a dataframe with relevant info, and label it:

```
data <- c(ybar, se, t, pt(-abs(t), df, lower.tail=FALSE))
names(data) <- c("Mean", "Std Error", "t_score", "p-value")
round(data, 4)
```

The results are as follows:

```
Mean Std Error   t_score   p-value
 98.4400    2.6186   -0.5957    0.7215
```

From this, we can draw a conclusion. Our p-value = 0.7215 which is greater than 0.05, which means we cannot reject the null hypothesis, as we cannot conclude that the mean IQ of the students in this school is significantly greater than 100.

We can test if our calculations are correct using the t.test function:

```
t.test(y,
       mu = 100,
       var.equal = FALSE,
       alternative = "greater",
       conf.level = .9)
```

We get the result:

```
One Sample t-test

data:  y
t = -0.59574, df = 24, p-value = 0.7215
alternative hypothesis: true mean is greater than 100
90 percent confidence interval:
 94.98915      Inf
sample estimates:
mean of x
    98.44
```

As the t and p-value given by the t.test function match the values that we calculated, we can conclude we have the correct result.

Question 2 (50 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`.

- **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)?**

  Region is a categorical variable, yet it is currently not structured as a factor. Instead, it is structured as an integer. The as.factor() function converts a variable into a factor. We use the $ symbol to tell R which variable in your dataset we want to change.

```
expenditure$Region<-as.factor(expenditure$Region)

expenditure$Region<-factor(expenditure$Region,
                           levels=c(1,2,3,4),
                           labels=c("North East", "North Central", "South
   ", "West"))
```

  We can double check the structure to see if this worked properly.

```
str(expenditure)
```

```
'data.frame': 50 obs. of  6 variables:
$ STATE : chr  "ME " "NH " "VT " "MA " ...
$ Y     : num  61 68 72 72 62 91 120 99 70 82 ...
$ X1    : num  1704 1885 1745 2394 1966 ...
$ X2    : num  388 272 397 458 157 162 494 153 152 187 ...
$ X3    : num  399 598 370 868 899 690 728 826 656 674 ...
$ Region: Factor w/ 4 levels "North East","North Central",..: 1 1 1 1 1 1 1 1 1 2
```

We can now use the pairs function to plot the relationships between our variables.

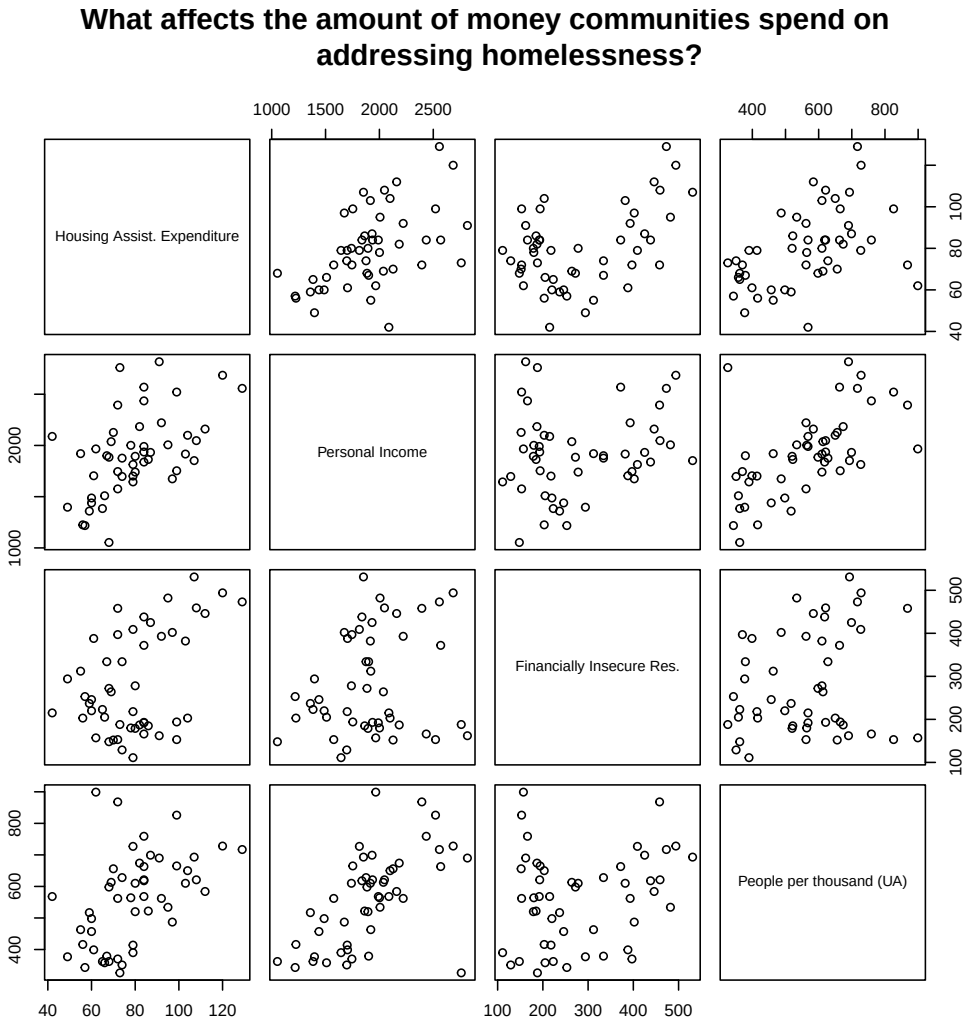
First, we create a new variable expenditure2 so that we are not hindered by our non-numeric variable STATES

```
expenditure2 <- expenditure[2:5]
```

