Homework assignment #4

Milica Cudina

2021-09-23

Cafe data

(2 points) First, you should read the data from our csv file "cafedata.csv" into a data.frame called `cafe.data`:

```
cafe.data<-read.csv("cafedata.csv")
```

If you want to see what your data.frame looks like, you can click on it in the **Global environment** in the upper right pane. The data.frame will get displayed in the upper left pane.

(2 points) Now, you interested in the types and names of the variables in your data.frame. What do you run?

```
ls.str(cafe.data)
```

```
## Bread.Sand.Sold :  chr [1:48] "5" "6" "8" "4" "3" "7" "6" "0" "3" "2" "3" "4" "9" "1" "3" "8" ...
## Bread.Sand.Waste :  chr [1:48] "3" "8" "2" "2" "0" "1" "6" "0" "4" "6" "7" "4" "1" "1" "5" "0" ...
## Chips :  chr [1:48] "12" "0" "0" "20" "0" "4" "2" "20" "3" "16" "2" "9" "13" "10" ...
## Coffees :  chr [1:48] "41" "33" "34" "27" "20" "23" "32" "31" "30" "27" "30" "27" ...
## Cookies.Sold :  chr [1:48] "5" "1" "1" "3" "3" "5" "10" "0" "3" "6" "5" "4" "13" "1" "8" ...
## Cookies.Waste :  chr [1:48] "3" "6" "0" "1" "0" "0" "0" "0" "2" "0" "0" "1" "0" "1" "0" "1" ...
## Day.Code :  int [1:48] 2 3 4 5 1 2 3 4 5 1 ...
## Day.of.Week :  chr [1:48] "Tue" "Wed" "Thu" "Fri" "Mon" "Tue" "Wed" "Thu" "Fri" "Mon" ...
## Fruit.Cup.Sold :  chr [1:48] "1" "0" "0" "3" "2" "2" "2" "0" "1" "2" "1" "2" "3" "0" "1" "2" ...
## Fruit.Cup.Waste :  chr [1:48] "4" "3" "3" "0" "0" "0" "0" "0" "1" "0" "0" "0" "0" "0" "0" "0" ...
## Juices :  chr [1:48] "8" "0" "13" "0" "5" "4" "5" "6" "4" "7" "0" "6" "6" "1" "2" ...
## Max.Daily.Temperature..F. :  int [1:48] 36 34 39 40 36 26 34 33 20 37 ...
## Muffins.Sold :  chr [1:48] "5" "3" "4" "5" "8" "1" "6" "6" "0" "3" "5" "4" "14" "2" "2" ...
## Muffins.Waste :  chr [1:48] "1" "5" "0" "0" "0" "0" "0" "1" "4" "0" "0" "1" "0" "0" "0" "0" ...
## Sales :  num [1:48] 200 196 103 163 102 ...
## Sodas :  chr [1:48] "20" "13" "23" "13" "13" "33" "15" "27" "12" "19" "33" "20" ...
## t :  int [1:48] 1 2 3 4 5 6 7 8 9 10 ...
## Total.Items.Wasted :  chr [1:48] "16" "39" "5" "10" "0" "4" "9" "1" "20" "9" "9" "6" "1" "4" "7" ...
## Total.Soda.and.Coffee :  chr [1:48] "61" "46" "57" "40" "33" "56" "47" "58" "42" "46" "63" "47" ...
## Wraps.Sold :  chr [1:48] "25" "7" "14" "5" "10" "5" "19" "7" "4" "13" "10" "15" "13" "6" ...
## Wraps.Waste :  chr [1:48] "5" "17" "0" "7" "0" "3" "3" "0" "9" "3" "2" "0" "0" "2" "2" ...
```

