Monday Lab Review Worksheet

BSS Stat 20

2022-07-11

NOTICE

Below are some relevant questions which have been pulled from previous exams and some problems I selected from our problem sets and lectures. Much of the material from last semester is available to you at the course Github, linked here.

I suggest you use the chalkboards in the room with your groups to work out the problems and take turns writing! This is a great way to study the material.

There is no order to the material on this worksheet, though I have put the relevant week next to each problem. Do whatever you and you lab group are most uncomfortable with first. *There's no pressure to finish all of these, either!* I just put a lot of exercises here so you guys would have a lot to choose from when it comes to review.

Questions

(Week 1) Marvel Cinematic Universe films.

The data frame below contains information on Marvel Cinematic Universe films through the Infinity saga (a movie storyline spanning from Ironman in 2008 to Endgame in 2019). Box office totals are given in millions of US Dollars.

		Length				Gross	
	Title	Hrs	Mins	Release Date	Opening Wknd US	US	World
1	Iron Man	2	6	5/2/2008	98.62	319.03	585.8
2	The Incredible Hulk	1	52	6/12/2008	55.41	134.81	264.77
3	Iron Man 2	2	4	5/7/2010	128.12	312.43	623.93
4	Thor	1	55	5/6/2011	65.72	181.03	449.33
5	Captain America: The First Avenger	2	4	7/22/2011	65.06	176.65	370.57
6	Marvel's The Avengers	2	23	5/4/2012	207.44	623.36	1518.82
7	Iron Man 3	2	10	5/3/2013	174.14	409.01	1214.81
8	Thor: The Dark World	1	52	11/8/2013	85.74	206.36	644.78
9	Captain America: The Winder Soldier	2	16	4/4/2014	95.02	259.77	714.42
10	Guardians of the Galaxy	2	1	8/1/2014	94.32	333.72	773.34
...	...	...	...	...	...	...	...
22	Avengers: Endgame	3	1	4/26/2019	357.12	858.37	2797.8
23	Spiderman: Far from Home	2	9	7/2/2019	92.58	390.53	1131.93

How many observations and how many variables does this data frame have? What is the observational unit (what each row corresponds to)?

(Week 1) Smoking habits of UK residents.

A survey was conducted to study the smoking habits of 1,691 UK residents. Below is a data frame displaying a portion of the data collected in this survey. A cell with NA indicates that data for that variable was not available for a given respondent.[1]

						amount	
	sex	age	marital_status	gross_income	smoke	weekend	weekday
1	Female	61	Married	2,600 to 5,200	No	NA	NA
2	Female	61	Divorced	10,400 to 15,600	Yes	5	4
3	Female	69	Widowed	5,200 to 10,400	No	NA	NA
4	Female	50	Married	5,200 to 10,400	No	NA	NA
5	Male	31	Single	10,400 to 15,600	Yes	10	20
...	...	...	...	...	...	NA	NA
1691	Male	49	Divorced	Above 36,400	Yes	15	10

a. What does each row of the data frame represent?

b. How many participants were included in the survey?

c. Identify the type of each variable in the Taxonomy of Data.

[1]The `smoking` data used in this exercise can be found in the **openintro** R package.

(Week 1) Views on immigration.

Nine-hundred and ten (910) randomly sampled registered voters from Tampa, FL were asked if they thought workers who have illegally entered the US should be (i) allowed to keep their jobs and apply for US citizenship, (ii) allowed to keep their jobs as temporary guest workers but not allowed to apply for US citizenship, or (iii) lose their jobs and have to leave the country. The results of the survey by political ideology are shown below.

Response	Conservative	Liberal	Moderate
Apply for citizenship	57	101	120
Guest worker	121	28	113
Leave the country	179	45	126
Not sure	15	1	4

When answering these questions, also try to think about what type of proportion needs to be calculated: conditional, joint, or marginal.

a. What percent of these Tampa, FL voters identify themselves as conservatives?