Use the pairs function to plot 'expenditure2' i.e. columns 2 to 5.

```
pairs(expenditure2,
      # Change points by group
      # Change labels
       labels = c("Housing Assist. Expenditure", "Personal Income", "
   Financially Insecure Res.",
                  "People per thousand (UA)"),
      main = "What affects the amount of money communities spend on
      addressing homelessness?")
```

Figure 2: Homelessness



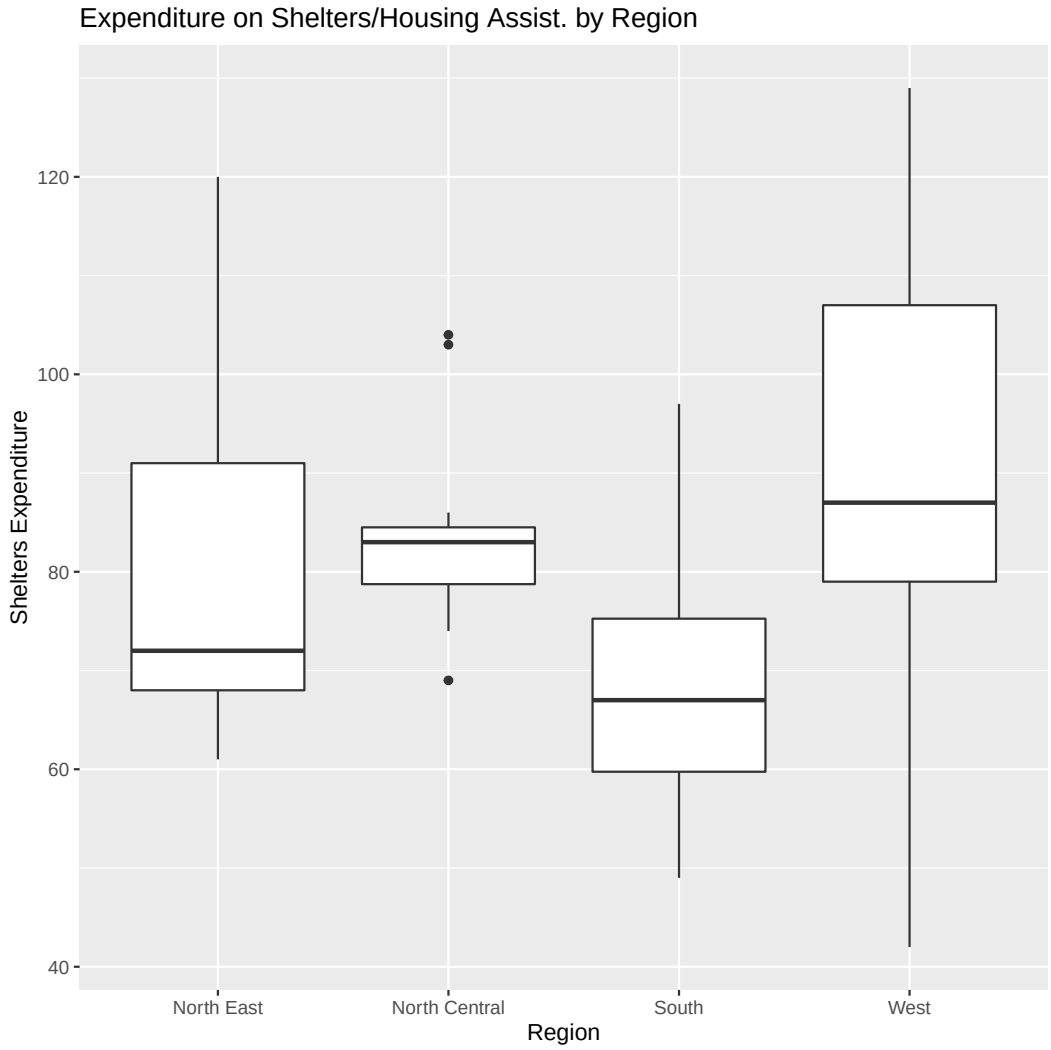
- There is a positive correlation between Housing Assistance Expenditure and Personal Income which would indicate that Personal Income has some kind of effect on Housing Assistance Expenditure.
- There is no correlation between Housing Assistance Expenditure and Financially Insecure Residents. The number of financially insecure residents will not have an effect of Housing Assistance expenditure.
- There is no correlation between People per Thousand in Urban Areas and Housing Assistance Expenditure, so we do not expect this to affect Housing Assistance Expenditure.

