Fundamentals of Statistics III

Moving Beyond Mean and SD

DATA SCIENCE INSTITUTE

What will we learn

- Quantiles and Box-Whisker Plot
- Histogram
- Measure of Skewness

Dissect Data with Quantiles

- Quartiles divide the distribution into 4 equal parts.

  Q1: Lower Quartile(25% observations are below Q1)

  Q3: Upper Quartile (25% observations are above Q3)

  Q2 is same as median

- Deciles divide the distribution into 10 equal parts.

  5^{th} Decile is same as median

- Percentiles divide the distribution into 100 equal parts.

  50^{th} percentile is same as median

  75^{th} percentile is same as Q3

DATA SCIENCE INSTITUTE

Quantiles in R

 # Import basic_salary2 data and store in object salary

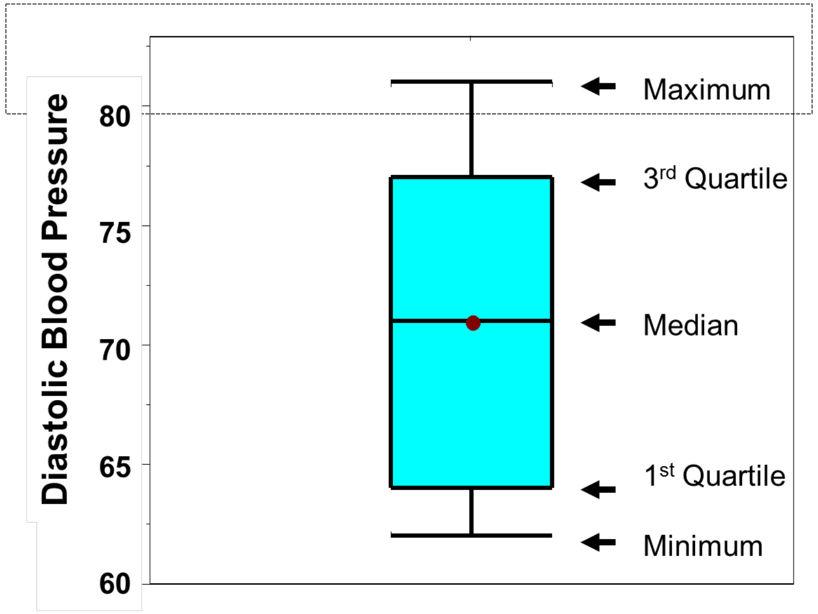
```
> quantile(salary$ba,na.rm=T)

   0%   25%   50%   75%  100%
10940 13785 16230 19305 29080

> quantile(salary$ba,prob=c(0.1,0.5,0.8),na.rm=T)

  10%   50%   80%
13084 16230 20280
```

Box-Whisker Plot

- Box and Whisker plot summarizes data graphically using 5 measures: Minimum,Q1,Q2,Q3 and Maximum.

- The body of the box goes from the first quartile (Q1) to the third quartile (Q3).
- The whiskers go from Q1 to smallest non outlier and Q3 to highest non outlier data points.

- The distribution is considered symmetric if median is at the center of the box and whiskers have same length



DATA SCIENCE INSTITUTE

Defining Outliers

- An outlier is an observation that lies an abnormal distance from other values in a random sample from a population.

- What will be considered abnormal? Before abnormal observations can be singled out, it is necessary to characterize normal observations.