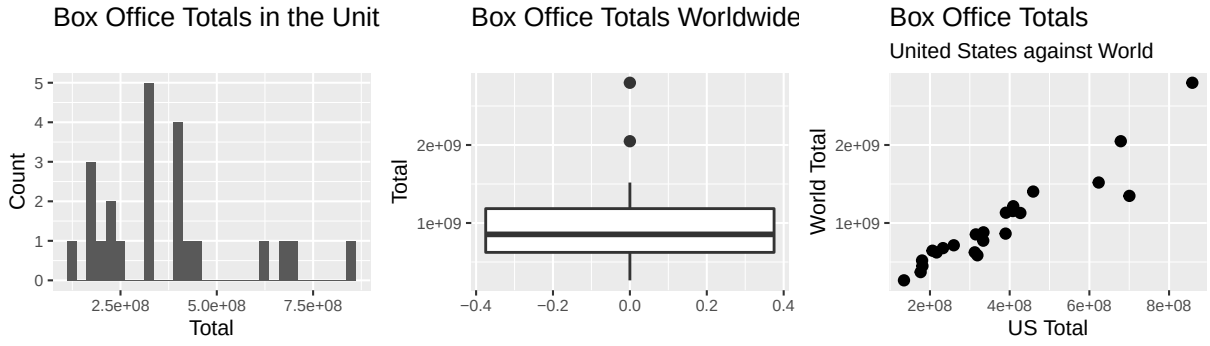
b. What percent of these Tampa, FL voters are in favor of the citizenship option?

c. What percent of these Tampa, FL voters identify themselves as conservatives and are in favor of the citizenship option?

d. What percent of these Tampa, FL voters who identify themselves as conservatives are also in favor of the citizenship option? What percent of moderates share this view? What percent of liberals share this view?

(Week 2) - Marvel Cinematic Universe films.

The following three plots come from a data set called `mcu_films` inside the `openintro` package. Please write out the `ggplot2` code that will produce each one.



(Week 2) - NBA

Solutions - See bCourses. This problem was included in Problem Set 2.

The hometown Golden State Warriors won another NBA Championship recently with their victory against the Boston Celtics. Some would say that the Warriors recent run of success is all the more impressive

considering that the league that they play in during the season and during the playoffs, the Western Conference, has generally had the more talented teams than its counterpart, the Eastern Conference. However, the Eastern Conference has gotten much stronger in recent years and arguably now has the better players among the two leagues. How did the two conferences fare in this year's playoffs?

To think about this question, we can pull individual player per game statistics during the 2022 NBA playoffs from Basketball Reference (linked here). We can read the data in with the following code:

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

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

The following line of code also adds a `Conference` column which takes the values of `Eastern` and `Western` depending on the team which each player is a part of.

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

Let's look at a few of Dean Oliver's famous "Four Factors" to see the state of parity between the two conferences. We will examine turnovers per game `TOV` and effective field goal percentage `eFG%`. The latter statistic accounts for the fact that a 3-point shot is worth more than a 2-point shot.

---

part a

First, trim the data to only include players who have started more than 67 percent of the games they played in. (Use the provided link to find the right columns to perform the trimming on if you are unsure just by looking at the data set which columns to work with). Save the result.

part b

Now, return a data frame with summary statistics for turnovers per game, by conference. Include the mean, median, IQR, standard deviation and median absolute deviation.

part c

Create a boxplot(s) of turnovers per game, separated by conference.

part d

Based off your results of *part b* and *part c*, what is your conclusion about the distributions of turnovers for game for players in the Eastern Conference versus those in the Western conference?

part e

Repeat parts b, c, and d for effective field goal percentage. **NOTE: you will need to surround `eFG%` with the back-tick when accessing it in your pipeline.** This is because `%` is a special character in R.

(Week 3) - Roulette

- Roulette is a very simple but popular casino game.
- You spin a wheel with 38 colored slices:
  - 18 red, 18 black, 2 green
- You pick either red or black. If the wheel lands on your color, you win!

