HW7

Rhea Toves

10/19/2021

–

Problem 1 - This question is related to your personal data set project. The purpose is to help you narrow down your potential data topics and sources by identifying some specific research questions you might be able to answer with your dataset.

(A) Write the broad topic area that you are interested in. For example, individual basketball statistics or national election data or salmon aquaculture.

- I'm interested in a lot of topics for this research project. I want to look into data from the music app, Spotify, specifically their growth throughout the years, top genres, how top genres vary in different countries/states, top artists, amount of listens throughout the years, etc.

(B) Based on your topic from part (a), list three specific questions that you might like to be able to answer with your data.

- 1. How many listens have there been from 2015-2021? (2) What is the top genre listened to? How does this vary in the U.S.? (3) How much revenue did Spotify generate in 2020?

(C) Based on the questions you listed in part (b), list one potential source that might have relevant data to address them.

- (https://www.businessofapps.com/data/spotify-statistics/ (https://www.businessofapps.com/data/spotify-statistics/)) This article includes information on Spotify's users, revenue, annual users, users by region, subscribers, etc.

–

Problem 2 - In your own words, provide definitions of the following:

(A) Random Variable - a random variable is a variable; a numerical outcome of a random event.

(B) Probability Distribution - probability distribution displays the probabilities of all possibe outcomes in an event.

(C) Sample Space - sample space is a set that holds all possible outcomes in an event.

(D) Bernoulli Trial - this is an experiment with only two possible outcomes. For example, a coin flipping experiment would be a Bernoulli Trial.

(E) Random Sample - a random sample is an "X" number of variables or objects being chosen from a sample space or set.

–

Problem 3 - Provide short answers to the following:

(A) Write the formula for the probability density function for $N(-3, 4)$.

- f(x)= $1/\sqrt{(2\pi(4))}e^{(-x-(-3))}2/(2(4))$

(B) What is the probability that a number drawn from $N(-3, 4)$ is greater than -3?

- P(X > -3) = 0.5

(C) For a probability distribution over the real line, if you are told that the probability of a draw from that distribution being less than 5 is $.2$ and the probability of a draw being less than 7 is $.25$, what is the probability of a number between 5 and 7 being drawn?

- P(X > 5) = 0.2, P(X > 7) = 0.25 then P(5 < X < 7) = 0.25 - 0.2 = 0.05.

(D) Use the pnorm function to compute the probability that a draw from $N(-3, 4)$ is between 5 and 7.

```
## [1] 0.0165
```

_

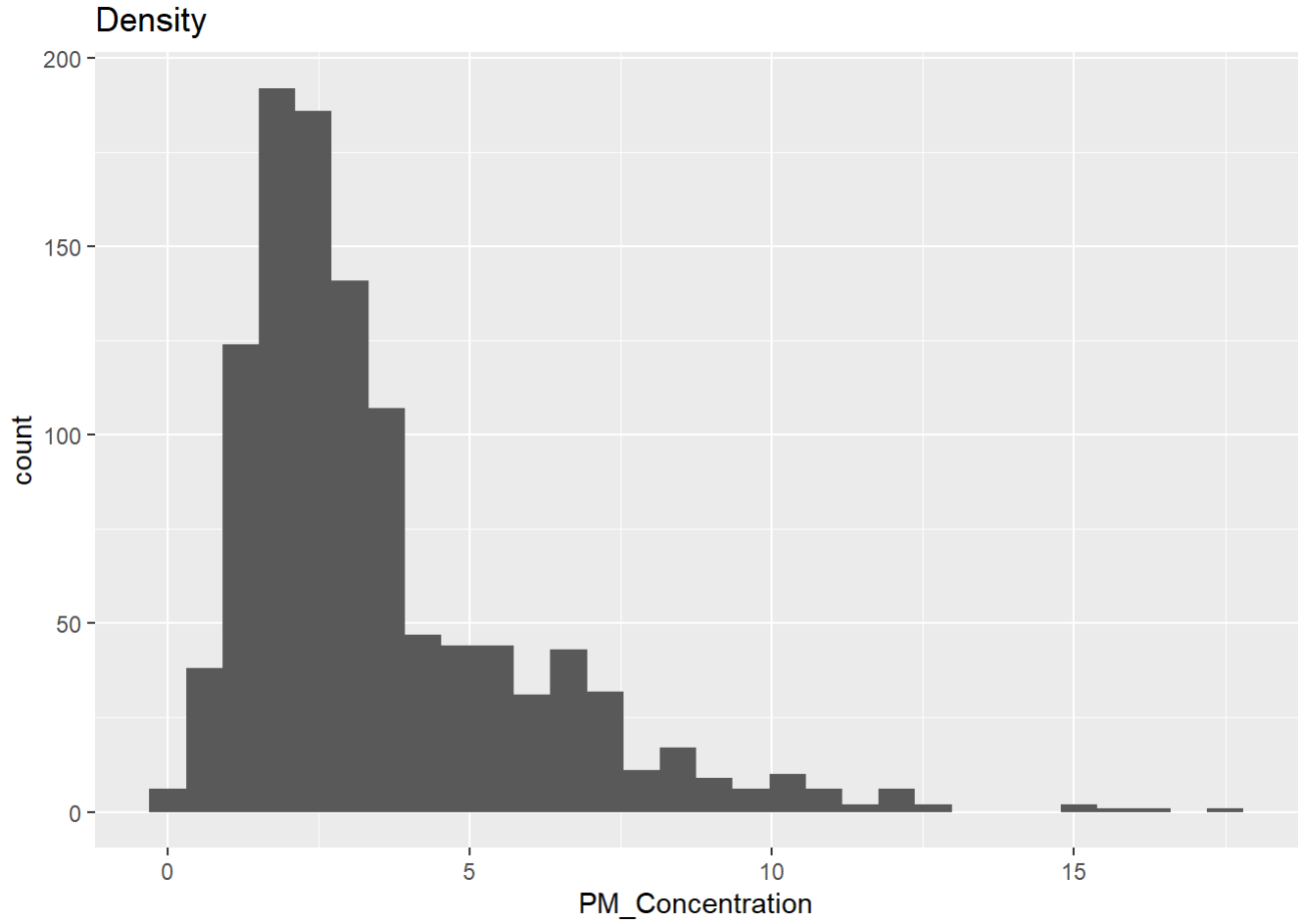
Problem 4 - Load the Air Quality Data into R as a dataframe.

(A) Compute the mean and standard deviation of the PM_Concentration column

```
## [1] 3.52624
```

```
## [1] 2.423968
```

(B) Make a density histogram of the PM_Concentration column and overlay a plot of the normal distribution with mean and standard deviation from part (B).



(C) Is the normal distribution a good fit for this data? Why or why not?

- The normal distribution is a good fit for this data because this curve is similar to the density graph.