# DATA ANALYSIS USING R

*Himank Jain*

*February 8, 2019*

## Contents

This is an R Markdown document consisting basic methods oF *data analysis, stastical inFerence, data visualization* and some other data inbuilt Function oF R. A large part oF the data and methods used in this document were taken From Foundation OF Data Analysis part-I From **edx** https://courses.edx.org/courses/course-v1:UTAustinX+UT.7.11x+2T2017/course/

## UniVariate Data:

Univariate means "one variable" (one type oF data) Example: You weigh the pups and get these results: 1,3,4,6,8,10,12 The one variable is Puppy Weight

```
pups=c(1,3,4,6,8,10,10)
```

### Measure OF Center In Univariate Data

- Mean:The arithmetic mean is the central value oF a discrete set oF numbers: speciFically, the sum oF the values divided by the number oF values.

$$\bar{x}(m) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^m\right)^{\frac{1}{m}}$$

For our puppy weights data set mean is:

```
mean(pups)
```

```
## [1] 6
```

*Median:The median is the value separating the higher halF From the lower halF oF a data sample. For a data set, it may be thought oF as the "middle" value.

$$\text{median}(a) = \frac{a_{\lceil \#x \div 2 \rceil} + a_{\lceil \#x \div 2+1 \rceil}}{2}$$

For our puppy weights data set median is:

```
median(pups)
```

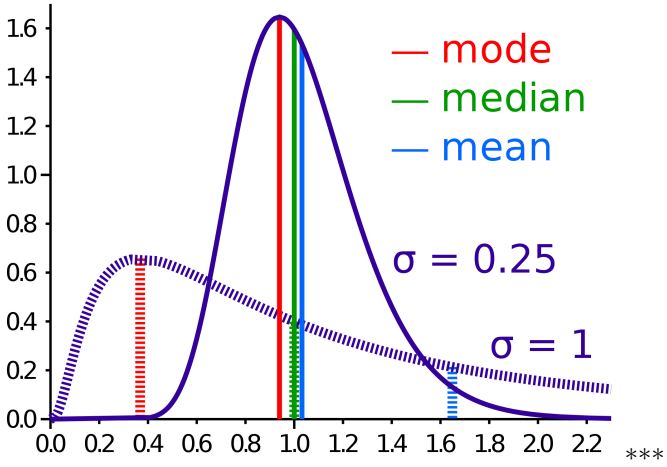
```
## [1] 6
```

*Mode:The mode oF a set oF data values is the value that appears most oFten. For our puppy weights data set mode is:

```
mode(pups)
```

```
## [1] "numeric"
```

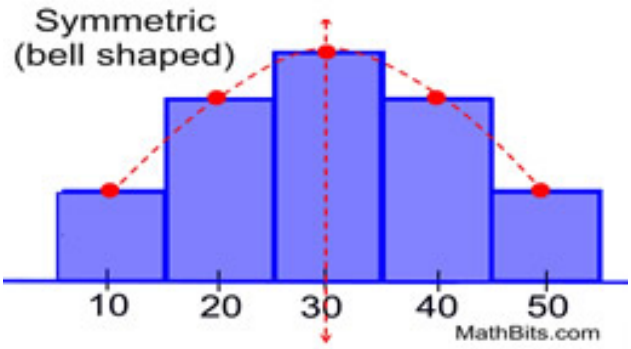
---



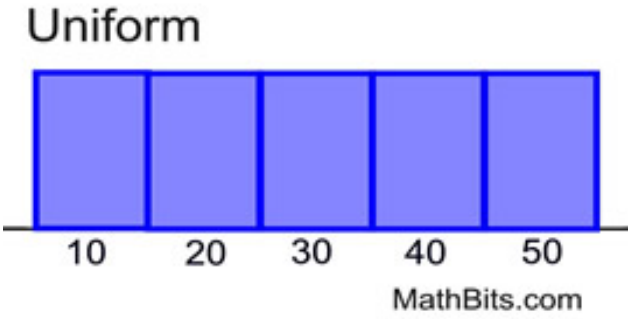
***

## Shapes OF Distribuitions:

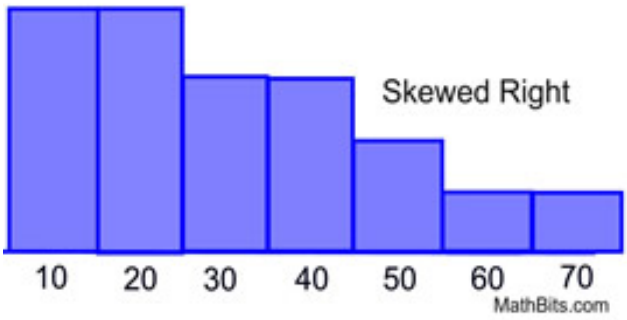
**1.Symmetric(Bell shaped)**: Unimode. Occurs in normal distributions. - when graphed, a vertical line drawn at the center will Form mirror images, with the leFt halF oF the graph being the mirror image oF the right halF oF the graph.



