

# Setting up your machine

- Download a copy of [R](#) on your local computer from the Comprehensive R Archive Network (CRAN). You can choose between binaries for Linux, Mac and Windows.
- Install one of R's integrated development environment (IDE), [RStudio](#), which makes R coding much easier and faster as it allows you to type multiple lines of code, handle plots, install and maintain packages and navigate your programming environment much more productively.

<https://cran.r-project.org/doc/manuals/R-intro.pdf>

- Download contents located in the Github

<https://github.com/beveratraining/Introduction-to-Statistical-Analysis-in-R>



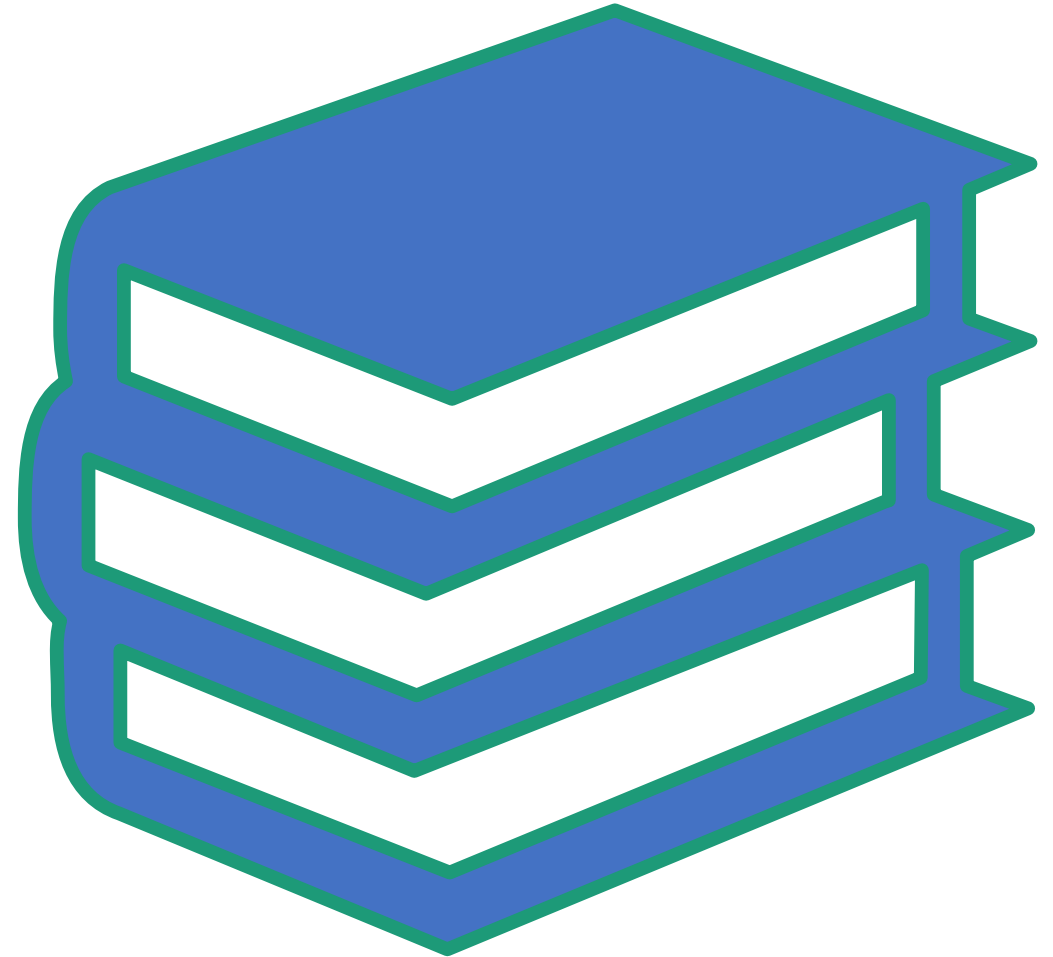
# Statistical Analysis in R Participant's Guide

Instructor: Brian Ashford  
Yvonne Phillips

# About the course

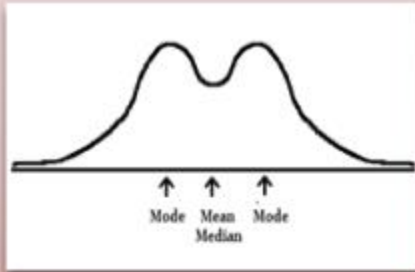
---

The *Statistics in R* course was written to provide resource material that is useful for everyday statistical project work and prepared the student for higher level statistical learning courses in Machine Learning. The content is ordered along the lines of many popular statistics books. The *Statistics in R* course includes some theory to introduce statistical concept but is mainly a hands-on workshop with the intent of the course to provide the class participant with R script programs to demonstrate important topics and concepts covered in a statistics course.



