Dealing with Data II Homework 2 Solutions

Joshua D. Ingram

2022-09-13

R Reading

1)

Please read and study the R handout on "Describing the Distribution of Quantitative Variables", which has many of the commands we used in class for histograms and boxplots.

2)

Please read and study the R handout on "Describing the Distribution of Categorical Variables", Section 3, which starts on page 7. This section has all the commands for creating contingency tables and obtaining the marginal and conditional distributions, as we discussed in class.

Textbook Reading

1)

Please read Section 3.1 and Section 3.2 of the textbook. Pay special attention to the paragraph on the difference and ratio of proportions, which we will go over in class on Tuesday.

Exercises

1)

In class, we looked at Freshmen status vs. political view for a survey of 70 students. For this exercise, lets analyze the relationship between gender identity (for these data, students identified as either male or female) and political view.

a. Load the dataset (which is under Files -> Datasets) into R and create a contingency table

```
# import ClassSurvey.csv (nead `readr` package imported to use read_csv())
survey <- read_csv("https://img1.wsimg.com/blobby/go/bbca5dba-4947-4587-b40a-db346c01b1b3/downloads/Cla
```

Answer:

```
## Rows: 70 Columns: 8
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (4): Frosh, Chocolate, Gender Identity, Political View
## dbl (4): Student, Textbook Costs ($), Streaming (min/day), Social Media (min...
##
## i Use `spec()` to retrieve the full column specification for this data.
```

```
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# create contingency table with `table()` and use `addmargins()` to get marginal counts
addmargins(table(survey$`Gender Identity`, survey$`Political View`))
```

```
##
##          Conservative Liberal Moderate Sum
##   Female            2      26        2  30
##   Male              5      21       14  40
##   Sum               7      47       16  70
```

b. What proportion of students identify as female and liberal?

Answer: 0.3714 of students in the survey identify as female and liberal.

```
prop.table(table(survey$`Gender Identity`, survey$`Political View`))
```

```
##
##          Conservative    Liberal   Moderate
##   Female   0.02857143 0.37142857 0.02857143
##   Male     0.07142857 0.30000000 0.20000000
```

c. The question above is a question about which distribution: joint, marginal, or conditional?

Answer: The question above refers to the joint distribution.

d. Find the marginal distribution of political view and write a one sentence statement about it.

```
# conditional distribution of political view
marginal <- addmargins(prop.table(table(survey$`Gender Identity`, survey$`Political View`)))[3,]
marginal
```

Answer:

```
## Conservative      Liberal     Moderate          Sum
##    0.1000000    0.6714286    0.2285714    1.0000000
```

Identifying as liberal is most common, followed by moderate, then conservative.

e. Find the conditional distribution of political view, given gender identity. Obtain a side-by-side bar chart in R.

