# Problem Set 1

Ciara O'Flaherty

Due: October 3, 2021

## 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 in .pdf form.
- This problem set is due before 14:59 on Monday October 3, 2021. No late assignments will be accepted.
- Total available points for this homework is 100.

## Question 1 (50 points): Education

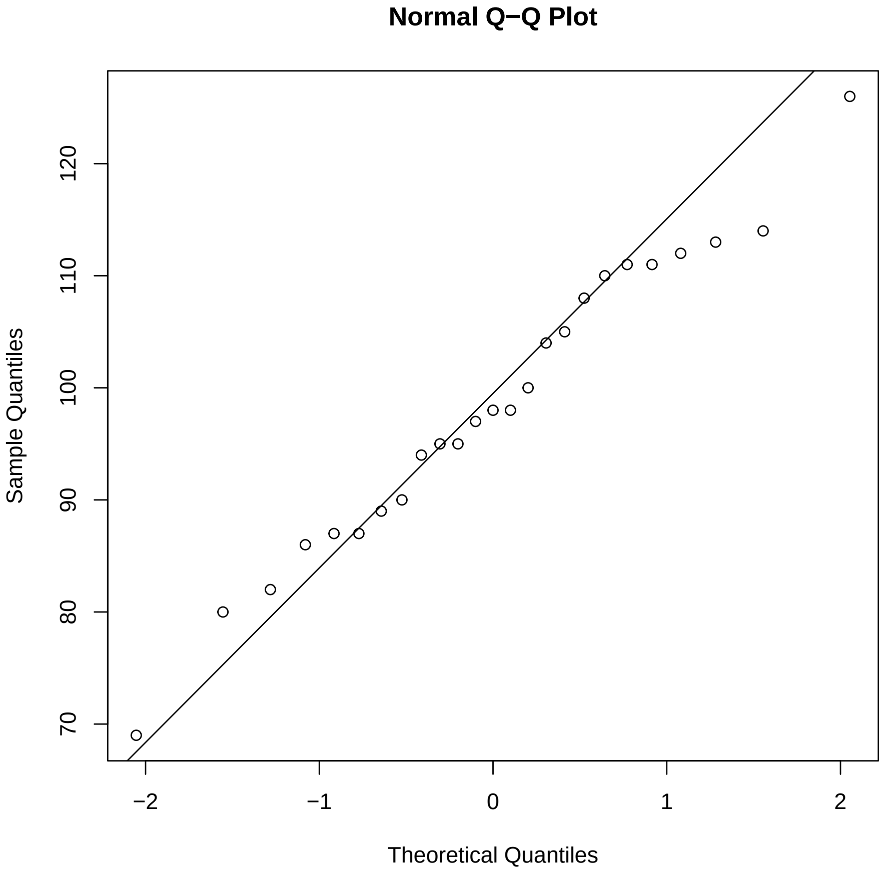
"A school counselor was curious about the average of IQ of the students in her school and took a random sample of 25 students' IQ scores..."

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

### Q1: Answer 1

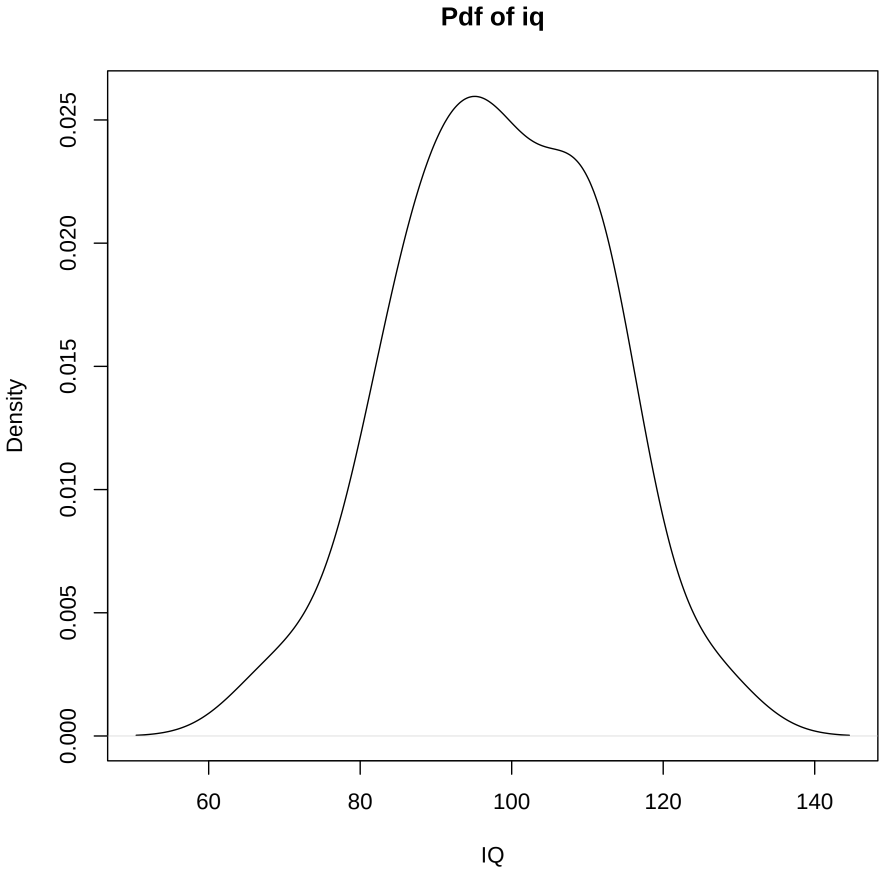
A preliminary inspection of the data revealed an apparently non-normal distribution of IQ. Upon plotting the data (Figure 1) against the expected values for a normal distribution it revealed that our sample data probably includes more extreme values than would be expected if it were truly normally distributed. The data may possibly be 'uniformly' distributed. However a plot of the density of the data (Figure 2) suggests the data is in

