Homework 1

Daniel Caley

7/7/2021

Contents

Homework Question List

The homework for week one is exercises 1, 3, and 4 on page 20.

Question 1

Using the material from this chapter and possibly other information that you look up, write a brief definition of these terms in your own words: mean, median, mode, variance, standard deviation, histogram, normal distribution, and Poisson distribution.

- **Mean:** The average from a vector or column. The calculation is derived by taking the sum of a vector or column and divide the vector by the number of observations in the data set.
- **Median:** The middle number in a sorted dataset. For example if we have 1, 2, 3, 4, 5. the middle number would be 3.
- **Mode:** The value that appears most in a vector or column.
- **Standard Deviation:** The squared root of the variances. In the async, which I think is so important is, an important note was that we want standard deviation expressed in the same units as the original measure.
- **Histogram:** The frequency of a vector which represents the distribution of the data. Graphically this will be shown as Bars.
- **Normal Distribution:** Data is continuous and bell curve.
- **Poisson Distribution:** random continuous data that follows a pattern in which most of the observations decrease around lambda. For example a lambda of 1 has many observations for 1 and 2 but decreases after 1. For lambda of 5 observations left and right begin to decrease.

Question 2

Part 1

Write the equations, using the appropriate Greek letters, for the population mean and population standard deviation. Explain briefly what each Greek letter means.

Mean:

$$\frac{\sum_{i=1}^{n} x_i}{n}$$

- n = the size of the population
- xi = each value from the population
- = summing or totally a set of numbers

Standard Deviation:

$$\sqrt{\frac{\sum(x_i -)^2}{n}}$$

- n = the size of the population
- Xi = each value from the population
- μ = the population mean
- = summing or totally a set of numbers

Part 2

The R environment offers about 20 different kinds of statistical distributions. Choose any one of these distributions other than the normal distribution or the Poisson distribution. (The help system in R can assist you with finding a description of these distributions and their commands: type "?distributions" at the command line. For a hint about one distribution you might choose to study, read the beginning of the next chapter!) Write some R code that generates 100 random points in that distribution.

Make sure to use the technique shown just above that begins with assigning the 100 points to a vector that can be reused for all of the other commands.

```
?distribution

random_points <- runif(100, 1, 5)

exponential_distribution <- dexp(random_points, rate = 1)
exponential_distribution
```

