Problem Set 1 Answers

Doireanna Craven

Due: October 1, 2021

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.

   Using the code below I found the mean and standard deviation of the sample, and then calcuated the estimated standard error. We know that the population distribution of IQ scores will be normal. Since the sample size is small, we can assume the distribution of sample means will be a t-distribution, with n-1 = 24 degress of freedom. This will give us a wider margin of error compared to a using normal curve, which compensates for the uncertainty due to the small sample size. The code returns the appropriate t-score, which tells us that 90% of sample means of size 25 will lie within 1.71 standard errors of the sample mean. (I checked this against stats table also for my own understanding.) This gives a 90% confidence interval for the average IQ in the school of 94-103.

```
n <- length(y)                                  # Sample size
mean_y <- mean(y)                               # Mean of sample
sd_y <- sd(y)                                   # Standard dev of sample
se_y <- (sd_y/sqrt(n))                          # Estimated SE of sample
t90 <- qt(0.05, 24, lower.tail = FALSE)         # t-0.05 =  1.711 from tables
lower_90 <- mean_y - (t90*se_y)                 # 1.711 se's below mean
upper_90 <- mean_y + (t90*se_y)                 # 1.711 se's above mean
CI90_for_students <- c(lower_90, upper_90)
```

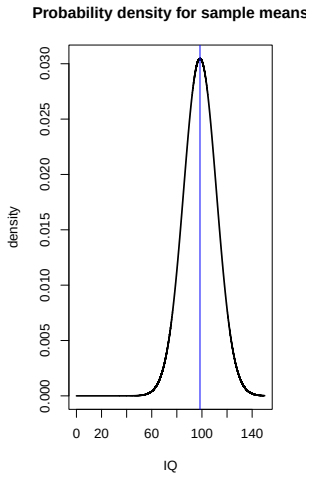
R has a built-in fuction that will generate a 95% confidence interval. Since it's such short code (below) I have run that also as a sense check, and I can confirm that my 90% CI fits within in.

```
t.test(y)
```

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.

Before we conduct a hypothesis test (see next page), we can say straight away from the 90% CI interval that this sample has not provided evidence that the average IQ score in the school is above the national average. I have plotted the probability density of the sample means (code below), including a blue vertical line indicating the sample mean. We can see that the sample means would be spread both above and below 100, so we would not be surprised either way.

```
IQ <- seq(0, 150, by=0.001)                          # Input vector
plot(IQ, dnorm(x=x.range, mean=mean_y, sd=sd_y),  # I have used dnorm
     type="l",                                       # Specify line type
     main="Probability density for sample means", # Title
     ylab="density",                                 # Label y-axis
     lwd=2,                                          # Specify type of line
     xaxt="n")
axis(1, at=seq(0,150,by=20), labels=seq(0,150,by=20)) # x-axis up in 20's
abline(v=mean_y, col="blue")
# Add a blue line at the sample mean
```



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

Null hypothesis: Average IQ in the school is less than or equal to 100; alternative hypothesis: average IQ in the school is greater than 100. The code below calculates the probability that the average IQ in the school is less than or equal to 100. The test statistic is the standardised variance from the sample mean to 100. The test assumes a t-distribution with 24 degrees of freedom.The p-value returned is 0.722. Since the p-value is larger than 0.05 we have no evidence to reject the null hypothesis, as expected from the discussion above. Means of sample size 25 from the school would be expected to be at or below the national average with a probability of 72%. A 72:28 split is not going to result in surprise either way.

```
test_stat <- (mean_y - 100)/se_y
P_value <- pt(abs(test_stat), 24, lower.tail = TRUE)
```

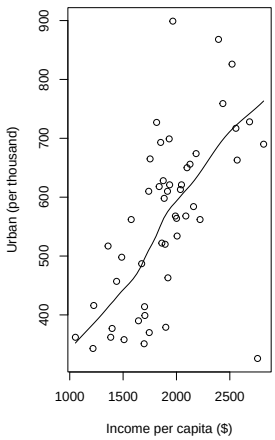
Question 2 (50 points): Political Economy

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

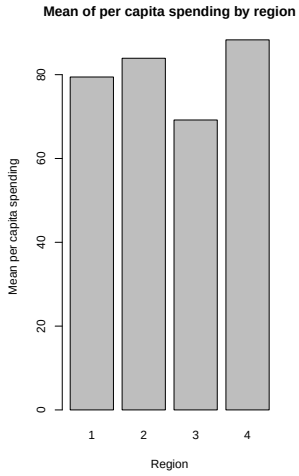
- Please plot the relationships among *Y*, *X1*, *X2*, and *X3*? What are the correlations among them (you just need to describe the graph and the relationships among them)?

Y is moderately positively correlated to *X1*, with a correlation coefficient R = 0.53, and is weakly positively correlated to *X2* and *X3*, with R = 0.45 and R = 0.46 respectively.
X1 is weakly positively correlated to *X2* with R = 0.21, and moderately correlated to *X3* with R = 0.6.
X2 and *X3* are moderately positively correlated with R = 0.53.
The strongest correlation was seen between per capita income and numbers in urban residences. I have included code below used to plot a scatter of their correlation, and to determine the correlation coefficient.

```
scatter.smooth(expenditure$X1, expenditure$X3, ylab="Urban (per thousand)",
    xlab="Income per capita")
cor(expenditure$X1, expenditure$X3)
```



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



Region 4 has the highest mean per capital expenditure. This data was compiled by creating a dataframe of the means of Y, and the regions 1-4, using this code:

```
Y_means_by_region <- aggregate(expenditure$Y, by = list(expenditure$
    Region), FUN = mean)
barplot(Y_means_by_region$x,
        main = "Mean of per capita spending by region",
        names.arg = Y_means_by_region$Group.1,
        ylab = "Mean per capita spending",
        xlab = "Region")
```

- 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.