S520 Problem Set 2 Instructor's Solutions

Spring 2023 STAT-S 520

January 24th, 2023

1.

a. Note that $S = A \cup A^c$ and by definition A and A^c are disjoint events. Using the finite additivity property and $P(S) = 1$ we get:

$$\begin{aligned} P(A) + P(A^c) &= P(A \cup A^c) \\ &= P(S) \\ &= 1 \end{aligned} \tag{1}$$

Hence

$$P(A) = 1 - P(A^c) \quad \square$$

Also observe that the empty set, $\emptyset$, is the complement of S; therefore $P(\emptyset) = 1 - P(S) = 0$

b. If $A \subset B$, then $B = A \cup (A^c \cap B)$, two disjoint events. Hence

$$\begin{aligned} P(A) + P(A^c \cap B) &= P\Big(A \cup (A^c \cap B)\Big) \\ &= P(B) \end{aligned} \tag{2}$$

And since $P(A^c \cap B) >= 0$ (probability cannot be negative) then $P(A) \leq P(B) \quad \square$

c. Let's rewrite the probabilities of A and B, each as the probability of the union of two disjoint events, as follows:

$$P(A) = P\Big((A \cap B^c) \cup (A \cap B)\Big) \tag{3}$$

$$P(B) = P\Big((A^c \cap B) \cup (A \cap B)\Big) \tag{4}$$

We can then sum (3) and (4) and use the finite additivity property to get

$$\begin{aligned} P(A) + P(B) &= P\Big((A \cap B^c) \cup (A \cap B)\Big) + P\Big((A^c \cap B) \cup (A \cap B)\Big) \\ &= P(A \cap B^c) + P(A \cap B) + P(A^c \cap B) + P(A \cap B) \\ &= \Big(P(A \cap B^c) + P(A \cap B) + P(A^c \cap B)\Big) + P(A \cap B) \\ &= P\Big((A \cap B^c) \cup (A \cap B) \cup (A^c \cap B)\Big) + P(A \cap B) \\ &= P(A \cup B) + P(A \cap B) \end{aligned} \tag{5}$$

where the last equality holds because $A \cup B = (A \cap B^c) \cup (A \cap B) \cup (A^c \cap B)$, i.e. the union of A and B is the union of those three disjoint events (It's easier to see this on a Venn diagram). Subtracting $P(A \cap B)$ from both sides of (5) we get

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad \square \tag{6}$$

2.

We are tossing a fair coin five times. So each time we toss a coin we have 2 possible outcomes. So the total number of possible outcomes in tossing 5 coins will be $\#S = 2^5 = 32$.

a. The number of ways exactly 4 coins show heads among 5 tosses is the same as choosing 4 places among 5 places. Hence the probability will be:

```
choose(5,4) / 2^5
```

```
## [1] 0.15625
```

b. To have more heads than tails, consider the possibilities:

- 5 heads, 0 tails
- 4 heads, 1 tails
- 3 heads, 2 tails

This can be done by

```
(choose(5,5) + choose(5,4) + choose(5,3)) / 2^5
```

```
## [1] 0.5
```

The result is not surprising. Observe that the outcomes that are not part of B are

- 2 heads, 3 tails
- 1 heads, 4 tails
- 0 heads, 5 tails

which by symmetry, should be exactly half of them.

c. Similar to the problem in b:

```
(choose(5,3) + choose(5,4) + choose(5,5)) / 2^5
```

```
## [1] 0.5
```

d. The outcomes in A are not outcomes in D (if we get four head, we cannot have at least three tails). So, if $s \in A$ then $s \notin D$. By definition, $D \subset A^c$ so $A^c \cup D = A^c$. Therefore,

$$P(A^c \cup D) = P(A^c) = 1 - P(A)$$

```
1 - 0.15625
```

```
## [1] 0.84375
```

e. B and D are disjoint events, and together they cover every possible outcome. Either property would help obtain the result: $P(B \cup D) = P(B) + P(D) = 0.5 + 0.5 = 1$ or $P(B \cup D) = P(S) = 1$

3.

a. We are given that $P(A) = 0.6, P(B) = 0.7$ and $P(A^c \cap B^c) = 0.12$. Observe that if A and B would be disjoint, then $P(A \cup B) = P(A) + P(B) = 0.6 + 0.7 = 1.3$. Since no probability can be greater than 1, this is not possible and A and B are not disjoint.