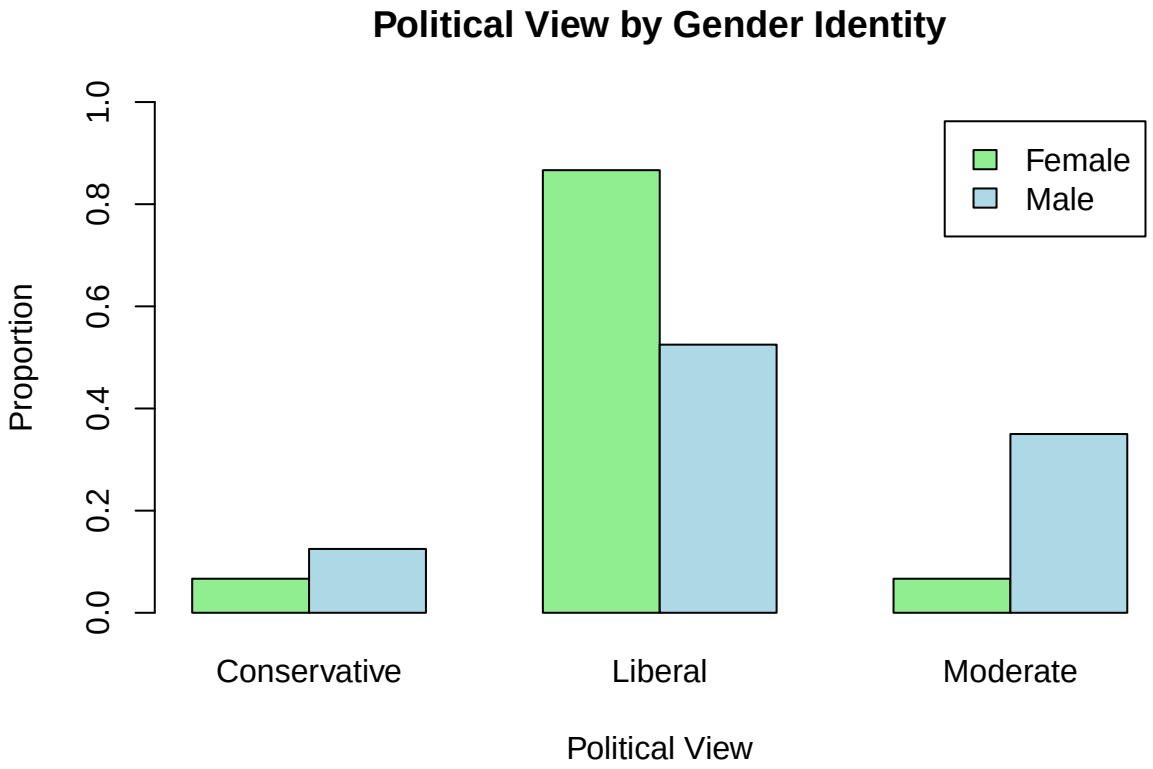
```
# conditional distribution of political view, given gender identity
conditional <- prop.table(table(survey$`Gender Identity`, survey$`Political View`), 1)
conditional
```

Answer:

```
##
##          Conservative    Liberal   Moderate
##   Female   0.06666667 0.86666667 0.06666667
##   Male     0.12500000 0.52500000 0.35000000
```

```
barplot(conditional,
        beside = T,
        col = c("lightgreen", "lightblue"),
        main = "Political View by Gender Identity",
```

```
        xlab = "Political View",
        ylab = "Proportion",
        legend = TRUE,
        ylim = c(0,1))
```



f. Using the conditional distributions, write two sentences about the relationship between gender and political view.

Answer: Out of the 3 political views, identifying as liberal is the most common regardless of gender identity. However, the proportion of females that identify as liberal versus males is greater with a difference in proportions of about 0.34, and there is a greater proportion of males that identify as conservative or moderate than there are females.

g. Do you think gender and political view are associated. Why or why not?

Answer: Yes. The proportion of females that have liberal political views is 0.34 higher than males, which is large. Additionally, the proportion of males identify as moderate is 0.28 considerably higher than females with moderate view. Further, the proportion of males identifying as conservative is $0.125/.06667 = 1.875$ times higher than that of females.

2) Textbook Exercise 3.8 - Eyewitness Testimony

To evaluate how eyewitness testimonies may affect jurors in reaching a verdict, researchers recorded two mock murder trials: one with weak and one with strong eyewitness testimonies. They then asked jurors to come to a verdict of either guilty or not guilty based on the version of the recording they were shown. You can enter the counts from the contingency table below in the *Association Between Two Categorical Variables* app to answer the following questions.

a. How many jurors were shown the recording with the strong and how many hurors were shown the recording with the weak testimony?

```
table(eyewitness$Testimony)
```

Answer:

```
##
## strong   weak
##    190    183
```

190 jurors were shown the recording with the strong testimony, and 183 jurors were shown the recording with the weak testimony.

b. What is the response variable, and what is the explanatory variable?

Answer: The response variable is the verdict of the juror. The explanatory variable is the strength of the testimony given to the jurors.