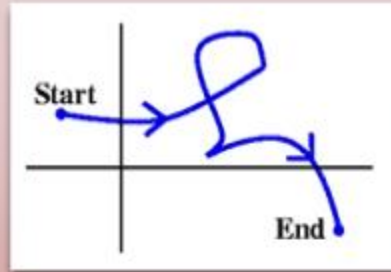
# Agenda

- Community Agreement
- Article Read
- Review - Data Types
- Descriptive Statistics vs. Inferential Statistics
- Review - Descriptive Statistics
- Central Tendency and Dispersion
- Statistical theory
- Frequency Distributions
- Statistical Distributions
- Sampling
- Confidence Intervals
- Hypothesis Testing
- Chi Squared
- z test
- t test
- F test
- Correlation
- Analysis of Variance



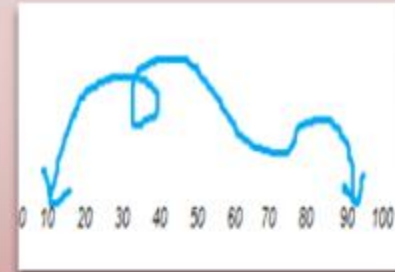
### Central Tendency

Mean, Median, Mode, Outliers



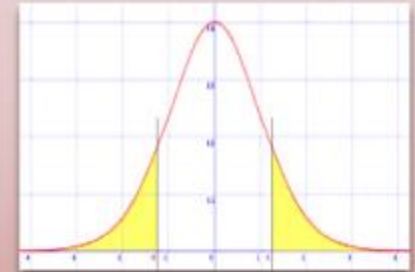
### Measures of Spread

Range, Standard deviation, Variance, Quartiles



### Percentiles

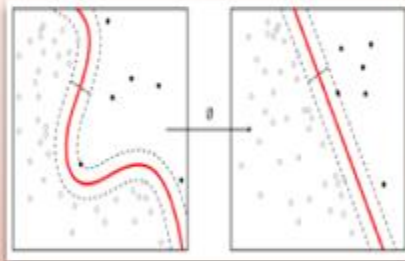
Position of data, percentile rank, percentile range



### Probability Distributions:

Uniform, normal (Gaussian), Poisson

## Basic Probability and Statistics



### Dimensionality reduction

Pruning, PCA



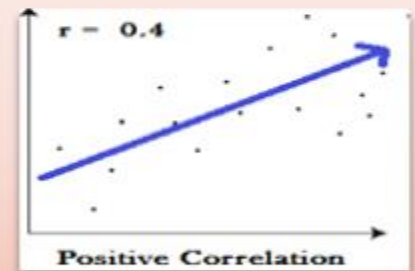
### Sampling

SRS, Reservoir, Undersampling, Oversampling,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

### Bayesian statistics

Measuring belief or confidence



### Covariance & correlation

How data is related

## More Advanced Probability and Statistics









# Our Community Circle Agreement

---

1. If I ask a question of you, you have the right to PASS.
2. Together, we know a lot.
- 3.
- 4.

# Read article:

Introduction to Research Statistical Analysis.pdf - Adobe Acrobat Reader DC

File Edit View Sign Window Help

Home Tools Introduction to Res... x

71 (1 of 5) 116%

**Education**

## Introduction to Research Statistical Analysis: An Overview of the Basics

Christian Vandever<sup>1</sup>

### Abstract

#### Description

This article covers many statistical ideas essential to research statistical analysis. Sample size is explained through the concepts of statistical significance level and power. Variable types and definitions are included to clarify necessities for how the analysis will be interpreted. Categorical and quantitative variable types are defined, as well as response and predictor variables. Statistical tests described include t-tests, ANOVA and chi-square tests. Multiple regression is also explored for both logistic and linear regression. Finally, the most common statistics produced by these methods are explored.