b. Draw Venn diagrams if it's not easy to visualize the following results:

$$P(A \cup B) = 1 - P(A^c \cap B^c) = 1 - 0.12 = 0.88$$
$$P(A \cap B) = P(A) + P(B) - P(A \cup B) = 0.6 + 0.7 + 0.88 = 0.42$$

$$P(B) = P(A \cap B) + P(A^c \cap B)$$

so

$$P(A^c \cap B) = P(B) - P(A \cap B) = 0.7 - 0.42 = 0.28$$

and finally

$$P(A \cup B^c) = 1 - P(A^c \cap B) = 1 - 0.28 = 0.72$$

c. Since

$$P(A) \cdot P(B) = 0.6 \cdot 0.7 = 0.42 = P(A \cap B)$$

A and B are independent.

d. Since A and B are independent $P(A|B) = P(A) = 0.6$. We can also find this result using the definition of conditional probability:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.42}{0.7} = 0.6$$

4.

a.

```
?fandango
```

```
## starting httpd help server ... done
```

The data frame has 146 rows, each represeting a movie and 23 columns. It contains films and film ratings from different sources such as Rotten Tomatoes (users and critics), Metacritic (users and critics), IMDb, and Fandango (users and critics).

b. One way to do this would be:

```r
rt <- fandango$rottentomatoes
mc <- fandango$metacritic
sum(rt)
```

```
## [1] 8884
```

```r
mean(rt)
```

```
## [1] 60.84932
```

```r
median(rt)
```

```
## [1] 63.5
```

```r
min(rt)
```

```
## [1] 5
```

```r
max(rt)
```

```
## [1] 100
```

```r
sum(mc)
```

```
## [1] 8586
```

```r
mean(mc)
```

```
## [1] 58.80822
```

```r
median(mc)
```

```
## [1] 59
```

```r
min(mc)
```

```
## [1] 13
```

```r
max(mc)
```

```
## [1] 94
```

We could also simply use the function `summary()`

```
summary(rt)
```

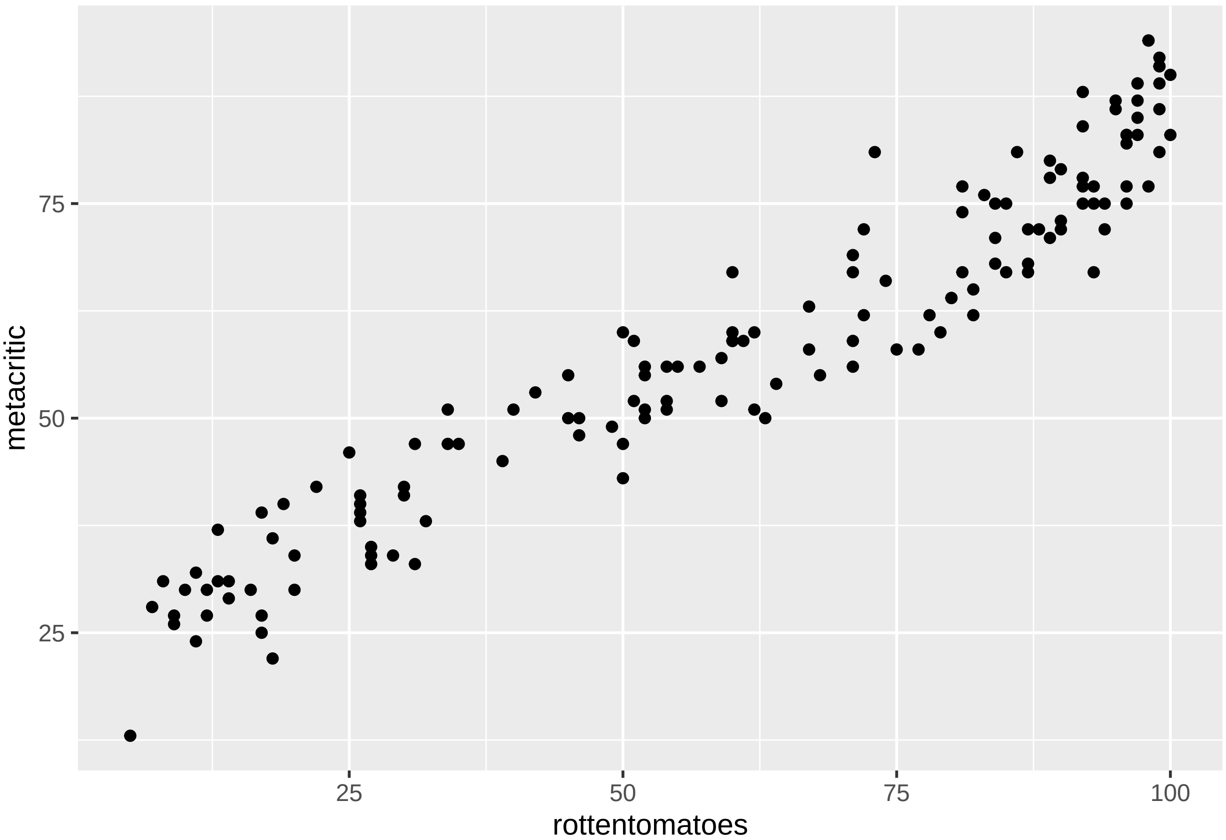
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    5.00   31.25   63.50   60.85   89.00  100.00
```

