STAT 140: Homework 3

YOUR NAME HERE

9/25/2019

BOOK EXERCISES (OpenIntro Statistics, Fourth Edition)

2.1) **Mammal life spans.** Data were collected on life spans (in years) and gestation lengths (in days) for 62 mammals. A scatterplot of life span versus length of gestation is shown on pg. 56.

(a) What type of an association is apparent between life span and length of gestation?

(b) What type of association would you expect to see if the axes of the plot were reversed, i.e. if we plotted length of gestation versus life span?

(c) Are life span and length of gestation independent? Explain your reasoning.

2.6) **Sleeping in college.** A recent article in a college newspaper stated that college students get an average of 5.5 hrs of sleep each night. A student who was skeptical about this value decided to conduct a survey by randomly sampling 25 students. On average, the sampled students slept 6.25 hours per night. Identify which value represents the sample mean and which value represents the claimed population mean.

2.7) **Days off at a mining plant.** Workers at a particular mining site receive an average of 35 days paid vacation, which is lower than the national average. The manager of this plant is under pressure from a local union to increase the amount of paid time off. However, he does not want to give more days off to workers because that would be costly. Instead he decides he should fire 10 employees in such a way as to raise the average number of days off reported by his employees. In order to achieve his goal, should he fire employees who have the most number of days off, least number of days off, or those who have about the average number of days off?

2.8) **Medians and IQRs.** For each part, compare distribution (1) and (2) based on their medians and IQRs. You do not need to calculate these statistics; simply state how the medians and IQRs compare. Make sure to explain your reasoning. Refer to the problem in the book for the distributions (pg. 57).

(a)

(b)

(c)

(d)

2.9) **Means and SDs.** For each part, compare distribution (1) and (2) based on their means and standard deviations. You do not need to calculate these statistics; simply state how the means and standard deviations compare. Make sure to explain your reasoning. *Hint:* It may be useful to sketch dot plots of the distributions. You do not need to include these in your answer, however. Refer to the problem in the book for the distributions (pg. 57).

(a)

(b)

(c)

(d)

2.15) **Distributions and appropriate statistics, Part I.** For each of the following, state whether you expect the distribution to be symmetric, right skewed, or left skewed. Also specify whether the mean or median would best represent a typical observation in the data, and whether the variability of observations would be best represented using the standard deviation or IQR. Explain your reasoning.

(a) Number of pets per household.

(b) Distance to work, i.e. number of miles between work and home.

(c) Heights of adult males.

2.18) **Midrange.** The *midrange* of a distribution is defined as the average of the maximum and the minimum of that distribution. Is this statistic robust to outliers and extreme skew? Explain your reasoning. *Hint:* You may want to explore specific examples to build your intuition.

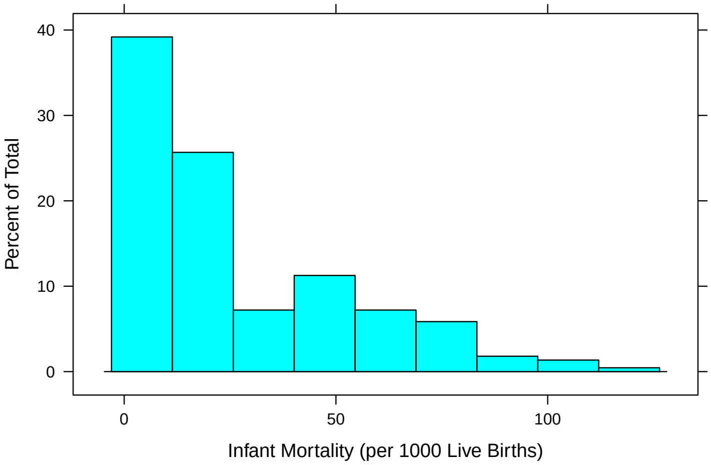
2.22) **View on Immigration.** 910 randomly sampled registered voters from Tampa, FL were asked if they thought workers who have illegally enetered the US should be (i) allowed to keep their jobs and apply for US citizenship, (ii) allowed to keep their jobs as temporary guest workers but not allowed to apply for US citizenship, or (iii) lose their jobs and have to leave the country. The results of the survey by political ideology are shown on pg. 69.

- What percent of these Tampa, FL voters identify themselves as conservatives?
- What percent of these Tampa, FL voters are in favor of the citizenship option?
- What percent of these Tampa, FL voters identify themselves as conservatives and are in favor of the citizenship option?
- What percent of these Tampa, FL voters who identify themselves as conservatives are also in favor of the citizenship option? What percent of moderates share this view? What percent of liberals share this view?
- Do political ideology and views on immigration appear to be independent? Explain your reasoning.

2.28) **Infant mortality.** The infant mortality rate is defined as the number of infant deaths per 1,000 live births. This rate is often used as an indicator of the level of health in a country. The relative frequency histogram below shows the distribution of estimated infant death rates for 224 countries for which such data were available in 2014.