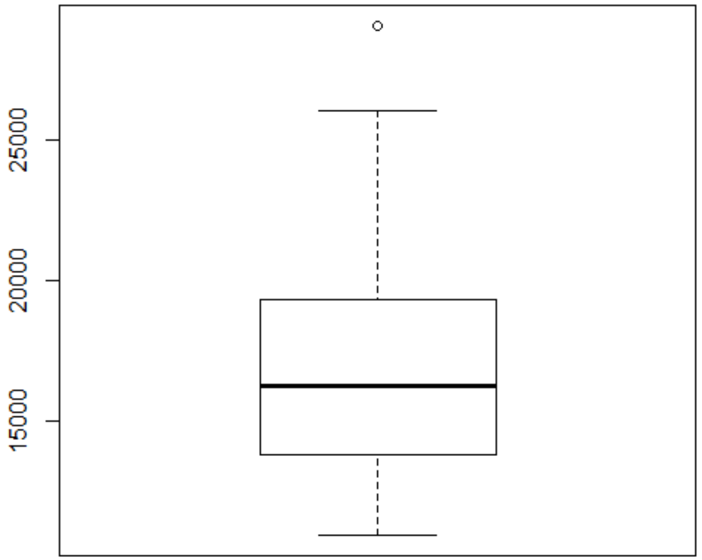
- Non-Outlier observation is

$$\geq Q1 - 1.5 * IQR \text{ and } \leq Q3 + 1.5 * IQR$$

where IQR: Inter-quartile Range =Q3-Q1

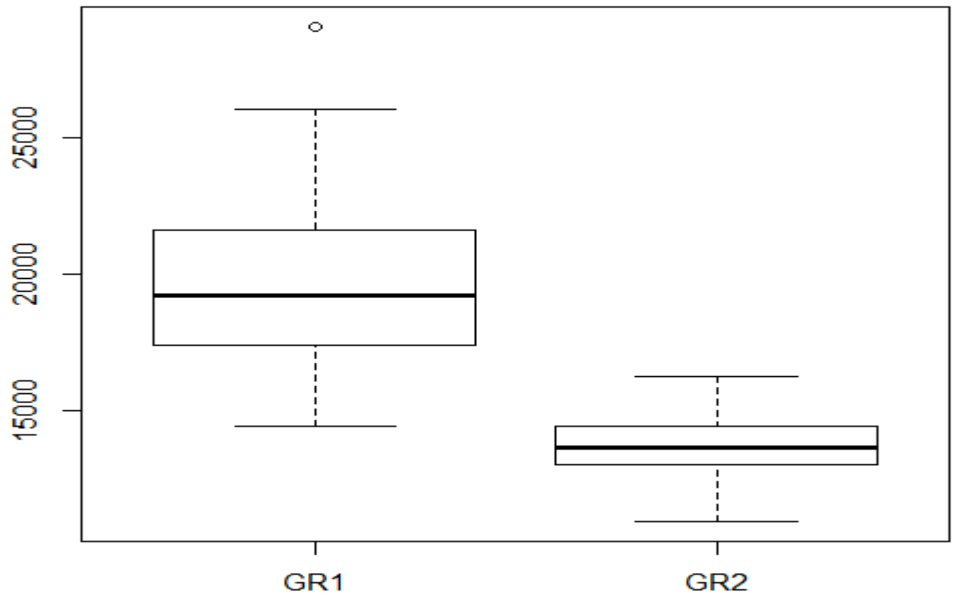
Box-Whisker Plot in R

```
> boxplot(salary$ba)
```



Box-Whisker Plot by Grade

```
> boxplot(ba~Grade,data=salary)
```

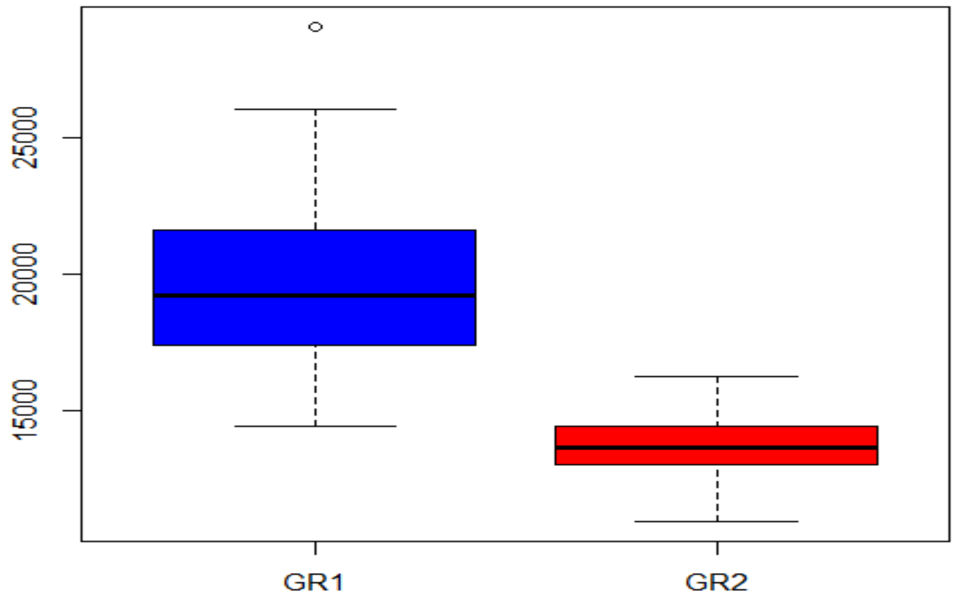


The distribution of ”ba” in G

DATA SCIENCE INSTITUTE

Adding Colours

```
> boxplot(ba~Grade,data=salary,col=(c("blue","red")))
```



DATA SCIENCE INSTITUTE

Display Observation Number on Outliers

boxplot() provides outlier values which can be accessed as follows:

