# MIS 64036: Business Analytics

## Lecture III

Rouzbeh Razavi, PhD

# Agenda

- Introduction
- Measures of Central Tendency
- Measures of Dispersion
- Measures of Skewness
- Practice in R
- Measures of Dependence
- Practice in R
- Normal Distribution
- Examples in R

# Agenda

- **Introduction**
- Measures of Central Tendency
- Measures of Dispersion
- Measures of Skewness
- Practice in R
- Measures of Dependence
- Practice in R
- Normal Distribution
- Examples in R

# Why Statistics?

Many studies generate large numbers of data points, and to make sense of all that data, statistics are used to summarize the data, providing a better understanding of overall tendencies within the distributions of scores.

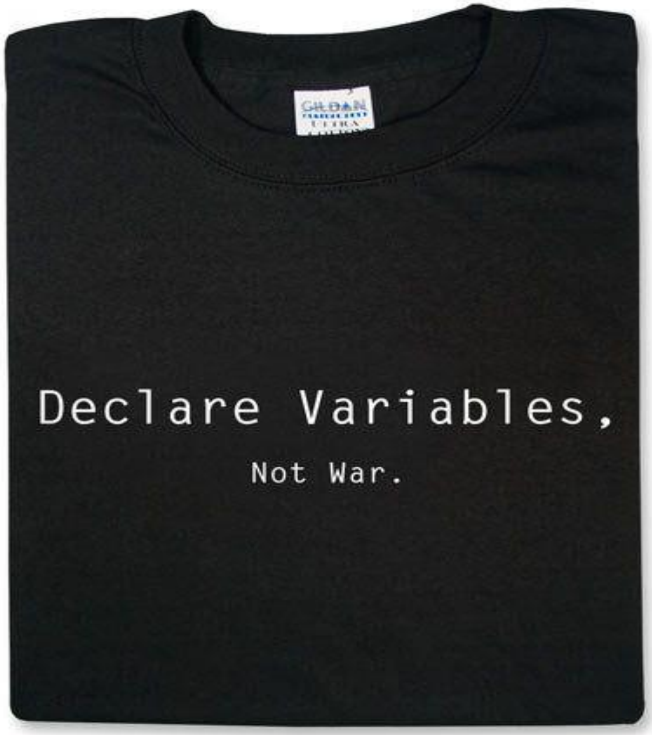
# Type of Variables

- Nominal scale assign numbers to attribute to name the category. The numbers have no meaning by themselves, e.g. DRG code.

- Ordinal scale assign numbers so that more of an attribute has higher values, e.g. Severity.

- Ratio/interval scale are normal numerical quantities that can be continues or discrete, e.g. Age, Height, Number of Kids

# You Do Not Need To Declare Variables in R Though



# Types of Statistics

## Two types of statistics

1. **Descriptive** (which summarize some characteristic of a sample)
    - Measures of central tendency (univariate)
    - Measures of dispersion (univariate)
    - Measures of skewness (univariate)
    - Measures of dependence (bivariate)

2. **Inferential** (which test for significant differences between groups and/or significant relationships among variables within the sample

# Agenda

- Introduction
- **Measures of Central Tendency**
- Measures of Dispersion
- Measures of Skewness
- Practice in R
- Measures of Dependence
- Practice in R
- Normal Distribution
- Examples in R

# Measures of Central Tendency

Example: Score={7,1, 6, 3, 7 }

- Mean is the arithmetic average.
  Mean(Score)=(7+1+6+3+7)/5=4.8

- Median is the halfway point in a data set. To calculate median, arrange data in order and find the middle point.

  Score_sorted={1, 3, 6, 7, 7 } so Mean(Score)=6

- Mode is the most frequent value observed is the mode.

  Mode(Score) =7 (repeated twice)

# Measures of Central Tendency

Console ~/

```
> Score=c(7,1, 6, 3, 7)
> mean(Score)
[1] 4.8
> median(Score)
[1] 6
> install.packages("modeest")
Installing package into 'C:/Users/lenovo pc/Documents/R/win-library/3.1'
(as 'lib' is unspecified)
Warning in install.packages :
  package 'modeest' is in use and will not be installed
> mlv(Score, method = "mfv")
Mode (most likely value): 7
Bickel's modal skewness: -0.6
Call: mlv.default(x = Score, method = "mfv")
>
```

# Measures of Central Tendency: Breaking Ties

Example: Score={7,1, 6, 3, 7, 3} #added a new element 3 at the end

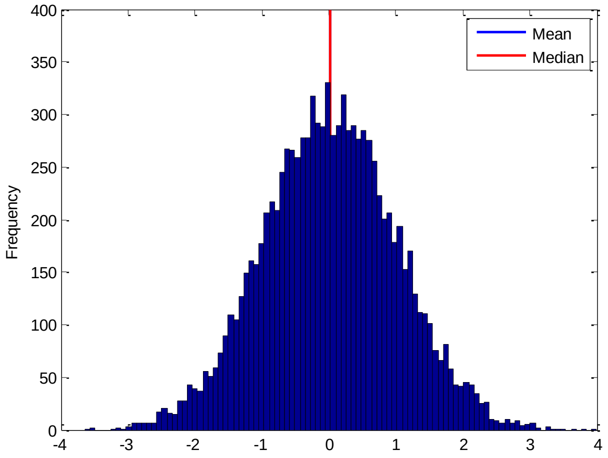
- Median: what is the median now? Score_sorted={1, 3, 3, 6, 7, 7 }
  Answer: Depends on the algorithm implementation for breaking ties. In most cases is either the average (4.5) or the smaller value

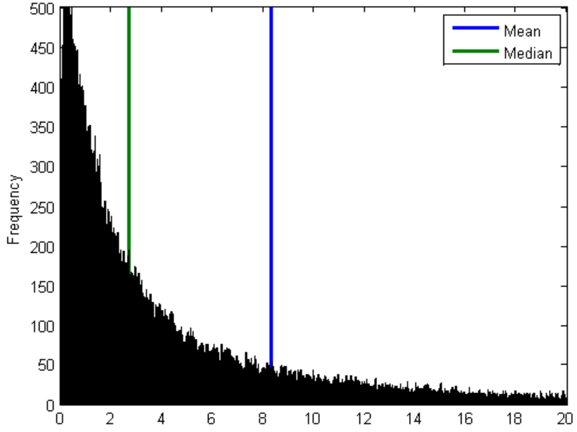
- Mode: what is the mode now? 3 or 7 (both appeared twice)

Answer: For any data set there is only one mean and median but there may be many modes.

# Measures of Central Tendency

- If the recorded values for a variable form a symmetric distribution, the median and mean are identical.
- In skewed data, the mean lies further toward the skew than the median. As such, the median is unaffected by outliers, making it a better measure of central tendency.



# Agenda

- Introduction
- Measures of Central Tendency
- **Measures of Dispersion**
- Measures of Skewness
- Practice in R
- Measures of Dependence
- Practice in R
- Normal Distribution
- Examples in R

# Measures of Dispersion

- Range: The spread, or the distance, between the lowest and highest values of a variable. i.e. range(a)=max(a)-min(a)

- Variance: The expectation of the squared deviation of a variable from its mean. Sample variance is defined as:

$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$$

while the population variance is defined in terms of the population mean μ and population size N:

$$\sigma^2 = \frac{\sum(X - \mu)^2}{N}$$

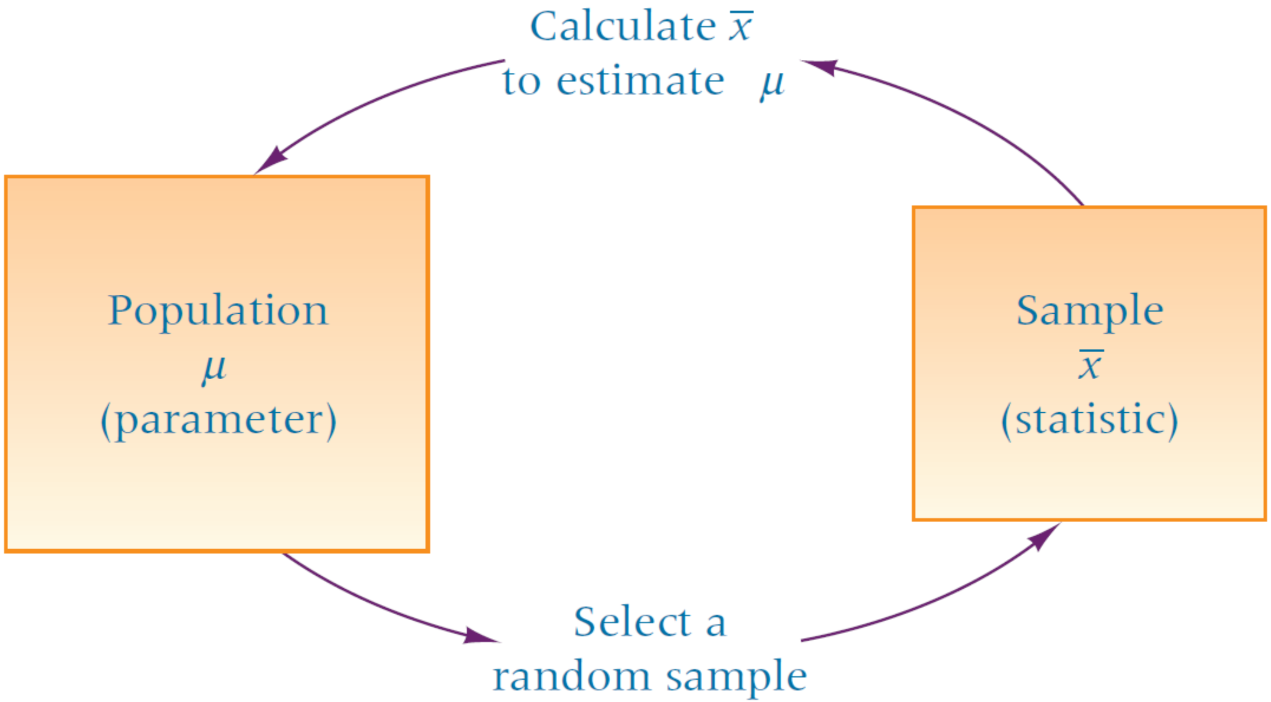
# Population vs. Sample

- The primary task of inferential statistics (or estimating or forecasting) is making an opinion about something by using only an **incomplete sample of data**.

- A **population** is defined as all members (e.g. occurrences, prices, annual returns) of a specified group. Population is the whole group.

- A **sample** is a **part of a population** that is used to describe the characteristics (e.g. mean) of the whole population. The size of a sample can be less than 1%, or 10%, or 60% of the population, but it is never the whole population.

# Population vs. Sample



# Measures of Dispersion

- Standard Division (σ ): The square root of the variance. It is expressed in the original units of measurement and represents the average amount of dispersion in a sample

$$\sigma = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$$

- Interquartile range (IQR) is a measure of variability and is equal to the difference between 75th and 25th percentiles, or between upper and lower quartiles, IQR = Q3 − Q1 based on dividing a data set into quartiles.