Figure 1: QQNorm Plot



fact approximately normally distributed, with very sharp slopes on both sides, 1 clear peak and 1 smaller one. Both tails are quite narrow, therefore we wouldn't actually expect many extreme values for IQ to appear, whether extremely high or extremely low. We can thus assess our ability to infer information about the school population from our sample. The Central Limit Theorem states the mean of the sampling distribution of the sample means is same as the population mean $\mu$, and that the standard error is equal to the population variance, or the population standard deviation $\sigma$ divided by the square root of the sample size n. Therefore, we don't need to bootstrap - rather we can infer the parameters. there is some doubt about the reliability of the Central Limit Theorem method (https://www.scirp.org/journal/paperinformation.aspx?paperid=76758), but for our purposes here i.e. a fairly normal distribution, the error should not be too great. Because the population is assumed normal, even though the sample size is ¡30, the theorem should still hold true (source: https://bit.ly/3CpQGyO). We can thus proceed to constructing our confidence intervals using this assumed normal distribution.

Figure 2: Density Plot



Because n is less than 30, it is usually best to use a t-distribution for such a small sample size. Constructing intervals as I did below for a t-distribution gave a slightly different interval.

As the p-value ($\alpha$) is less than 0.05, we find sufficient evidence in the t-test to reject the null hypothesis that the true population mean ($\mu$) equals 0, therefore the confidence interval is applicable and we would expect 90 per cent of sample means to fall within the two values of 93.95993 and 102.9201.

```
CI_lower <- qnorm(0.05,
mean = mean(iqdata$iq),
sd = (sd(iqdata$iq)/sqrt(length(iqdata$iq))) # standard error
# or population variance equation = denominator
)

CI_upper <- qnorm(0.95,
mean = mean(iqdata$iq),
sd = (sd(iqdata$iq)/sqrt(length(iqdata$iq)))
)
# sd of the sample here is the best estimate of sigma (CLT)

matrix(c(CI_lower, CI_upper), ncol = 2,
dimnames = list("",c("Lower", "Upper")))
```

Figure 3: Construction of 90 per cent Confidence Interval (normal)

```
   Lower     Upper
   94.13283  102.7472

    # so 90% of the samples means of our data would be expected to lie between
    # the above values of \bar_{x} i.e. between 94.13283 and 102.7472
```

Figure 4: Matrix of 90 per cent Confidence Interval (normal)

```
# manually
se <- sd(iqdata$iq)/sqrt(length(iqdata$iq))  # making standard error vector

t_score <- qt(.05, df = length(iqdata$iq)-1, lower.tail = FALSE)
# qt(x,df)  -inverse probability cumulative density
# lower.tail = FALSE - assumes importance of the area to the right of the
# lower tail -> one-sided
CI_lower_t <- mean(iqdata$iq) - (se * t_score)
CI_upper_t <- mean(iqdata$iq) + (se * t_score)
# formula: http://statisticslectures.com/topics/ciindependentsamplest/

# call
CI_lower_t # 93.95993
CI_upper_t # 102.9201

# Checking manual work with R’s t.test function
t.test(iqdata$iq, conf.level = 0.9,
alternative = "two.sided"
)
# returns: CI_Low -> 93.95993 CI_Upp -> 102.92007, p value -> 2.2e-16
# i.e. \alpha < 0.05 ergo we reject H0 that the true population mean is
# equal to 0.
# the ’more correct’ confidence interval is the one using the t-distribution
```

Figure 5: Construction of 90 per cent Confidence Interval (t-distribution)

# Q1: Answer 2

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 $\alpha = 0.05$.

