Monday Lab Review Worksheet Answers

BSS Stat 20

2022-07-11

(Week 1) - MCU Films

How many observations does this data frame have? - 23

How many variables does this data frame have? - five or six, depending on how time is coded. If five, students might say that time is coded as minutes.

What is the unit of observation? - A single movie within the MCU Universe.

(Week 1) - Smoking Habits

What does each row of the data frame represent? - A single UK resident.

How many participants were included in the survey?- 1,693.

Identify the type of each variable in the Taxonomy of Data. - `sex` is categorical nominal; `age` is numerical continuous; `marital_status` is categorical nominal; `gross_income` is categorical ordinal; `smoke` is categorical nomimal (could be justified as ordinal); `amount` is numerical continuous (both weekend and weekday).

(Week 1-2) - Views on Immigration

There are 910 total survey participants.

part a - Marginal Proportions

```
(57 + 121 + 179 + 15)/910
```

```
## [1] 0.4087912
```

part b - Marginal Proportions

```
(57 + 101 + 120)/910
```

```
## [1] 0.3054945
```

part c - Joint Proportions

```
(57)/910
```

```
## [1] 0.06263736
```

part d - Conditional Proportions

Conservatives Watch the wording difference between this subpart and **part c**; *and* versus *also are*.

```
(57)/(57+121+179+15)
```

```
## [1] 0.1532258
```

```
(101)/(101+28+45+1)
```

Liberals

```
## [1] 0.5771429
```

```
(120)/(120+113+126+4)
```

Moderates

```
## [1] 0.3305785
```

(Week 2) - MCU Films

The correct code is:

Plot 1

```
library(tidyverse)
ggplot(mcu_films, aes(x = gross_us)) +
    geom_histogram() +
      theme_gray(base_size = 8) +
  ggtitle("Box Office Totals in the United States") +
  xlab("Total") + ylab("Count")
```

Plot 2

```
p2 <- ggplot(mcu_films, aes(y = gross_world)) +
    geom_boxplot() +
      theme_gray(base_size = 8) +
  ggtitle("Box Office Totals Worldwide") +
  ylab("Total")
```

Plot 3

```
p3 <- ggplot(mcu_films, aes(x = gross_us,
                            y = gross_world)) +
    geom_point() +
      theme_gray(base_size = 8) +
  ggtitle("Box Office Totals", subtitle = "United States against World") +
  xlab("US Total") + ylab("World Total")
```

(Week 2) - NBA

First, students need to run the code provided to them.

```
library(rvest)
url <- "https://www.basketball-reference.com/playoffs/NBA_2022_per_game.html"

NBA <- (read_html(url) %>% html_table())[[1]] %>% filter(Tm != "Tm")
```

```
NBA <- NBA %>% mutate(Conference =
                 ifelse(Tm %in% c("ATL","TOR", "MIA", "BOS", "CHI",
                                  "BRK", "MIL", "PHI"), "Eastern",
                                   "Western"))
```

part a

The correct columns to work with are Games (G) and Games Started (GS). Note, we need to do some data type changing. Here is *a* possible solution.

```
NBA_3 <- NBA %>% mutate(G = as.numeric(G)) %>% mutate(GS = as.numeric(GS)) %>%
  mutate(G_GS_Ratio = GS/G) %>% filter(G_GS_Ratio > .67) %>%
  select(-G_GS_Ratio)
```

part b

Here is *a* possible solution. Again we need to change data types.

