Problem Set 1

Imelda Finn, 22334657

Due: October 3, 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:

```
iqData <- 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.

```
# calculate sample statistics
    # capture the number of observations
    n <- length(iqData)

    # calculate mean
    iqSum <- sum(iqData)           # sum of IQ scores
    iqMean <-iqSum / n             # mean IQ score for sample

    # calculate variance and standard deviation
    iqVar <- sum((iqData - iqMean)^2)/(n-1)
    iqSD <-sqrt(iqVar)
```

The code for the t-test at 90% is:

```
t.val <- qt(alphaVal/2, df = n-1, lower.tail = FALSE)

CI_lower <- iqMean - t.val * iqse
CI_upper <- iqMean + t.val * iqse
```

The result was:

Our Confidence interval for the IQ of the students in the sample is:

$93.96 <$ mean IQ < 102.92

with a confidence level of 90%. [1]

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 $\alpha = 0.05$.

   (a) The number of observations is 25, which isn't ideal for t-test statistics (we would prefer at least 30 observations).

   (b) H_0: the average iq score in the sample is less than the population average ie $\mu_O \leq \mu$

   (c) Hα: the average iq score in the sample is less than or equal to the population average ie $\mu_O > \mu$

   (d) Calculate test statistic
$$TS = \frac{\bar{Y} - \mu_0}{\sigma_{\bar{Y}}}$$

   (e) Calculate p-value
$$p = Pr(Z \leq -|\frac{\bar{Y} - \mu_O}{\sigma_{\bar{Y}}}|)$$

   (f) if p $\leq \alpha = 0.05$, we reject the null hypothesis

Code

```
# Test our hypothesis
   alphaVal <- 0.05
   popMean <- 100
   #get test statistic
   testStatistic <- (iqMean - popMean) / iqse
```

[1] A Z-test gave a 90% CI of $94.13 < \mu < 102.75$.

```
pValue <- pnorm(-abs(testStatistic))

# calculate t-test p-value
t_pValue <- pt(abs(testStatistic), df = n-1, lower.tail = FALSE)
```

Results p-value for normal distribution is 0.276

t	df	p-value
-0.5957439	24	0.2784617

The p-value is greater than 5%, so we cannot reject the null hypothesis. The data does not support the suggestion that the school IQ scores are greater than the population average.

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)?

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

- 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.

Appendix - R code

```r
#####################################################################
# Imelda Finn, 22334657
# POP77003 - Stats I
# clear global .envir, load libraries, set wd
#####################################################################

# remove objects
rm(list=ls())

# detach all libraries
detachAllPackages <- function() {
  basic.packages <- c("package:stats", "package:graphics", "package:grDevices", "package:utils", "package:datasets", "package:methods", "package:base")
  package.list <- search()[ifelse(unlist(gregexpr("package:", search()))==1, TRUE, FALSE)]
  package.list <- setdiff(package.list, basic.packages)
  if (length(package.list)>0)  for (package in package.list) detach(package, character.only=TRUE)
}
detachAllPackages()

# load libraries
pkgTest <- function(pkg){
  new.pkg <- pkg[!(pkg %in% installed.packages()[,  "Package"])]
  if (length(new.pkg))
    install.packages(new.pkg,  dependencies = TRUE)
  sapply(pkg,  require,  character.only = TRUE)
}

# load necessary packages
lapply(c("ggplot2", "stargazer", "tidyverse", "stringr"),  pkgTest)

# set working directory to current parent folder
setwd(dirname(rstudioapi::getActiveDocumentContext()$path))

######################
# Problem 1
######################

# load data as vector  - in .tex file - update if move from 38
iqData <- 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)

## Save our data to a .csv file in the data directory
write.csv(iqData,
          file = "Data/iq.csv",
          row.names = FALSE)

# Explore data
summary(iqData)
```

```r
str(iqData)
head(iqData)

# look at sampling from sample
meanIQ <- vector("double", length = 1000)
for (i in 1:1000) {
  meanIQ[i] <- mean(sample(iqData, 25, replace=TRUE))
}
summary(meanIQ)
boxplot(meanIQ, iqData, xlab=c("averaged vs original sample"))


# Visually inspect the data
hist(iqData, breaks = 10, main = "Histogram of IQ", xlab = "IQ")

plot(density(iqData), main = "PDF of IQ", xlab = "IQ")