#### Keywords

statistical analysis; sample size; power; t-test; anova; chi-square; regression

#### Introduction

Statistical analysis is necessary for any research project seeking to make quantitative conclusions. The following is a primer for research-based statistical analysis. It is intended to be a high-level overview of appropriate statistical testing, while not diving too deep into any specific methodology. Some of the

Author affiliations are listed at the end of this article.

Correspondence to:  
Christian Vandever  
HCA Healthcare Graduate  
Medical Education  
2000 Health Park Drive  
Brentwood TN, 37027  
([Christian.Vandever@hca-healthcare.com](mailto:Christian.Vandever@hca-healthcare.com))

Search 'Measure'

Edit PDF

Export PDF

**Adobe Export PDF**

Convert PDF Files to Word or Excel Online

Select PDF File

Introductio...nalysis.pdf

Convert to

Microsoft Word (\*.docx)

Document Language:  
English (U.S.) [Change](#)

Create, edit and sign PDF forms & agreements

Start Free Trial



## Types

Converting between common data types in R. Can always go from a higher value in the table to a lower value.

as.logical	TRUE, FALSE, TRUE	Boolean values (TRUE or FALSE).
as.numeric	1, 0, 1	Integers or floating point numbers.
as.character	'1', '0', '1'	Character strings. Generally preferred to factors.
as.factor	'1', '0', '1', Levels: '1', '0'	Character strings with preset levels. Needed for some statistical models.

## Maths Functions

log(x)	Natural log.	sum(x)	Sum.
exp(x)	Exponential.	mean(x)	Mean.
max(x)	Largest element.	median(x)	Median.
min(x)	Smallest element.	quantile(x)	Percentage quantiles.
round(x, n)	Round to n decimal places.	rank(x)	Rank of elements.
signif(x, n)	Round to n significant figures.	var(x)	The variance.
cor(x, y)	Correlation.	sd(x)	The standard deviation.

## Variable Assignment

```
> a <- 'apple'
> a
[1] 'apple'
```







## The Environment

ls()	List all variables in the environment.
rm(x)	Remove x from the environment.
rm(list = ls())	Remove all variables from the environment.

You can use the environment panel in RStudio to browse variables in your environment.




## Matrices

```
m <- matrix(x, nrow = 3, ncol = 3)
# Create a matrix from x.
```

 m[2, ]	Select a row	 t(m)	Transpose
 m[, 1]	Select a column	 m %*% n	Matrix Multiplication
 m[2, 3]	Select an element	 solve(m, n)	Find x in: m * x = n

## Lists

```
l <- list(x = 1:5, y = c('a', 'b'))
# A list is a collection of elements which can be of different types.
```

 l[[2]]	Second element of l.	 l[[1]]	New list with only the first element.	 l\$x	Element named x.	 l['y']	New list with only element named y.
--	----------------------	--	---------------------------------------	--	------------------	--	-------------------------------------




Also see the **dplyr** package.

## Data Frames



```
df <- data.frame(x = 1:3, y = c('a', 'b', 'c'))
# A special case of a list where all elements are the same length.
```

x	y
1	a
2	b
3	c

### Matrix subsetting

df[, 2]	
df[2, ]	
df[2, 2]	

### List subsetting

df\$x		df[[2]]	
Understanding a data frame			
View(df)	See the full data frame.		
head(df)	See the first 6 rows.		

nrow(df)  
Number of rows.

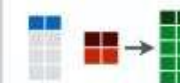
ncol(df)  
Number of columns.

dim(df)  
Number of columns and rows.

cbind - Bind columns.



rbind - Bind rows.



## Strings

Also see the **stringr** package.

paste(x, y, sep = ' ')	Join multiple vectors together.
paste(x, collapse = ' ')	Join elements of a vector together.
grep(pattern, x)	Find regular expression matches in x.
gsub(pattern, replace, x)	Replace matches in x with a string.
toupper(x)	Convert to uppercase.
tolower(x)	Convert to lowercase.
nchar(x)	Number of characters in a string.

## Factors

factor(x)	Turn a vector into a factor. Can set the levels of the factor and the order.
cut(x, breaks = 4)	Turn a numeric vector into a factor by 'cutting' into sections.