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

We can make an attractive boxplot using the qplot() function and specify the type of plot with geom() = boxplot

```
RegionPlot2 = qplot(Region, Y, data=expenditure, geom=("boxplot"),
                    main="Expenditure on Shelters/Housing Assist. by
   Region",
                    xlab="Region", ylab="Shelters Expenditure")
RegionPlot2
```

Figure 3: Expenditure By Region



We can see that the highest median for per capita spending on Housing Assistance is in the Wetst, and as the median is usually quite close to the mean, we can expect that the West will have the hight average per capita spending on Housing Assistance.

To confirm this, we can find the mean per capita expenditure on housing assistance

```
RegionMeans = with(expenditure, by(Y, Region, mean))
RegionMeans
```

```
 Region: North East
[1] 79.44444
------------------------------------------------------------------------
Region: North Central
[1] 83.91667
------------------------------------------------------------------------
Region: South
[1] 69.1875
------------------------------------------------------------------------
Region: West
[1] 88.30769
```

We can see from the above results that the Region with the highest average per capita spending on housing assistance is the West.

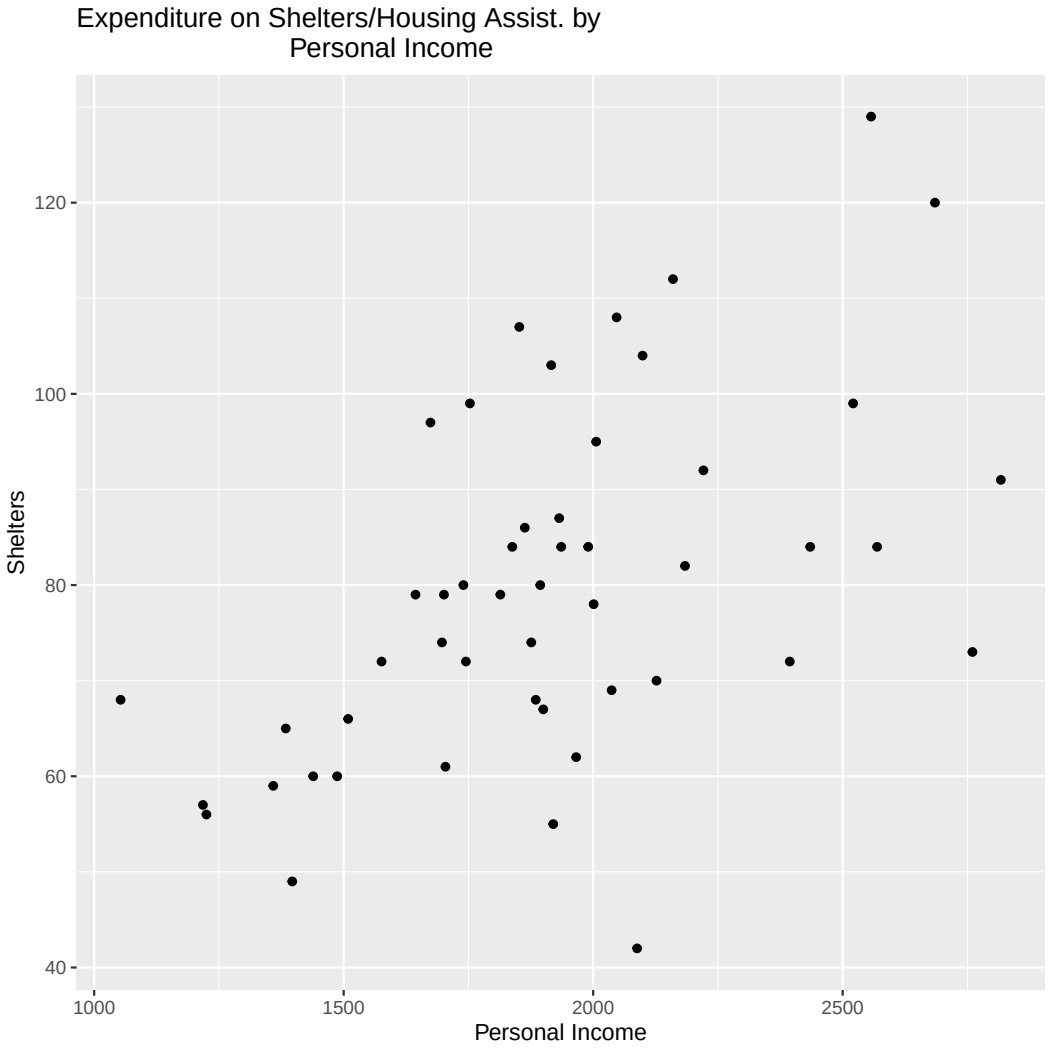
- **Please plot the relationship between *Y* and *X1*? Describe this graph and the relationship. Reproduce the above graph including one more variable *Region* and display different regions with different types of symbols and colors.**

  Make a more attractive scatter plot using the qplot funtion and specify type of plot with geom = point

```
PersonalIncomePlot2 <- qplot(X1, Y, data=expenditure, geom=("point"),