## Answer

The null hypothesis for this test is that the differences of the two means (the sample mean and the national average) is less than or equal to 0, i.e. is not greather than 0 - $H_0:\ \bar{x} - 100 \leq 0;\ H_A:\ \bar{x} - 100 > 0$ OR $H_0:\ \bar{x} \leq 1000;\ H_A:\ \bar{x} > 100$ (Figure 6)

```
iqdata %>%
select(iq) %>%
t.test(mu=100, alternative = "greater", var.equal = FALSE)
# H0: x bar </= 100, HA: x bar > 100
# also cannot assume that the variance of the 2 samples are equal and
# cannot test - do not have the national sample
```

Figure 6: t-Test

The resulting p value is 0.7215, for which $\alpha > 0.05$, therefore we fail to reject H0 that $\mu \geq 100$, and don't observe enough evidence that the teacher's students have a higher than national average IQ.

# 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..."

- 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)?

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

# Answers

- 1. As we can see from the distribution of the points in Figure 7, and the positively sloped line through the data, there seems to be a quite loose correlation between income per capita and housing assistance expenditure per capita, with quite a large standard error buffer. We might cautiously say that an increase in per capita income is correlated with a slight increase in per capita housing assistance expenditure. Higher income and a high proportion of urban residency appear as though they might be somewhat correlated, as we can see from the size and position of the points on the scatterplot, although more analysis would doubtless be needed. Light coloured points, i.e. those indicating a high proportion of financial insecurity in the state seem more prominent towards the top of the scatter plot, suggesting potentially that as financial insecurity increases, so too does state expenditure on housing assistance, a result we might expect. There are also virtually very few small, light points, suggesting that the proportion of urban residents and the proportion of those who are financially insecure may be somewhat positively correlated. In this way, we can take preliminary leads on our data set from a very simple and brief glimpse at a scatterplot of this type.

- 2. Figure 8 below shows that the mean of the Western Region appears to be the highest average per capita expenditure on housing assistance. To determine whether this difference from the other means is statistically significant, we would need to perform a null hypothesis test.

- 3. As we saw from Figure 7, Expenditure and Income have a loose, inconsistent positive slope, suggesting slight correlation, with a large buffer area for standard error. What this, upon further analysis, could suggest is that as the average income of the state increases, so does the expenditure of the state on housing assistance. Whether this correlation is meaningful would have to be determined by other statistical tests. When Region is added to the plot, we observe a certain amount of grouping - most notably that states in the South are predominantly gathered in the lower left quadrant of the plot, giving more of a correlated look to the entire plot. Without the inclusion of the Southern states, the remainder of the data may appear essentially randomly arranged, suggesting that while correlation might appear in the overall plot, this is mostly thanks to the suggested correlation in the Southern states. This particular and unique correlation would be worth investigating alone, but cannot necessarily be generalised to the rest of the data set.

Figure 7: Housing Assistance Expenditure against Income, by Financial Insecurity and Proportion of Urban Residency

