Homework_1

Warren Geither

8/29/2020

Problem 1

In class we discussed the Binomial distribution. Another discrete distribution is the Poisson distribution.

a.) Look online and in your book collection to read about the Poisson distribution. Is f(y|lambda) a pdf or pmf? - pmf

b.) What is the expected value E(y)? Prove it. - see attached pdf

c.) What is the variance var(y)? Prove it. - see attached pdf

d.) The number of mistakes a letter stuffing robot makes in an hour can be considered to be a Poisson random variable. Let the random variable Y have a Poisson distribution with parameter lambda=2.3.

- What is the probability that the robot makes 2 mistakes in an hour?

```r
dpois(2, lambda = 2.3)
```

```
## [1] 0.2651846
```

- What is the probability the robot makes at least 2 mistakes in an hour?

```r
1 - sum(dpois(0:2, lambda = 2.3))
```

```
## [1] 0.4039612
```

- What is the the probability that the robot makes 2.3 mistakes in an hour?
  - the domain is the Natural number union {0}, so we cant put 2.3 in the pmf

```r
dpois(2.3, lambda = 2.3)
```

```
## Warning in dpois(2.3, lambda = 2.3): non-integer x = 2.300000
```

```
## [1] 0
```

- What is the probability that the robot makes 0 mistakes in an hour?

```r
dpois(0, lambda = 2.3)
```

```
## [1] 0.1002588
```

- What is the probability that the robot makes less than 6 mistakes in an hour?

```r
sum(dpois(0:5, lambda = 2.3))
```

```
## [1] 0.9700243
```

e.) The following data measure the number of children arriving at a local daycare between the hours of 8-9 AM for 25 days. Assume these data arise from a Poisson distribution. - Based on the data, there are an average of (approximately) 7 children per day dropped off at the daycare between 8-9AM

```
y<-c(11,7,2,7,4,8,13,3,6,6,15,8,2,4,5,11,11,4,9,3,9,8,5,9,6)

est_lambda = sum(y)/25

print(est_lambda)
```

```
## [1] 7.04
```

Problem 3

The following data set includes measurements on: X, temperature (degrees Fahrenheit), and y, percent butterfat for 10 cows across 20 consecutive days. These data are courtesy of Dr. Jason Osborne at NCSU.

a.) Conduct a descriptive analysis of these data using graphics and summary statistics. Write a few sentences to interpret your analysis in the context of the problem.