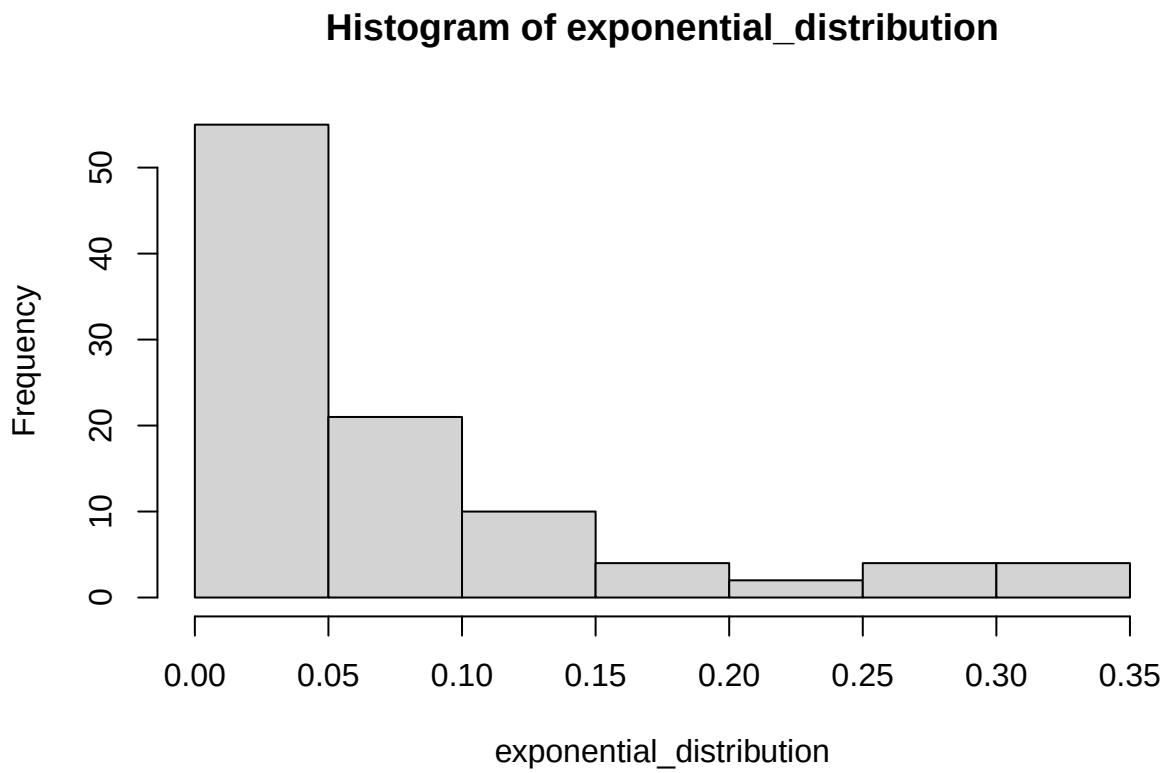
```
##   [1] 0.010003052 0.085948115 0.055432207 0.026330562 0.023211439 0.018245969
##   [7] 0.065709717 0.026662600 0.010545404 0.073538179 0.009559841 0.191472836
##  [13] 0.007242257 0.150295434 0.073005141 0.024901689 0.037326609 0.061480314
##  [19] 0.036274279 0.016847092 0.014918772 0.016620511 0.250240871 0.127178973
##  [25] 0.044687663 0.045020013 0.011487083 0.008765168 0.033378059 0.022248472
##  [31] 0.314649038 0.029996776 0.082286874 0.081881427 0.051606716 0.019365774
##  [37] 0.008479049 0.023112698 0.087648548 0.027627629 0.035624931 0.028820408
##  [43] 0.156095778 0.101001403 0.205224454 0.105637581 0.007831439 0.010228886
```

```
##  [49] 0.007580691 0.165322711 0.048223685 0.007056114 0.034266520 0.318616503
##  [55] 0.038011782 0.045444101 0.118058759 0.009880716 0.017703850 0.010888561
##  [61] 0.132847731 0.028314790 0.017472699 0.256415882 0.095870542 0.072802033
##  [67] 0.116539509 0.304410281 0.022031829 0.314657649 0.034831819 0.010710013
##  [73] 0.123935377 0.090100687 0.038021547 0.200582972 0.007139506 0.120170053
##  [79] 0.016189419 0.017385206 0.046879441 0.087308098 0.076727525 0.057339139
##  [85] 0.118304020 0.012609232 0.080114317 0.039357108 0.106281778 0.030121540
##  [91] 0.268931444 0.018459584 0.054893963 0.063758570 0.050134383 0.009361530
##  [97] 0.264641253 0.028143508 0.097952104 0.039216821
```

Part 3

Displays a histogram of those 100 points

```
hist(exponential_distribution)
```



Part 4

Calculates the mean of those points

```
mean(exponential_distribution)
```

```
## [1] 0.07347717
```

Part 5

Calculates the standard deviation.

```
sd(exponential_distribution)
```

```
## [1] 0.0784928
```

Question 3

Part 1

Use the data() function to get a list of the data sets that are included with the basic installation of R: just type "data()" at the command line and press enter. Choose a data set from the list that contains at least one numeric variable—for example, the Biochemical Oxygen Demand (BOD) data set. Use the summary() command to summarize the variables in the data set you selected—for example, summary(BOD).

```
data(Orange)
summary(Orange)
```

```
##  Tree       age         circumference
##  3:7   Min.   : 118.0   Min.   : 30.0
##  1:7   1st Qu.: 484.0   1st Qu.: 65.5
##  5:7   Median :1004.0   Median :115.0
##  2:7   Mean   : 922.1   Mean   :115.9
##  4:7   3rd Qu.:1372.0   3rd Qu.:161.5
##        Max.   :1582.0   Max.   :214.0
```

Part 2

Write a brief description of the mean and median of each numeric variable in the data set. Make sure you define what a "mean" and a "median" are, that is, the technical definition and practical meaning of each of these quantities.

Mean: Take the average of a given vector by totally the observation by the number of the observations. For the orange data set the mean age is 2.5 year old and the mean circumference of an orange tree is 4.5 inches. For the age the data set is in days so I wouldn't say 922 days old as that's not how we describe the age of something or someone. For the circumference we could say millimeters but most the time people like talking in inches.