- Now imagine you bet $1 on a particular spin. If you win, you get your dollar back, plus an additional dollar (you win one dollar). Otherwise, you have lost a dollar.

part a

Create a random variable that records your winnings from one game.

part b

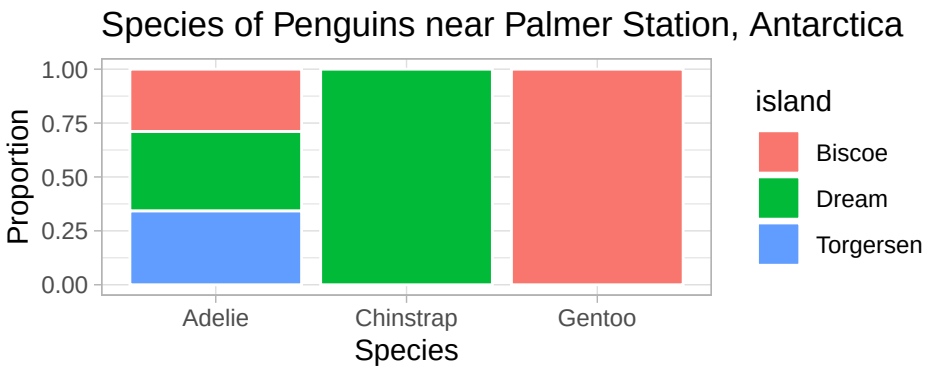
Calculate the expected value and variance of this random variable.

part c

Most of the time people will be betting more than a dollar. Now we will bet $100 instead. With a modified random variable from part 1, repeat part 2.

(Week 3) - Penguins

The following plot was made using the `penguins` data set within the `palmerpenguins` R package.



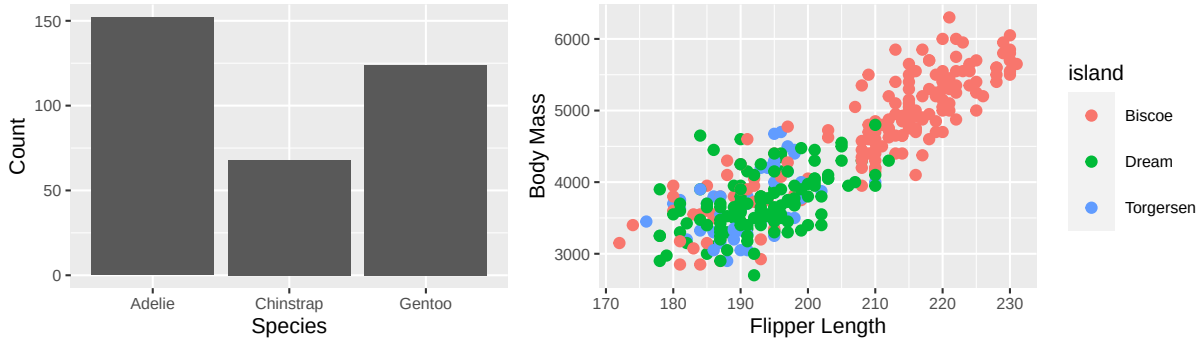
Which of the following are generalizations that could be made based off the above bar chart? **Select all that apply**.

a. All Chinstrap penguins near Palmer Station, Antarctica are found on Dream island.

b. All Gentoo penguins sampled near Palmer Station, Antarctica are found on Biscoe island.

c. Adelie penguins have the largest diaspora (the spread across islands) out of any of the penguin speices.