```
summary(mc)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   13.00   43.50   59.00   58.81   75.00   94.00
```

c.

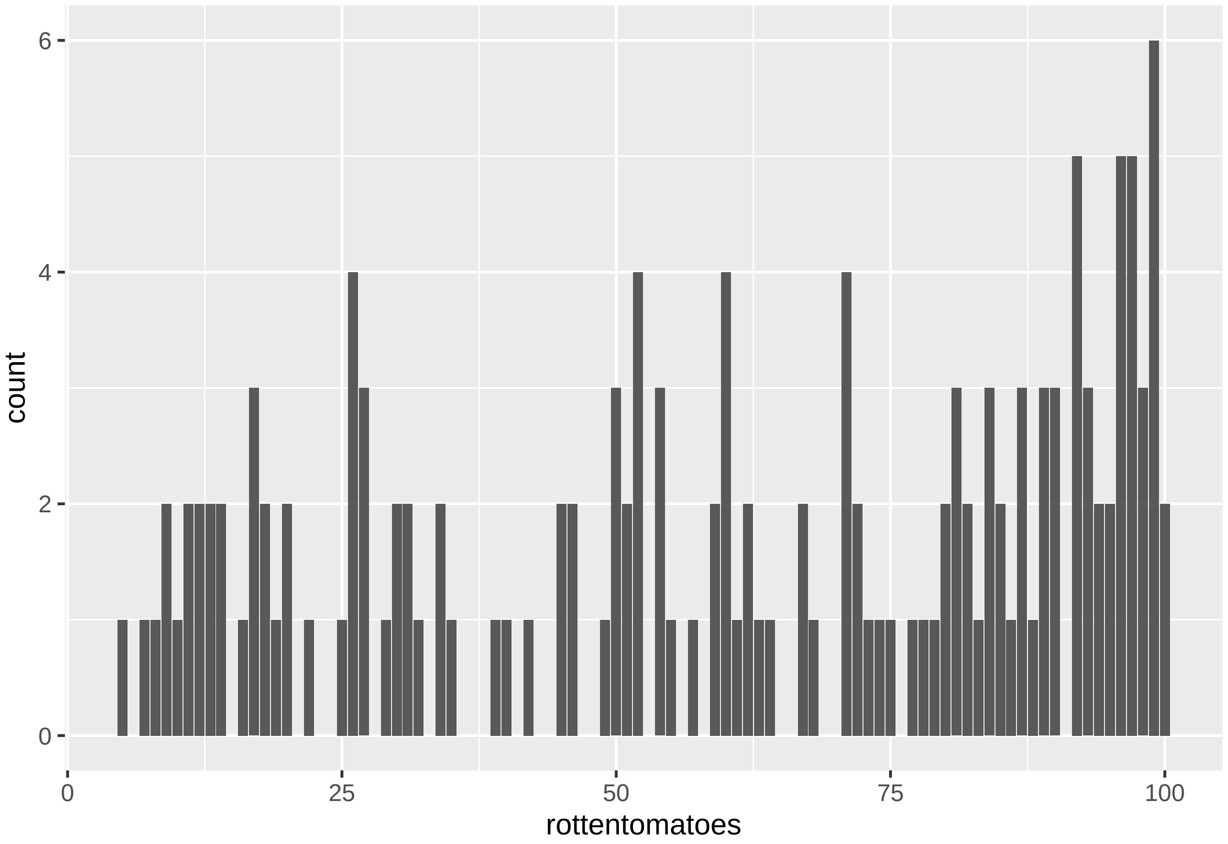
```
#scatterplot for rottentomatoes against metacritic