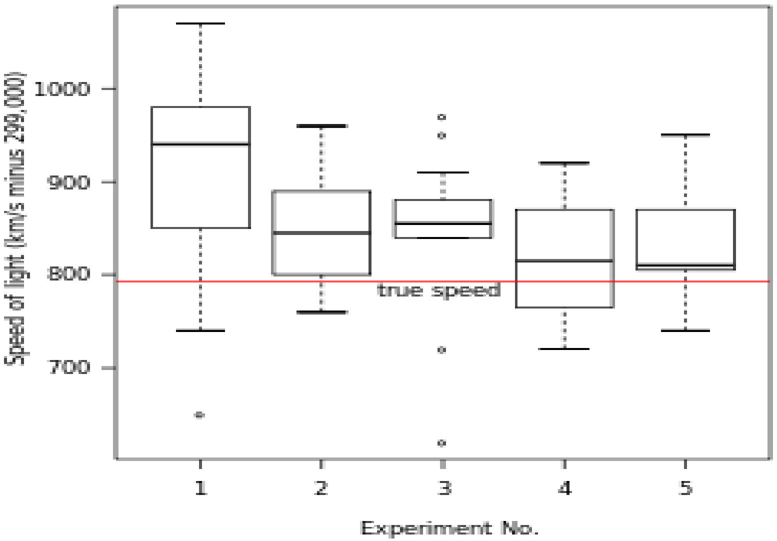
# Interquartile Range (IQR) : Example

<table>
<tr><th>i</th><th>x[i]</th><th>Median</th><th>Quartile</th></tr>
<tr><td>1</td><td>7</td><td rowspan="13">$Q_2$=87<br>(median of whole table)</td><td rowspan="7">$Q_1$=31<br>(median of upper half, from row 1 to7)</td></tr>
<tr><td>2</td><td>7</td></tr>
<tr><td>3</td><td>31</td></tr>
<tr><td>4</td><td>31</td></tr>
<tr><td>5</td><td>47</td></tr>
<tr><td>6</td><td>75</td></tr>
<tr><td>7</td><td>87</td></tr>
<tr><td>8</td><td>115</td><td rowspan="6">$Q_3$=119<br>(median of lower half, from row 7 to 13)</td></tr>
<tr><td>9</td><td>116</td></tr>
<tr><td>10</td><td>119</td></tr>
<tr><td>11</td><td>119</td></tr>
<tr><td>12</td><td>155</td></tr>
<tr><td>13</td><td>177</td></tr>
</table>

For the data in this table the interquartile range is IQR = $Q_3 - Q_1$ = 119 - 31 = 88.

# Visualizing Dispersion: Boxplot

- Boxplot shows groups of numerical data through their quartiles. Box plots may also have lines (whiskers) indicating variability outside the upper and lower quartiles. Outliers may be plotted as individual points.

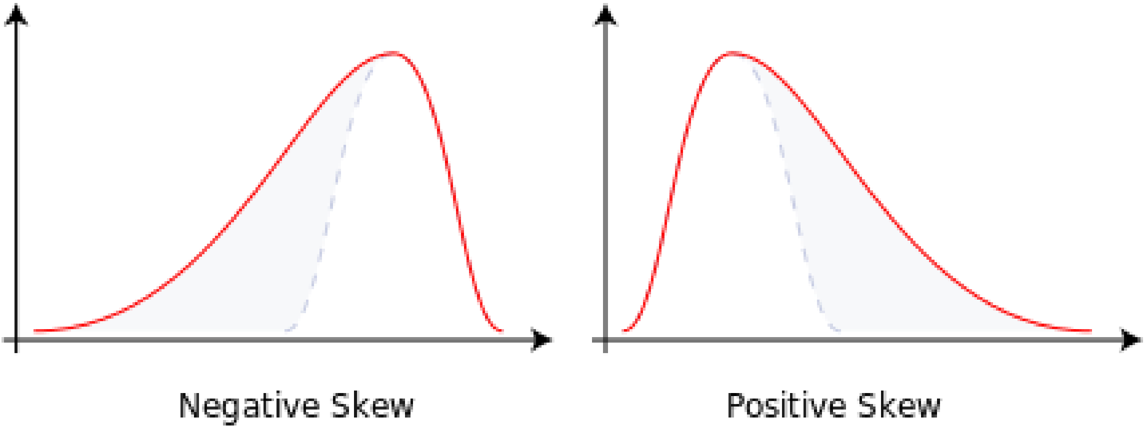


# Agenda

- Introduction
- Measures of Central Tendency
- Measures of Dispersion
- **Measures of Skewness**
- Practice in R
- Measures of Dependence
- Practice in R
- Normal Distribution
- Examples in R

# Measures of Skewness

- Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. The skewness value can be positive or negative.



# Measures of Skewness

- Population Skewness is can be expressed as:

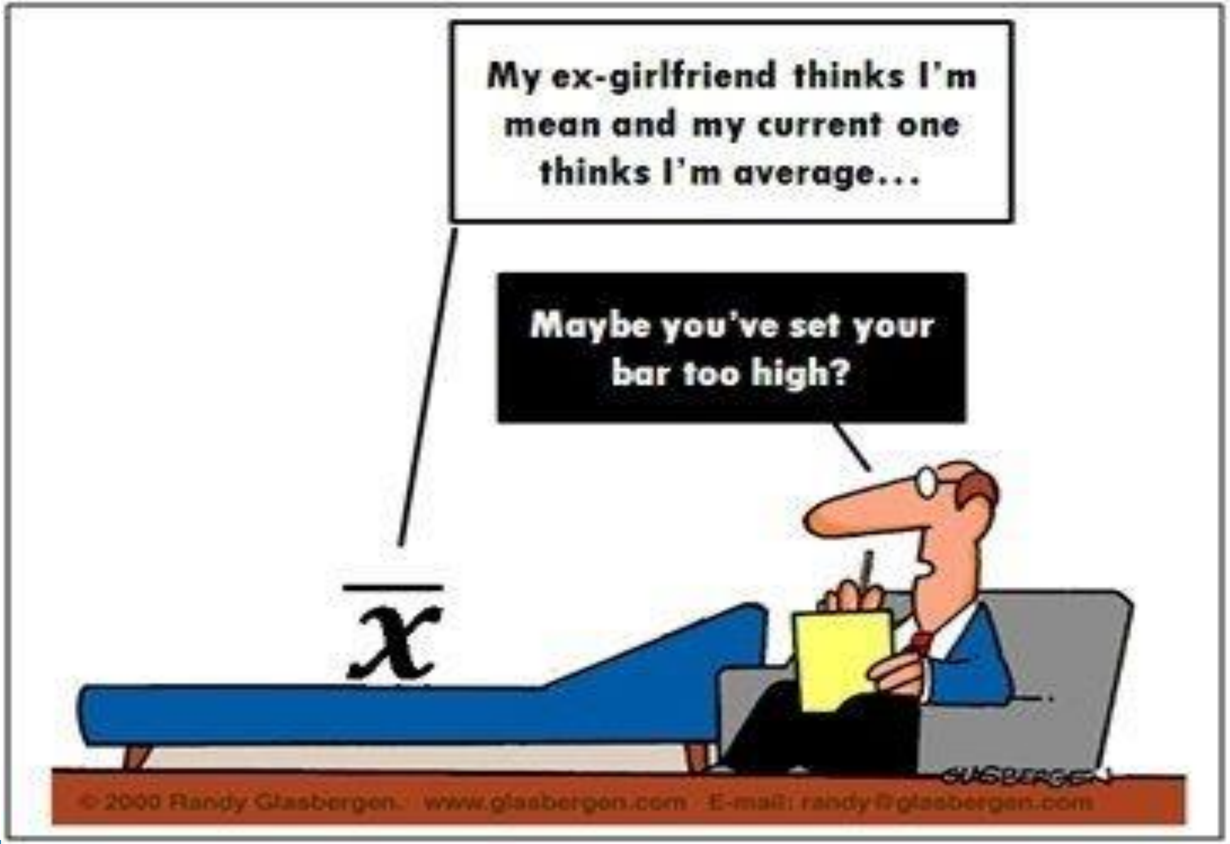
$$\gamma_1 = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$$

If the distribution is symmetric around mean the expected sum of X- μ will be zero

- Sample Skewness can be estimated as:

$$b_1 = \frac{m_3}{s^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}^{\,3}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^{3/2}}$$

# Mean and Average Are the Same !



# Agenda

- Introduction
- Measures of Central Tendency
- Measures of Dispersion
- Measures of Skewness
- **Practice in R**
- Measures of Dependence
- Practice in R
- Normal Distribution
- Examples in R

KENT STATE UNIVERSITY

# Practice in R

Script is available on the course website “Course Content\Scripts\Descriptive_Stats_1.r”

```
#read CSV file, by default the top column is the name of attributes
# hashtags are for comments!
MyData=read.csv('F:\\Kent Teaching\\Datasets\\Occupancy_Detection.csv'
head(MyData)                   #show the top fewlines
head(MyData$Temperature)       # MyData$Temperature is the Temperature
mean(MyData$Temperature)       #get the mean of MyData$Temperature
mean(MyData[,1])               #get the mean of the first column of MyD
sd(MyData$Temperature)         #get the standard deviation of MyData$Te
var(MyData$Temperature)        #get the variance of MyData$Temperature
quantile(MyData$Temperature)   #get the quantile of MyData$Temperature
#install.packages('moments')    #install the 'moments' package which ha
library(moments)
skewness(MyData$Temperature)
range(MyData$Temperature)
summary(MyData$Temperature)    #5-number summary (min, Q1, Median, Q3,
sapply(MyData,mean)
sapply(MyData,var)
sapply(MyData,quantile)
boxplot(MyData[,c(1,2)])
```

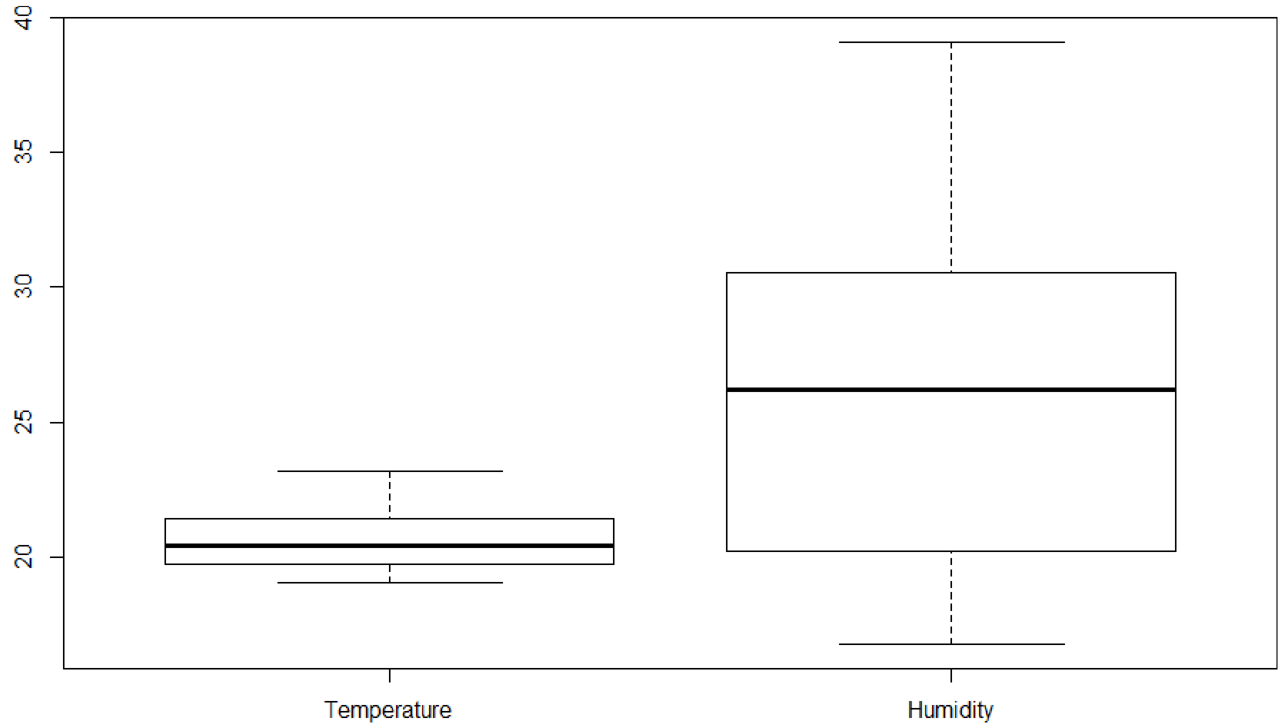
# Practice in R

```
> #read CSV file, by default the top column is the name of attributes
> # hashtags are for comments!
> MyData=read.csv('F:\\Kent Teaching\\Datasets\\Occupancy_Detection.csv');
> head(MyData)                      #show the top fewlines
  Temperature Humidity Light    CO2 HumidityRatio Occupancy
1       23.18  27.2720 426.0 721.25   0.004792988         1
2       23.15  27.2675 429.5 714.00   0.004783441         1
3       23.15  27.2450 426.0 713.50   0.004779464         1
4       23.15  27.2000 426.0 708.25   0.004771509         1
5       23.10  27.2000 426.0 704.50   0.004756993         1
6       23.10  27.2000 419.0 701.00   0.004756993         1
> head(MyData$Temperature)          # MyData$Temperature is the Temperature attribute
[1] 23.18 23.15 23.15 23.15 23.10 23.10
> mean(MyData$Temperature)          #get the mean of MyData$Temperature
[1] 20.61908
> mean(MyData[,1])                  #get the mean of the first column of MyData (whic is the
same as MyData$Temperature)
[1] 20.61908
> sd(MyData$Temperature)            #get the standard deviation of MyData$Temperature
[1] 1.016916
> var(MyData$Temperature)           #get the variance of MyData$Temperature
[1] 1.034119
> quantile(MyData$Temperature)      #get the quantile of MyData$Temperature
   0%   25%   50%   75%  100%
19.00 19.70 20.39 21.39 23.18
```