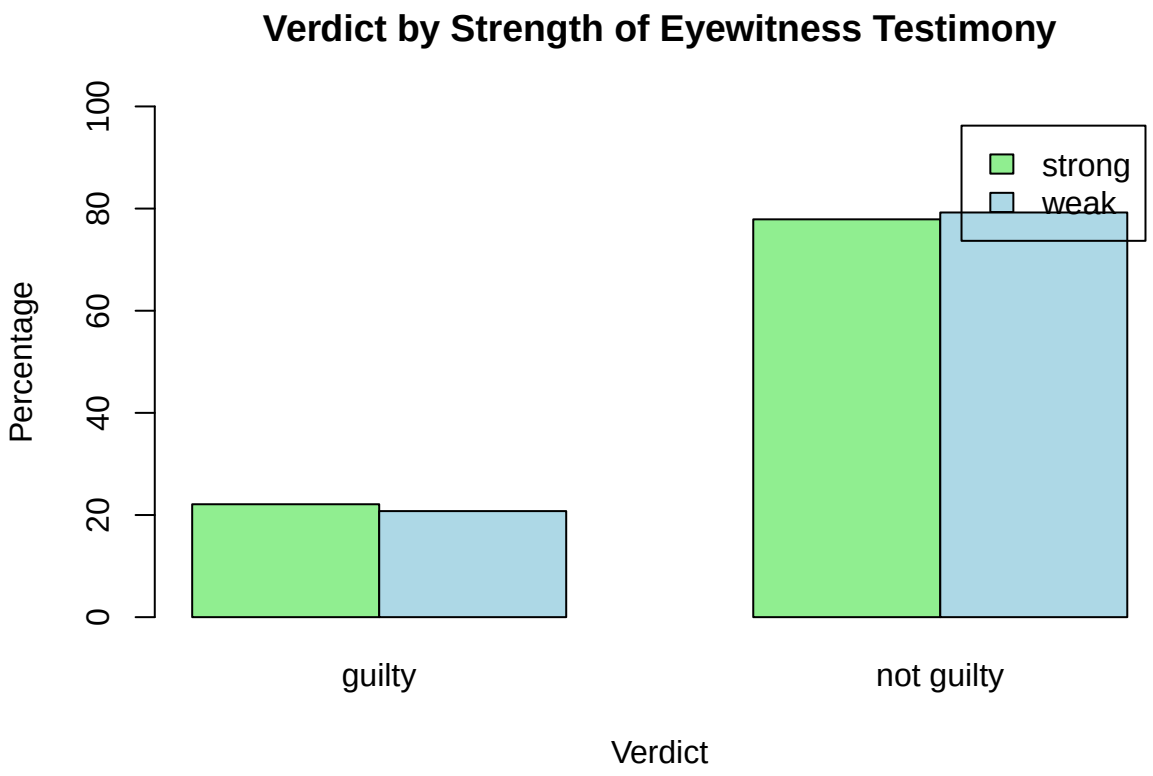
c. Find the conditional proportions (convert them to percentages) of a guilty verdict given eyewitness testimony. Interpret and sketch a plot comparing the conditional proportions.

```
# conditional distribution of political view, given gender identity
conditional <- prop.table(table(eyewitness$Testimony, eyewitness$Verdict), 1)*100
conditional
```

Answer:

```
##
##            guilty not guilty
##   strong 22.10526   77.89474
##   weak   20.76503   79.23497
```

```
barplot(conditional,
        beside = T,
        col = c("lightgreen", "lightblue"),
        main = "Verdict by Strength of Eyewitness Testimony",
        xlab = "Verdict",
        ylab = "Percentage",
        legend = TRUE,
        ylim = c(0,100))
```



d. Find the difference of the conditional proportions and interpret this difference in context.

Answer: The conditional proportion of guilty verdicts given strong eyewitness testimony is $22.10526/20.76503 = 6.5\%$ higher than guilty verdicts given weak eyewitness testimony.

e. Is it true that the proportion of jurors watching the recording with the strong testimony is 6% higher than the proportion of jurors watching the recording with the weak testimony? Explain.

Answer: This is true. The proportion of jurors watching the strong testimony is 1.88 *percentage points* higher than the proportion of jurors watching the recording of the weak testimony. Stated alternatively, by using the ratio, the proportion of jurors giving guilty verdicts while watching strong testimonies is 6.5% higher than the poportion of jurors giving guilty verdicts while watching weak testimonies.