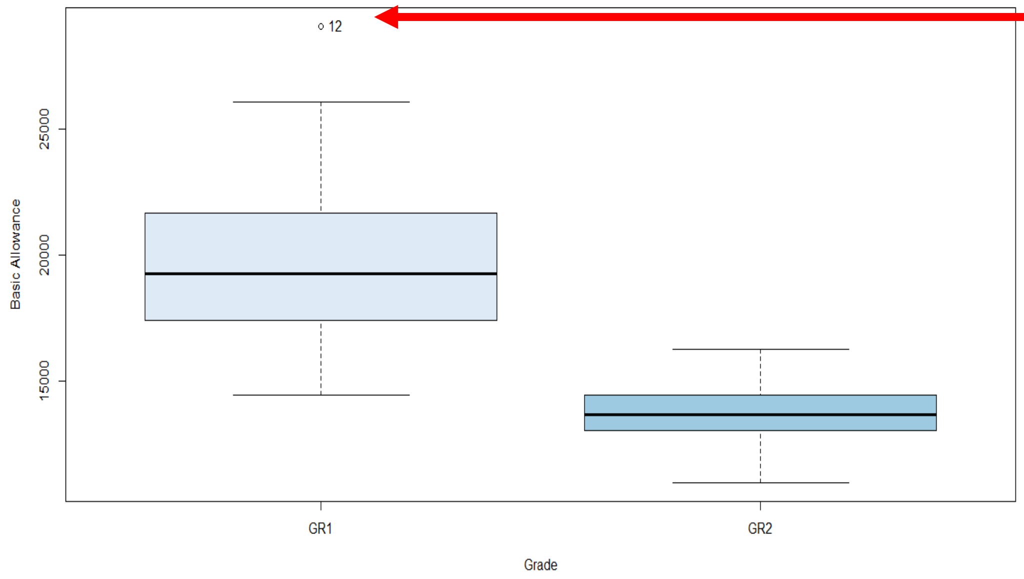
```
>box <- boxplot(ba~Grade,data=salary)
>box$out
[1] 29080
```

However, it is better to print observation numbers rather than values.

Boxplot() from package car solves this problem.

Display Observation Number on Outliers

```
> install.packages("car")

> library(car)


> Boxplot(ba~Grade,data=salary)
```

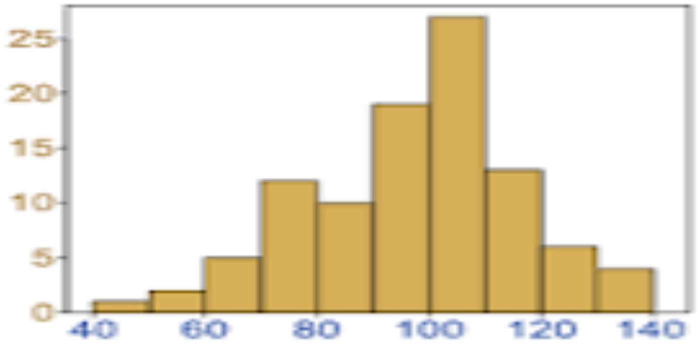


12^{th} observation has ba 29080

Histogram

- Histogram is useful in visualizing a distribution.

- To construct a histogram, the first step is to "bin" the range of values—that is, divide the entire range of values into a series of intervals—and then count how many values fall into each interval(frequency)