# Practice in R

```
> #install.packages('moments')   #install the 'moments' package which has an implemention
of Skewness statistics
> library(moments)
> skewness(MyData$Temperature)
[1] 0.4507854
> range(MyData$Temperature)
[1] 19.00 23.18
> summary(MyData$Temperature)   #5-number summary (min, Q1, Median, Q3, max) plus mean of
MyData$Temperature
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  19.00   19.70   20.39   20.62   21.39   23.18 
> sapply(MyData,mean)
  Temperature      Humidity         Light           CO2 HumidityRatio     Occupancy 
 2.061908e+01  2.573151e+01  1.195194e+02  6.065462e+02  3.862507e-03  2.123296e-01 
> sapply(MyData,var)
  Temperature      Humidity         Light           CO2 HumidityRatio     Occupancy 
 1.034119e+00  3.059430e+01  3.792982e+04  9.879761e+04  7.264687e-07  1.672663e-01 
> sapply(MyData,quantile)
     Temperature Humidity    Light       CO2 HumidityRatio Occupancy
0%         19.00 16.74500    0.000  412.7500   0.002674127         0
25%        19.70 20.20000    0.000  439.0000   0.003078284         0
50%        20.39 26.22250    0.000  453.5000   0.003800770         0
75%        21.39 30.53333  256.375  638.8333   0.004351931         0
100%       23.18 39.11750 1546.333 2028.5000   0.006476013         1
> boxplot(MyData[,c(1,2)])
```

# Practice in R



# Practice in R

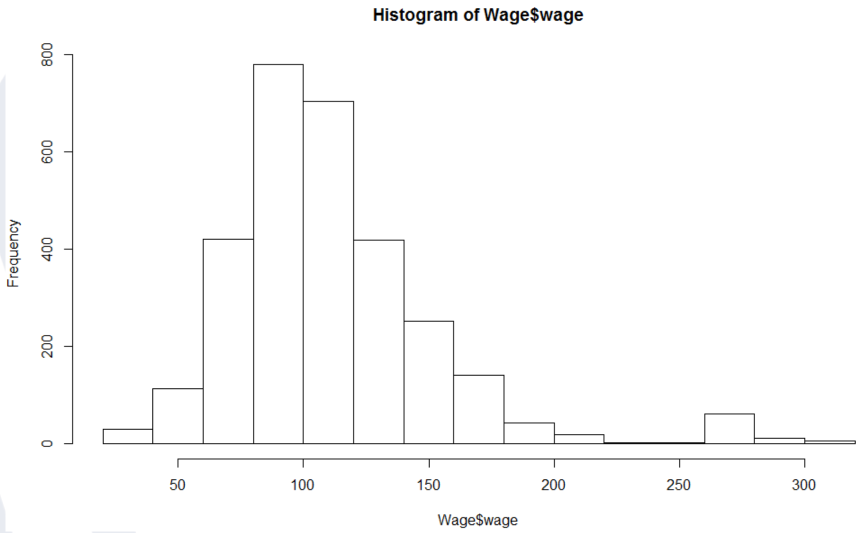
```
> install.packages('ISLR')
Installing package into 'C:/Users/lenovo pc/Documents/R/win-library/3.1'
(as 'lib' is unspecified)
trying URL 'http://cran.rstudio.com/bin/windows/contrib/3.1/ISLR_1.0.zip'
Content type 'application/zip' length 2912830 bytes (2.8 Mb)
opened URL
downloaded 2.8 Mb

package 'ISLR' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
 C:\Users\lenovo pc\AppData\Local\Temp\RtmpA3f3un\downloaded_packages
> library(ISLR)
Warning message:
package 'ISLR' was built under R version 3.1.3
> summary(Wage)
```

# Practice in R

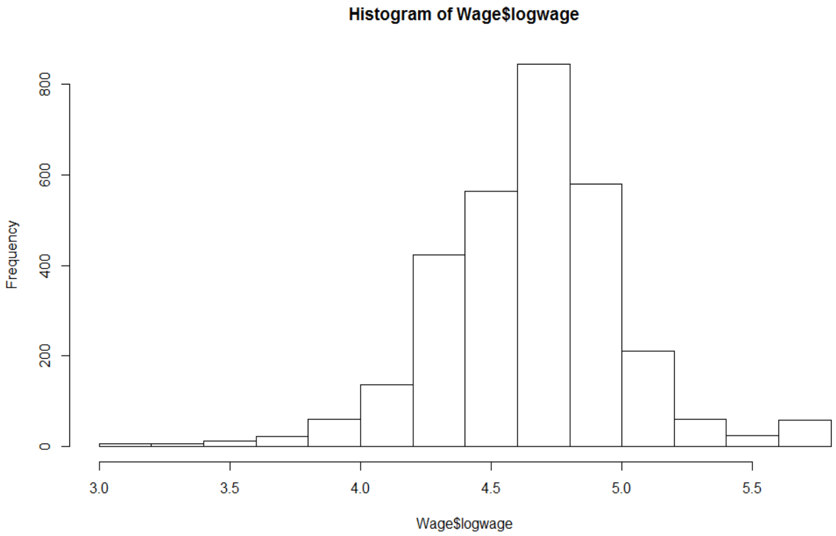
```
> summary(Wage)
      year           age             sex                    maritl    
 Min.   :2003   Min.   :18.00   1. Male  :3000   1. Never Married: 648  
 1st Qu.:2004   1st Qu.:33.75   2. Female:   0   2. Married      :2074  
 Median :2006   Median :42.00                    3. Widowed      :  19  
 Mean   :2006   Mean   :42.41                    4. Divorced     : 204  
 3rd Qu.:2008   3rd Qu.:51.00                    5. Separated    :  55  
 Max.   :2009   Max.   :80.00                                           

       race                   education                     region    
 1. White:2480   1. < HS Grad        :268   2. Middle Atlantic   :3000  
 2. Black: 293   2. HS Grad          :971   1. New England       :   0  
 3. Asian: 190   3. Some College     :650   3. East North Central:   0  
 4. Other:  37   4. College Grad     :685   4. West North Central:   0  
                 5. Advanced Degree:426     5. South Atlantic    :   0  
                                            6. East South Central:   0  
                                            (Other)              :   0  
           jobclass               health      health_ins      logwage     
 1. Industrial :1544   1. <=Good     : 858   1. Yes:2083   Min.   :3.000  
 2. Information:1456   2. >=Very Good:2142   2. No : 917   1st Qu.:4.447  
                                                           Median :4.653  
                                                           Mean   :4.654  
                                                           3rd Qu.:4.857  
                                                           Max.   :5.763  

      wage       
 Min.   : 20.09  
 1st Qu.: 85.38  
 Median :104.92  
 Mean   :111.70  
 3rd Qu.:128.68  
 Max.   :318.34  
```

# Practice in R

```
> hist(Wage$wage)
> library(moments)
> skewness(Wage$wage)
[1] 1.681489
```




```
> hist(Wage$logwage)
> skewness(Wage$logwage)
[1] -0.1235535
```



# Practice in R

```r
#%%%%%%%%%%%%%%%%%%%%%%%%%% Effect of Missing Values %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

MyData=read.csv('F:\\Kent Teaching\\Datasets\\Occupancy_Detection_Missing.csv');
mean(MyData$Temperature)
var(MyData$Temperature)
mean(MyData$Temperature,na.rm=TRUE) #This removes missing values before calcuatig stat
var(MyData$Temperature,,na.rm=TRUE)
```

22:2 (Top Level)

Console ~/

```
> MyData=read.csv('F:\\Kent Teaching\\Datasets\\Occupancy_Detection_Missing.csv');
> mean(MyData$Temperature)
[1] NA
> var(MyData$Temperature)
[1] NA
> mean(MyData$Temperature,na.rm=TRUE) #This removes missing values before calcuatig stat
[1] 20.61826
> var(MyData$Temperature,,na.rm=TRUE)
[1] 1.033083
```

# Agenda

- Introduction
- Measures of Central Tendency
- Measures of Dispersion
- Measures of Skewness
- Practice in R
- **Measures of Dependence**
- Practice in R
- Normal Distribution
- Examples in R

# Measures of Dependence

- Correlation is a statistical technique used to determine the degree to which two variables are related

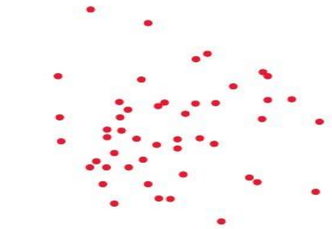
- Correlations are useful because they can indicate a predictive relationship that can be exploited in practice.

- Two types of correlation statistics are discussed in this course:
  - Pearson Correlation Coefficient
  - Spearman Rank Correlation Coefficient

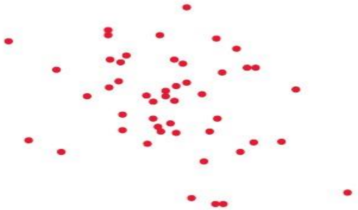
# Pearson Correlation Coefficient

- The Pearson correlation coefficient is a scale free measure of the linear correlation between two variables X and Y.
- It has a value between +1 and −1, where 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation.
- The correlation coefficient is symmetrical with respect to X and Y.
- The correlation coefficient is independent of the choice of origin and scale of measurement of the variables