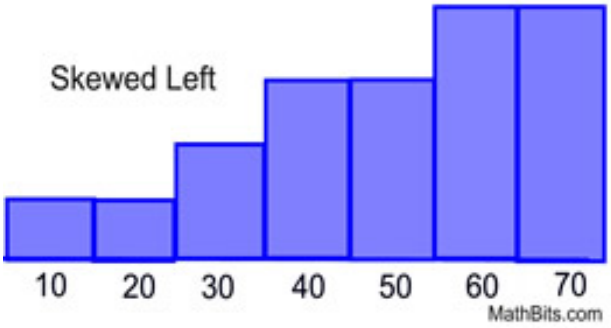

**2.UniForm**: The data is spread equally across the range.



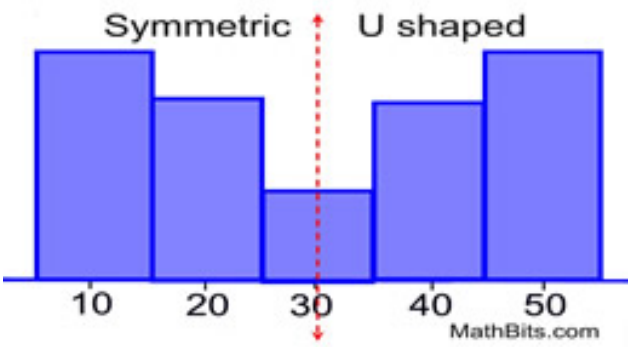

**3.Right Skewed**(positively skewed): Fewer data plots are Found to the right oF the graph (toward the larger numeric values).




**4.LeFt Skewed** (negatively skewed): Fewer data plots are Found to the leFt oF the graph (toward the smaller numeric values).




**5.Bimodal**: Usually has two modes.



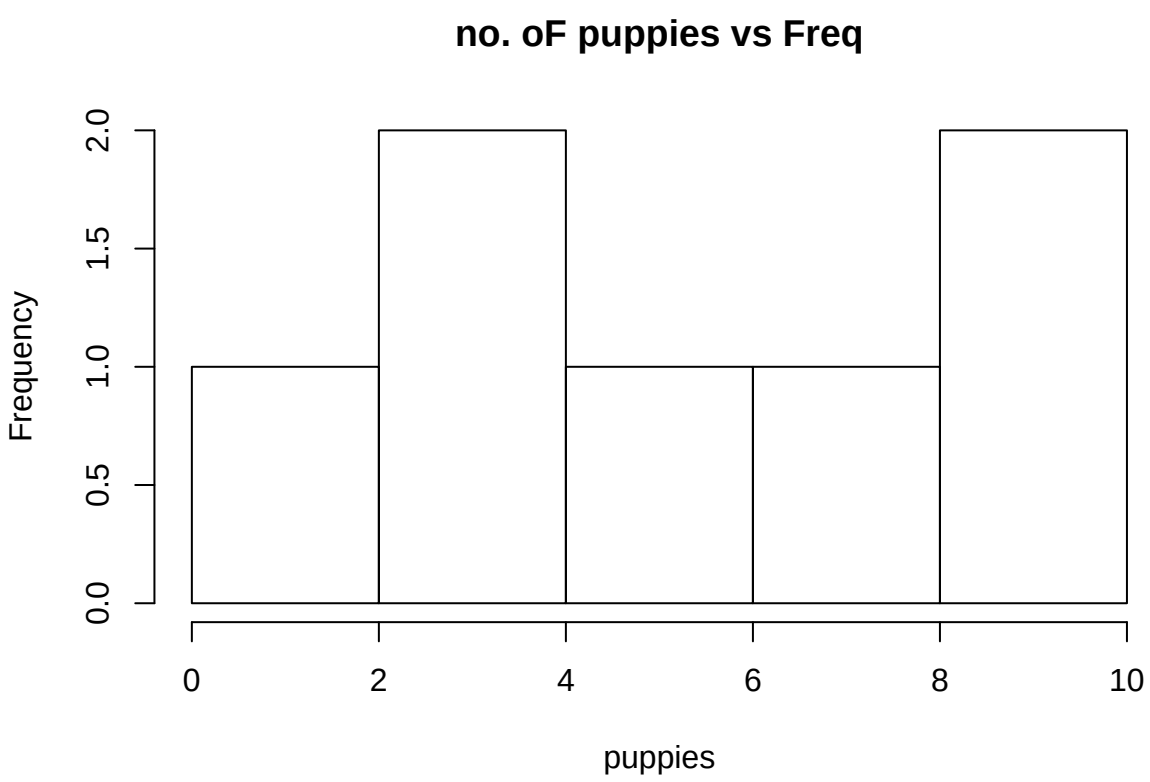

## Creating a Frequency table For A variable:

```
ptable=table(pups)
ptable
```

```
## pups
##  1  3  4  6  8 10
##  1  1  1  1  1  2
```

