Final Exam (Practice Questions) Template

Question 1: Use the built-in data set USArrests to answer this question.

(a) What is the variable Murder being measured in the data set?

(b) What type of variable is this?

(c) What is the most appropriate type of graph to visualize the distribution of this variable?

(d) Graph the distribution of the variable (using the type of graph that you identified in part (c)). Your graph should include:

- a main title.
- x-axis title.
- scales on the x and y-axis which fully extend from at least the min value to at least the max value.

(e) Describe the shape of the distribution (that is, symmetric, left-skewed, right-skewed).

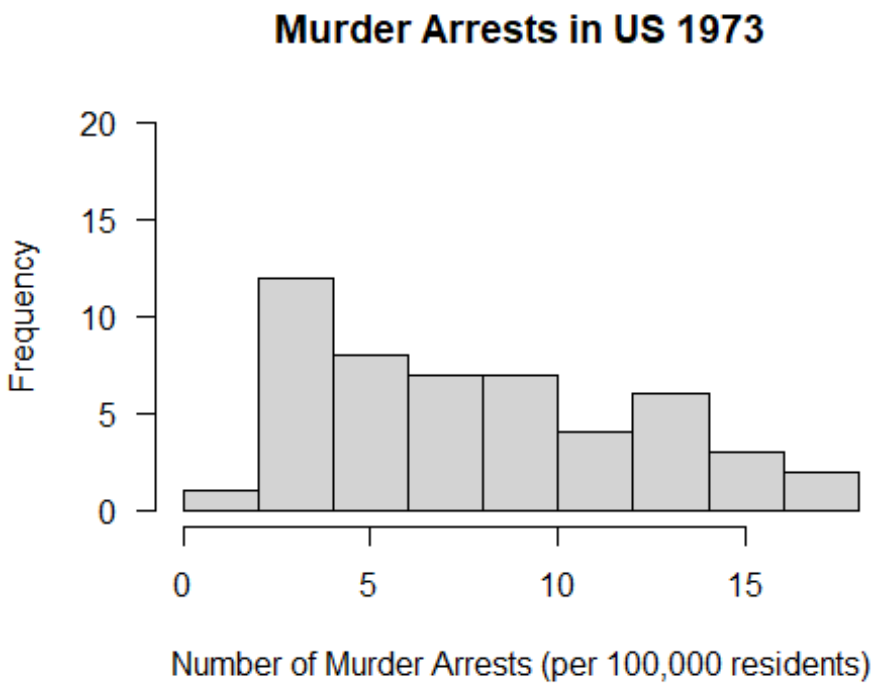
(f) What is an appropriate statistic to measure the center of the distribution? Why?

(g) Compute the observed value of this statistic.

(h) What is an appropriate statistic to measure the spread of the distribution? Why?

(i) Compute the observed value of this statistic.

```
# (a) Answer below:
?USArrests

## starting httpd help server ... done

# The variable being measured is the number of murder arrests per 100000
residents# for each of the 50 US states in 1973.

# (b) Answer below:
# numeric

# (c) Answer below:
# histogram

# (d) Answer below:
hist(USArrests$Murder, main="Murder Arrests in US 1973",
 xlab = "Number of Murder Arrests (per 100,000 residents)",
 ylim=c(0,20),las=1)
```




```
# (e) Answer below:
# This is a right skewed distribution.

# (f) Answer below:
# median

# (g) Answer below:
median(USArrests$Murder)

## [1] 7.25

# (h) Answer below:
# Since we used median to measure the center, the quartiles Q1 and
# Q3 are appropriate for measuring spread.

# (i) Answer below:
quantile(USArrests$Murder,c(0.25,0.75))

##    25%    75%
##  4.075 11.250
```

Question 2: Suppose you take a random sample of size 100 of a normally distributed variable Z. The sample mean is 126 and the sample standard deviation is 18.

(a) Between what range of values should approximately 70% of the observations lie?