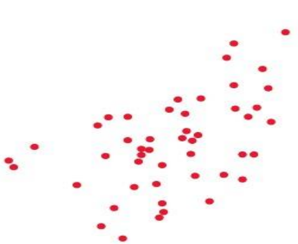
# Pearson Correlation Coefficient



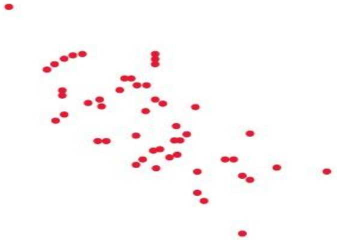
Correlation $r = 0$



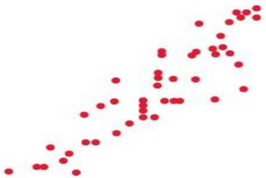
Correlation $r = -0.3$



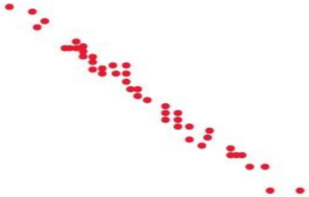
Correlation $r = 0.5$



Correlation $r = -0.7$

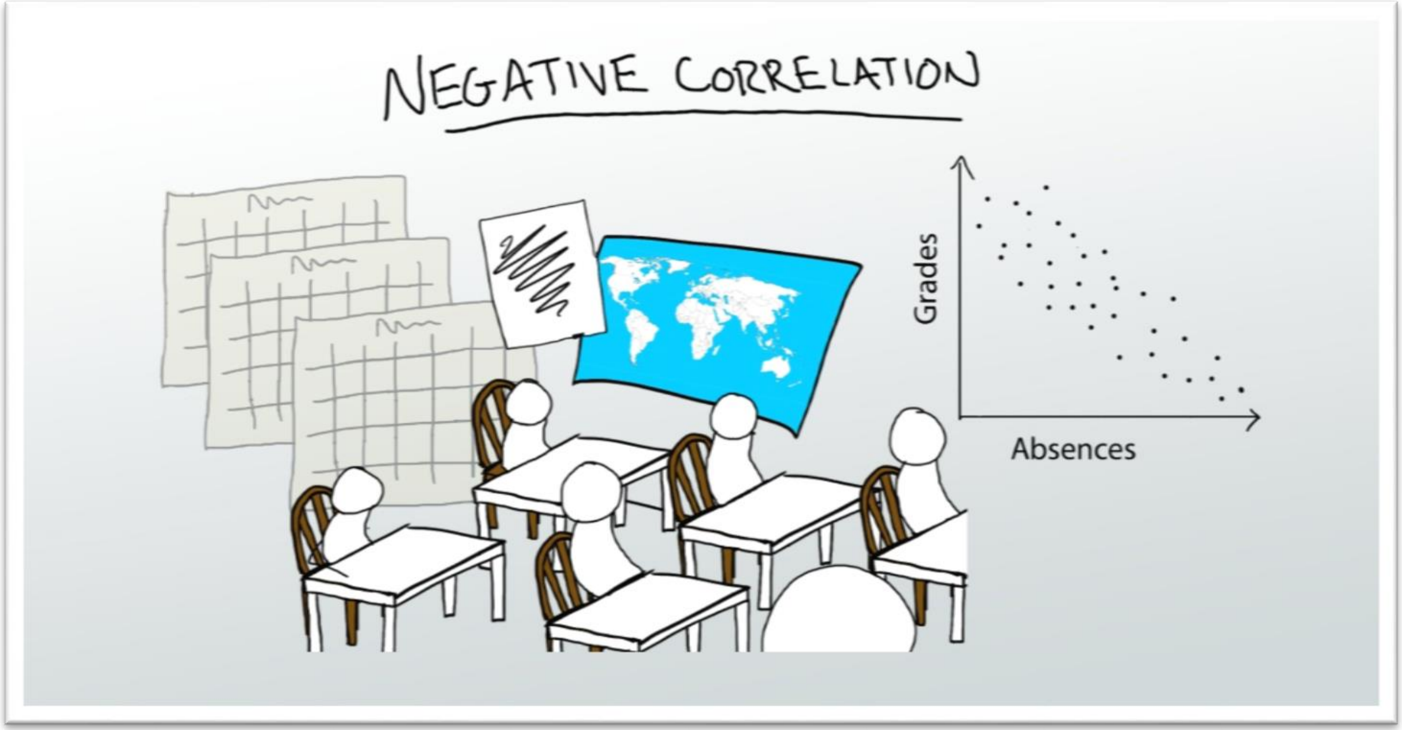


Correlation $r = 0.9$



Correlation $r = -0.99$

# We Can Confirm This One at the End of the Semester !



# Pearson Correlation Coefficient

If we have one dataset {x1,...,xn} containing n values and another dataset {y1,...,yn} containing n values then the Pearson correlation coefficient can be computed as

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where:

- $n$ is the number of samples
- $x_i, y_i$ are the single samples indexed with i
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (the sample mean); and analogously for $\bar{y}$

# Spearman Rank Correlation Coefficient

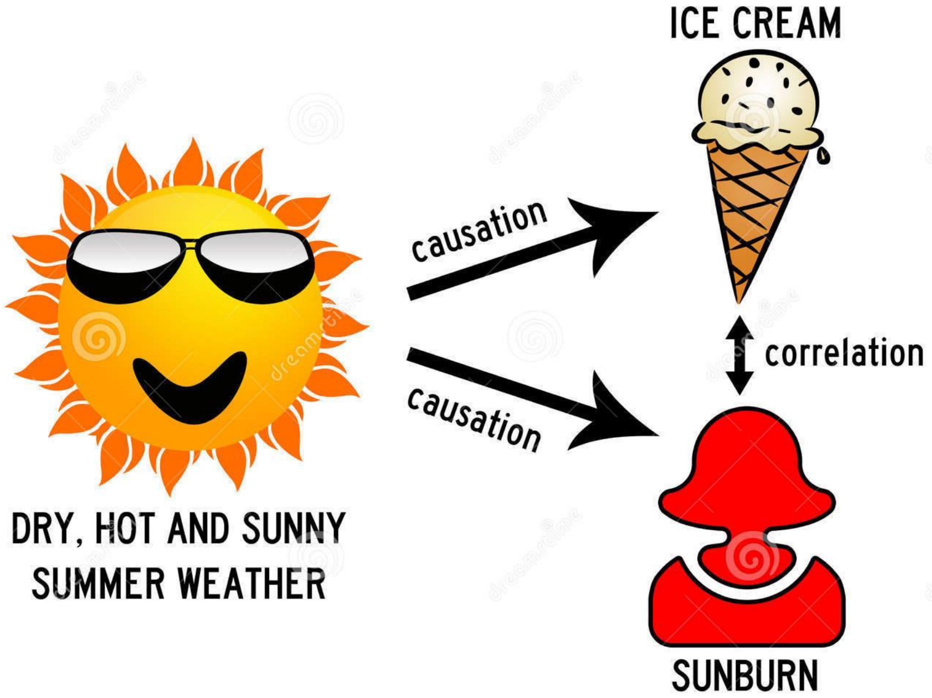
- The Spearman correlation between two variables is equal to the Pearson correlation between the rank values of those two variables; while Pearson's correlation assesses linear relationships, Spearman's correlation assesses monotonic relationships (whether linear or not).

- Spearman Rank can be applied to both ordinal (they are already ranked) and numerical (i.e. interval) variables. With numerical variables, we need to first sort them and use their ranking instead.

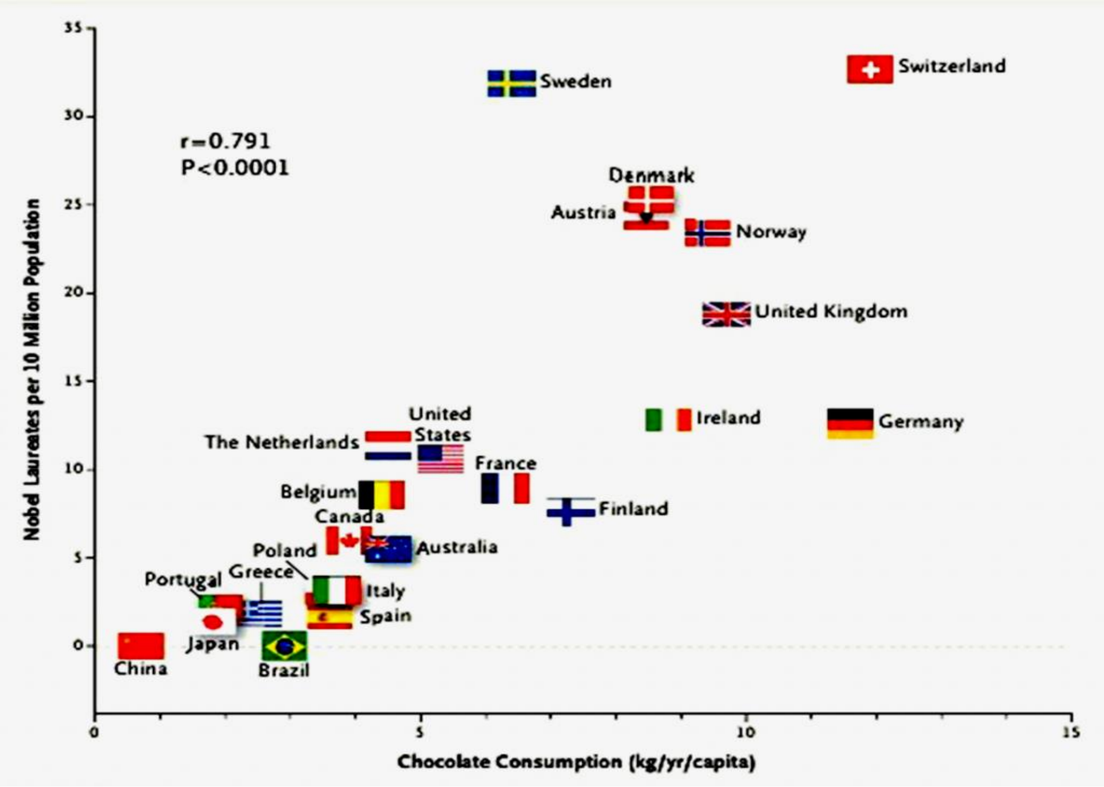
- Like Pearson correlation, the Spearman correlation is symmetric.

