Midterm 2

Parker DeBruyne - V00837207

21/03/2022

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

(a) Create a single table called hair_eye_totals which summarizes the total number of statistics students with each combination of hair and eye colour.

Note: The built-in data set consists of two tables with this information (one for women and one for men). The answer to part (a) is a single table combining the information from these two tables.

```
head(HairEyeColor)
```

```
## , , Sex = Male
##
##        Eye
## Hair    Brown Blue Hazel Green
##   Black    32   11    10     3
##   Brown    53   50    25    15
##   Red      10   10     7     7
##   Blond     3   30     5     8
##
## , , Sex = Female
##
##        Eye
## Hair    Brown Blue Hazel Green
##   Black    36    9     5     2
##   Brown    66   34    29    14
##   Red      16    7     7     7
##   Blond     4   64     5     8
```

```
HEC = HairEyeColor
#want total students of each combination (either gender)
hair_eye_totals = HEC[,,1]+HEC[,,2]
```

(b) Print out the hair_eye_totals table.

```
hair_eye_totals
```

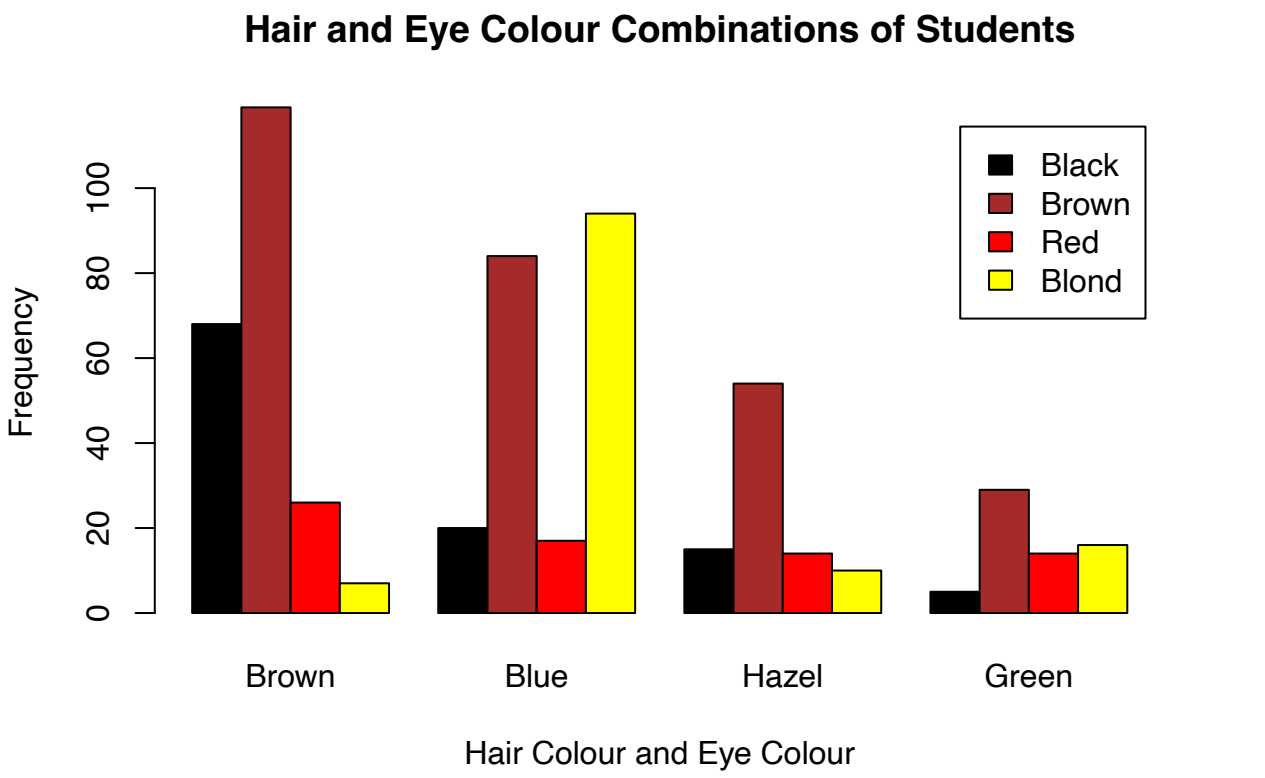
```
##        Eye
## Hair    Brown Blue Hazel Green
##   Black    68   20    15     5
##   Brown   119   84    54    29
##   Red      26   17    14    14
##   Blond     7   94    10    16
```