## Statistics

lm(y ~ x, data=df)	Linear model.	t.test(x, y)	Perform a t-test for difference between means.	prop.test	Test for a difference between proportions.
glm(y ~ x, data=df)	Generalised linear model.	pairwise.t.test	Perform a t-test for paired data.	aov	Analysis of variance.
summary	Get more detailed information out a model.				

## Distributions

	Random Variates	Density Function	Cumulative Distribution	Quantile
Normal	rnorm	dnorm	pnorm	qnorm
Poisson	rpois	dpois	ppois	qpois
Binomial	rbinom	dbinom	pbinom	qbinom
Uniform	runif	dunif	punif	qunif

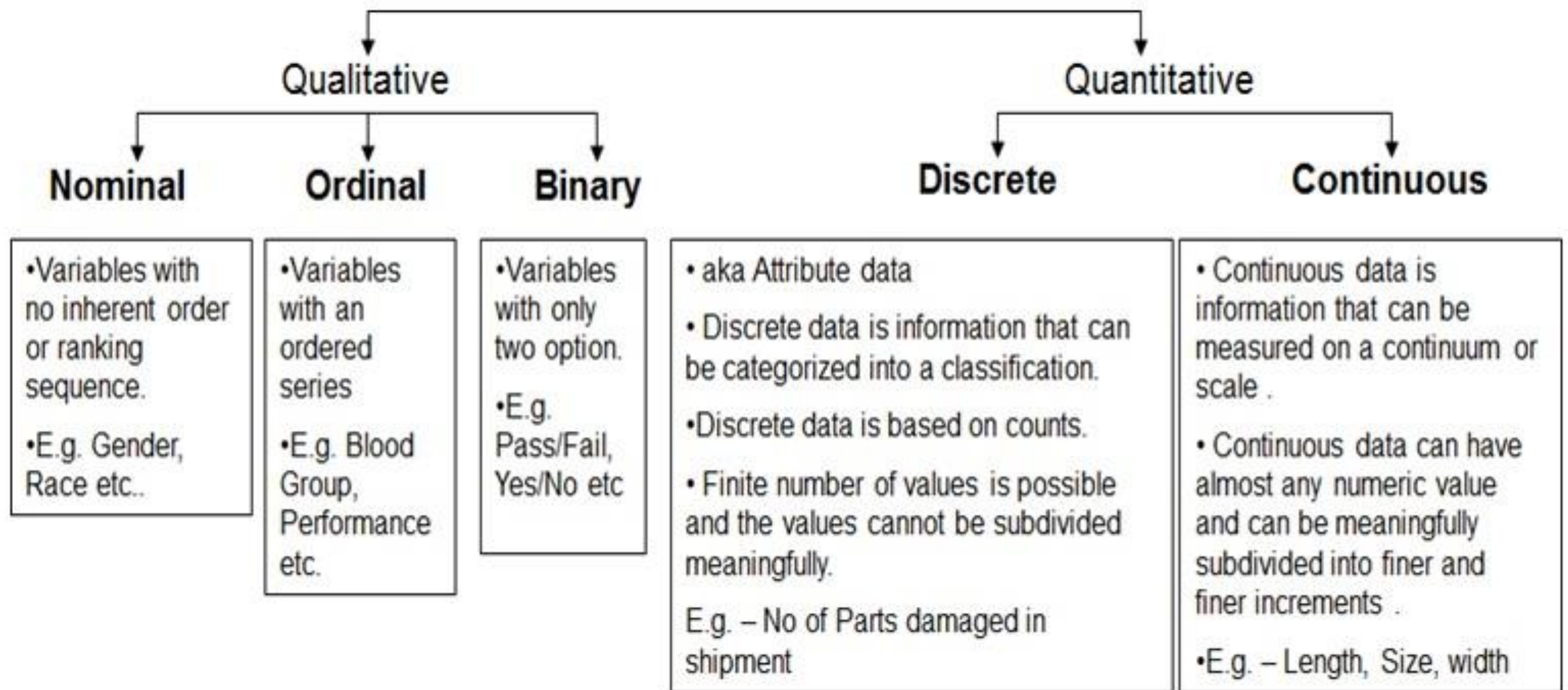
## Plotting

Also see the **ggplot2** package.

 plot(x)	Values of x in order.	 plot(x, y)	Values of x against y.	 hist(x)	Histogram of x.
---	-----------------------	--	------------------------	---	-----------------

## Dates

See the **lubridate** package.