# Correlation Versus Causation

Click on the image to watch the video

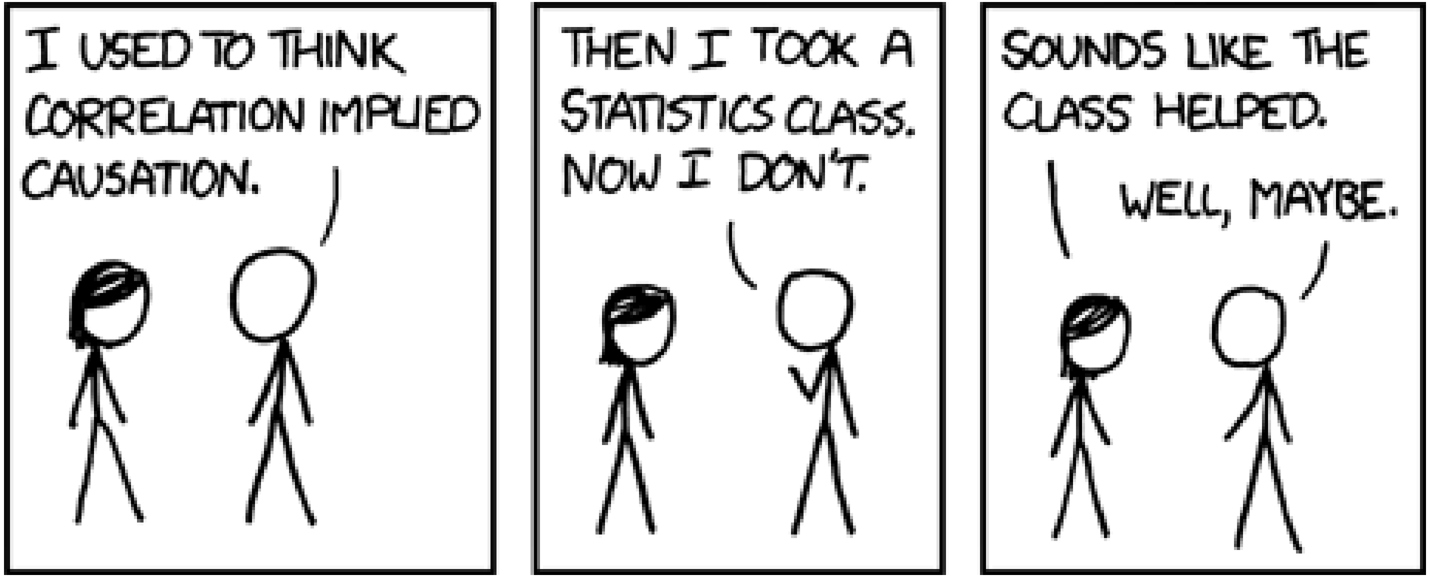


# Correlation Versus Causation



# Correlation Versus Causation

# Correlation Versus Causation



# Agenda

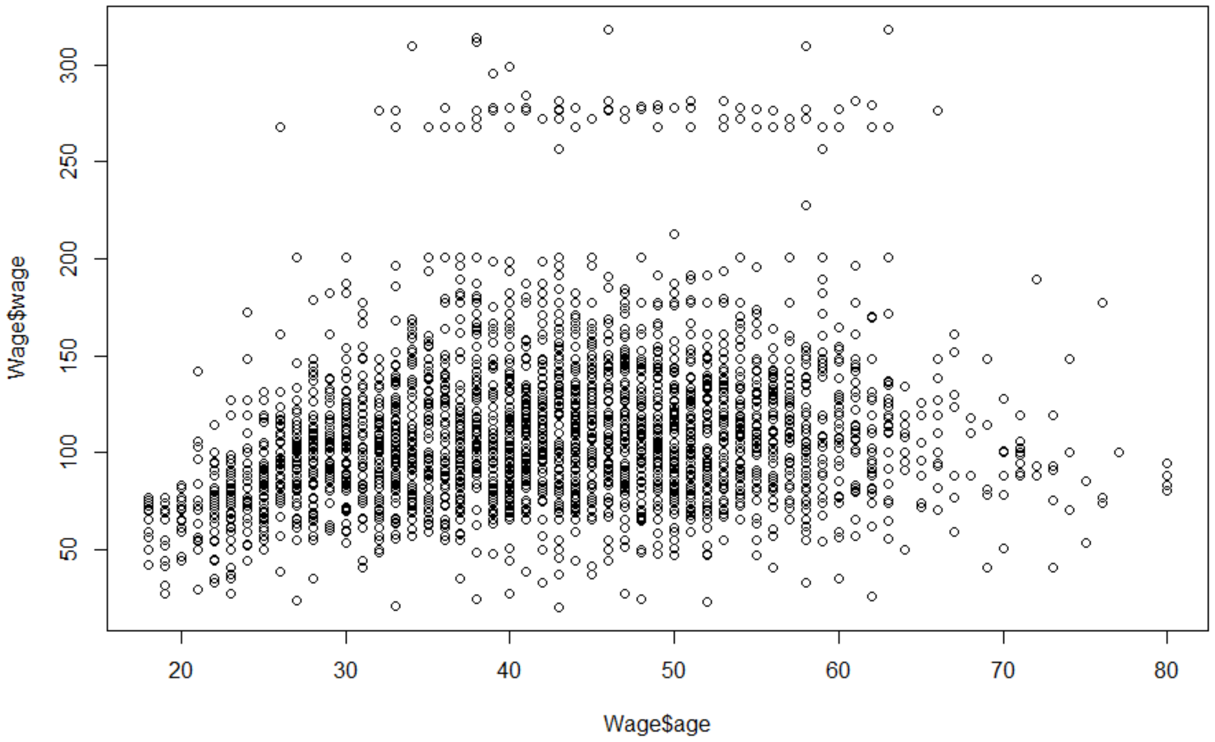
- Introduction
- Measures of Central Tendency
- Measures of Dispersion
- Measures of Skewness
- Practice in R
- Measures of Dependence
- **Practice in R**
- Normal Distribution
- Examples in R

# Examples in R

Script is available on the course website "Course Content\Scripts\Descriptive_Stats_2.r"

```
> library(ISLR)
> plot(Wage$age,Wage$wage)
> cor(Wage$age,Wage$wage,method = 'pearson') #Pearson Correlation
[1] 0.1956372
> cor(Wage$age,Wage$wage)    #defualt method is 'Pearson' so can be omitted
[1] 0.1956372
> cor(Wage$wage,Wage$age)                   #Correlation is symetric
[1] 0.1956372
> cor((Wage$age*10+1),(Wage$wage*0.2)+7) #Correlation is scale free
[1] 0.1956372
> ################################################################################
>
> #lets calucate the correlation step by step (i.e. no use of cor function)
> numerator=sum((Wage$age-mean(Wage$age))*(Wage$wage-mean(Wage$wage)))
> denominator=sqrt(sum((Wage$age-mean(Wage$age))^2))*sqrt(sum((Wage$wage-mean(Wage$wage))^2))
> numerator/denominator
[1] 0.1956372
>
> ####################################################################################
```

# Examples in R



# Examples in R

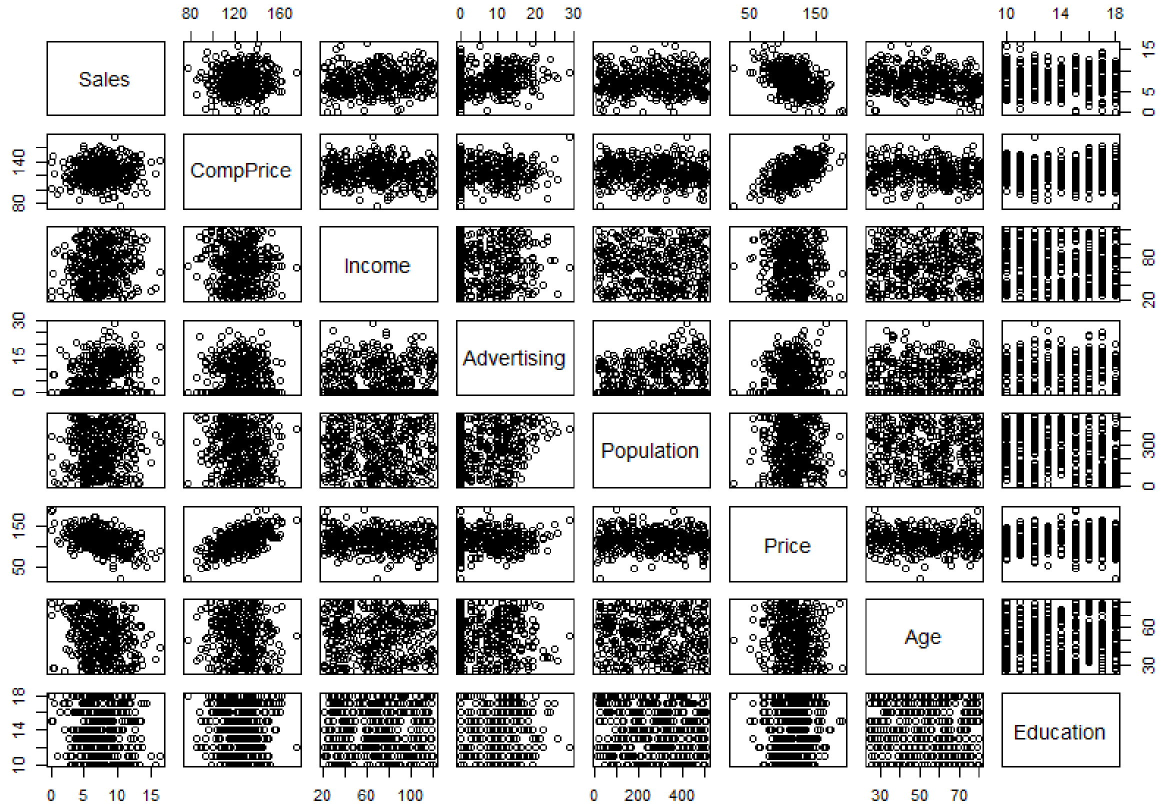
Console ~/

```
> cor(Wage$age,Wage$wage,method = 'spearman') #Spearman Correlation Coefficient
[1] 0.2298977
> levels(Wage$education)      #This shows differenet levels of 'education' variable
[1] "1. < HS Grad"        "2. HS Grad"          "3. Some College"     "4. College Grad"
[5] "5. Advanced Degree"
> head(as.numeric(Wage$education)) #This converts the variable to numeric (< HS Grad --> 1 etc)
[1] 1 4 3 4 2 4
> #head() can be used to show the top 6 records.
> cor(as.numeric(Wage$education),Wage$wage,method = 'spearman') #Spearman Correlation Coefficient
[1] 0.5031817
>
> ###########################################################################
> # A simple example to show the difference between Pearson and spearman co
>
> A=c(1,2,3,4,5);
> B=c(11,10,14,15,17);
> C=c(11,10,14,15,1000);
>
> cor(A,B,method = 'spearman')
[1] 0.9
> #The Spearman Correlation Coefficient dosent change as long as the rank orders are the same
> cor(A,C,method = 'spearman')
[1] 0.9
>
> #... but Pearson Correlation does!
> cor(A,B,method = 'pearson')
[1] 0.9329962
> cor(A,C,method = 'pearson')
[1] 0.7099633
```

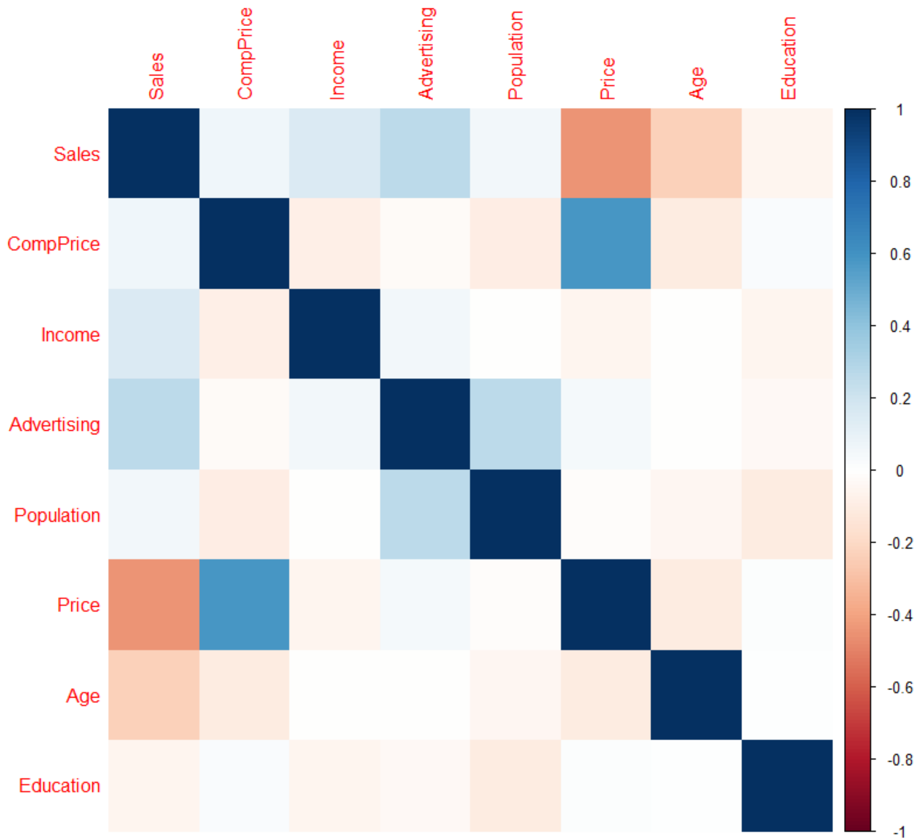
# Examples in R

```
> summary(Carseats) #summary of car seat dataset
     Sales          CompPrice       Income        Advertising       Population        Price      
 Min.   : 0.000   Min.   : 77   Min.   : 21.00   Min.   : 0.000   Min.   : 10.0   Min.   : 24.0  
 1st Qu.: 5.390   1st Qu.:115   1st Qu.: 42.75   1st Qu.: 0.000   1st Qu.:139.0   1st Qu.:100.0  
 Median : 7.490   Median :125   Median : 69.00   Median : 5.000   Median :272.0   Median :117.0  
 Mean   : 7.496   Mean   :125   Mean   : 68.66   Mean   : 6.635   Mean   :264.8   Mean   :115.8  
 3rd Qu.: 9.320   3rd Qu.:135   3rd Qu.: 91.00   3rd Qu.:12.000   3rd Qu.:398.5   3rd Qu.:131.0  
 Max.   :16.270   Max.   :175   Max.   :120.00   Max.   :29.000   Max.   :509.0   Max.   :191.0  
  ShelveLoc        Age          Education    Urban       US     
 Bad   : 96   Min.   :25.00   Min.   :10.0   No :118   No :142  
 Good  : 85   1st Qu.:39.75   1st Qu.:12.0   Yes:282   Yes:258  
 Medium:219   Median :54.50   Median :14.0                      
              Mean   :53.32   Mean   :13.9                      
              3rd Qu.:66.00   3rd Qu.:16.0                      
              Max.   :80.00   Max.   :18.0                      
> Carseat_num<-Carseats[,c(1:6,8,9)] #selecting numerical variables (columns 1-6,8,9)
> pairs(Carseat_num)
> install.packages('corrplot')
> library(corrplot)
> corrplot(cor(Carseat_num), method="color")
> 
```

# Examples in R



# Examples in R



# Agenda

- Introduction
- Measures of Central Tendency
- Measures of Dispersion
- Measures of Skewness
- Practice in R
- Measures of Dependence
- Practice in R
- **Normal Distribution**
- Examples in R