(c) Create a grouped bar plot which displays the information from the hair_eye_totals table. Your plot should include the following:

- a main title

- titles for the x-axis and y-axis
- colours to help differentiate the bars
- a legend to identify what each colour represents

```
barplot(hair_eye_totals,legend=rownames(hair_eye_totals),
        main="Hair and Eye Colour Combinations of Students",
        col = c("black","brown","red","yellow"),
        xlab ="Hair Colour and Eye Colour", ylab="Frequency",beside=TRUE)
```



(d) Create and print out a vector called percent_eye which contains the percent of statistics students with each eye colour (rounded to 2 decimal places). Show any additional code needed to create this vector.

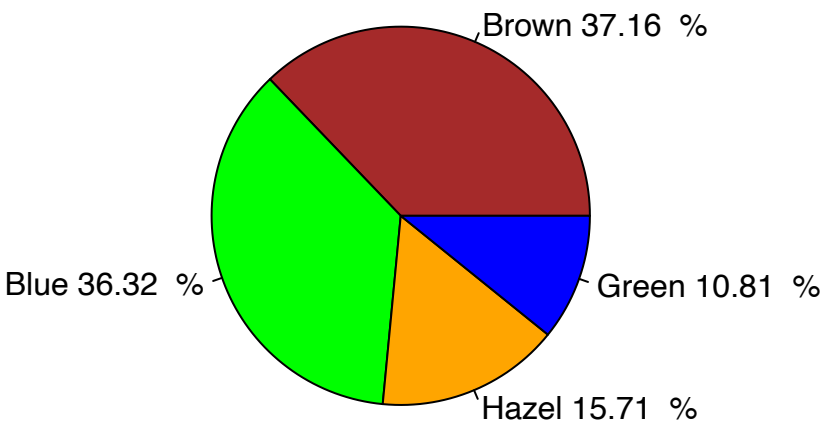
```
totals = colSums(hair_eye_totals)
percent_EC = round((totals/sum(totals))*100,2)
percent_EC
```

```
## Brown  Blue Hazel Green
## 37.16 36.32 15.71 10.81
```

(e) Create a pie chart displaying the information in the percent_eye vector. Your graph should include:

- a main title
- labels for each wedge displaying the eye colour
- a different colour for each eye colour
- the percentages displaying next to each wedge.

```
EC_labels = paste(names(percent_EC),percent_EC," %")
pie(percent_EC,labels=EC_labels,main="Eye Colours of Students ",
    col=c("brown","green","orange","blue"))
```




```
# (a) Answer below:

# (b) Answer below:

# (c) Answer below:

# (d) Answer below:

# (e) Answer below:
```

Question 2: Use the built-in data set USArrests to answer this question.

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

```
?USArrests

#Ans: Number of Murder arrests (per 100,000)
```

(b) What type of variable is this?

```
str(USArrests$Murder)
```

```
##  num [1:50] 13.2 10 8.1 8.8 9 7.9 3.3 5.9 15.4 17.4 ...
```

```
#Ans: Numeric
```