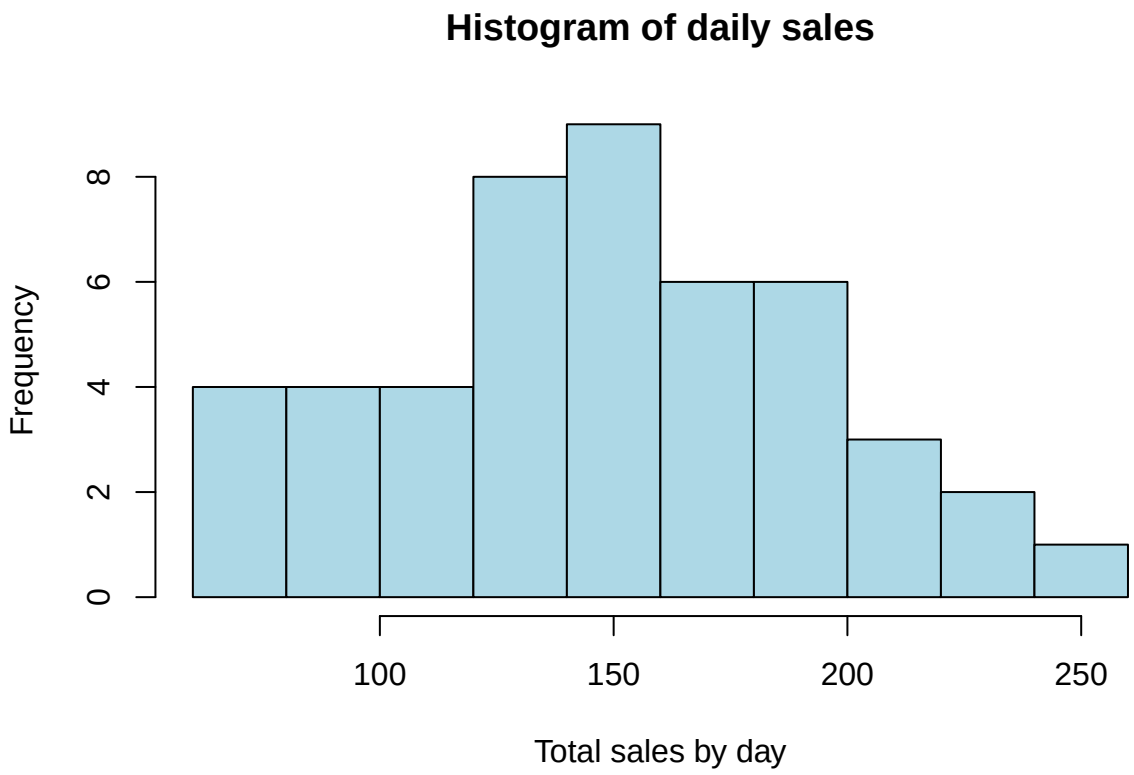
You see that the days/**cases** all have corresponding **rows**. They are labeled by the row indices. The **column** names stand for the variable names.

Then, you can do a bit of exploratory analysis.

Problem 1. (3 points)

Plot the histogram of daily sales amounts. Make sure that your plot has the main title and that the axes are also labeled.