ggplot(data = fandango, mapping = aes(rottentomatoes, metacritic)) + geom_point()
```



There is a strong linear positive relation between the rottentomatoes and metacritic ratings as observed from the scatterplot

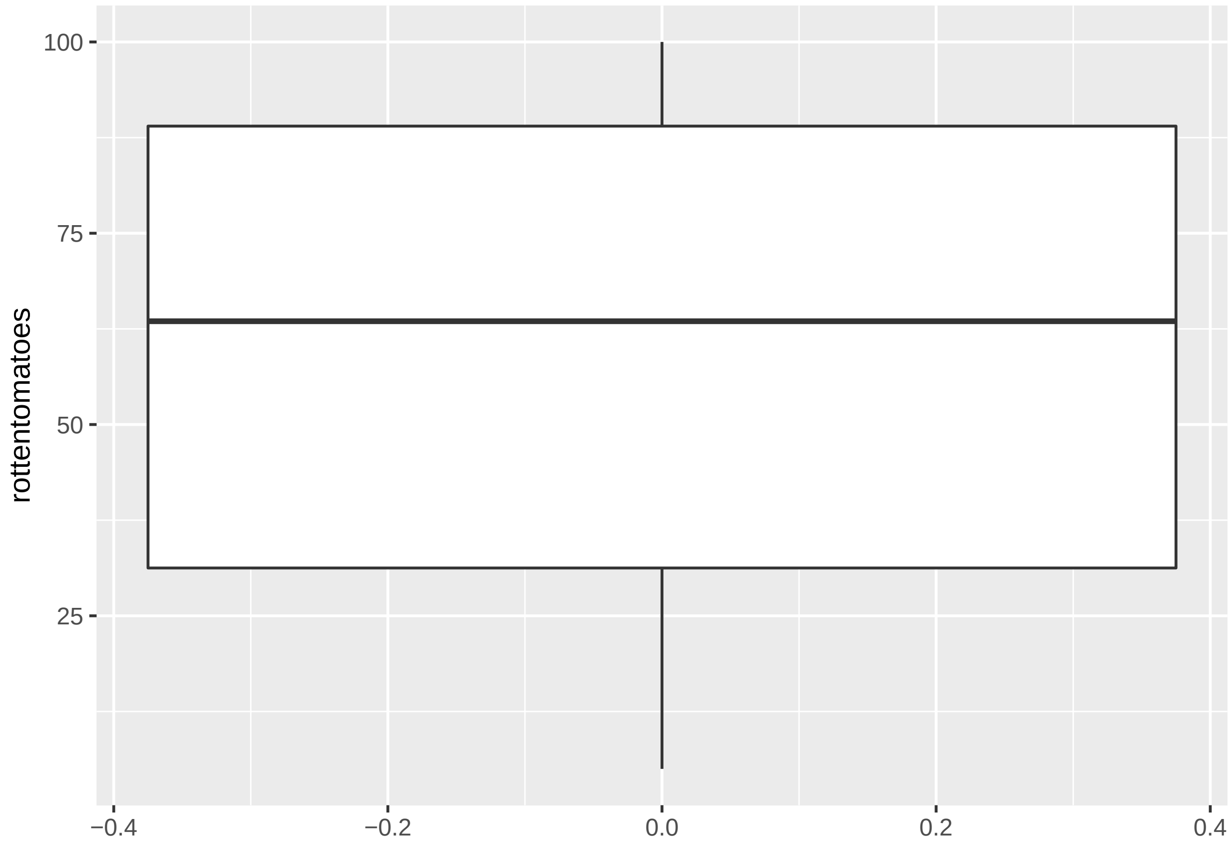
d.