(Week 2) - Penguins



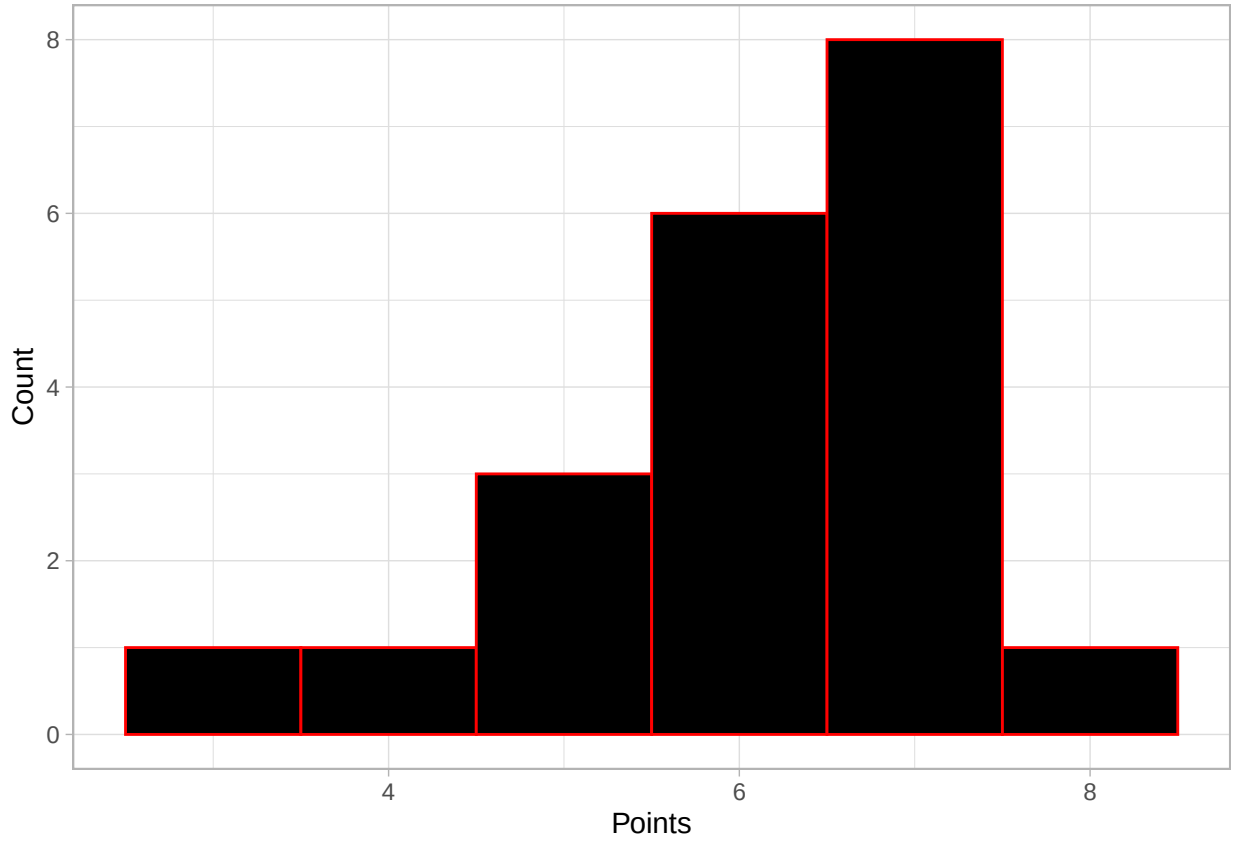
Here are some more plots made using the `palmerpenguins` data set. Reproduce these plots accurately with `ggplot()` code.

(Week 2) - Flights

Let's go back to the `flights` data set from Lab 2!

Create a histogram showing the distribution of departure delays for all flights. Describe in words the shape and modality of the distribution and, using numerical summaries, (i.e. summary statistics) its center and spread. Be sure to use measures of center and spread that are most appropriate for this type of distribution. If you want set the limits of the x-axis to focus on where most of the data lie, use the `xlim()` layer (**not required for you to know for the test**).

(Week 3) - Distributions



What named distribution (with parameters) best describes the distribution of these 20 realizations of a random variable?

a. $X \sim DiscreteUnif(a = 4, b = 8)$

b. $X \sim Binomial(n = 8, p = .8)$

c. $X \sim Bernoulli(p = .8)$

d. $X \sim Binomial(n = 8, p = .3)$