# Continuous Probability Distributions

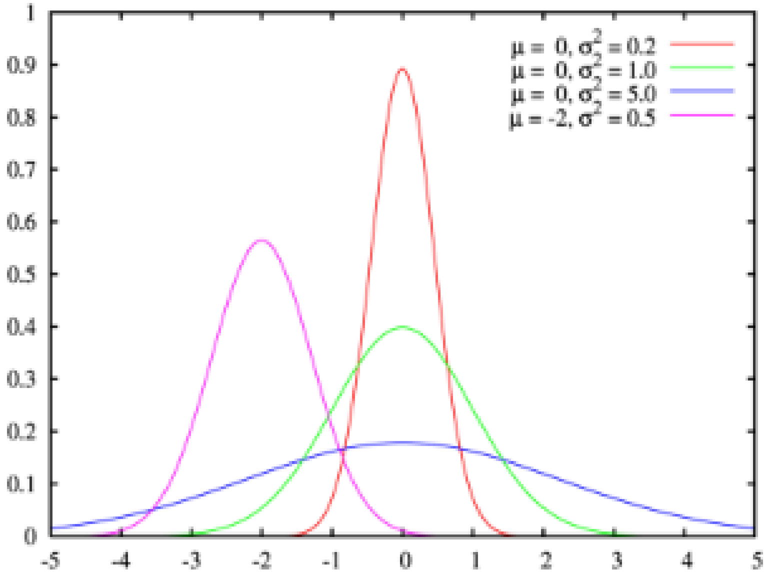
- Continuous Random Variable
  - Values from interval of numbers
  - Absence of gaps

- Continuous Probability Distribution
  - Distribution of continuous random variable

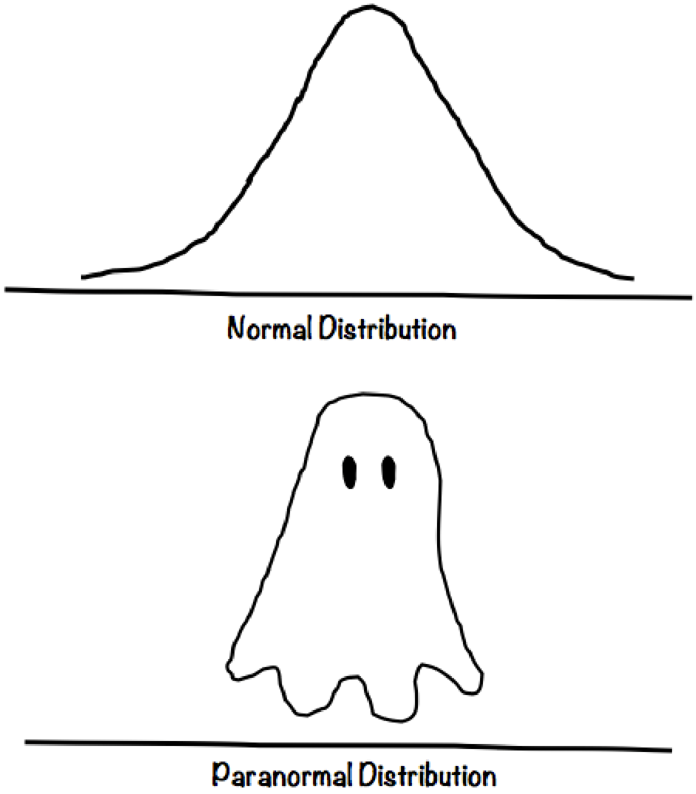
- Most Important Continuous Probability Distribution
  - The normal distribution

# Normal Distribution

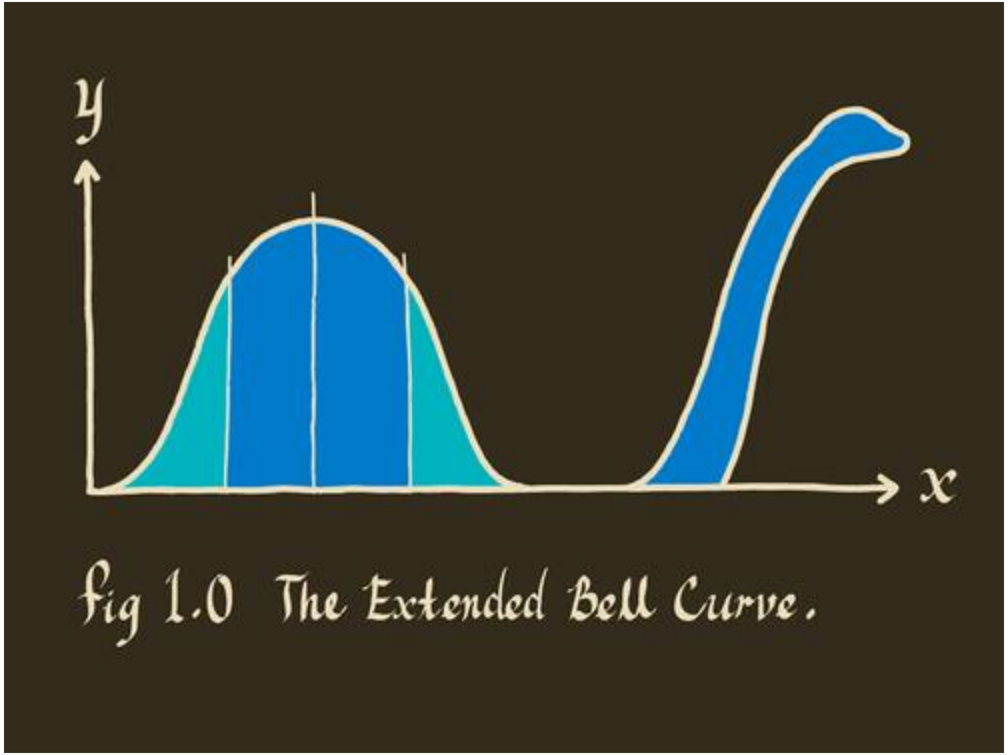
The normal distribution is a probability distribution that resembles a bell curve. A normal distribution has two parameters, a mean (denoted μ) and a standard deviation (denoted σ).



# Normal Versus Paranormal Distribution!



# …and an Extended Bell Curve



# Business Applications of Normal Distribution

The Normal (Gaussian) distribution has many business applications. e.g.:

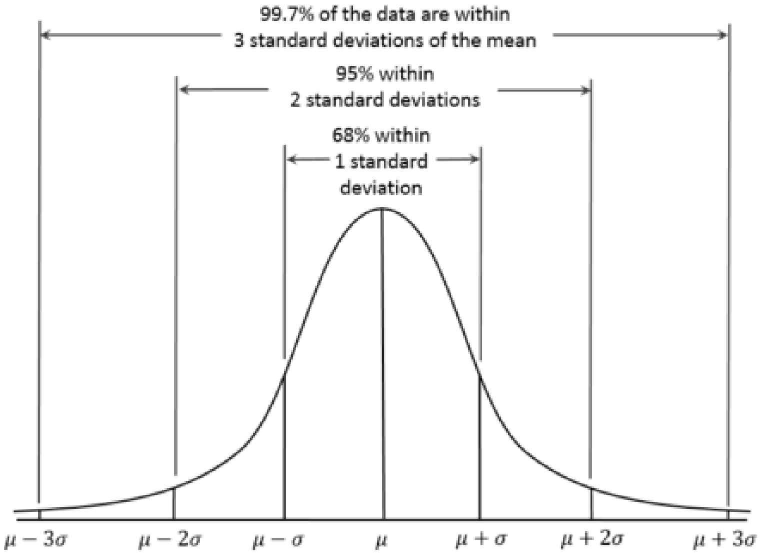
- In the field of operations management, results of many processes fall along the Normal distribution.
- The Normal Probability Distribution governs many aspects of human performance. e.g. employee performance.
- A diversified portfolio will typically have returns that fall in a Normal distribution.

<u>The Normal distribution is often a rough substitute for any distribution that is symmetrically distributed about an axis and is unimodal</u>

# 68–95–99.7 Rule

Empirical rule: If distribution is approx. bell-shaped:

- about 68% of data within 1 standard dev. of mean
- about 95% of data within 2 standard dev. of mean
- about 99.7% of data within 3 standard dev. of mean



# z-score (Standard score )

- The z-score is a dimensionless quantity obtained by subtracting the population mean from an individual raw score and then dividing the difference by the population standard deviation.

$$z = \frac{x - \mu}{\sigma}$$

- In simple words, z-score says how many standard deviation s we are away from the mean.

- Knowing the z-score, we can easily lookup the probability using tables or software packages.

- The probability is expressed as the left area to the left of the Z score. i.e. P(x<X)

# z-score (Standard score )

- You can find the Standard Normal Probabilities on the course website under Course content> Standard Normal Probabilities

- Example: find the accumulative probability associated to z=0.12

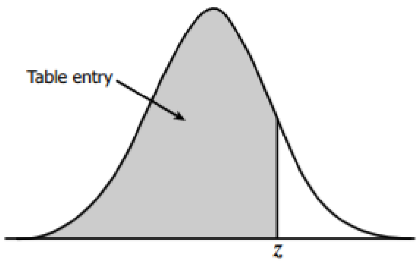
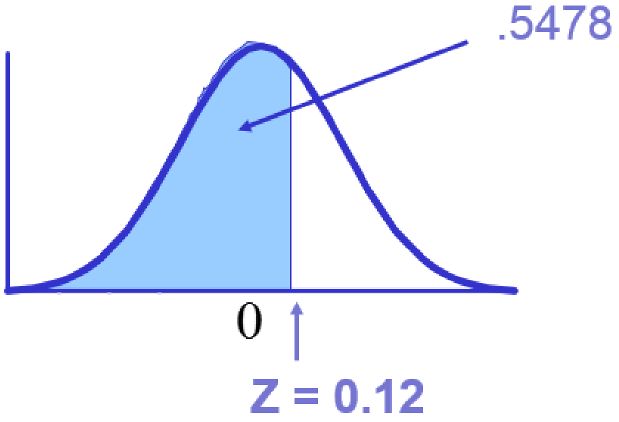



Table entry for $z$ is the area un
to the left of $z$.

| $z$ | .00 | .01 | .02 | .03 | .04 | .05 | .06 |
|---|---|---|---|---|---|---|---|
| 0.0 | .5000 | .5040 | .5080 | .5120 | .5160 | .5199 | .5239 |
| 0.1 | .5398 | .5438 | $\boxed{.5478}$ | .5517 | .5557 | .5596 | .5636 |
| 0.2 | .5793 | .5832 | .5871 | .5910 | .5948 | .5987 | .6026 |
| 0.3 | .6179 | .6217 | .6255 | .6293 | .6331 | .6368 | .6406 |
| 0.4 | .6554 | .6591 | .6628 | .6664 | .6700 | .6736 | .6772 |



# Example I

- The distribution of the budget needed to complete a specific project has found to have a normal shape with a mean of $235m and a standard deviation of $20m. What is the probability that the project can be competed with $266 ?

