HW7

Grace Okamoto

10/20/2021

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.

NEW DATA 115 REPOSITORY (because I panic deleted my other Github account after accidentally following someone while I was snooping on their profile) https://github.com/notcaughtyet/DATA115

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

Cybersecurity analysis

(B) Based on your topic from part (a), list three specific questions that you might like to be able to answer with your data. Examples like Who had the largest difference between free throw and three point averages in the NBA in last year's playoffs?'' orWhich Congressional candidate most outperformed their local poling in 2018?’’ or What is the percent increase in farmed to wild salmon consumed in the US between 1980 and 2010?'' are at the right level of specificity, while something likeWhat was the best basketball team in history?!?!?!?’’ is too general.

What keywords appear most frequently in solutions proposed to cybersecurity breach reports? Which domains are the most vulnerable gateways for cyber attacks? What kinds of data chunks are the most frequently targeted?

(C) Based on the questions you listed in part (b), list one potential source that might have relevant data to address them. This could be a website (https://www.sports-reference.com/), a government branch or organization (the USDA performs a national aquaculture census every five years and the UN conducts national level aquaculture sector overviews), or another organization (the MIT election data + science lab, 538, or the NYT). Note that you will not be required to use the source you list here for the actual project, this is just an opportunity to do a little exploring.

https://nesg.ugr.es/nesg-ugr16/

Problem 2

In your own words, provide definitions of the following:

(A) Random Variable

A variable whose value can be random within the constraint of the event.

(B) Probability Distribution

A function that plots the probability for any particular outcome.

(C) Sample Space

The set of all possible outcomes.

(D) Bernoulli Trial

A single binomial trial.

(E) Random Sample

An event chosen randomly from a population.

Problem 3

Provide short answers to the following:

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

$1/(\mathrm{sqrt}(2\mathrm{pi}(4^{2)))e^{((-(x-(-3))}2)/24}2)$

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

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.

```r
pnorm(7, -3, 4) - pnorm(5, -3, 4)
```

```
## [1] 0.01654047
```

Problem 5

(A) Load the Air Quality Data into R as a dataframe.

```r
PAC <- read.csv("Pullman_Air_Quality.csv")
```

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

```r
mean(PAC$PM_Concentration)
```

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

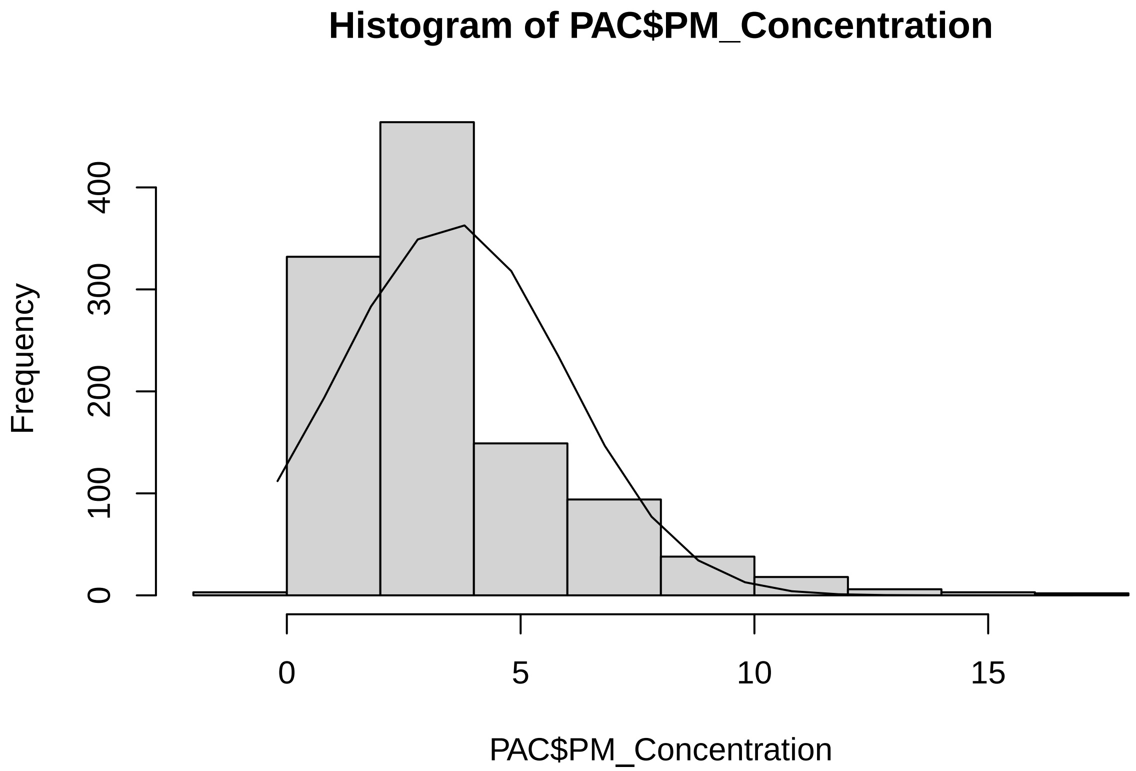
```r
sd(PAC$PM_Concentration)
```

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

(C) 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).

I'm going to be honest here, I really don't know what I did in the following chunk of code - it all came from https://www.statmethods.net/graphs/density.html. Thanks Google

```
h<-hist(PAC$PM_Concentration)
xfit<-seq(min(PAC$PM_Concentration),max(PAC$PM_Concentration))
yfit<-dnorm(xfit,mean=3.52624,sd=2.423968)
yfit <- yfit*diff(h$mids[1:2])*length(PAC$PM_Concentration)
lines(xfit, yfit)
```



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

I'd say it's an OK fit for the data, but certainly not perfect. For instance, the normal distribution tapers off toward the end, making the value appear as zero, whereas the histogram shows that values above 15 still appear occasionally. In addition, the normal distribution doesn't show just how high the frequency for some of the values are, particularly the peak of this graph.