Dealing with Data II Homework 1 Solutions

Joshua D. Ingram

2022-09-08

Reading

Reading Assignment 1

Read Chapter 2, excluding sections 2.6-2.7

Questions from Chapter 2

1) Exercise 2.14

What do alligators eat? The following bar chart is from a study investigating the factors that influence alligators' choice of food. For 219 alligators captured in four Florida lakes, researchers classified the primary food choice (in volume) found in the alligator's stomach in one of the categories—fish, invertebrate (snails, insects, crayfish), reptile (turtles, baby alligators), bird, or other (amphibian, mammal, plants). (Data available on the book's website.)

a. Is the primary food choice categorical or quantitative?

Answer: The primary food choice is categorical.

b. Which is the model category for primary food choice?

Answer: The model category is the category with the highest frequency in the data, which is the "fish" category.

c. About what percentage of alligators had fish as the primary food choice?

```
# pass the frequency table to prop.table to get proportions
prop.table(table(alligator$food))
```

Answer:

```
##
##         bird         fish invertebrate        other      reptile
##   0.05936073   0.42922374   0.27853881   0.14611872   0.08675799
```

About 42.92% of alligators in the study had fish as the primary food choice.

d. This type of bar chart, with categories listed in order of frequency, has a special name. What is it?

Answer: The bar chart given in *Example 2.14* is called a Pareto Chart.

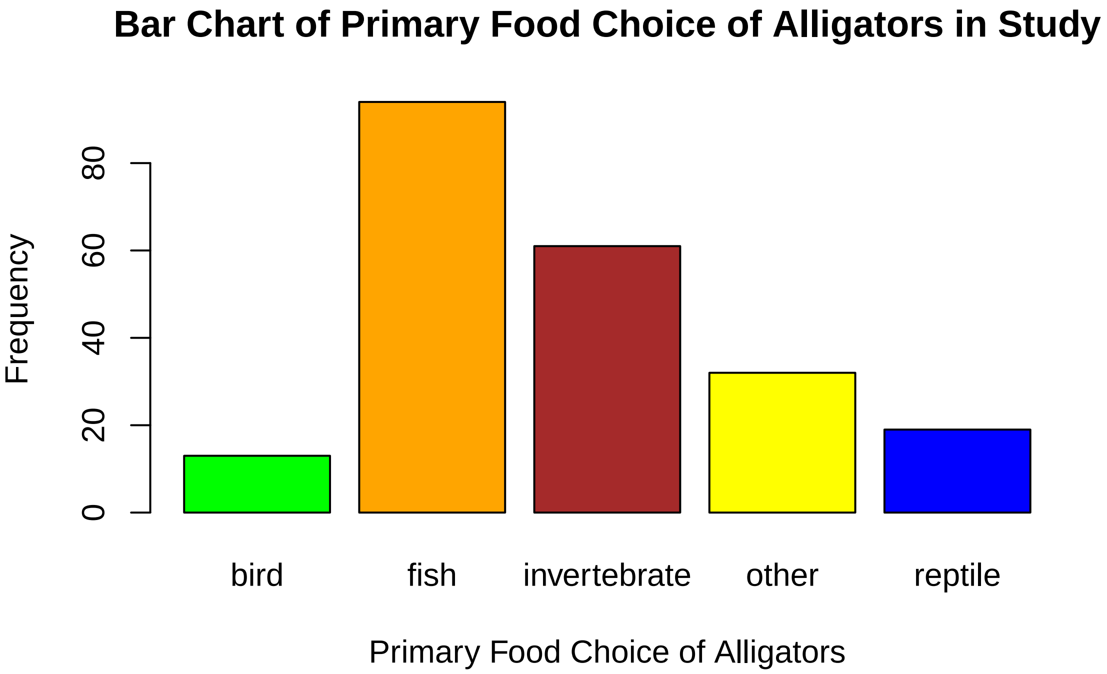
2) Construct a frequency table and bar chart in R. Write a one or two sentence statement about the distribution of alligator food choice

```
# save frequency table using the table() function
frequencies <- table(alligator$food)
print(frequencies)
```