- Even before starting, we know the expected probability should be greater than 0.5 since the allocated budget is greater than the mean (i.e. $z>0$)

- $z=(266-235)/20=1.55$

- Answer p=0.9394

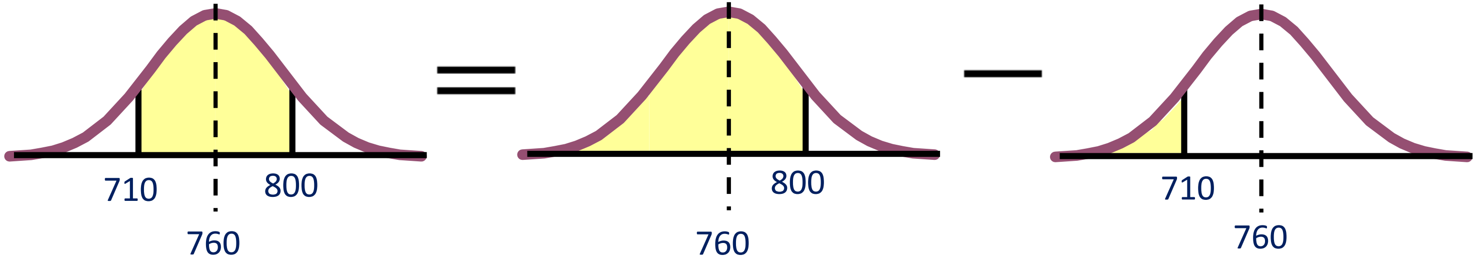
➔ Still 6% chance that you would need more.

| z | .00 | .01 | .02 | .03 | .04 | .05 |
|---|---|---|---|---|---|---|
| 0.0 | .5000 | .5040 | .5080 | .5120 | .5160 | .5199 |
| 0.1 | .5398 | .5438 | .5478 | .5517 | .5557 | .5596 |
| 0.2 | .5793 | .5832 | .5871 | .5910 | .5948 | .5987 |
| 0.3 | .6179 | .6217 | .6255 | .6293 | .6331 | .6368 |
| 0.4 | .6554 | .6591 | .6628 | .6664 | .6700 | .6736 |
| 0.5 | .6915 | .6950 | .6985 | .7019 | .7054 | .7088 |
| 0.6 | .7257 | .7291 | .7324 | .7357 | .7389 | .7422 |
| 0.7 | .7580 | .7611 | .7642 | .7673 | .7704 | .7734 |
| 0.8 | .7881 | .7910 | .7939 | .7967 | .7995 | .8023 |
| 0.9 | .8159 | .8186 | .8212 | .8238 | .8264 | .8289 |
| 1.0 | .8413 | .8438 | .8461 | .8485 | .8508 | .8531 |
| 1.1 | .8643 | .8665 | .8686 | .8708 | .8729 | .8749 |
| 1.2 | .8849 | .8869 | .8888 | .8907 | .8925 | .8944 |
| 1.3 | .9032 | .9049 | .9066 | .9082 | .9099 | .9115 |
| 1.4 | .9192 | .9207 | .9222 | .9236 | .9251 | .9265 |
| 1.5 | .9332 | .9345 | .9357 | .9370 | .9382 | $\boxed{.9394}$ |

# Example II

- The average price for a 42-inch TV at Best Buy seems to follow a normal distribution with mean 760\$ and standard deviation 145\$. What is the likelihood that a randomly selected TV has price a) between 710\$-800\$? B) above 730?

Part A) : $P(710<\text{Price}<800)= P(\text{Price}<800)- P(\text{Price}<710)$




…Continued

# Example II

P(Price<800)=?    z=(800-760)/145=0.27

| z | .00 | .01 | .02 | .03 | .04 | .05 | .06 | .07 |
|---|---|---|---|---|---|---|---|---|
| 0.0 | .5000 | .5040 | .5080 | .5120 | .5160 | .5199 | .5239 | .5279 |
| 0.1 | .5398 | .5438 | .5478 | .5517 | .5557 | .5596 | .5636 | .5675 |
| 0.2 | .5793 | .5832 | .5871 | .5910 | .5948 | .5987 | .6026 | $\boxed{.6064}$ |
| 0.3 | .6179 | .6217 | .6255 | .6293 | .6331 | .6368 | .6406 | .6443 |

P(Price<710)=?    z=(710-760)/145= -0.34,  After lookup
P(Price<710)=0.3669

P(710<Price<800)= 0.6064-0.3669=0.2395

Part B: P(Price>730)=1-P(Price≤730)

P(Price≤730) =? z=(730-760)/145=-0.21 (round up) ==> P(Price≤730) =0.4168

P(Price>730)=1-0.4168= 0.5832

# Example III

- An independent third-party consultancy has reviewed the performance of a large number of companies in the financial sector. Each reviewed company has received a score in range 0-10. The scores seems to follow a Gaussian distribution with mean of 7.55 and standard deviation of 1.2. The board of directors at "ABC Financials" wanted to assure that they are in top 10% of their industry. What should be their minimum score for this?

This is a reverse problem, let's see what minimum z-score they should have to be amongst top 10%? Being amongst top 10% means receiving a reviewing score greater than 90% of other complies. In other words we are looking for smallest z-score which assures that the accumulative probability is greater than 0.9. Looking at the table we have:

# Example III

| z | .00 | .01 | .02 | .03 | .04 | .05 | .06 | .07 | .08 | .09 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.0 | .5000 | .5040 | .5080 | .5120 | .5160 | .5199 | .5239 | .5279 | .5319 | .5359 |
| 0.1 | .5398 | .5438 | .5478 | .5517 | .5557 | .5596 | .5636 | .5675 | .5714 | .5753 |
| 0.2 | .5793 | .5832 | .5871 | .5910 | .5948 | .5987 | .6026 | .6064 | .6103 | .6141 |
| 0.3 | .6179 | .6217 | .6255 | .6293 | .6331 | .6368 | .6406 | .6443 | .6480 | .6517 |
| 0.4 | .6554 | .6591 | .6628 | .6664 | .6700 | .6736 | .6772 | .6808 | .6844 | .6879 |
| 0.5 | .6915 | .6950 | .6985 | .7019 | .7054 | .7088 | .7123 | .7157 | .7190 | .7224 |
| 0.6 | .7257 | .7291 | .7324 | .7357 | .7389 | .7422 | .7454 | .7486 | .7517 | .7549 |
| 0.7 | .7580 | .7611 | .7642 | .7673 | .7704 | .7734 | .7764 | .7794 | .7823 | .7852 |
| 0.8 | .7881 | .7910 | .7939 | .7967 | .7995 | .8023 | .8051 | .8078 | .8106 | .8133 |
| 0.9 | .8159 | .8186 | .8212 | .8238 | .8264 | .8289 | .8315 | .8340 | .8365 | .8389 |
| 1.0 | .8413 | .8438 | .8461 | .8485 | .8508 | .8531 | .8554 | .8577 | .8599 | .8621 |
| 1.1 | .8643 | .8665 | .8686 | .8708 | .8729 | .8749 | .8770 | .8790 | .8810 | .8830 |
| 1.2 | .8849 | .8869 | .8888 | .8907 | .8925 | .8944 | .8962 | .8980 | .8997 | $\boxed{.9015}$ |
| 1.3 | .9032 | .9049 | .9066 | .9082 | .9099 | .9115 | .9131 | .9147 | .9162 | .9177 |

$z=1.29$ (From Table) $\mu=7.55$ $\sigma=1.2$, $x=?$

$1.29=(x-7.55)/1.2 \rightarrow x=9.098$

ABC Financials score should be above 9.098 to place them amongst the top 10%.

# Example IV

- You have inspected a manufacturing line that produces a special plastic tubes. You have measured the length of large number of samples of the tubes and the distribution of the measurements was normal with the average length being 20.50m. Few days later, your manager asks you about the standard deviation of your measurements. You don't recall. But you clearly remember that 7% of samples where longer than 20.55m. Can you estimate the standard deviation without getting back to your measurement data?

First, try to solve it yourself without looking at the solution in the next slide.

# Example IV

What is the z-score of an observation that stands at 93-percentile? (i.e. only 7% of the observations are longer than that observation).

Lets check the table

$z=1.48$ (From Table)