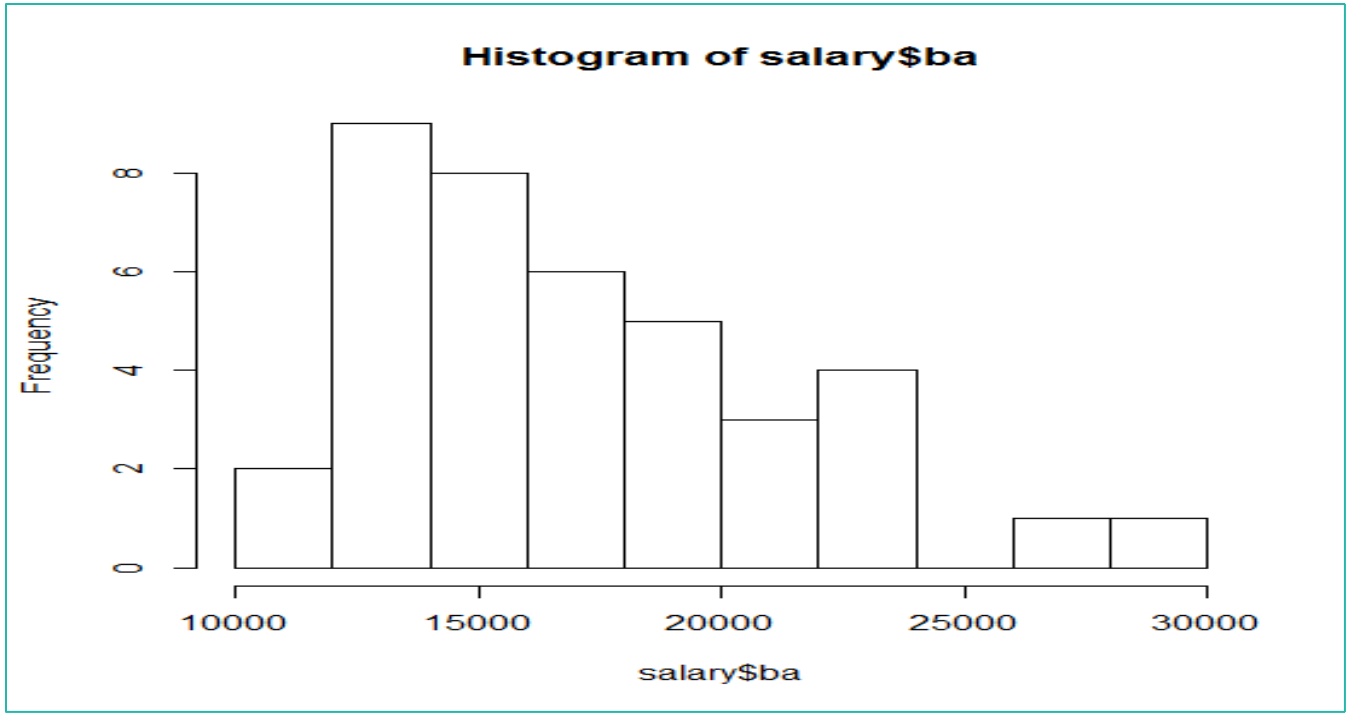


The number of bins *k* can be assigned directly or can be calculated from a suggested bin width *h* as: (max Y – min Y)/h where Y is a variable of interest.

DATA SCIENCE INSTITUTE

Histogram in R

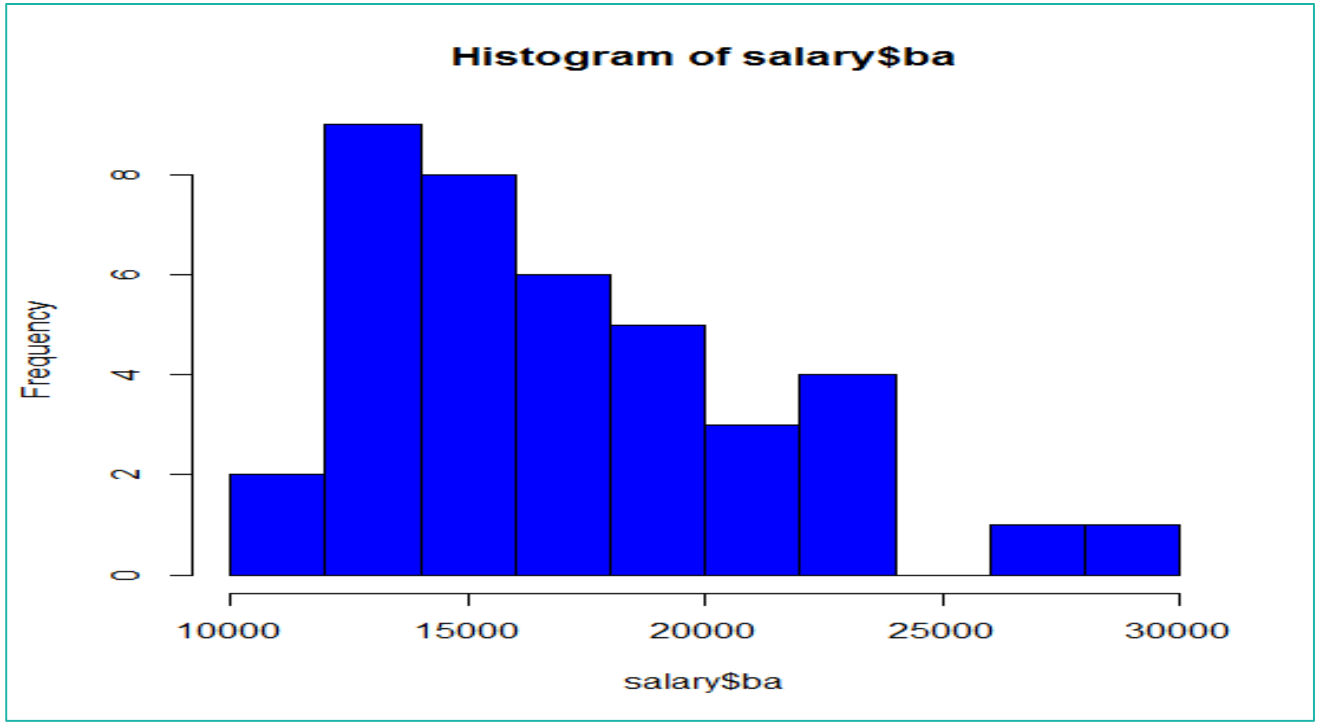
> hist(salary$ba)