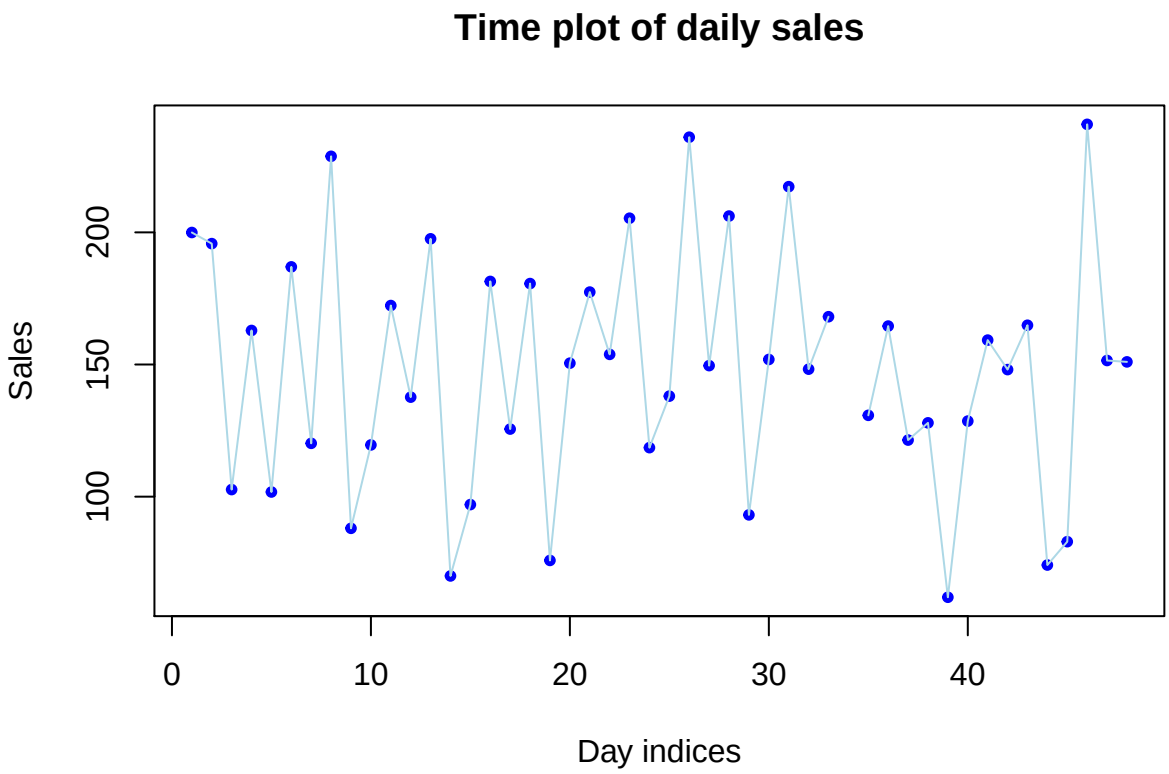
```
daily.sales<-cafe.data$Sales
hist(daily.sales,
     main="Histogram of daily sales",
     xlab="Total sales by day",
     ylab="Frequency",
     col="lightblue")
```



Problem 2. (3 points)

Draw the **time plot** on how the daily sales amounts evolved with time. Make sure that you connect the dots in your plot. Make sure that your plot has the main title and that the axes are also labeled.

```
days<-cafe.data$t
plot(days, daily.sales, pch=20, col="blue",
     main="Time plot of daily sales",
     xlab="Day indices",
     ylab="Sales")
lines(days, daily.sales, col="lightblue")
```



Problem 3. (2 points)

What is the maximum number of wraps sold on a Tuesday? Use `R` commands, please, do not just look it up in the spreadsheet - that defies the purpose! Be careful about the **type** of data your number of wraps is recorded as. *Hint:* Typing `?numeric` in the RStudio console might help.

```
tue.wraps<-as.numeric(cafe.data$Wraps.Sold[cafe.data$Day.of.Week=="Tue"])
max(tue.wraps)
```

```
## [1] 25
```

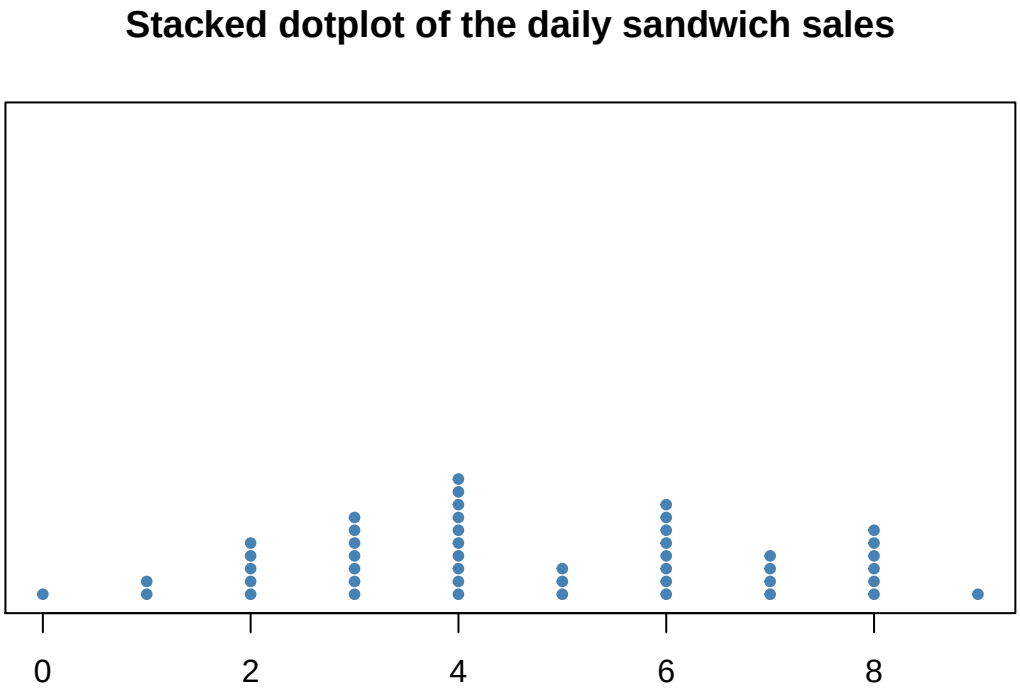
Problem 4. (3 points)