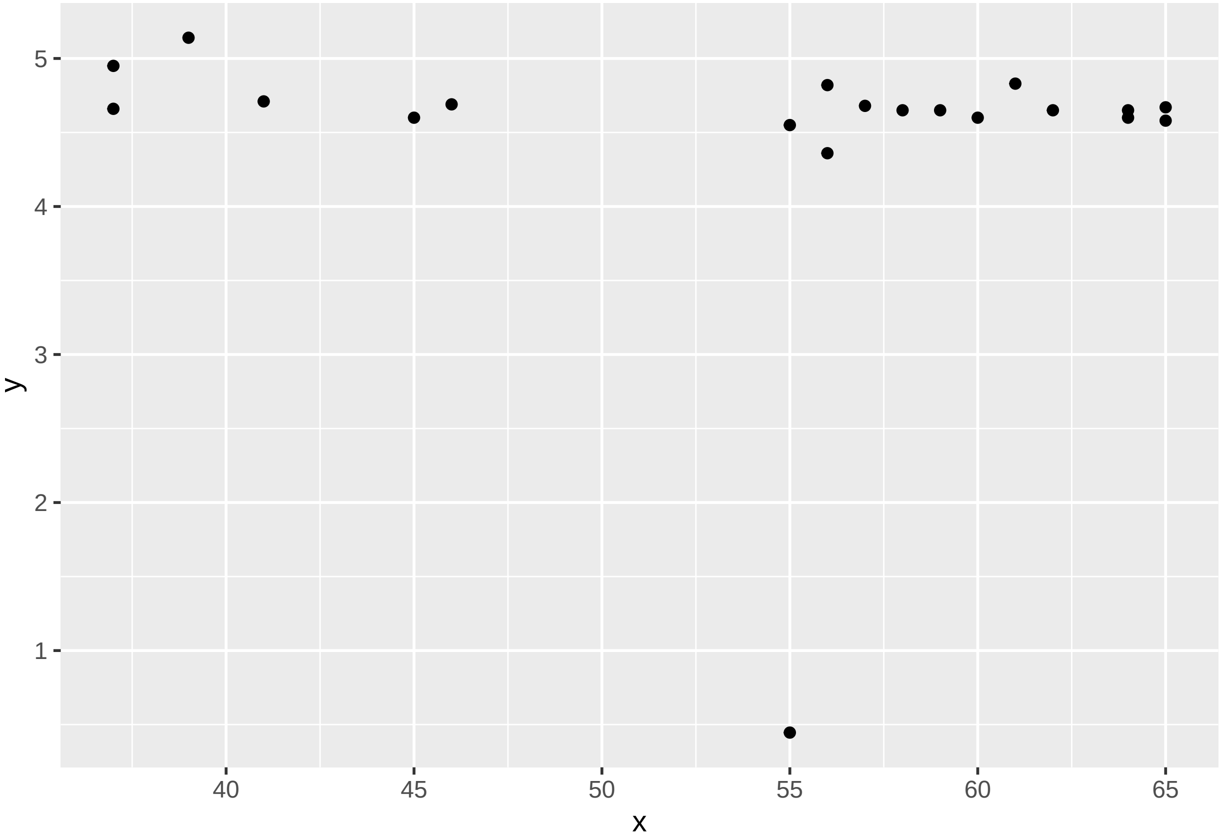
```
# temperature vector
x <- c(64,65,65,64,61,55,39,41,46,59,56,56,62,37,37,45,57,58,60,55)

# percent of butterfat
y <- c(4.65,4.58,4.67,4.60,4.83,4.55,5.14,4.71,4.69,4.65,4.36,4.82,4.65,4.66,4.95,4.60,4.68,4.65,4.6,.44

# bring in ggplot
library("ggplot2")

# create dataframe
df = data.frame("temp" = x, "pfb" = y)

# display data
print(df)
```

```
##    temp   pfb
## 1    64 4.650
## 2    65 4.580
## 3    65 4.670
## 4    64 4.600
## 5    61 4.830
## 6    55 4.550
## 7    39 5.140
## 8    41 4.710
## 9    46 4.690
## 10   59 4.650
## 11   56 4.360
## 12   56 4.820
## 13   62 4.650
## 14   37 4.660
## 15   37 4.950
## 16   45 4.600
## 17   57 4.680
## 18   58 4.650
## 19   60 4.600
## 20   55 0.446
```

```
# run summary stats
summary(df)
```

```
##       temp            pfb       
##  Min.   :37.00   Min.   :0.446  
##  1st Qu.:45.75   1st Qu.:4.600  
##  Median :56.50   Median :4.650  
##  Mean   :54.10   Mean   :4.474  
##  3rd Qu.:61.25   3rd Qu.:4.695  
##  Max.   :65.00   Max.   :5.140  
```

```
# scatterpolot for temp vs. pbf
ggplot(data = df) +
  geom_point(mapping = aes(x,y))
```



The temperature is between 37 and 65 with a mean of 54.10. There is 1 outlier value for the percent butterfat .446% but the first quartile starting at 4.6% This could be do to some kind of measurement error and we may want to remove the data for further analysis. Looking at the scatterplot, we can see what appears to be a small negative correlation between percent butterfat and temperature.

b.) Compute and report Pearson's sample correlation coefficient for these data. What two aspects of the relationship between temperature and butterfat are described by this correlation? Write a few sentences to interpret this statistic.

```
# calculate pearson's coorelation coefficient
cor(x,y, method="pearson")
```

```
## [1] -0.09911088
```

- This coefficient describes the strength and direction of the linear relationship between these variables. They are negatively correlated (as temperature rises, percent of butterfat goes down), however the relationship is not very strong since it is in between -1 and 0.

c.) Describe whether the assumptions underlying your analyses in parts (a) and (b) are valid. Do any aspects of this analysis concern you? - The outlier concerns me, but that could be removed.

Problem 4

Are the following quantities statistics or parameters? Explain your reasoning.

a.) Joseph wonders what the true average Intellectual quotient (IQ) score is among all first year graduate students in the United States. - Parmaeter. Since its the true average of the whole population.

b.) Based on an online survey completed by shoppers of a pet supply chain, Jennifer calculates that 32% of respondents are current or previous cat owners. - statistics, because its based on a sample, not all of her customers did the survey (probably)

c.) After a remarkable effort, the Census Bureau successfully counts every single person in the country. Is this count a statistic or a parameter? - parameter. Since its the true number of people in the US

Problem 5

Determine the type of the following variables (cf. Figure 1.3 in course notes). (a) Tree diameter - Continuous (b) Salary - Discrete (c) Salary ranges - Ordinal (d) A cow's annual milk yield - Continuous (e) Number of defective light bulbs out of 100 - Count (f) Presence of a species in an ecosystem - Binary (g) Soybean lines - Nominal (h) Total cholesterol level - Continuous (i) Presence of larvae in a cm3 of soil - Binary (j) Number of flaws on a silicon wafer - Count (k) Fat content of milk - Continuous