The properties of four scales of measurement in comparison table illustrated below:

Scales	←—— PROPERTIES ——→			
	Distinctiveness	Ordering in magnitude	Equal intervals	Absolute zero point
Nominal	✓	×	×	×
Ordinal	✓	✓	×	×
Interval	✓	✓	✓	×
Ratio	✓	✓	✓	✓

A summary of these four levels is given below :

Scale	Nature	Statistics
1. Nominal	Equivalence	Frequency Distribution ; Mode
2. Ordinal	Equivalence ; Greater than	Median : Percentile ; Rank-correlation.
3. Interval	Equivalence ; Greater than Ratio between two intervals	Mean, S.D., Pearson $r$ : Multiple $r$
4. Ratio	Equivalence ; Greater than ; known ratio.	Geometric Mean ; Coefficient of Variation.



# Data Types in R

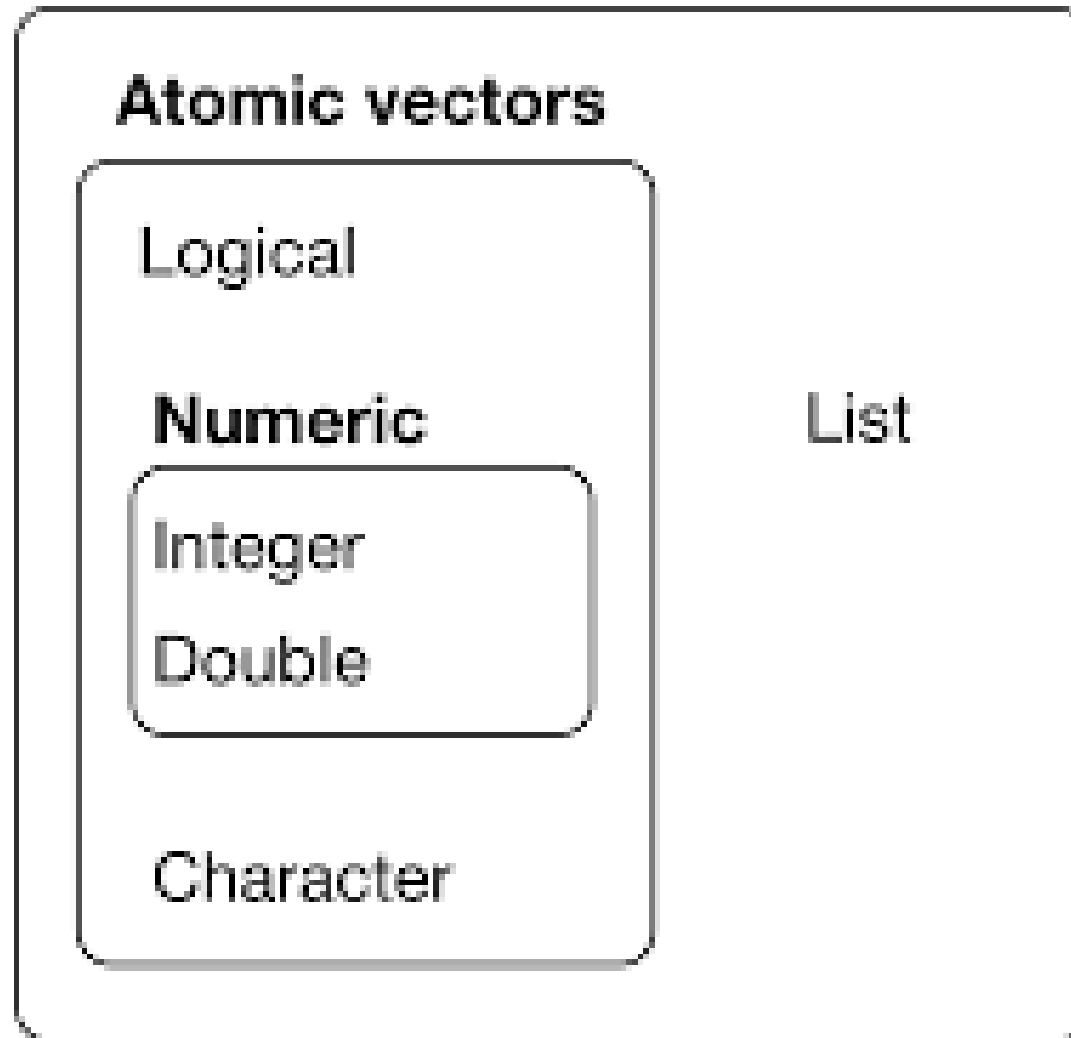
```
> typeof(letters)
```

```
> typeof(1:10)
```

```
> x <- list("a", "b", 1:10)
```

```
> length(x)
```