Create a stacked dot plot of the number of sandwiches sold daily. Be careful about the **type** of data your number of sandwiches is recorded as. *Hint:* Typing `?numeric` in the RStudio console might help.

```
sands<-as.numeric(cafe.data$Bread.Sand.Sold)
```

```
## Warning: NAs introduced by coercion
```

```
stripchart(sands, at=0, pch=20, method="stack", col="steelblue",
           main="Stacked dotplot of the daily sandwich sales")
```



Problem 5. (5 points)

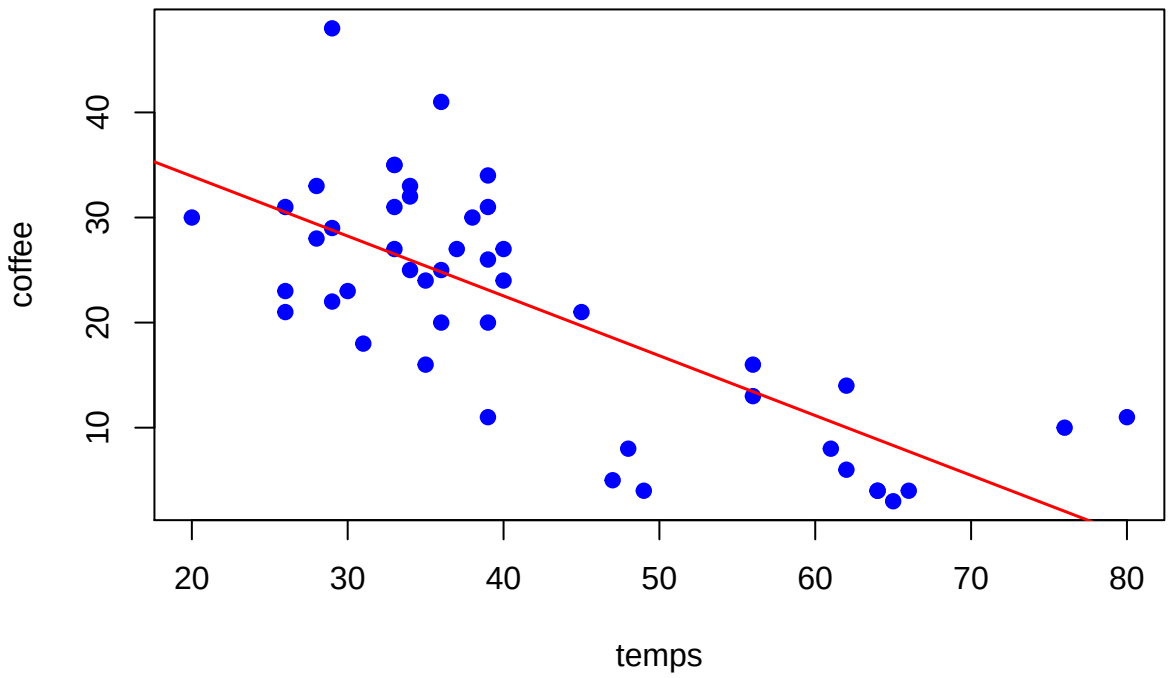
What else could we ask? Think a question the cafe manager might be interested in. Identify the relevant variable(s) from `cafe.data`, chose an appropriate plot, plot it, and comment on what you notice.