(b) Between what range of values should approximately 80% of the observations lie?

(c) What is the estimated standard error for the sample mean?

(d) What is the critical value for an 86% confidence interval for the mean?

(e) Determine an 86% confidence interval for the mean.

```
# (a) Answer below:
qnorm(c(0.15, 0.85), mean = 126, sd = 18)

## [1] 107.3442 144.6558

# (b) Answer below:
qnorm(c(0.1, 0.9), mean = 126, sd = 18)

## [1] 102.9321 149.0679

# (c) Answer below:
ese = 18/sqrt(100)
ese

## [1] 1.8

# (d) Answer below:
cv = qnorm(1 - (1-.86)/2)
cv

## [1] 1.475791

# (e) Answer below:
c(126 - cv*ese, 126 + cv*ese)

## [1] 123.3436 128.6564
```

Question 3: Consider the gapminder data set that we worked with in class. We will need this data set to answer this question.

(a) Either load the data set into R by typing in library(gapminder) or download the gapminder.csv file from Brightspace and read the data into R, saving it as gapminder.

(b) Suppose you are looking to explore the relationship between the population and Life Expectancy. What type of graph should you use to visualize this relationship?

(c) Create a graph which visualizes the relationship between these two variables. Put Life Expectancy is on the x-axis. This graph does not need any titles.

(d) What is wrong with the graph?

(e) Create a vector which contains the populations recorded for Italy in the data set. Call this vector italy_pop.

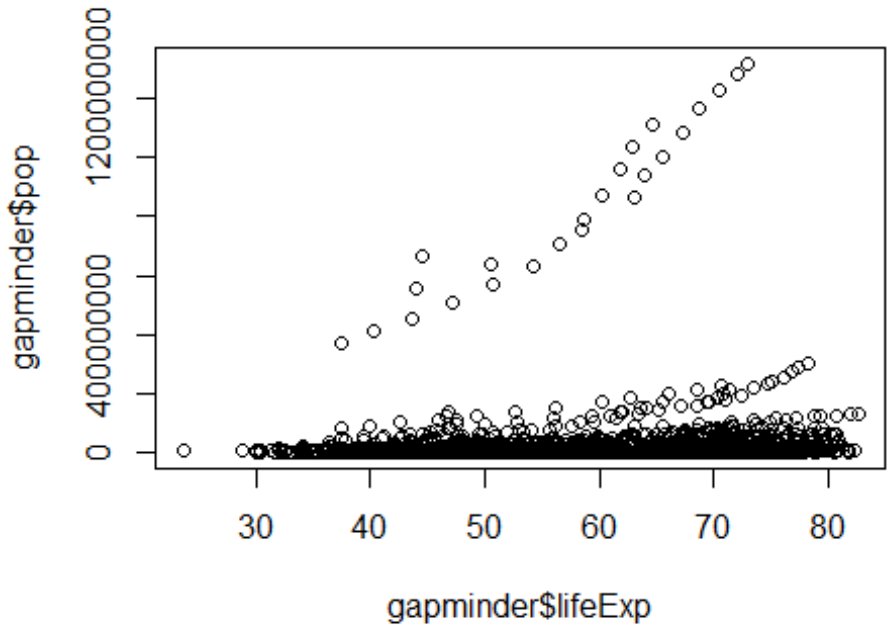
(f) Create a vector which contains the Life Expectancy for Italy in the data set. Call this vector italy_lifexp.

(g) Create a graph which visualizes the relationship between the population size (on y-axis) and Life Expectancy (on x-axis) for Italy. Your graph should include:

- a main title.
- a title for both the x-axis and the y-axis
- the scale should not be in scientific notation.