## Plotting Histogram For a Univariate Distribution:

```
hist(pups,xlab='puppies',ylab='Frequency',main='no. oF puppies vs Freq')
```




---

### Range and Quartiles:

*Range: The range is simply the diFFerence between the smallest value (minimum) and the largest value (maximum) in the data. In our puppies dataset range is:

```
range(pups)
```

```
## [1]  1 10
```

*Quartile: A quartile divides the data into Four approximately equal groups. The lower quartile,sometimes abbreviated as Q1 , is also know as the 25th percentile.The upper quartile, or Q3, is also know as the 75th percentile. We can get a summary oF our pups data in R using summary() Function which inlcudes quartiles,min,max mean,median,etc..

```
summary(pups)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0     3.5     6.0     6.0     9.0    10.0
```

*Interquartile Range: The interquartile range (IQR) is the range oF the data that contains the middle 50% oF cases. $IQR = Q3 - Q1$

## five number summary:

The five number summary is a numerical description of a data set comprised of the following measures: min,lower quartile,median,upper quartile,max.

```
fivenum(pups)
```

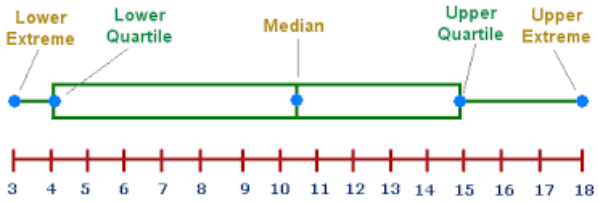
```
## [1]  1.0  3.5  6.0  9.0 10.0
```

---

---

## Box and whisker plots:

A box and whisker plot is a very convenient and informative way to diplay the info captured in the five number summary. A box and whisker plot shows the centers and spread of the values on a single quantative variable.

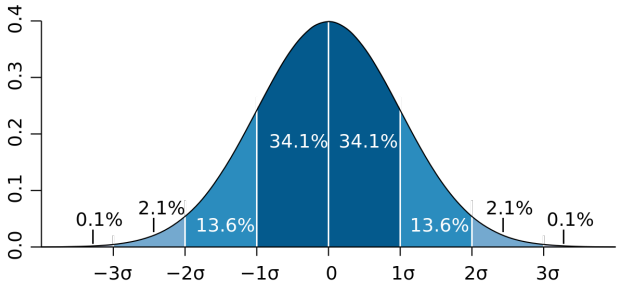



## Standard Deviation And Mean:

When data is normally Distributed, there are two preferred measures of center and spread. These are arithmetic mean and standard deviation.The **Standard Deviation** of a data set tells us how it is spread out. The larger the standard deviation is,the more spread out data is. A vertical line from inflectionn point to x-axis marks one standard deviation from the mean.Approx 68% of the data is located within one standard deviation of the mean. For out pups data std dev is:

```
sd(pups)
```

```
## [1] 3.511885
```




## Variance:

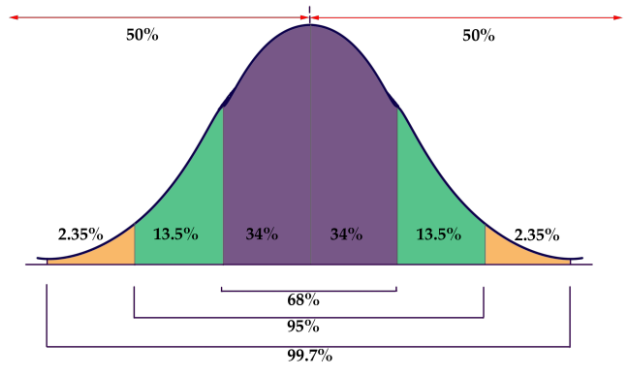
**Variance** is also a measure of spread. It is simply the square of Standar Deviation. For our pups data variance is:

```
var(pups)
```

```
## [1] 12.33333
```

### Emperical Rule:

Emperical Rule states that the percentages of data in a normal distribution within 1,2 and 3 standard deviations of the mean are approximately 68%,95% and 99.7%.




**Z-Score:**

A **z-score** is a measure of the number of standard deviations a particular data point is away from the mean.

$$z = Deviation/StandardDeviation$$

## Bivariate Data:

Bivariate Data is data set with two variables (quantative or categorical). **Correlation** measures the linear relationship between two quantative variables.