# Use a QQ plot to determine if our IQ variable is normally distributed
qqnorm(iqData)
qqline(iqData, distribution = qnorm)
# Sample values fall away from normal line at upper end

##---------------------------------------------------------------------------
# calculate sample statistics
# capture the number of observations
n <- length(iqData)

# calculate mean
iqSum <- sum(iqData)            # sum of IQ scores
iqMean <-iqSum / n              # mean IQ score for sample

# calculate variance and standard deviation
iqVar <- sum((iqData - iqMean)^2)/(n-1)
iqSD <-sqrt(iqVar)

iqse <- iqSD / sqrt(n)          # standard error of sample

##---------------------------------------------------------------------------
## Confidence Intervals
# Calculate 90 percent confidence intervals using normal distribution
# assuming  iqMean ~ N(mu, iqse)
alphaVal = 0.1
CI_lower <- qnorm(alphaVal/2, mean = iqMean, sd =  iqse)

CI_upper <- qnorm(1-alphaVal/2, mean = iqMean, sd = iqse)

# output
cat(str_glue("{(1-alphaVal)*100}% Confidence Intervals, two-sided z-test"))
matrix(c(CI_lower, CI_upper), ncol = 2,
       dimnames = list("",c("Lower", "Upper")))

```

```r
# Calculate 90 percent confidence intervals using t-test distribution
# degrees of freedom = n-1 = 24 - should be >30
t.val <- qt(alphaVal/2, df = n-1, lower.tail = FALSE)

CI_lower <- iqMean - t.val * iqse
CI_upper <- iqMean + t.val * iqse

# calculate using t-test
cat(str_glue("{(1-alphaVal)*100}% Confidence Intervals, two-sided t-test"))
matrix(c(CI_lower, CI_upper), ncol = 2,
       dimnames = list("",c("Lower", "Upper")))

# t-test results in (slightly) wider confidence interval

# Check our working
#t.test(iqData, conf.level = 1-alphaVal, alternative = "two.sided")

cat("Our Confidence interval for the IQ of the students in the sample is: ")
cat(str_glue("  {round(CI_lower,2)} < mean IQ < {round(CI_upper,2)} "))
cat(str_glue("with a confidence level of {(1-alphaVal)*100}%"))


##------------------------------------------------------------------------------
## Hypothesis Testing
# Wrangling our data
class(iqData) # What class of vector is our IQ variable? - numeric

# Hypothesis test:
# H0 : average IQ of students in school is less than or equal to  national average
# Ha : average IQ of students in school is greater than national average

# alpha = 0.05, 1-tail test, single population

# don't have variance of population, only have mean to compare against

# Test our hypothesis
alphaVal <- 0.05
popMean <- 100
#get test statistic
testStatistic <- (iqMean - popMean) / iqse

pValue <- pnorm(-abs(testStatistic))

# calculate t-test p-value
t_pValue <- pt(abs(testStatistic), df = n-1, lower.tail = FALSE)

matrix(c(testStatistic, n-1, t_pValue  ), ncol = 3,
       dimnames = list("",c("t", "df", "p-value")))
cat(str_glue("p-value for normal distribution is {round(pValue,3)}"))

```

```r

t.test( iqData  ,
       mu = 100, # population mean
       var.equal = TRUE, # The default is FALSE - don't have var for popn
       alternative = "less", # H0: sample mean > population mean
       conf.level = .95) #


# How do we interpret the output?
# for confidence level of 95%, we cannot reject the hypothesis (p-value >
    alpha)


#####################
# Problem 2
#####################
# function to save output to a file that you can read in later to your docs
output_stargazer <- function(outputFile, appendVal=TRUE, ...) {
  output <- capture.output(stargazer(...))
  cat(paste(output, collapse = "\n"), "\n", file=outputFile, append=appendVal)
}

# read in expenditure data
expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/StatsI_
    Fall2022/main/datasets/expenditure.txt", header=T)
#expenditure <- read.table("../../datasets/expenditure.txt", header=T)

# 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
data_headers <- c("State", "$ExpenditurePC", "$IncomePC", "FInsecureResidents"
    ,
               "UrbanResidents", "Region")
regions <- c("Northeast","North Central", "South", "West")
names(expenditure)
#colnames(expenditure) <- data_headers