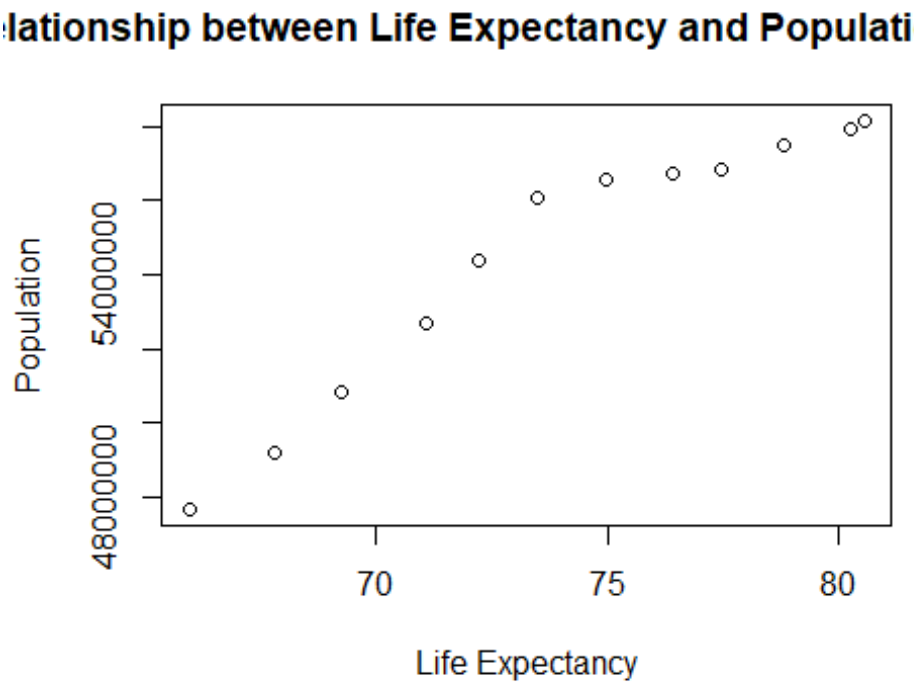
(h) Describe the direction and form of the relationship.

```
# (a) Answer below:
library(gapminder)

# (b) Answer below:
# You should use a scatterplot to try and visualize the relationship
# between the two variables.

# (c) Answer below:
plot(gapminder$lifeExp,gapminder$pop)
```




```
# (d) Answer below:
# The scatterplot is crowded and hard to read. It also includes several
# points for each country and year so there is a lot of duplication in the
# plot.

# (e) Answer below:
italy_index = which(gapminder$country=="Italy")
italy_pop = gapminder$pop[italy_index]
```

```r
# or you can use filter function to save italy_index
# library(dbplyr)
# italy_index= gapminder %>% filter(gapminder$country=="Italy")

# (f) Answer below:
italy_lifexp = gapminder$lifeExp[italy_index]

# (g) Answer below:
options(scipen=999)
plot(italy_lifexp,italy_pop,
 main="Relationship between Life Expectancy and Population in Italy", xlab =
"Life Expectancy", ylab="Population")
```




```r
# (h) Answer below:
# The direction of the relationship is positive and the form appears
# mostly linear or (possibly) exponential.
```

Question 4: We will again use the data from the built-in data vector USArrests$Murder.

(a) Create a variable n which equals the sample size for the variable.

(b) Bootstrap 10000 sample means and save the bootstrapped means to a vector called mean_Murder.

(c) Plot the sampling distribution of the sample mean (with probability = TRUE) and plot an estimated density curve on the same graph. Your plot should include the following:

- a main title
- a title for the x-axis
- a density curve which is a different colour than your plot.