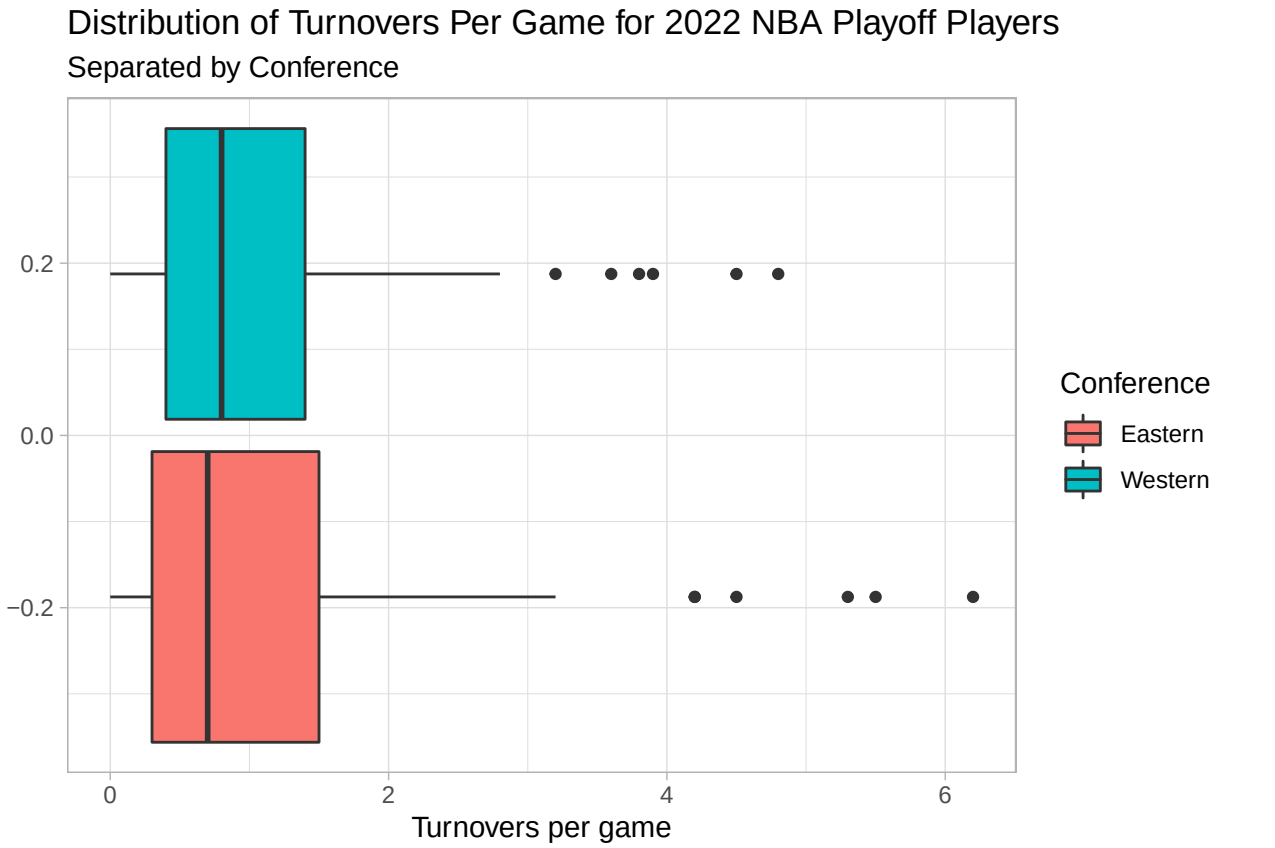
```
NBA_3 %>% mutate(TOV = as.numeric(TOV)) %>%  group_by(Conference) %>%
  summarise(Mean = mean(TOV), Median = median(TOV),
            IQR = IQR(TOV), SD = sd(TOV), MAD = mad(TOV))
```

```
## # A tibble: 2 x 6
##   Conference  Mean Median   IQR    SD   MAD
##   <chr>      <dbl>  <dbl> <dbl> <dbl> <dbl>
## 1 Eastern     2.11   1.75  1.5   1.46  1.11
## 2 Western     1.93   1.6   1.35  1.13  1.19
```

part c

Here is *a* possible solution.

```
NBA %>% mutate(TOV = as.numeric(TOV)) %>%
  ggplot() + geom_boxplot(mapping = aes(x = TOV, fill = Conference)) +
  ggtitle("Distribution of Turnovers Per Game for 2022 NBA Playoff Players",
          subtitle = "Separated by Conference") + xlab("Turnovers per game") + theme_light()
```



part d

This is *a* possible answer.