## Vectors



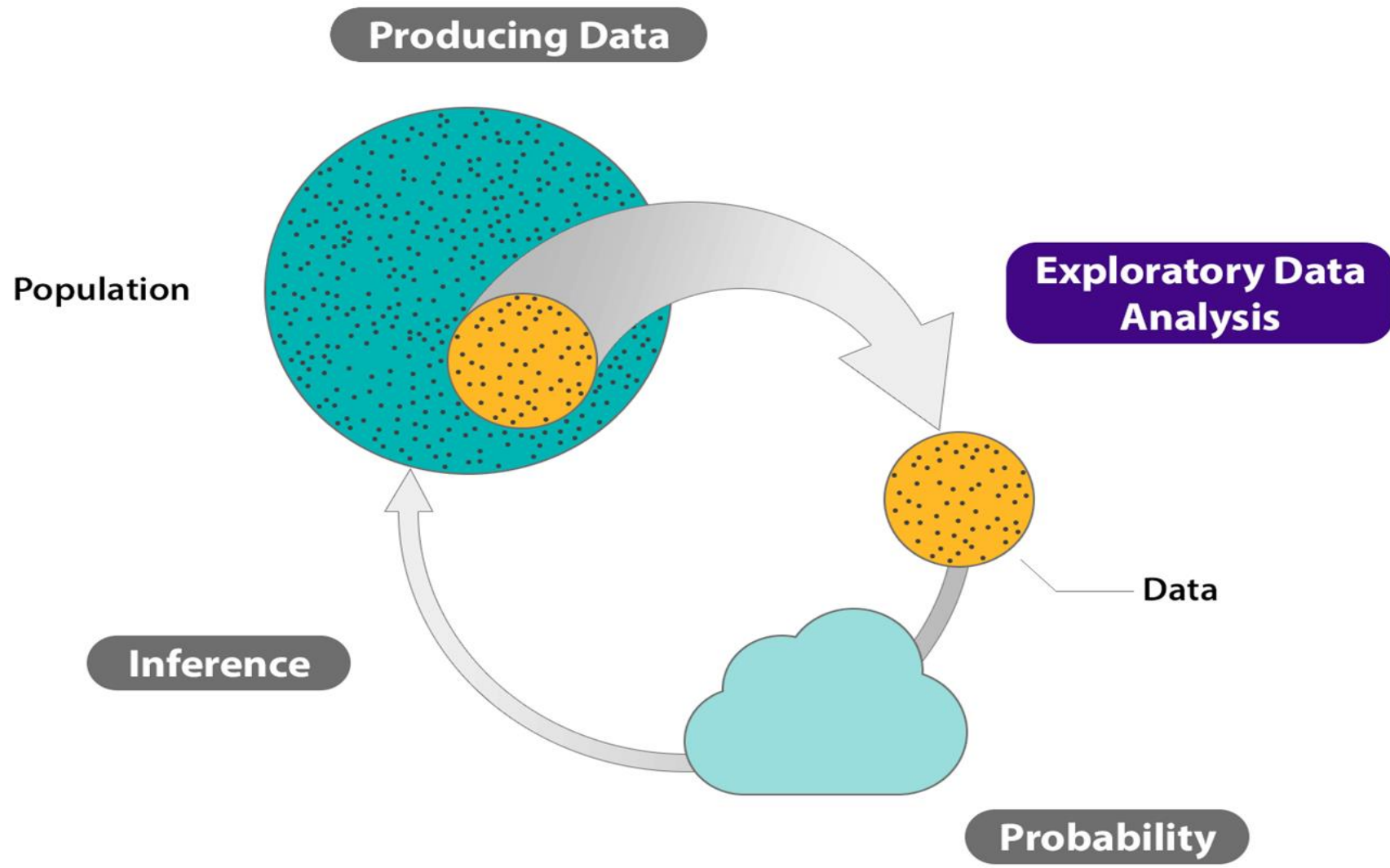
**NULL**

Logical Operators return values inside the vector based on logical conditions.