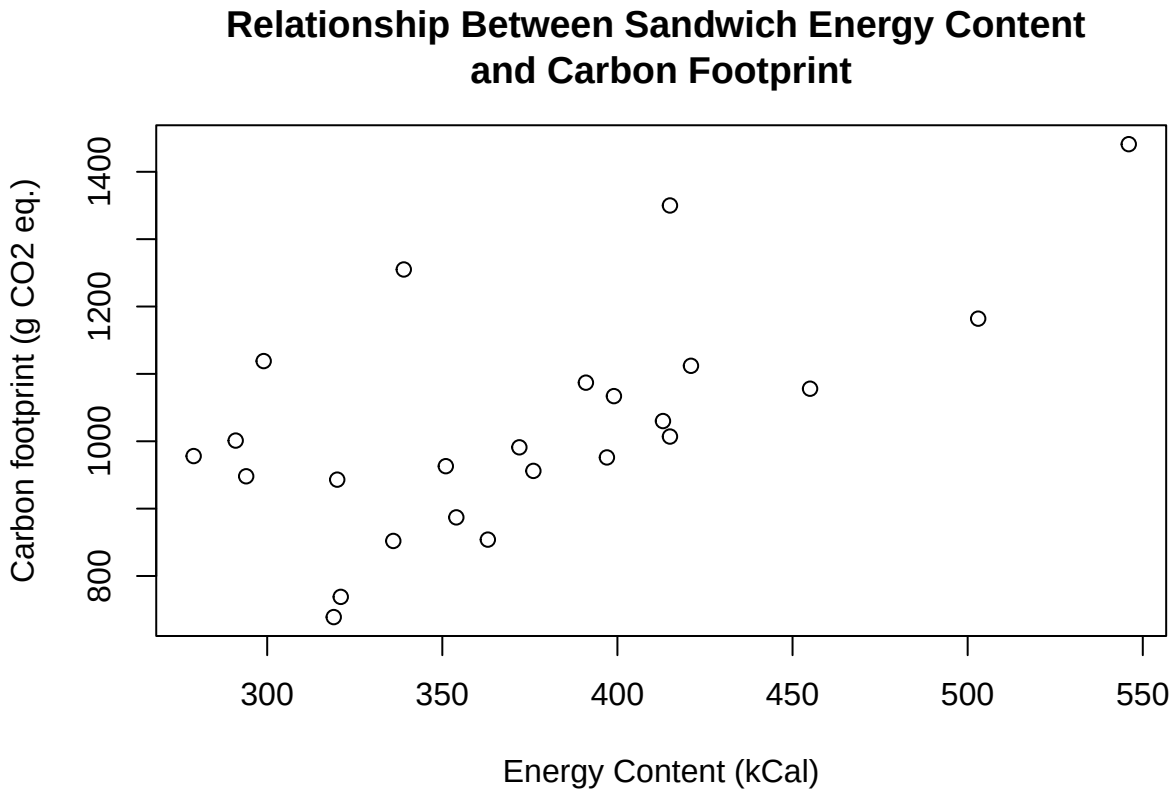
3) Textbook Exercise 3.16 - Carbon Footprint of Sandwiches

The article "Understanding the impact on climate change of convenience food: Carbon footprint of sandwiches" analyzed data of 24 commercially available sandwiches with regard to their carbon footprint and how the carbon footprint depends on the energy content of the sandwich. The data are available on the book's website and are pre-loaded in the *Scatterplots & Correlation* app. Describe the relationship between energy content and carbon footprint by

a. Visualizing the relationship with a scatterplot

```
# Using base r
plot(sandwiches$`EnergyContent (kCal)`, sandwiches$`Carbon footprint (g CO2 eq.)`,
     main = "Relationship Between Sandwich Energy Content \nand Carbon Footprint",
     xlab = "Energy Content (kCal)",
     ylab = "Carbon footprint (g CO2 eq.)")
```