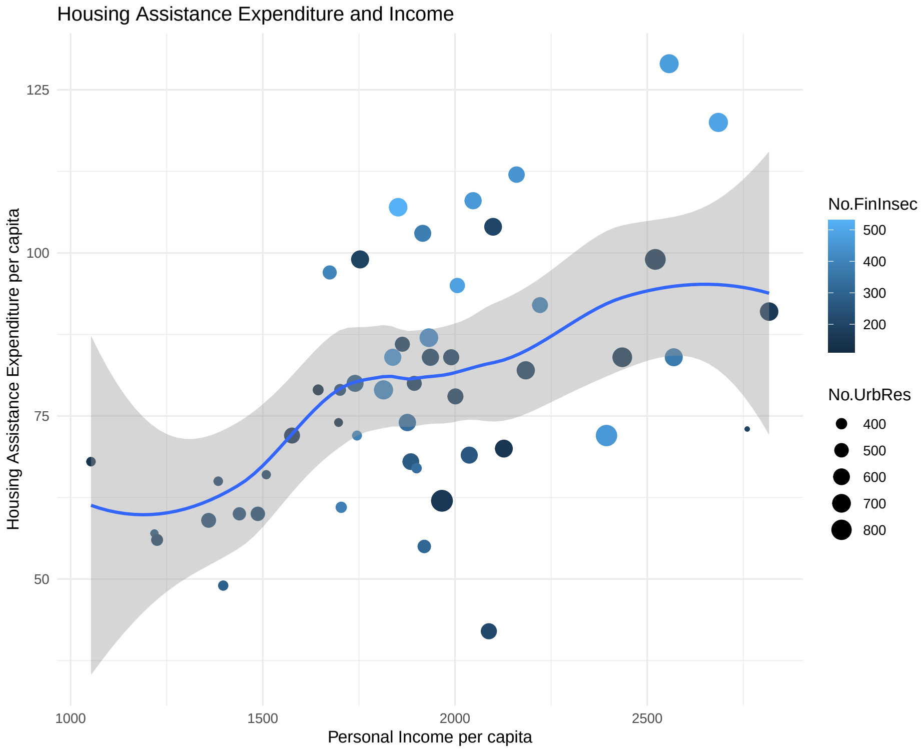


Figure 8: Housing Assistance Expenditure against Region

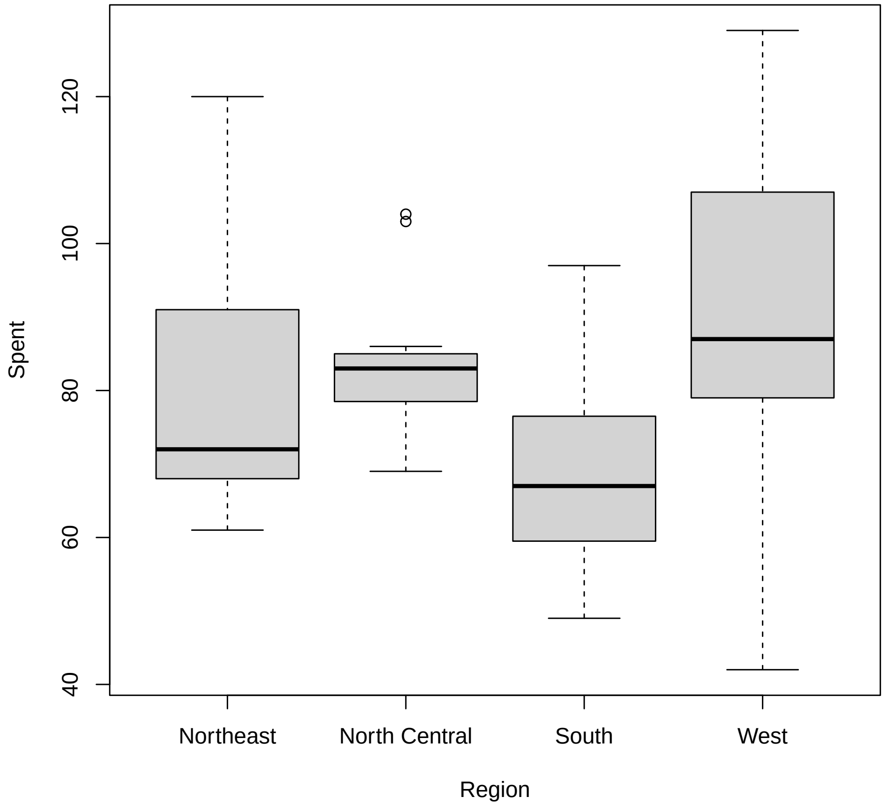


Figure 9: Housing Assistance Expenditure against Income

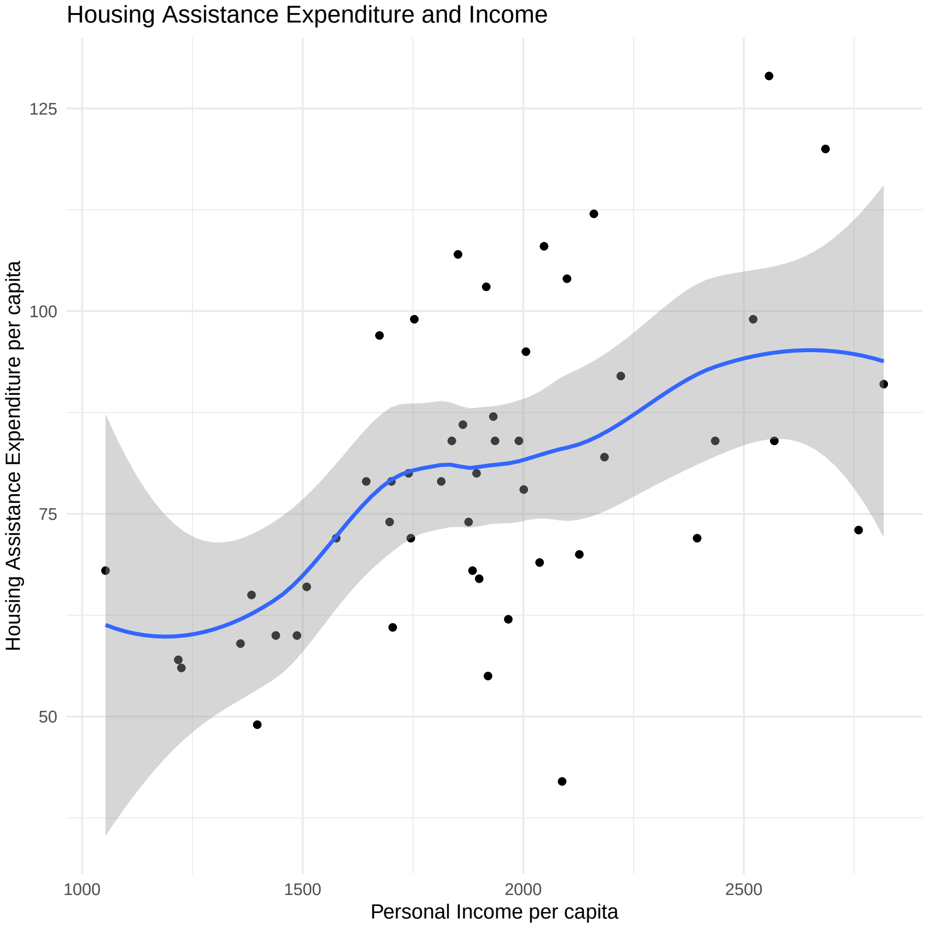