One possible solution: I was interested in the relationship between the temperature and the coffee sales, so I decided to draw a scatterplot.

```
temps<-cafe.data$Max.Daily.Temperature..F.
coffee<-as.numeric(cafe.data$Coffees)
```

```
## Warning: NAs introduced by coercion
```

```
plot(temps, coffee, pch=19, col="blue")
# I can see that there is an association judging from the shape of my scatterplot
# for those of you who know what the least-squares line is,
# this is how you add it to your scatterplot
abline(lm(coffee~temps), col="red", lwd=1.5)
```



Problems

Problem 6. (15 points)

Source: Problem 2.1.5. from Pitman.

Given that there are 12 heads in 20 independent tosses, find the probability that at least two of the first five tosses landed heads.

Note: You cannot solve this kind of a problem simply using `R`*. You have to use analytic methods. It is OK to leave binomial coefficients in your answer; simplify the remainder of your expression as much as you can.*

Solution: The number of heads in 20 has the binomial distribution $Binomial(n = 20, p = 1/2)$. Let

$$E = \{\text{there are 12 heads}\}, \quad F = \{\text{at least two of the first five tosses landed heads}\}.$$

Immediately $\mathbb{P}[E] = \binom{20}{12} \cdot \frac{1}{2^{20}}$. We need to find

$$\mathbb{P}[F|E] = \frac{\mathbb{P}[E \cap F]}{\mathbb{P}[E]} = \frac{\mathbb{P}[E] - \mathbb{P}[E \cap F^c]}{\mathbb{P}[E]} = 1 - \frac{\mathbb{P}[E \cap F^c]}{\mathbb{P}[E]}.$$

Let

$$H_0 = \{\text{there are no heads in the first five tosses}\},$$
$$H_1 = \{\text{there is exactly one head in the first five tosses}\}.$$

Then, $F^c = H_0 \cup H_1$, and H_0 and H_1 are mutually exclusive. Also, with

$$G_0 = \{\text{there are exactly 12 heads in last 15 tosses}\},$$
$$G_1 = \{\text{there are exactly 11 heads in last 15 tosses}\},$$

we have, due to independence of trials,

$$\mathbb{P}[E \cap H_0] = \mathbb{P}[H_0 \cap G_0] = \mathbb{P}[H_0]\mathbb{P}[G_0],$$
$$\mathbb{P}[E \cap H_1] = \mathbb{P}[H_1 \cap G_1] = \mathbb{P}[H_1]\mathbb{P}[G_1].$$

Moreover, the number of heads in the first five trials is $Binomial(n = 5, p = 1/2)$, and the number of heads in the remaining 15 trials is $Binomial(n = 15, p = 1/2)$. So,

$$\mathbb{P}[E \cap H_0] = \binom{5}{0}\binom{15}{12} \cdot \frac{1}{2^{20}},$$
$$\mathbb{P}[E \cap H_1] = \binom{5}{1}\binom{15}{11} \cdot \frac{1}{2^{20}}.$$

and

$$\mathbb{P}[F|E] = 1 - \frac{\binom{15}{12}}{\binom{20}{12}} - 5 \cdot \frac{\binom{15}{11}}{\binom{20}{12}}.$$

Note that the fact that the coin is (by default) fair is completely irrelevant in the above calculation.

Problem 7. (5 points)

A pair of dice is thrown. Find the probability that the sum of the outcomes is 10 or greater **if** a 5 appears on the first die.

Solution: Let A_i denote the event that i was the outcome on the first die for $i = 1, 2, \ldots 6$. Let E denote the event that the sum of the outcomes on both of the dies was greater than or equal to 10. Formally,

$$\begin{aligned} E &= \{(i,j) : 1 \le i,j \le 6 \text{ and } i+j \ge 10\} \\ &= \{(i,j) : 1 \le i,j \le 6 \text{ and } i+j = 10\} \cup \{(i,j) : 1 \le i,j \le 6 \text{ and } i+j = 11\} \\ &\quad \cup \{(i,j) : 1 \le i,j \le 6 \text{ and } i+j = 12\}. \end{aligned} \tag{1}$$