Answer:

```
##
##         bird         fish invertebrate        other      reptile
##           13           94           61           32           19
```

```
# colors for the bar chart
colors <- c("green", "orange", "brown", "yellow", "blue")
# bar chart using the frequencies given by the table function
barplot(frequencies,
        main = "Bar Chart of Primary Food Choice of Alligators in Study",
        xlab = "Primary Food Choice of Alligators",
        ylab = "Frequency",
        col = colors)
```



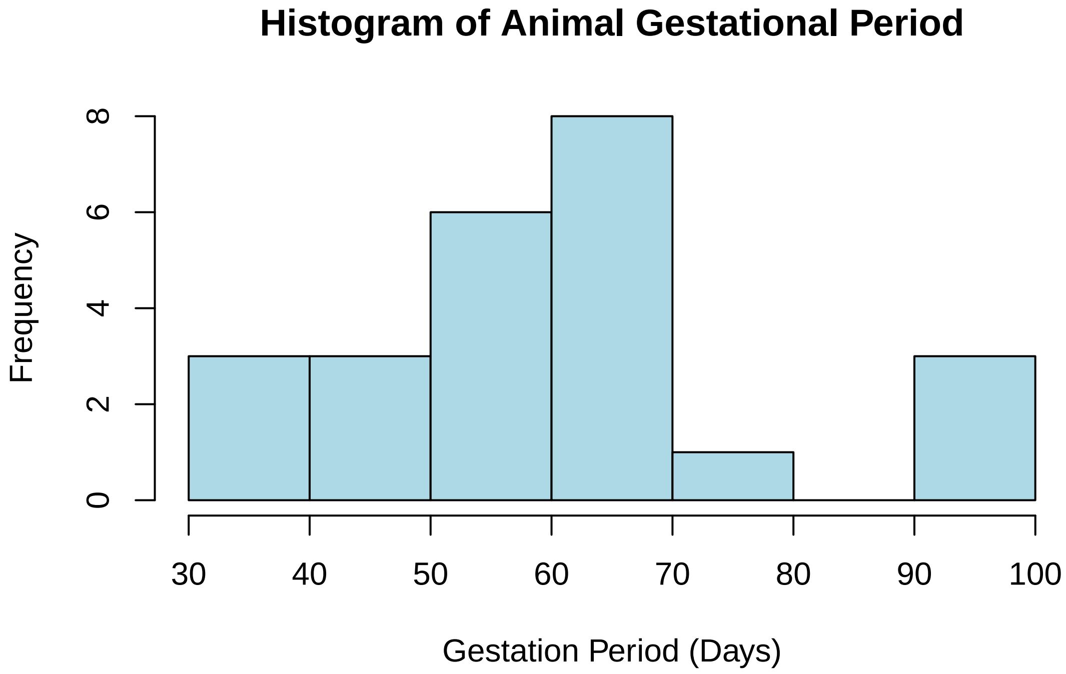
Fish are the most common food choice of alligators making up 42.92% of the food choices, followed by invertebrate as the second most common food choice making up 27.85% of the observed food choices. Birds are the least common food choice, being about 5.94% of the observed alligator food choices in the study.

3) Exercise 2.26

Gestational period. The Animals data set from the book's website contains data on the length of the gestational period (in days) of 21 animals. (*Source*: Wildlife Conservation Society)

a. Using software, obtain a histogram of the gestational period.

```r
# hist function creates a histogram of the data
hist(animals$`Gestation (days)`,
     main = "Histogram of Animal Gestational Period",
     xlab = "Gestation Period (Days)",
     col = "lightblue")
```



Answer:

b. Do you see any observation that is unusual? Which animal is it?

```r
# select rows where gestation period is greater than 80 days
animals[which(animals$`Gestation (days)` > 80), ]
```

Answer:

```
## # A tibble: 3 x 8
##   Animal Family `Common name` `Gestation (da~` `Longevity (yr~` `Birth Weight ~`
##    <dbl> <chr>  <chr>                    <dbl>            <dbl>            <dbl>
## 1      9 Felid~ Cougar                      92             23.8              400
## 2     11 Felid~ Jaguar                      99             28                820
## 3     12 Felid~ Leopard                     97             27.3              550
## # ... with 2 more variables: `Adult Weight (g)` <dbl>,
## #   `Growth Rate (1/days)` <dbl>
```