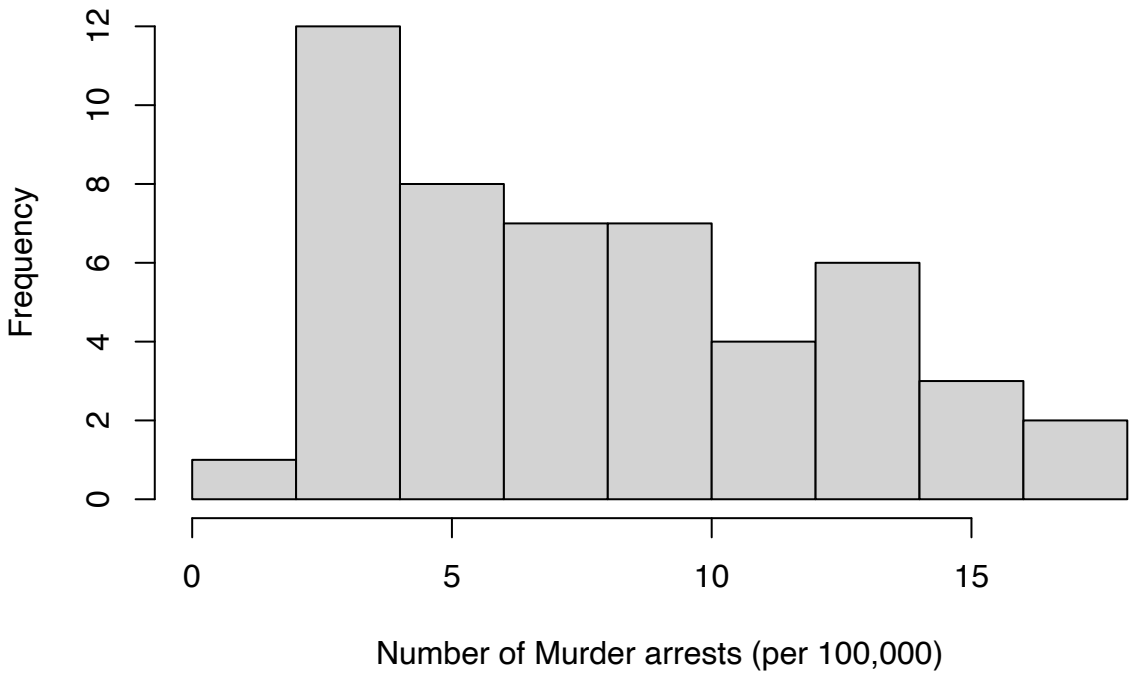
(c) What is the most appropriate type of graph to visualize the distribution of this variable?

```
#Ans: Histogram
```

(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.

```
hist(USArrests$Murder, main="Frequency of Values",
     xlab="Number of Murder arrests (per 100,000)")
```



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

```
#Ans: The distribution is not symmetric and is right skewed.
```

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

```
# Median, because we do not have a normal distribution.
```

(g) Compute the observed value of this statistic.

```
median(USArrests$Murder)
```

```
## [1] 7.25
```

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

```
# Ans: The 1st and 3rd quantiles since we are using the median
#      as the center of our distribution.
```

(i) Compute the observed value of this statistic.

```
quantile(USArrests$Murder,c(0.25,0.75))
```

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

```
# (a) Answer below:

# (b) Answer below:

# (c) Answer below:

# (d) Answer below:
```

```
# (e) Answer below:

# (f) Answer below:

# (g) Answer below:

# (h) Answer below:

# (i) Answer below:
```

Question 3: 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?

```
#68-95-99.7
mu = 126
sig = 18

diff = (100-70)/2
first = diff
last = 100-diff
qnorm(first/100, mean = mu, sd = sig)
```

```
## [1] 107.3442
```

```
qnorm(last/100, mean = mu, sd = sig)
```

```
## [1] 144.6558
```

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

```
mu = 126
sig = 18

diff = (100-88)/2
first = diff
last = 100-diff
qnorm(first/100, mean = mu, sd = sig)
```

```
## [1] 98.01408
```

```
qnorm(last/100, mean = mu, sd = sig)
```

```
## [1] 153.9859
```

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

```
n=100
std_err = sig/sqrt(n)
std_err
```

```
## [1] 1.8
```

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

```
diff_86 = (100-86)/2
interval_86 = c(diff_86, 100-diff_86)
```

```
crit_val_86 = qnorm(interval_86/100, mean=0, sd=1)
print(paste("Critical Value estimate for an 86% confidence interval =", round(crit_val_86[2], 2)))
```

```
## [1] "Critical Value estimate for an 86% confidence interval = 1.48"
```

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

```
confidence_int_86 = c(mu - crit_val_86[2]*std_err, mu + crit_val_86[2]*std_err)

print(paste("Confidence interval for 86% = (",
            round(confidence_int_86[1], 2),
            ",",
            round(confidence_int_86[2], 2),
            ")"))
```

```
## [1] "Confidence interval for 86% = ( 123.34 , 128.66 )"
```

```
# (a) Answer below:

# (b) Answer below:

# (c) Answer below:

# (d) Answer below:

# (e) Answer below:
```

Question 4: 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.