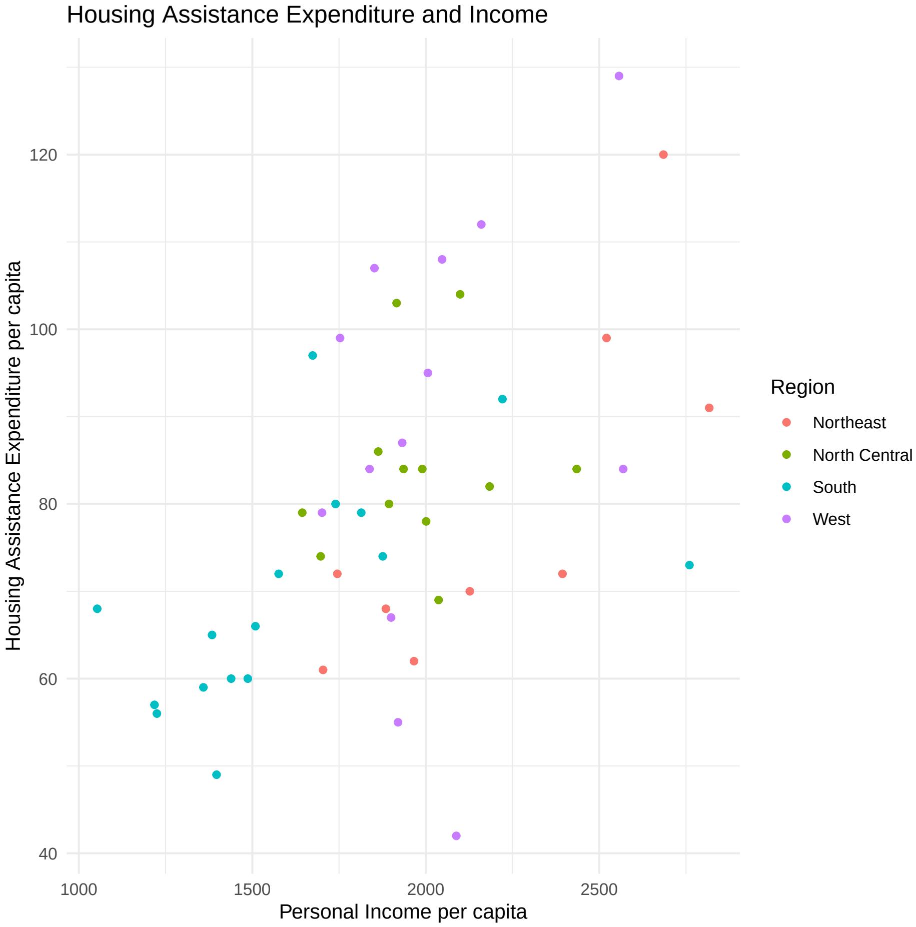


Figure 10: Housing Assistance Expenditure against Income by Region


Housing Assistance Expenditure and Income
Housing Assistance Expenditure per capita
Personal Income per capita
Region
Northeast
North Central
South
West