Median: The middle number in a dataset. The median age is 2.75 year old and the median circumference is 4.5 inches.

```
# Mean
mean(Orange$age)
```

```
## [1] 922.1429
```

```
mean(Orange$circumference)
```

```
## [1] 115.8571
```

```
# Median
median(Orange$age)
```

```
## [1] 1004
```

```
median(Orange$circumference)
```

```
## [1] 115
```

Question 4

Part 1

As in the previous exercise, use the data() function to get a list of the data sets that are included with the basic installation of R. Choose a data set that includes just one variable, for example, the LakeHuron data set (levels of Lake Huron in the years 1875 through 1972). Use the hist() command to create a histogram of the variable—for example, hist(LakeHuron).

```
MyBJsales <- BJsales

# Looking at summary statistics of the data
summary(MyBJsales)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   198.6   212.6   220.7   230.0   254.7   263.3
```

```
# Looking at the Standard Deviation
sd(MyBJsales)
```

```
## [1] 21.47969
```

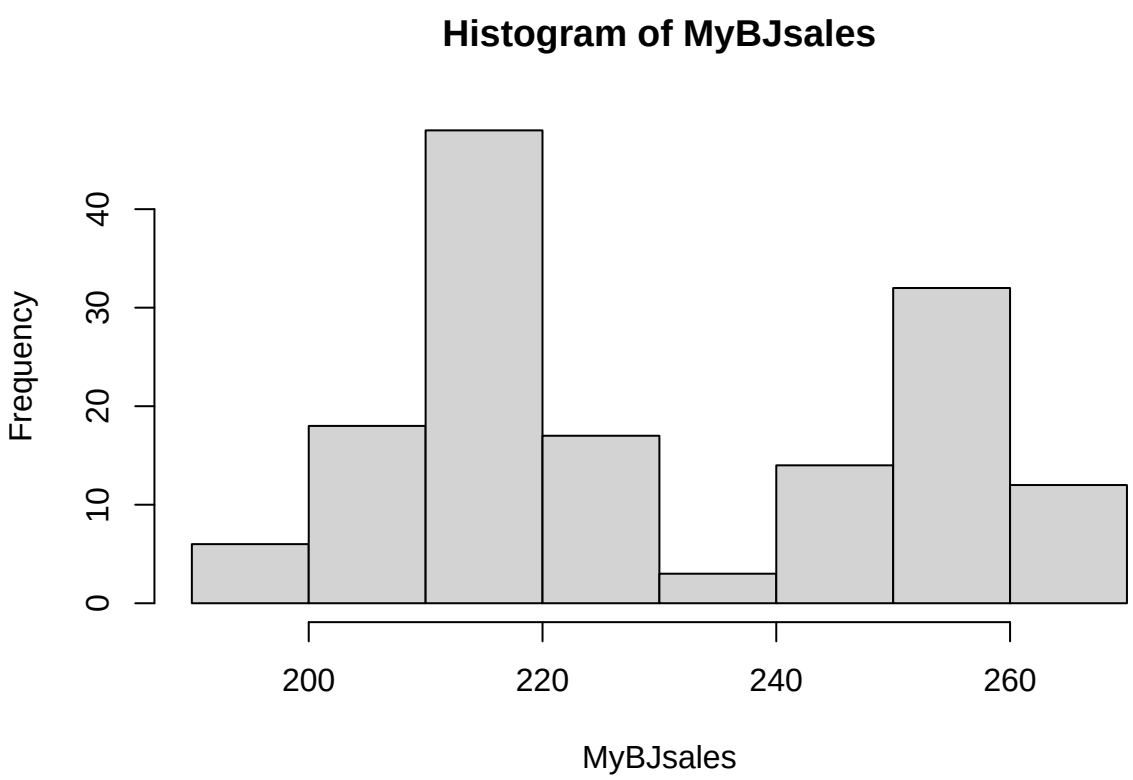
```
# Looking at the average Sales
mean(MyBJsales)
```

```
## [1] 229.978
```

```
# Looking at the Median Sales
median(MyBJsales)
```

```
## [1] 220.65
```

```
hist(MyBJsales)
```



Part 2

Describe the shape of the histogram in words. Which of the distribution types do you think these data fit most closely (e.g., normal, Poisson). Speculate on why your selected data may fit that distribution.

The shape of the histogram is Bimodel which has two peaks instead of a normal distribution or poisson that has one peak. When reading more about Bimodel's this can happen when two different types of cohorts exist in the data. For example height of men and women. When separating the two we would see only 1 peak. In my data BJ sales probably had leadership change that created two peaks. When leaders take over the potential of a drop can happen in the immediate and bounce back in a few years or when a new leader comes in. The Standard Deviation is 21.48, the mean is 229.98, and the median is 220.65.