$x=20.55$

$\mu=20.50$

$\sigma=?$

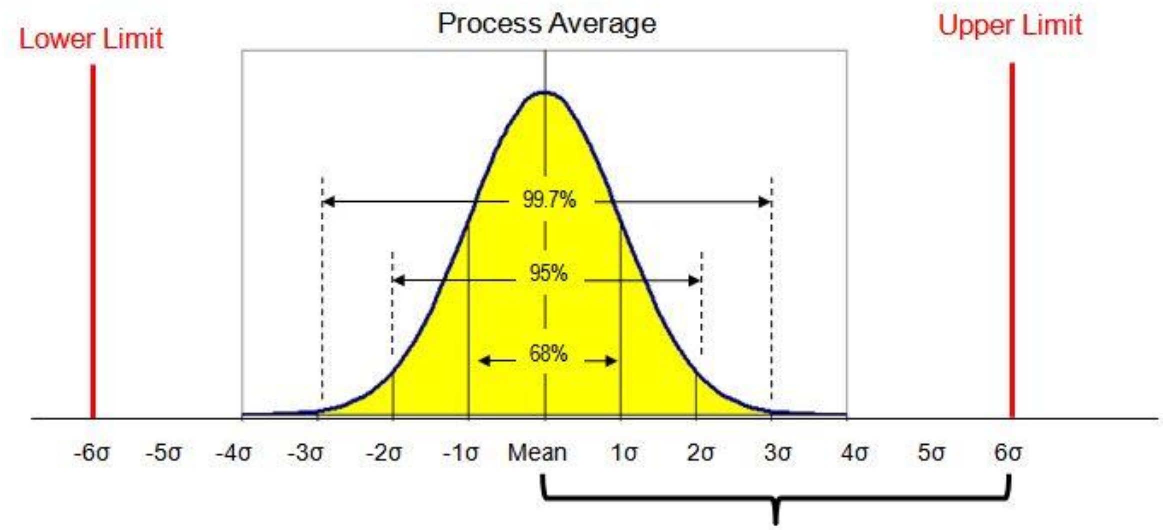
$1.48=(20.55-20.50)/\sigma$

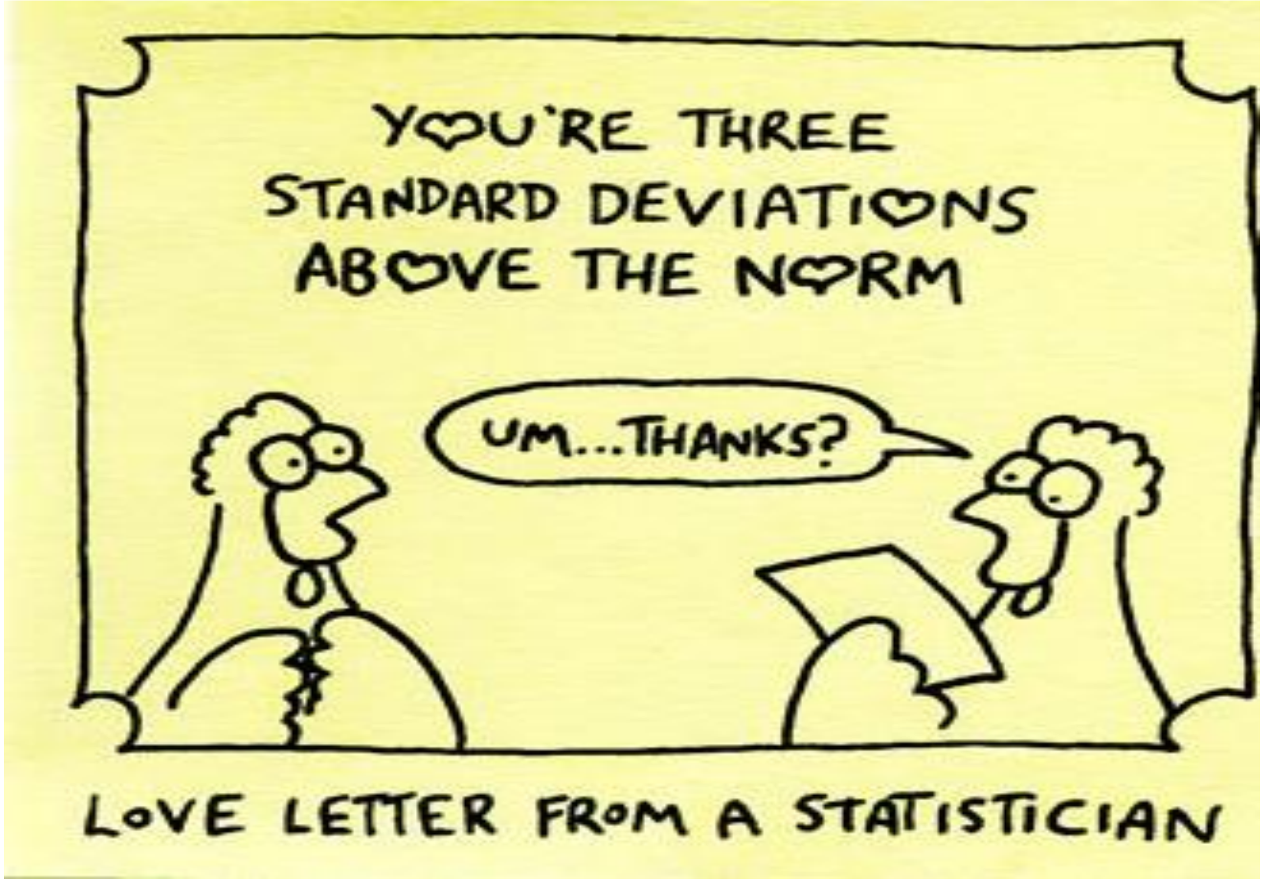
$\sigma=0.0338$

| $z$ | .00 | .01 | .02 | .03 | .04 | .05 | .06 | .07 | .08 | .09 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.0 | .5000 | .5040 | .5080 | .5120 | .5160 | .5199 | .5239 | .5279 | .5319 | .5359 |
| 0.1 | .5398 | .5438 | .5478 | .5517 | .5557 | .5596 | .5636 | .5675 | .5714 | .5753 |
| 0.2 | .5793 | .5832 | .5871 | .5910 | .5948 | .5987 | .6026 | .6064 | .6103 | .6141 |
| 0.3 | .6179 | .6217 | .6255 | .6293 | .6331 | .6368 | .6406 | .6443 | .6480 | .6517 |
| 0.4 | .6554 | .6591 | .6628 | .6664 | .6700 | .6736 | .6772 | .6808 | .6844 | .6879 |
| 0.5 | .6915 | .6950 | .6985 | .7019 | .7054 | .7088 | .7123 | .7157 | .7190 | .7224 |
| 0.6 | .7257 | .7291 | .7324 | .7357 | .7389 | .7422 | .7454 | .7486 | .7517 | .7549 |
| 0.7 | .7580 | .7611 | .7642 | .7673 | .7704 | .7734 | .7764 | .7794 | .7823 | .7852 |
| 0.8 | .7881 | .7910 | .7939 | .7967 | .7995 | .8023 | .8051 | .8078 | .8106 | .8133 |
| 0.9 | .8159 | .8186 | .8212 | .8238 | .8264 | .8289 | .8315 | .8340 | .8365 | .8389 |
| 1.0 | .8413 | .8438 | .8461 | .8485 | .8508 | .8531 | .8554 | .8577 | .8599 | .8621 |
| 1.1 | .8643 | .8665 | .8686 | .8708 | .8729 | .8749 | .8770 | .8790 | .8810 | .8830 |
| 1.2 | .8849 | .8869 | .8888 | .8907 | .8925 | .8944 | .8962 | .8980 | .8997 | .9015 |
| 1.3 | .9032 | .9049 | .9066 | .9082 | .9099 | .9115 | .9131 | .9147 | .9162 | .9177 |
| 1.4 | .9192 | .9207 | .9222 | .9236 | .9251 | .9265 | .9279 | .9292 | $\boxed{.9306}$ | .9319 |
| 1.5 | .9332 | .9345 | .9357 | .9370 | .9382 | .9394 | .9406 | .9418 | .9429 | .9441 |

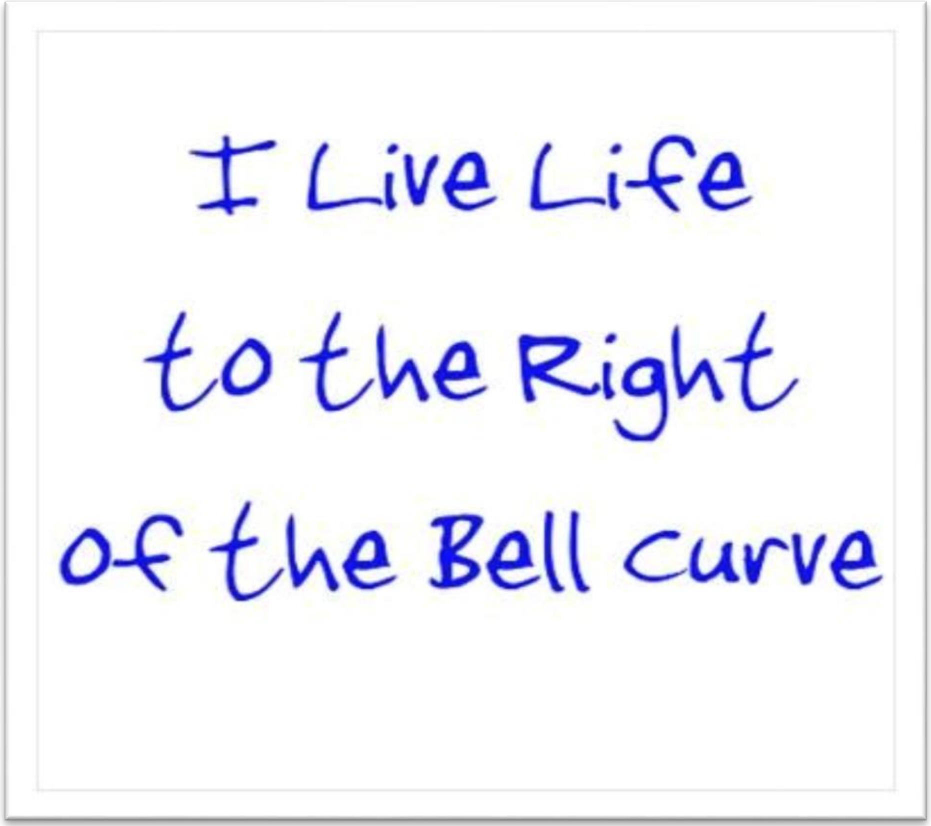
# Six Sigma (6σ)

- Six Sigma (6σ) is a set of techniques and tools for process improvement. A six sigma process is one in which 99.99966% of all opportunities to produce some feature of a part are statistically expected to be free of defects (3.4 defective features per million opportunities).




YOU'RE THREE STANDARD DEVIATIONS ABOVE THE NORM
UM...THANKS?
LOVE LETTER FROM A STATISTICIAN

# Living More Than an Average Life …



# Agenda

- Introduction
- Measures of Central Tendency
- Measures of Dispersion
- Measures of Skewness
- Practice in R
- Measures of Dependence
- Practice in R
- Normal Distribution
- **Examples in R**

# Examples in R

Script is available on the course website "Course Content\Scripts\Descriptive_Stats_3.r"

Console ~/

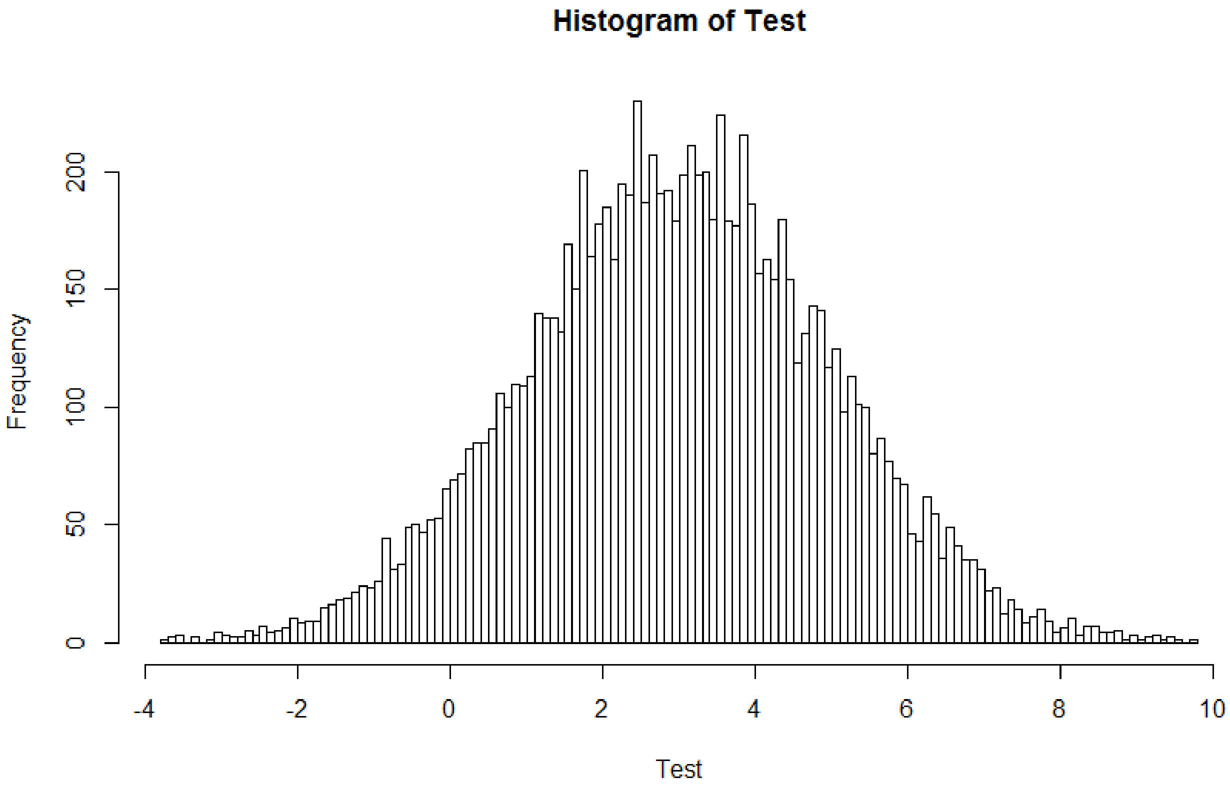
```
> #pnorm(z) returns the area under the curve from -inf to
> #z (cumulative probability) of the pdf of the normal distribution where z is a Z-score
> #i.e. does the job of Table lookup
> pnorm(0)
[1] 0.5
> pnorm(0.12) #example in our slides
[1] 0.5477584
>
> #examination of 68-95-99.7 rule
> pnorm(1)-pnorm(-1)  #Recall that 68% of data  within 1 standard dev. of mean
[1] 0.6826895
> pnorm(2)-pnorm(-2)  #Recall that 95% of data  within 2 standard dev. of mean
[1] 0.9544997
> pnorm(3)-pnorm(-3)  #Recall that 99.7% of data  within 2 standard dev. of mean
[1] 0.9973002
>
> #pnorm() function allows you to be lazy and to provide mean and sd as well instead of z-score
> #example x=6, mu=3, sd=2 so z=(6-3)/2=1.5 so you can use pnorm(1.5) OR pnorm(6,mean=3,sd=2)
> pnorm(1.5)
[1] 0.9331928
> pnorm(6,mean=3,sd=2)
[1] 0.9331928
> pnorm(266,mean=235,sd=20) #Example I in our slides
[1] 0.9394292
> pnorm(800,mean=760,sd=146)-pnorm(710,mean=760,sd=146) #Example II part A
[1] 0.241947
> #The small difference in answers is because of the rounding errors
```

# Examples in R

Console ~/

```
> #pnorm by defult returns the area under the lower tail i.e. P[X = x] but you can change this.
> pnorm(1.5,lower.tail = FALSE) #This is the upper
[1] 0.0668072
> # the sum is obviously 1
> pnorm(1.5,lower.tail = FALSE)+pnorm(1.5,lower.tail = TRUE)
[1] 1
> pnorm(730,mean=760,sd=146,lower.tail = FALSE) #Example II part B
[1] 0.5814012
> #The small difference in answers is because of the rounding errors
>
> ################################################################
> #qnorm is the  is the inverse of pnorm. The idea behind qnorm is that you give it a probability,
> #and it returns the z-score whose cumulative distribution matches the probability.
> #No need for reverse table reading!
>
> qnorm(0.5)
[1] 0
> #Similarly, you can define mean and sd so it returns the observatin value instead of the z-score.
> qnorm(0.9,mean=7.55,sd=1.2) #Example III
[1] 9.087862
> qnorm(0.93) #Part of Example IV
[1] 1.475791
> ##############################
> #rnorm() generates random numbers that follow normal distribution, defult mu=1,sd=1
> Test<-rnorm(10000,mean=3,sd=2) # 10000 random numbers with normal distribution, mean=3,sd=2
> hist(Test)
> hist(Test,n=100) #use 100 split bins i.e. higher resolution
```

# Examples in R



# What We Covered Today

- Introduction
- Measures of Central Tendency
- Measures of Dispersion
- Measures of Skewness
- Practice in R
- Measures of Dependence
- Practice in R
- Normal Distribution
- Examples in R


ANY
QUESTIONS
?