IST772

Parin Patel
Homework 1
10/6/2019

Attribution Statement:

1. Homework 1 by Parin Patel: I did this homework by myself, with help from the book and the professor. I also utilized the two rdocumentation.org websites for background information on the UKgas dataset and the USArrests datasets.
    a. Link for UKgas: https://www.rdocumentation.org/packages/datasets/versions/3.6.1/topics/UKgas
    b. Link for USArrests:
        i. https://www.rdocumentation.org/packages/datasets/versions/3.6.1/topics/USArrests

Exercises:

1. Definitions of the following terms in my own words:
    a. Mean : The average of a set of numbers. This is the sum of the numbers added and then divided by the count.
    b. Median: The middle number in a set of numbers when they are arranged from lowest to highest.
    c. Mode: The number that occurs the most.
    d. Variance: On average, how far each number in a set is from the mean.
    e. Standard deviation: the square root of the variance. How far a group of numbers are from the mean.
    f. Histogram: a graphical representation of the distribution of range of numbers (data points) in a bar-format. It is used to show frequency within the data points, and the intervals between each bar on the graph is consistent.
    g. Normal distribution: a distribution that results in an symmetrical, bell-shaped density curve on a graph . The curve is centered around the mean, with the highest point, the peak, being at the mean (therefore, the middle of the curve).
    h. Poisson distribution: a distribution based on a discrete probability that shows how likely an event will occur within a certain, specified time period. Its used often for independent, but consistent occurrences, like customers dining in a restaurant on a Friday night.
2. **Note: Question #2 was not required for the homework according to the syllabus. Completed it for my own reference and practice. – thanks!**
    a. Population mean:

IST772

Parin Patel
Homework 1
10/6/2019

i. $\mu = \frac{(\Sigma\ X_1)}{N}$

1. μ = "mu". Symbol for population mean.
2. Σ = "sigma. Amount of all the variance of a set of values.
3. $(\Sigma\ X_1)$ = the sum of all the data values
4. N = the number of all the data items.

b. Population standard deviation:

i. $\sigma = \sqrt{\frac{\Sigma\ (x-\mu)^2}{N}}$

1. σ = lower case sigma. Represents population variance.
2. Σ = "sigma. Amount of all the variance of a set of values.
3. X = individual value
4. μ = average of the population
5. $\Sigma\ (x-\ \mu)^2$ = sum of all the squared deviations.
6. N = the number of all the data items.

3. The dataset used was "USArrests". Below is the R code for the calling the dataset and using the summary() command.

```
#3. Data for US arrests
data("USArrests")
head(USArrests)
  #List of variables: Murder, Assaults, UrbanPop, Rape
summary(USArrests)
```

Here is the output for the command:

IST772

Parin Patel
Homework 1
10/6/2019

```
> summary(USArrests)
     Murder          Assault
 Min.   : 0.800   Min.   : 45.0
 1st Qu.: 4.075   1st Qu.:109.0
 Median : 7.250   Median :159.0
 Mean   : 7.788   Mean   :170.8
 3rd Qu.:11.250   3rd Qu.:249.0
 Max.   :17.400   Max.   :337.0
    UrbanPop          Rape
 Min.   :32.00   Min.   : 7.30
 1st Qu.:54.50   1st Qu.:15.07
 Median :66.00   Median :20.10
 Mean   :65.54   Mean   :21.23
 3rd Qu.:77.75   3rd Qu.:26.18
 Max.   :91.00   Max.   :46.00
```

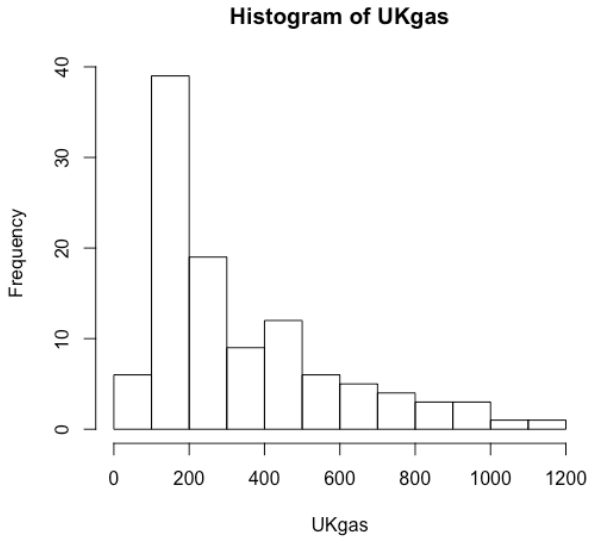
The numeric variables in this dataset are Murder , Assaults, UrbanPop (Urban Population) and Rape. The mean is the average number in the set of data. The median is the middle number in the dataset. For a symmetrical dataset, you can expect the distribution to peak around this number. The mean murders in the United States in 1975 was 7.788. This means that, on average, there was about 8 murders that year amongst all 50 states. In that same year, the most median number of murders between all 50 states was 7. This means that the median is 7 murders. Since the mean and median are the same, we can expect the distribution to be symmetric and unimodal. The mean number of assaults in the United States in 1975 was 171. In that year, the middle number of assaults in the US was 159. In this case, since the mean is greater than the median, it can be expected that the distribution is skewed to the right. The mean Urban population in 1975 was 66 people per 100,000 residents. The middle number of people in an urban population was 66. Since the median is slightly larger than the mean, then the distribution can be expected to be slightly skewed to the left. Finally ,for the last numeric variable, the mean rape in 1975 was 21 people. The middle value for this distribution is 20. Therefore, you can expect the distribution to peak at around 20-21. In addition, since the mean is slightly higher than the median, distribution is skewed slightly to the right.

IST772

Parin Patel
Homework 1
10/6/2019

4. UKgas dataset was chosen to because it includes just one variable. The dataset represents the quarterly UK gas consumption from Q1 of 1960 to Q4 of 1984. The numerical values are represented in millions of therms.

   In the histogram below, the distribution is positively skewed to the right. This means that it is likely that the mean of this dataset is greater than the median. Or that it is greater than the mode. The highest point is centered around 200 therms. The Poission distribution would fit best for this dataset. This is because the dataset can be modeled to understand how often a person in the UK is going to get gas between Q1 of 1960 and Q4 of 1984. Since the situation would be counting amount of consumption within a certain time, it would fit the requirement. Also, instance in this dataset is independent of the other instances. Overall, utilization of the Poission distribution would determine the probability of gas being consumed from Q1 of 1960 to Q4 of 1984.




```
#4. Dataset for the UkGas.

data()
data("UKgas")
#review
summary(UKgas)
#create histogram
hist(UKgas)


```

IST772

Parin Patel
Homework 1
10/6/2019

Here is the output for the command:

```
> #4. Dataset for the UkGas.
> data("UKgas")
> #review
> summary(UKgas)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   84.8   153.3   220.9   337.6   469.9  1163.9 
> #create histogram
> hist(UKgas)
> 
```

Final Code:

```
#2.R code for Binomial Distribution:
set.seed(120)
binomDis<-rbinom(100,100,0.2)
hist(binomDis)
mean(binomDis)
  #[1] 20.35
sd(binomDis)
  #[1] 4.063573



#3. Data for US arrests
data("USArrests")
head(USArrests)
  #List of variables: Murder, Assaults, UrbanPop, Rape
summary(USArrests)


#4. Dataset for the UkGas.
data("UKgas")
```

IST772

Parin Patel
Homework 1
10/6/2019

```
#review
summary(UKgas)
#create histogram
hist(UKgas)
```