Operator	Description
<	Less than
<=	Less than or equal to
>	Greater than
>=	Greater than or equal to
==	Exactly equal to
!=	Not equal to
!x	Not x
x	y
x & y	x AND y
isTRUE(x)	Test if X is TRUE

You can add many conditional statements, but we need to include them in a parenthesis. Follow this structure to create a conditional statement:

```
variable_name[(conditional_statement)]
```





# Two Branches of Statistics

## Descriptive Statistics

- as a science, involves the collection, organization, summarization, and presentation of data
- involves raw data, as well as graphs, tables, and numerical summaries
- “Just the facts”
- Refer to sample without making any assumptions about the population

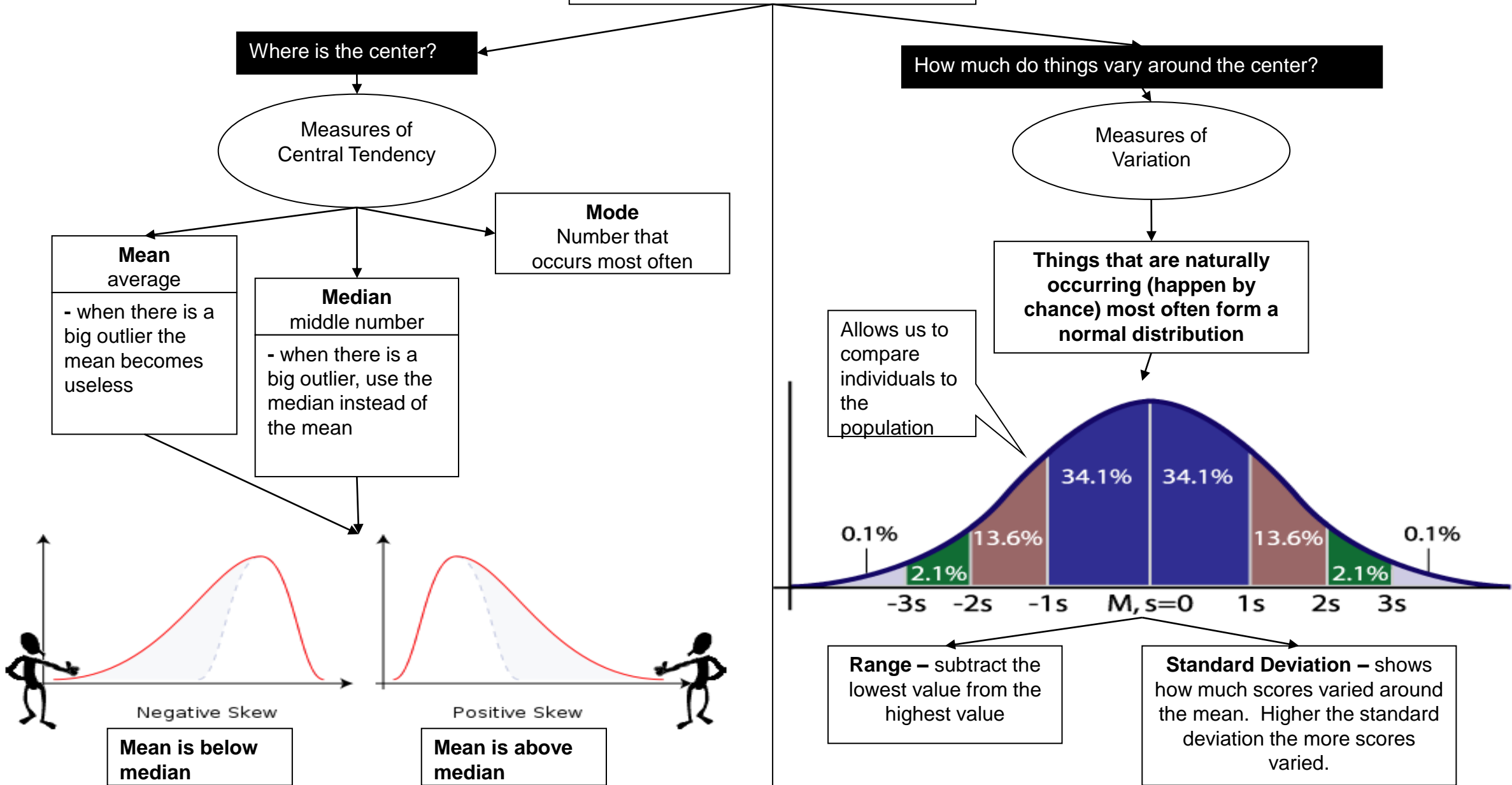
## Inferential Statistics

- as a science, involves using descriptive statistics to estimate population parameters
- deals with interpretation of the information collected
- usually used in conjunction with descriptive statistics within a statistical study

# Common Forms of Descriptive Statistics

1. Summary statistics
  - a. Measures of central tendency
  - b. Measures of dispersion
2. Graphs
  - a. Visualize data with Boxplots, histograms, stem and leaf plots and scatter plots
3. Tables
  - a. Frequency Tables

# Descriptive Statistics





# Measure of Central Tendency

Mean	Average value
Median	Middle value
Mode	Most frequent value

When to use mean, median and mode?