Overall, there is not that much of a difference. The Median for the Western conference is higher, but the Eastern conference distribution is a little more variable, but overall both distributions exhibit a right skew and their medians are relatively close to one another.

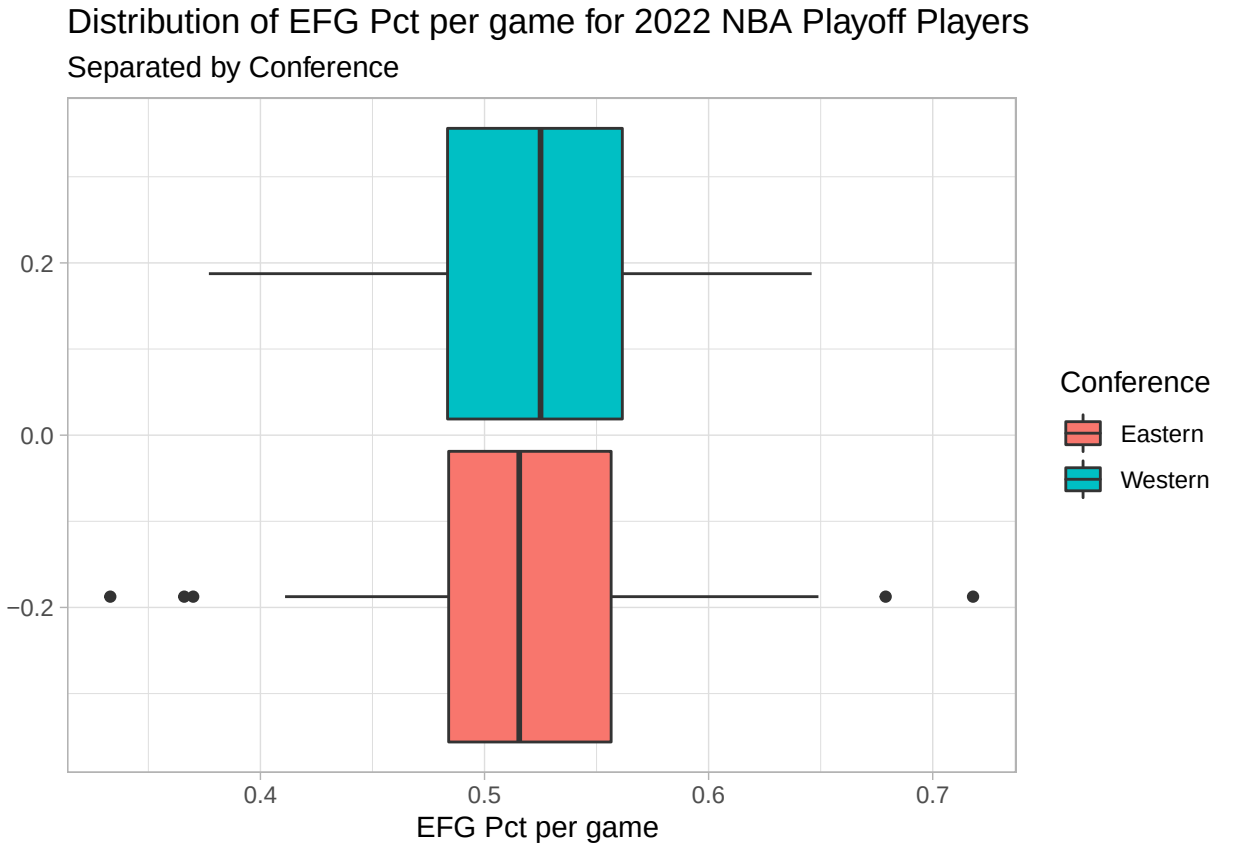
part e

Here is *a* possible solution (code-wise).

```
NBA_3 %>% mutate(`eFG%` = as.numeric(`eFG%`)) %>%  group_by(Conference) %>%
  summarise(Mean = mean(`eFG%`), Median = median(`eFG%`),
            IQR = IQR(`eFG%`), SD = sd(`eFG%`), MAD = mad(`eFG%`))
```

```
## # A tibble: 2 x 6
##   Conference  Mean Median    IQR     SD    MAD
##   <chr>      <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
## 1 Eastern    0.521  0.516 0.0725 0.0824 0.0630
## 2 Western    0.523  0.525 0.078  0.0672 0.0578
```

```
NBA_3 %>% mutate(`eFG%` = as.numeric(`eFG%`)) %>%
  ggplot() + geom_boxplot(mapping = aes(x = `eFG%`, fill = Conference)) +
  ggtitle("Distribution of EFG Pct per game for 2022 NBA Playoff Players",
          subtitle = "Separated by Conference") + xlab("EFG Pct per game ") + theme_light()
```



This is *a* possible answer to the written portion.

We see that the variability in the two distributions is pretty similar, save for some outliers in the Eastern conference on both sides. However, it looks like the median eFG% is larger in the Western Conference. This might suggest that the Western Conference is a *slightly* better conference when it comes to shooters.

(Week 3) - Roulette

part a

I would create a random variable X that takes two outcomes: 1 and -1, with probabilities 18/38 and 20/38, respectively. Note that it does not matter which color you choose; the probabilities remain the same.

part b.

The Expected Value is given by

$$\mathbb{E}(X) = [1 \cdot P(X = 1)] + [-1 \cdot P(X = -1)] = \frac{18}{38} - \frac{20}{38} = -\frac{2}{38} \approx -5cents$$

.