# Inspect the data
head(expenditure)
str(expenditure)
summary(expenditure)

#investigate spending on Housing assistance

# Visualise
hist(expenditure$Y,
     #breaks = 12,
     main = "Histogram of spending on HA ",
```

```r
      xlab = "$, per capita"
)

plot(density(expenditure$Y),
      main = "PDF of spending on HA ",
      xlab = "$, per capita"
)

pairs(~Y + X1 + X2 + X3, expenditure)

qqnorm(expenditure$Y)
qqline(expenditure$Y,
        distribution = qnorm)


# create plots of Y and Xn
onefile <- TRUE
#pdf( file = if(onefile) "expenditure_plots.pdf" else "expenditure_plots%03d.
    pdf")
#pdf("plot_example.pdf" )

ggplot(expenditure) +
  geom_point(aes( Y, X1), colour = "blue") +
  geom_smooth(aes( Y, X1))

ggplot(expenditure) +
  geom_point(aes( Y, X2), colour = "blue") +
  geom_smooth(aes( Y, X2))

ggplot(expenditure) +
  geom_point(aes( Y, X3), colour = "blue", ) +
  geom_smooth(aes( Y, X3), colour = "red")


ggplot(expenditure) +
  geom_point(aes( STATE, Y), colour = "green") +
  geom_point(aes( STATE, X1), colour = "blue")

ggplot(expenditure) +
  geom_point(aes( STATE, X1/Y), colour = "green")


ggplot(expenditure) +
  geom_point(aes( Y, X1), colour = "blue")

ggplot(expenditure) +
  geom_point(aes( Y, X2), colour = "green")

ggplot(expenditure) +
  geom_point(aes( Y, X3)) +
  geom_smooth(aes( Y, X3))
```

```
#main = "Income per capita vs spending on HA "
ggplot(expenditure) +
  geom_point(aes( Y, X1), colour = "blue") +
  geom_smooth(aes( Y, X1))

#dev.off() # close pdf file


# regional expenditure on housing assistance

ggplot(expenditure) +
  geom_point(aes( Y, X2, colour = factor(Region))) +
  geom_smooth(aes( Y, X2))
#logarithmic scale

#factor(expenditure$Region) <- regions
ggplot(expenditure) +
  geom_point(aes(  Region, Y,colours = factor(Region)))
# more spread in r4, least in r2

# can see eg that no crossover in interquartile ranges
boxplot(expenditure$Y ~ expenditure$Region, # here we use formula notation to group
        main = "Boxplot of per capita spending on HA by Region",
        names=regions,
        ylab = "$",
        xlab = "")


regional_mean_table <-expenditure %>% # Tidyverse method for grouping
  group_by(Region) %>%
  summarise(mean = mean(Y))

regional_mean_table <- cbind(regional_mean_table, regions)

output_stargazer("regional_means.tex", appendVal = FALSE, regional_mean_table[, -1])   # file fragment

ggplot(re_means) +
  geom_point(aes(regions, mean, colour = regions), size=3)

# West region has highest per capita mean expenditure on housing assistance

#---------------------------------------------------------------------------
# look at income vs expenditure on HA, by region
#factor(expenditure$Region) <- regions
ggplot(expenditure) +
  geom_point(aes( Y, X1)) +
  geom_smooth(aes(Y, X1))

```

```
ggplot(expenditure) +
  geom_point(aes( Y, X1, colour= regions[Region], shape= regions[Region]))

ggplot(expenditure) +
  geom_point(aes( Y, X1, colour= regions[Region], shape= regions[Region])) +
  geom_smooth(aes(Y, X1))

ggplot(expenditure) +
  geom_point(aes( Y, X1, colour= factor(Region), shape= factor(Region))) +
  geom_smooth(aes( Y, X1, colour = factor(Region)))




ggplot(data = expenditure) +
  geom_point(mapping = aes(x = Y, y = X1)) +
  facet_wrap(~ Region, nrow = 2)


##  try - todo
mat <- as.matrix(with(expenditure, table(Y, Region)))


barplot(height = mat,
        beside = TRUE,
        legend.text = TRUE,
        args.legend = list(x = "topleft",
                           cex = 0.4,
                           box.col = "white"))


# run an example regression, to show how to save table


lm(Y~X1, data=expenditure)
lm(Y~X2, data=expenditure)
lm(Y~X3, data=expenditure)

# execute function and check ls() to make sure it worked
ls()
```