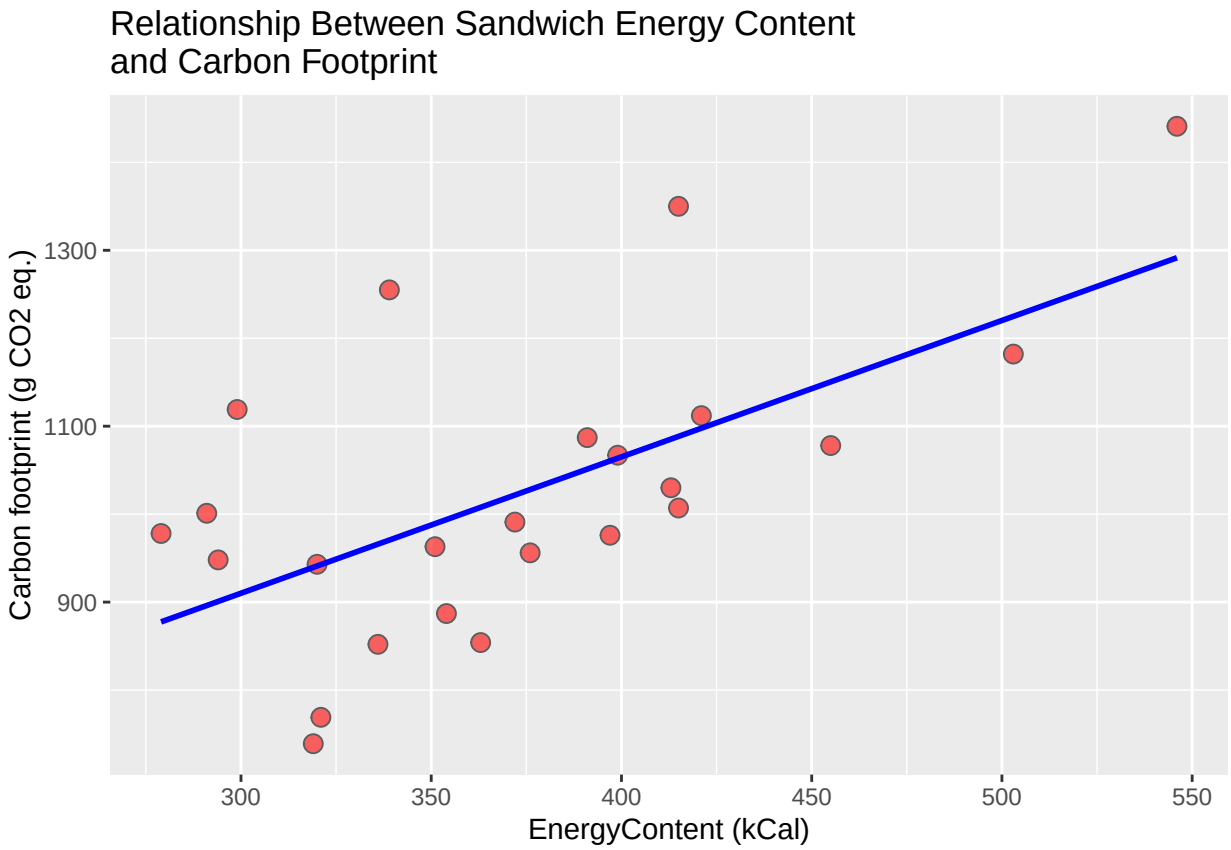
```
## Another way of getting a nice scatterplot
library(ggplot2) #make sure you have installed the ggplot2 library first
```



Answer:

```
ggplot(data=sandwiches,
       aes(
         x = `EnergyContent (kCal)`,
         y = `Carbon footprint (g CO2 eq.)`
       )
) +
  geom_point(alpha=0.6, size=3, shape=21, fill="red", color="black") +
  #geom_smooth(se=FALSE, color="green") +
  geom_smooth(method="lm", color="blue", se=FALSE) +
  labs(title = "Relationship Between Sandwich Energy Content \nand Carbon Footprint")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



There is a positive relationship between the energy content of the sandwich and its carbon footprint.

b. Write a paragraph about the relationship that quotes the correlation coefficient

```
cor(sandwiches$`EnergyContent (kCal)`, sandwiches$`Carbon footprint (g CO2 eq.)`)
```

Answer:

```
## [1] 0.6208991
```

The relationship between a sandwiches energy content and its carbon footprint is approximately linear, with a moderately strong positive relationship with a correlation coefficient of 0.621. As the energy content of the sandwich increases, we expect to observe an increase in the carbon footprint to produce the sandwich.