- Estimate Q1, the median, and Q3 from the histogram.
- Would you expect the mean of this data set to be smaller or larger than the median? Explain your reasoning.
- Estimating Q1, the median, and Q3 from a histogram is hard. Use R to find out what these values are exactly for the variable `infMortRate` from the infMortRate dataframe. You should refer to the R Lab for Chapter 2 for a helpful function that will provide all three of these numbers.
- We have studied another plot that summarizes a single numeric variable that is more convenient when we want to identify the Q1, median, and Q3. Use R to create the appropriate plot. You should refer to the R Lab for Chapter 2 for a helpful function to generate this plot.

2.30) **A new statistic.** The statistic $\frac{\bar{x}}{median}$ can be used as a measure of skewness. Suppose we have a distribution where all observations are great than 0, $x_i > 0$. What is the expected shape of the distribution under the following conditions? Explain your reasoning.



(a) $\frac{\bar{x}}{median} = 1$

(b) $\frac{\bar{x}}{median} < 1$

(c) $\frac{\bar{x}}{median} > 1$

Stats scores. The final exam scores of twenty introductory statistics students are stored as a vector called `stats_scores` in the R chunk below. In the chunk below,

- Create a boxplot of the distribution of these scores using R. Are there any apparent outliers?
- Create a histogram of the distribution of these scores using R. What is the shape of this histogram? Do you expect the mean or median to be larger? Why?

```r
## store stats scores in a vector call stats_scores
stats_scores <- c(57,66,69,71,72,73,74,77,78,78,79,79,81,81,82,83,83,88,89,94)
```

R EXERCISES

For the following exercises, you will be exploring a dataset collected on the game Pokemon Go. As noted in the dataset description on OpenIntro, "A key part of Pokémon Go is using evolutions to get stronger Pokemon, and a deeper understanding of evolutions is key to being the greatest Pokemon Go player of all time. This data set covers 75 Pokemon evolutions spread across four species. A wide set of variables are provided, allowing a deeper dive into what characteristics are important in predicting a Pokemon's final combat power (CP)."

```r
## source pokemon go data from OpenIntro website
source("http://www.openintro.org/stat/data/pokemon.R")

## load ggplot2 package
library(ggplot2)
```

1) **Exercise 1.** Suppose we are interested in the relationship between pre-evolution combat power (`cp`) and post-evolution combat power (`cp_new`) for 75 Pokemon evolutions.

- Pre-evolution combat power (`cp`) and post-evolution combat power (`cp_new`) are what types of variables? Feel free to investigate them in an R chunk.
- Make an appropriate plot to examine the relationship between these two variables. First, make this plot using the simple plot functions. Then, make the same plot using ggplot(). In both cases, fix your axis labels such that the x-axis label is "Pre-evolution Combat Power" and the y-axis label is "Post-evolution Combat Power". Refer to the R Lab from Chapter 2 if you do not remember how to do this.

```
## Simple plot
```

```
## ggplot2
```

- Describe the relationship between these two variables. Does there appear to be an association?
- There are four species represented in this data set that may be important when considering the relationship between pre-evolution combat power (`cp`) and post-evolution combat power (`cp_new`). In a new chunk, copy and paste your ggplot code from the previous chunk and adapt it so that each dot gets a color corresponding to species. There should be four colors in total. Refer to the R Lab from Chapter 2 if you do not remember how to do this.

```
## ggplot2 with color for species
```

2. **Exercise 2.** Now suppose we are interested just the pre-evolution combat power (`cp`) variable. We wish to examine the distribution of this variable.

- Make a frequency histogram for pre-evolution combat power. Do this using the hist() function. Be sure to use an appropriate argument for the breaks option, which determines the size of the bins for your histogram.
- Make a frequency histogram for pre-evolution combat power. Do this using the ggplot() function. Be sure to use an appropriate quantity for binwidth (an argument within geom_histogram) so that your histogram has bins that are reasonably sized.
- What do you expect the relationship between the mean and median will be for this histogram? Which would you recommend as a measure of center?
- Would you recommend using the standard deviation or the IQR as a measure of spread for these data? Why?

3. **Exercise 3.** The book R for Data Science can be a great resource for you as you learn more about what we can do in R. While a lot of it is advanced, it does have great examples for data visualizations. Open the PDF of the book in your browser. For this exercise, you will use the table of contents: Click on 3. Data Visualization and then on 3.8 Position Adjustments. Using the pokemon data, and specifically the variables `attack_weak_type` and `attack_weak_type_new`, generate the same bar plots that were generated for the diamonds data set in the book example. You should make 7 plots in total (I have already included the R chunks for your convenience). Pay attention to the subtle differences in the code that let you change how these plots look.

- For these first two plots, choose either of the two variables to depict in your bar chart. What is the difference between these two plots? What piece of the code accomplished this change?
- For the rest of the plots, you will be using both variables. Identify the plots that we have studied in class. What are the differences among these five plots? What pieces of code accomplished these changes?
- Based on the last three plots you generated, what is the default argument for position?