```
#barplot
ggplot(fandango, mapping = aes(rottentomatoes)) +
  geom_bar()
```



There are observations for very low values (around 5) to highest (at 100). There are a few more observations for higher values than other values, relatively speaking.

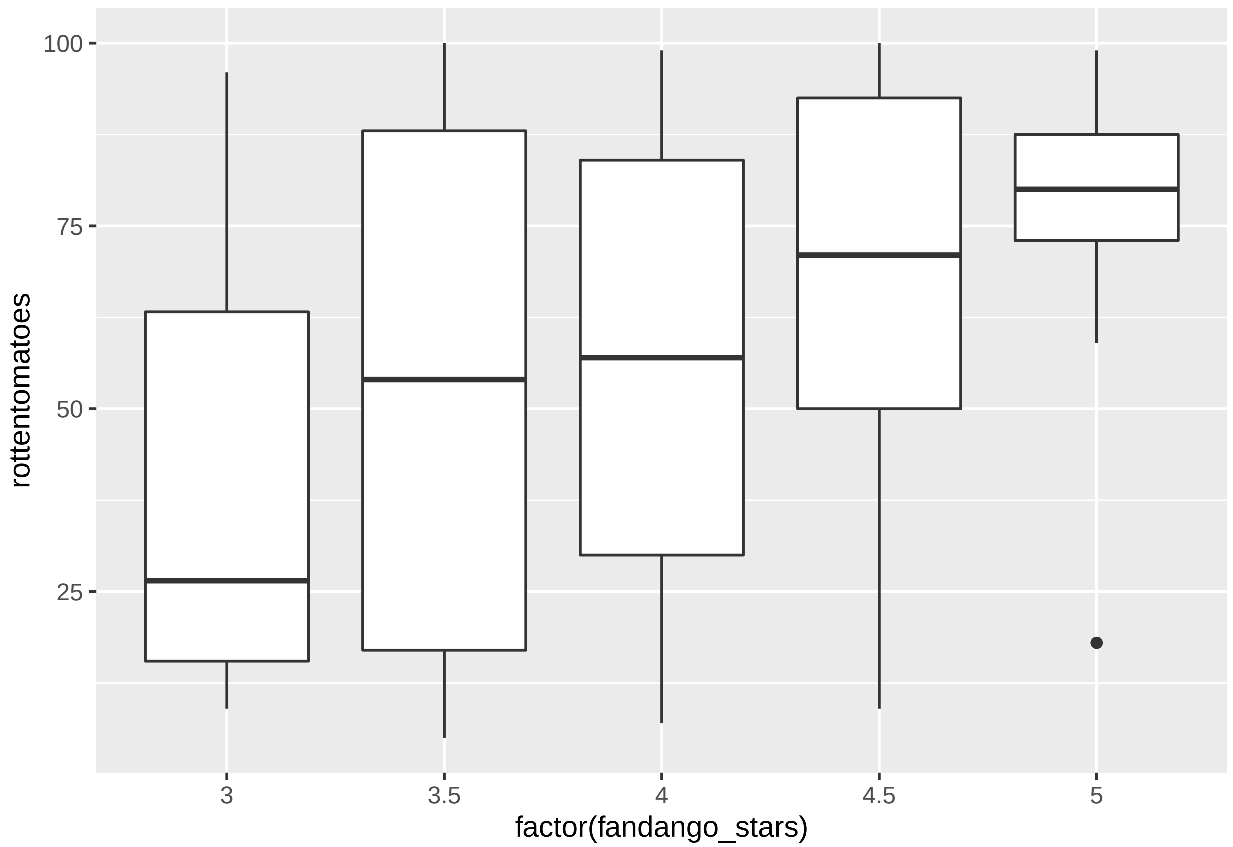
```
#boxplot
ggplot(fandango, mapping = aes(y = rottentomatoes)) +
  geom_boxplot()
```



The boxplot matches the summary of values obtained in part b. the height of the lower and upper edges of the box correspond to the first and third quartile values, respectively; the height of the horizontal line inside the box corresponds to the median; and the lower and upper heights of the whiskers represent the minimum and maximum values, respectively. The distribution of these data is slightly left skewed.

e.

```
#side-by-side boxplot of rottentomatoes scores split by fandango_stars
ggplot(fandango, mapping = aes(x = factor(fandango_stars),y=rottentomatoes)) +
  geom_boxplot()
```



While the overall association between fandango_stars and rottentomatoes is positive, but it's not too strong. For example, it's interesting to see movies that have receive 3.5 starts from fandango vary greatly for rottentomatoes.