We want to find the probability $\mathbb{P}[E|A_5]$. Directly from the definition of conditional probability, we get

$$\mathbb{P}[E|A_5] = \frac{\mathbb{P}[E \cap A_5]}{\mathbb{P}[A_5]}.$$

From the representation in (1), we get that

$$\begin{aligned} \mathbb{P}[E \cap A_5] &= \mathbb{P}[\{(i,j) : 1 \le i,j \le 6 \text{ and } i+j = 10\} \cap \{(i,j) : i = 5 \text{ and } 1 \le j \le 6\}] \\ &\quad + \mathbb{P}[\{(i,j) : 1 \le i,j \le 6 \text{ and } i+j = 11\} \cap \{(i,j) : i = 5 \text{ and } 1 \le j \le 6\}] \\ &\quad + \mathbb{P}[\{(i,j) : 1 \le i,j \le 6 \text{ and } i+j = 12\} \cap \{(i,j) : i = 5 \text{ and } 1 \le j \le 6\}] \\ &= \mathbb{P}[\{(i,j) : i = 5 \text{ and } i+j = 10\}] \\ &\quad + \mathbb{P}[\{(i,j) : i = 5 \text{ and } i+j = 11\}] \\ &\quad + \mathbb{P}[\{(i,j) : i = 5 \text{ and } i+j = 12\}] \\ &= \mathbb{P}[\{(5,5)\}] + \mathbb{P}[\{(5,6)\}] + \mathbb{P}[\emptyset] \\ &= \frac{1}{36} + \frac{1}{36} + 0 = \frac{1}{18}. \end{aligned}$$

On the other hand, $\mathbb{P}[A_5] = \frac{1}{6}$, and so $\mathbb{P}[E|A_5] = \frac{1}{3}$.

Note: The above solution is deliberately ridiculously formal. You could have solved this problem correctly by straightforward counting of "good" outcomes quite fast and in many ways.

Problem 8. (5 points)

Remember that The University of Texas has a subscription to *The Economist* and that you can study and enjoy their charts and articles whenever you want.

Consider the following charts and choose which of the offered statements is/are **correct**. **Justify your response!** Even if correct, answers without a justification will receive zero credit.

```
knitr::include_graphics("visas.png")
```

a. The shortest number of work days needed to purchase a visa is for Europe.
b. The more affluent the region, the longer one needs to work to purchase the visa.
c. One has to work more than three times as long in Sub-Saharan Africa than in the Middle East to purchase a visa.
d. One has to work longer in Oceania than in Asia to purchase a visa.
e. None of the above statements are correct.

Solution: Statement a. is **not** correct since the bar corresponding to North America in the bar graph on the right-hand side is actually the shortest.

Statement b. is **not** correct, since the trend line in the scatterplot on the left-hand side has a negative slope.

Statement c. is **correct** since the bar corresponding to Sub-Saharan Africa is longer than 15 and the bar corresponding to the Middle East is shorter than 5.

Statement d. is **not** correct since the bar corresponding to Oceania is shorter than the bar corresponding to Asia in the right-hand side graph.

Statement e. is **not** correct since statement c. is correct.

Problem 9. (5 points)

Two independent coins are tossed. If E_1 is the event "heads on first coin", E_2 the event "head on the second coin", and E_3 the event "the coins match; both are heads or tails". Which of the following statements is/are **not** true? **Justify your response!** Even if correct, answers without a justification will receive zero credit.

a. E_1 and E_2 are independent.
b. E_2 and E_3 are independent.
c. E_1 and E_3 are independent.
d. E_3 and E_1 are independent.
e. E_1, E_2 and E_3 are independent.

Solution: Statement a. is **correct** since we are given that the two coins are independent.

Statement b. is **correct**. Straight from the definitions of the events in the problem, we see that

$$\mathbb{P}[E_2] = \frac{1}{2} \quad \text{and} \quad \mathbb{P}[E_3] = \frac{1}{2}.$$

On the other hand,

$$\mathbb{P}[E_2 \cap E_3] = \mathbb{P}[HH] = \frac{1}{4}.$$

So,

$$\mathbb{P}[E_2]\mathbb{P}[E_3] = \frac{1}{4} = \mathbb{P}[E_2 \cap E_3].$$

Hence, statement b. is correct.

Statement c. is **correct** which can be shown using an argument analogous to the one in the justification of statement b.

Statement d. is **correct** since it is equivalent to statement c. which is already shown to be correct.

Statement e. is **not** correct since knowing that events E_1 and E_2 happened implies that the event HH happened which implies that E_3 happened.