The Variance (using the Computational Formula):

$$Var(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$$

and we have that since X takes either -1 or 1, squaring it means that X^2 will always take the value of 1. Hence we have the following calcuation of $\mathbb{E}X^2$:

$$\mathbb{E}(X^2) = [1 \cdot P(X^2 = 1)] + [1 \cdot P(X^2 = 1)] = 1$$

so that

$$Var(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = 1 - (-\frac{2}{38})^2 = 1 - \frac{4}{38^2} = \approx 0.997 dollars^2$$

part c

Note: We are not playing 100 times, just merely playing once with a larger bet (How would the calculation change if the former was the case?).

Let's use some of the rules for expected value and variance. Let $a = 100$ be a constant. Then we have that

$$\mathbb{E}(aX) = a \cdot \mathbb{E}(X) = 100 \cdot \mathbb{E}(X) \approx -500cents = -5dollars$$

.

$$Var(aX) = a^2 \cdot Var(X) = 1000 \cdot Var(X) \approx 1003 dollars^2$$

.

(Week 3) - Penguins

Choice *a.* and choice *c.* are examples of generalizations.

(Week 2) - Penguins

First plot:

```
p2 <- ggplot(penguins,
             aes(x = species)) +
  geom_bar() +
  xlab("Species") + ylab("Count") +
      theme_gray(base_size = 8)
```

Second plot:

```
p3 <- ggplot(penguins,
             aes(x = flipper_length_mm,
                 y = body_mass_g,
                 color = island)) +
  geom_point() + xlab("Flipper Length") + ylab("Body Mass") +
      theme_gray(base_size = 8)
```

(Week 2) - Flights

```
flights %>%
  ggplot(aes(x = dep_delay)) +
  geom_histogram(binwidth = 15) +
  xlim(c(-50, 100))
flights %>%
  summarise(med_delay = median(dep_delay, na.rm = T),
            iqr_delay = IQR(dep_delay, na.rm = T))
```

The statistical summaries should be median and IQR since we have a skewed distribution. The shape students should describe as unimodal and right skewed.

(Week 3) - Distributions

The answer is $X \sim Binomial(n = 8, p = .8)$