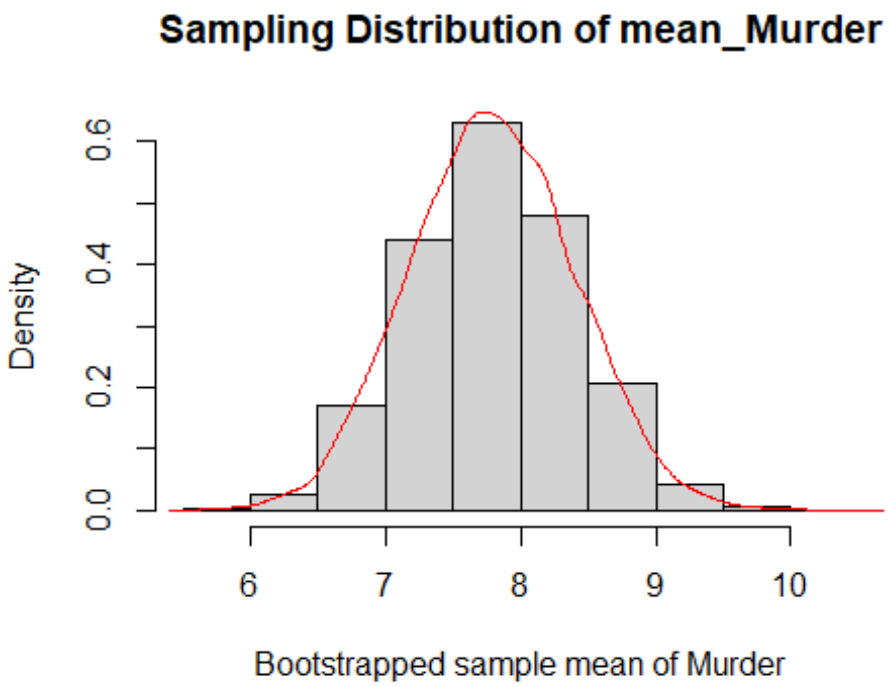
(d) Bootstrap 10000 sample 80th percentiles and save the bootstrapped 80th percentiles to a vector called percentile80_Murder.

(e) Plot the sampling distribution of the sample 80th percentile. Your plot should include the following:

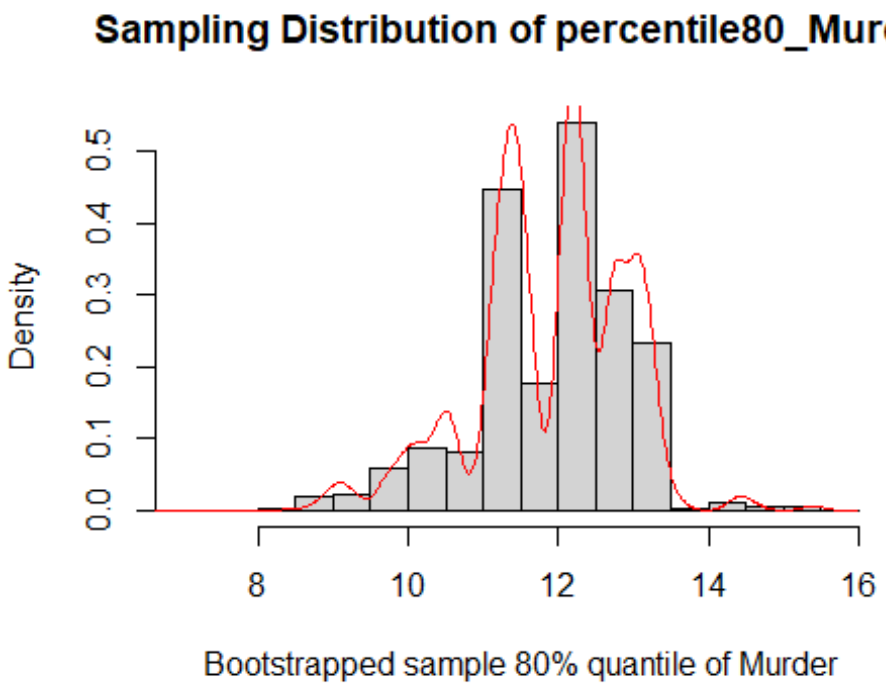
- a main title
- a title for the x-axis

(g) Compute a 96% confidence interval for the 80th percentile.

```
# (a) Answer below:
n = length(USArrests$Murder)
n

## [1] 50

# (b) Answer below:
mean_Murder = numeric()
for (i in 1:10000) {
 mean_Murder[i] = mean(sample(USArrests$Murder, n, replace = TRUE))
}

# (c) Answer below:
hist(mean_Murder, prob = TRUE, main = "Sampling Distribution of mean_Murder",
xlab ="Bootstrapped sample mean of Murder")
lines(density(mean_Murder), col = "red")
```




```r
# (d) Answer below:
percentile80_Murder = numeric()
for (i in 1:10000) {
 percentile80_Murder[i] = quantile(sample(USArrests$Murder, n, replace =
TRUE), .8)}

# (e) Answer below:
hist(percentile80_Murder, prob = TRUE, main = "Sampling Distribution of
percentile80_Murder", xlab = "Bootstrapped sample 80% quantile of Murder")
lines(density(percentile80_Murder), col = "red") # optional
```




```
# (g) Answer below:
quantile(percentile80_Murder, c(0.02, 0.98))

##    2%   98%
##  9.28 13.44
```

Question 5: Consider the realEstate.csv data set available in Brightspace. Read this data set into R and save it as RE.

(a) Create a table called storeCounts which contains the total number of houses in the data set near each number of convenience stores (that is, how many houses in the data set are near 0 stores, near 1 store, etc…).

(b) Print storeCounts.

(c) Create a bar graph which displays the information in the table.

(d) Create a vector called percents which contains the percent of houses in the data set near each number of convenience stores.

(e) Create a pie chart which displays the percents.

```
# (a) Answer below:
RE = read.csv(file.choose())
head(RE)

##   No X1.transaction.date X2.house.age
X3.distance.to.the.nearest.MRT.station
```

```
## 1  1            2012.917         32.0
84.87882
## 2  2            2012.917         19.5
306.59470
## 3  3            2013.583         13.3
561.98450
## 4  4            2013.500         13.3
561.98450
## 5  5            2012.833          5.0
390.56840
## 6  6            2012.667          7.1
2175.03000
##   X4.number.of.convenience.stores X5.latitude X6.longitude
## 1                              10    24.98298     121.5402
## 2                               9    24.98034     121.5395
## 3                               5    24.98746     121.5439
## 4                               5    24.98746     121.5439
## 5                               5    24.97937     121.5425
## 6                               3    24.96305     121.5125
##   Y.house.price.of.unit.area
## 1                       37.9
## 2                       42.2
## 3                       47.3
## 4                       54.8
## 5                       43.1
## 6                       32.1

# Part (a) answer below:
storeCounts = table(RE$X4.number.of.convenience.stores)

# Part (b) answer below:
storeCounts

##
##  0  1  2  3  4  5  6  7  8  9 10
## 67 46 24 46 31 67 37 31 30 25 10

# (c) Answer below:
RE = read.csv(file.choose())
# Part (a) answer below:
storeCounts = table(RE$X4.number.of.convenience.stores)
# Part (b) answer below:
storeCounts

##
##  0  1  2  3  4  5  6  7  8  9 10
## 67 46 24 46 31 67 37 31 30 25 10

# (d) Answer below:
totalHouses = sum(storeCounts)
percents = round(storeCounts/totalHouses*100,2)
```

```
# Print out percents
percents
```

```
##
##     0     1     2     3     4     5     6     7     8     9    10
## 16.18 11.11  5.80 11.11  7.49 16.18  8.94  7.49  7.25  6.04  2.42
```

```
# Check that percents adds up to 1
sum(percents)
```

```
## [1] 100.01
```

```
# (e) Answer below:
# create labels for the pie chart
pieLabels = paste(names(storeCounts),"store",percents,"%")
pie(percents,labels=pieLabels,
main="Number of Convenience stores near houses",
col = rainbow(length(pieLabels)))
```