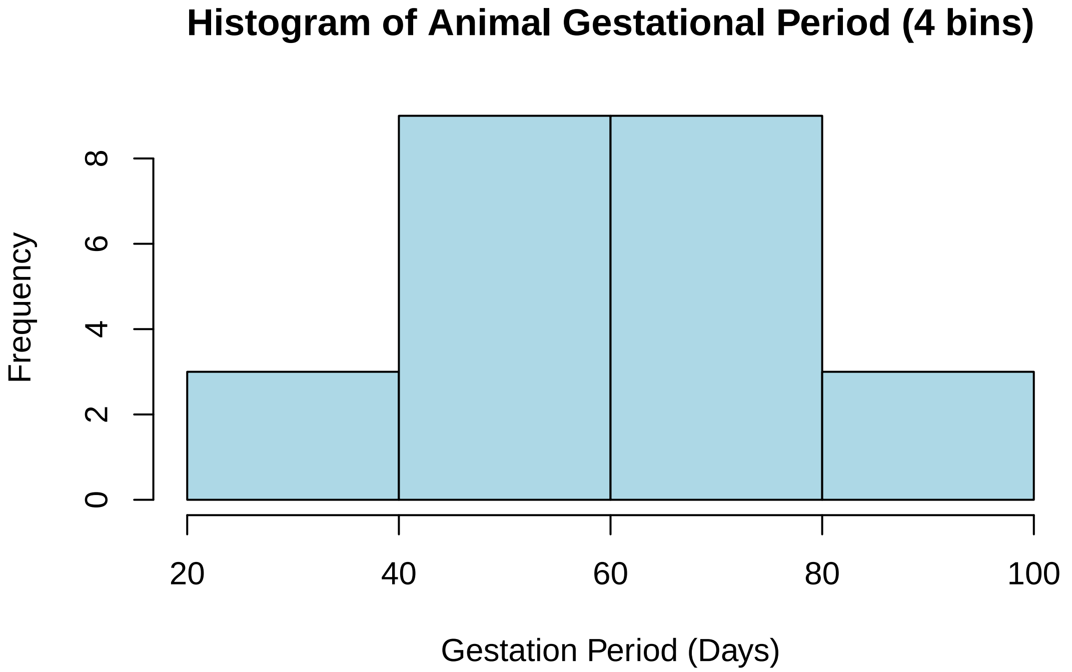
There are 3 outliers with unusually high observed gestation periods. These animals include the cougar (92 day gestation period), the leopard (97 day gestation period), and the jaguar (99 day gestation period).

c. Is the distribution right-skewed or left-skewed?

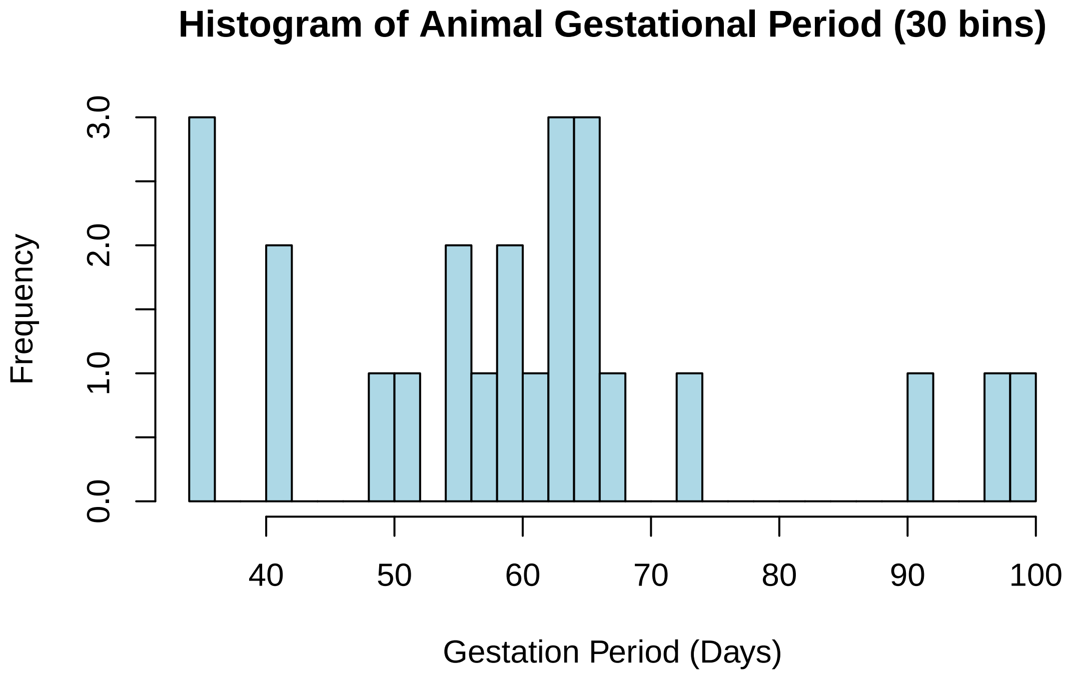
Answer: The distribution is right-skewed.