DATA SCIENCE INSTITUTE

Adding Colour in Histogram

```
> hist(salary$ba,col="blue")
```



Try this code

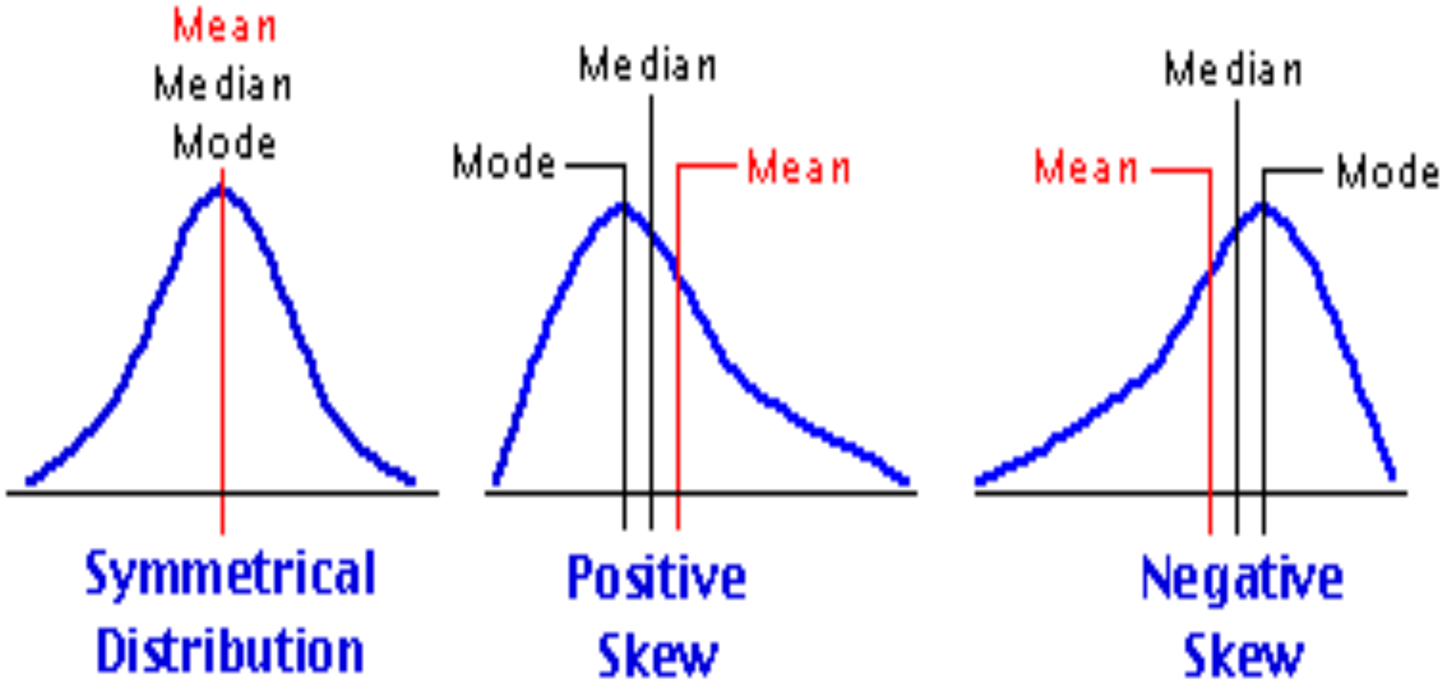
```
hist(salary$ba,col="blue",breaks=c(10000,15000,20000,25000,30000))
```

DATA SCIENCE INSTITUTE

What is Skewness?