                             main="Expenditure on Shelters/Housing Assist. by
                             Personal Income",
                             xlab="Personal Income", ylab="Shelters")

PersonalIncomePlot2
```

Figure 4: Effect of Personal Income on Expenditure on Shelters/Housing Assistance



We can see from the above graph that there is a relatively weak positive correlation between per capita expenditure on shelters/housing assistance in state and per capita personal income in state. This would imply that personal income does have an effect on expenditure on housing assistance, but a large change in Personal Income would result in just a small change in Housing Assistance. Further investigation would be required here.

To answer the second part of the question, we can use the pairs() function to reproduce the a graph of the relationship between our variables as we did above.

```
expenditure3 <- expenditure[2:3]

pairs(~ Y + X1, data = expenditure2)

with(expenditure3, pairs(~ Y + X1))
```

To include one more variable 'Region', create an object from the 6th column

```
Region <- expenditure[, 6]
```

To display different regions with different types of symbols and colors, simply include the arguments 'col' and 'pch'.

```
EffectByRegion <- pairs(expenditure3,
      col = c("red", "cornflowerblue", "purple")[Region],    # Change
   color by group
      pch = c(8, 18, 1)[Region],
      cex = 1,
      # Change points by group
      # Change labels
      labels = c("Housing Assist. Expenditure", "Personal Income", "
   Financially Insecure Res.",
                 "People per thousand (UA)"),
      main = "What affects the amount of money communities spend on
      addressing homelessness?")
```

Figure 5: Effect of Personal Income on Expenditure on Shelters/Housing Assistance by Region