d. Try to override the default setting and plot a histogram with only very few intervals and one with many intervals. Would you perfer either one to the histogram created in part a? Why or why not?

```
# histogram with few intervals
hist(animals$`Gestation (days)`,
     main = "Histogram of Animal Gestational Period (4 bins)",
     xlab = "Gestation Period (Days)",
     col = "lightblue",
     breaks = 4)
```



```
# histogram with many intervals
hist(animals$`Gestation (days)`,
     main = "Histogram of Animal Gestational Period (30 bins)",
     xlab = "Gestation Period (Days)",
     col = "lightblue",
     breaks = 30)
```



Answer: The histogram with only 3 intervals is too few, as we cannot observe any interesting shapes or trends in the data with such low resolution. The histogram with 30 bins is too high for a dataset of this size, as we have many 0 bin counts and too much noise. Ideally, the bin count would be the like in **part a**, or slightly more (8-12) so that we can observe the overall shape of the data without getting lost in the noise.

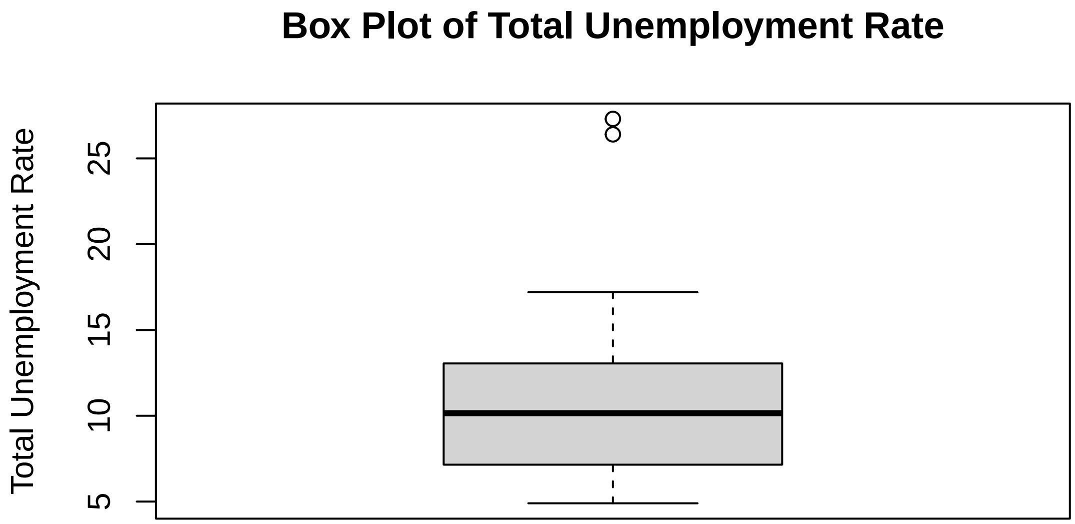
4) Exercise 2.62 - Youth Unemployment in the EU

The Youth Unemployment data file on the book's website and pre-loaded with the *Explore Quantitative Data* app contains 2013 unemployment rates in the 28 EU countries for people between 15 and 24 years of age. (The data are also shown in Exercise 2.65). Using the app or other software,

