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.

(b) Print out the hair_eye_totals table.

(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

(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.

(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.

```
# (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?

(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:

# (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?

(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:

# (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.

(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:


# (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:
```