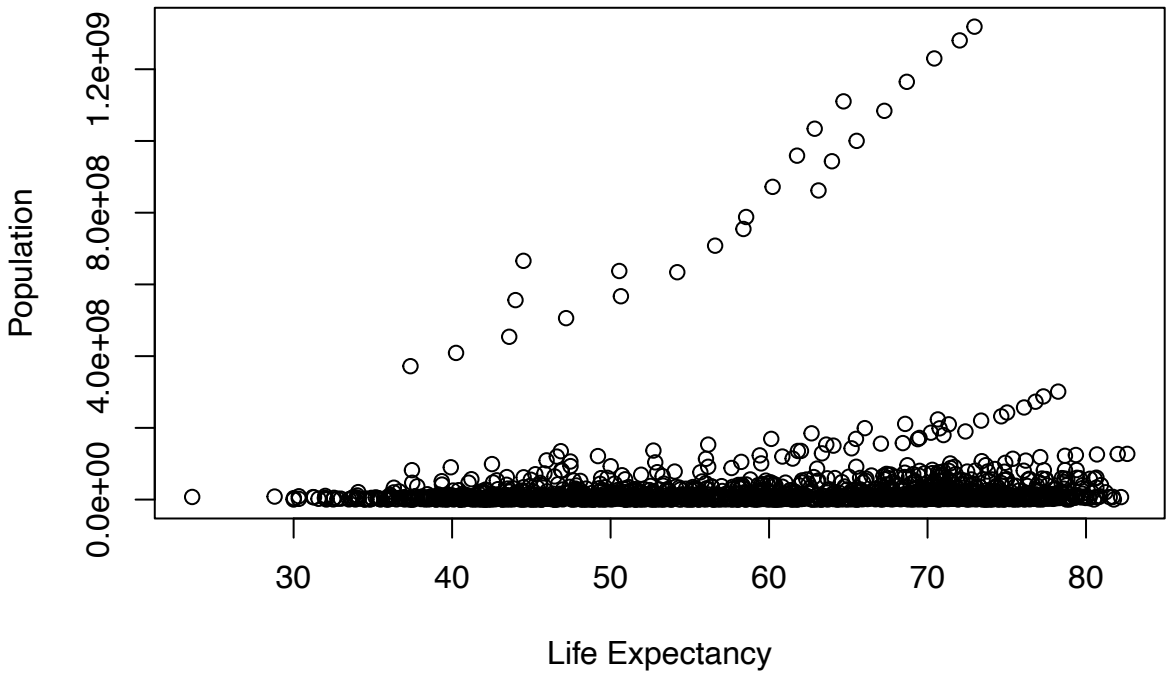
```
library(gapminder)
```

(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?

```
#Ans: Line Graph as it is a time series
```

(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.

```
plot(gapminder$lifeExp,gapminder$pop,
     xlab = "Life Expectancy",
     ylab = "Population")
```



(d) What is wrong with the graph?

```
# Ans: It is clustered near the bottom, with a positive relationship between outliers. This implies
```

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

```
# vitaly_index = which(gapminder$Country == "Italy")
# italy_population = gapminder$Country[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:


# (b) Answer below:


# (c) Answer below:


# (d) Answer below:


# (e) Answer below:
```

```
# (f) Answer below:


# (g) Answer below:


# (h) Answer below:
```

Question 5: 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) What kind of distribution does it look like? Was this what you expected? Explain.

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

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

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

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

```
# (a) Answer below:

# (b) Answer below:

# (c) Answer below:

# (d) Answer below:

# (e) Answer below:
```