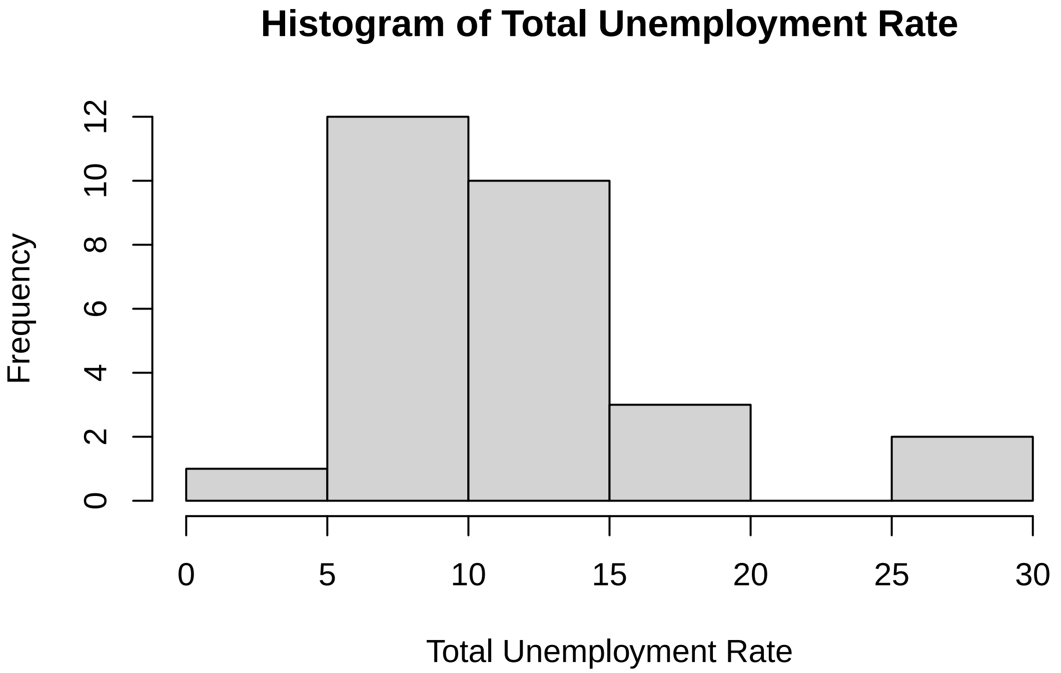
a. Construct an appropriate graph to visualize the distribution of the unemployment rate.

Answer: A box plot or histogram are appropriate graphs to visualize this sort of data.

```
# box plot
boxplot(unemployment$total,
        main = "Box Plot of Total Unemployment Rate",
        ylab = "Total Unemployment Rate")
```



```
# histogram
hist(unemployment$total,
     main = "Histogram of Total Unemployment Rate",
     xlab = "Total Unemployment Rate",
     breaks = 7)
```



b. Find the mean, median, and standard deviation.

```
# mean
mean(unemployment$total)
```

Answer:

```
## [1] 11.12857
```

```r
# median
median(unemployment$total)
```

```
## [1] 10.15
```

```r
# standard deviation
sd(unemployment$total)
```

```
## [1] 5.58383
```

Mean = 11.12857%

Mode = 10.15%

Standard Deviation = 5.58383%

c. Write a short paragraph summarizing the distribution of the youth unemployment rate, interpreting some of the above statistics in context.

```r

unemployment[which(unemployment$total > 20),]
```

Answer:

```
## # A tibble: 2 x 4
##   country total  male female
##   <chr>   <dbl> <dbl>  <dbl>
## 1 Greece   27.3  24.3   31.3
## 2 Spain    26.4  25.8   27
```

The average unemployment rates of people between the ages of 15 and 24 years in the EU countries in 2013 is 11.13%. We observe a right skewed distribution with two outliers in the data, Spain and Greece, that had relatively high unemployment rates of 26.4% and 27.3%, respectively. The unemployment rates vary, on average, by about 5.58%, with a median of 10.15% for the unemployment rate.