- Skewness is a measure of 'lack of symmetry' of the data.
- *positive skew*: The right tail is longer; the mass of the distribution is concentrated on the left
- *negative skew*: The left tail is longer; the mass of the distribution is concentrated on the right.
- If the distribution is symmetric, then the mean is equal to the median, and the distribution has zero skewness.
- The Normal distribution is symmetric distribution.

Visualizing Skewness



DATA SCIENCE INSTITUTE

How to Calculate Skewness?

- The Pearson measure of skewness is defined as

  (mean – mode) / standard deviation

- Another form of the Pearson measure of skewness is

  3 (mean – median) / standard deviation

- The Bowley coefficient of skewness is

$$\frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1} = \frac{Q_1 - 2Q_2 + Q_3}{Q_3 - Q_1},$$

DATA SCIENCE INSTITUTE

Skewness Based on Third Moment

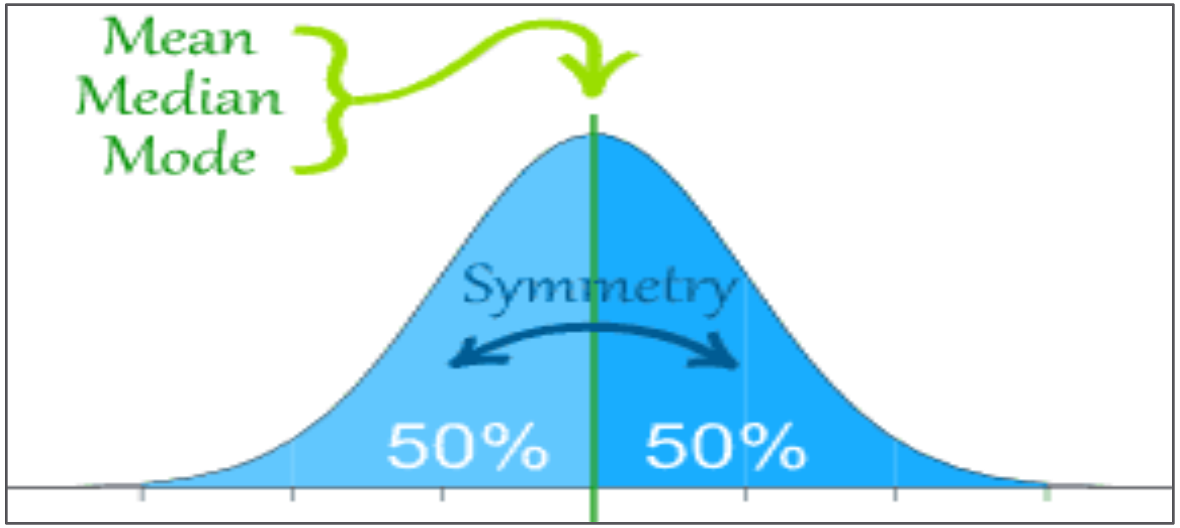
- The most widely used measure of skewness is based on the third moment.

$$\frac{n}{(n-1)(n-2)}\sum\left(\frac{x_i-\bar{x}}{s}\right)^3$$

- Any threshold or rule of thumb is arbitrary, but here is one: If the skewness is greater than 1.0 (or less than -1.0), the skewness is substantial and the distribution is far from symmetrical. Value ‘zero’ indicates symmetric distribution.

Normal Distribution

- Commonly used distribution for continuous variables.
- Also known as the Gaussian distribution.
- Normal curve is a symmetric bell-shaped curve.
- Many statistical methods assume that population is normally distributed.



Using R to Measure Skewness

```
> install.packages("e1071")
> library(e1071)

# Use basic_salary2 data

> skewness(salary$ba,na.rm=T,type=2)

[1] 0.9033507
```

Note that type=2 uses moment based formula as discussed.
Most softwares use the same formula.

DATA SCIENCE INSTITUTE

Skewness by Grade

```
> library(e1071)

> f<-function(x) skewness(x,na.rm=T,type=2)
> aggregate(ba~Grade,data=salary,FUN=f)

  Grade         ba
1   GR1 0.85500651
2   GR2 0.08682743
```