- Mean – When your data is not skewed i.e., normally distributed. In other words, there are no extreme values present in the data set.
- Median – When your data is skewed or you are dealing with ordinal (ordered categories) data (e.g., Likert scale 1. Strongly dislike 2. Dislike 3. Neutral 4. Like 5. Strongly like)
- Mode - When dealing with nominal (unordered categories) data.

# Measures of Dispersion:

Measures of variability gives how “spread out” the data

Range	Difference between max and min in a distribution
Interquartile range	Correspondes to the difference between the first and third quartiles
Standard Deviation	Average distance of scores in a distribution from their mean
Variance	Square of the standard deviation
Skewness	Degree to which scores in a distribution are spread out
Kurtosis	Flatness of peakness of the curve

# R Functions for Computing Descriptive Statistics

Description	R function
Mean	<code>mean()</code>
Standard deviation	<code>sd()</code>
Variance	<code>var()</code>
Minimum	<code>min()</code>
Maximum	<code>maximum()</code>
Median	<code>median()</code>
Range of values (minimum and maximum)	<code>range()</code>
Sample quantiles	<code>quantile()</code>
Generic function	<code>summary()</code>
Interquartile range	<code>IQR()</code>

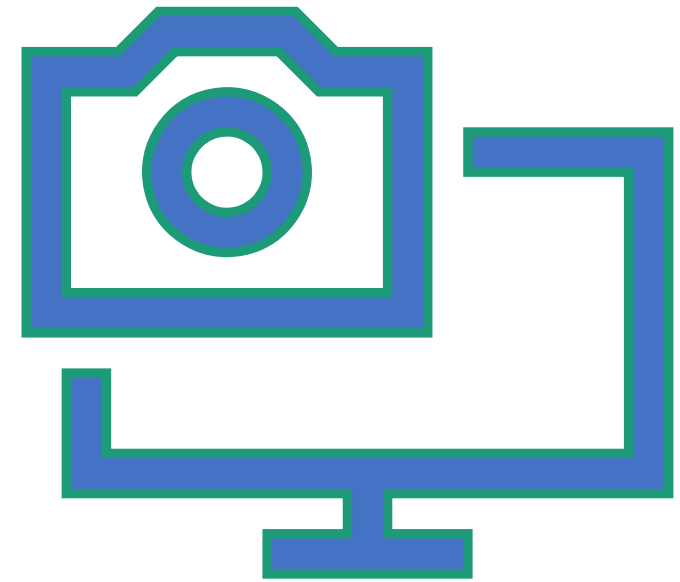


## Descriptive Statistics

Variable	<u>Obs</u>	Mean	<u>Std.Dev.</u>	Min	Max
price	74	6165.257	2949.496	3291	15906
mpg	74	21.297	5.786	12	41
rep78	69	3.406	.99	1	5
headroom	74	2.993	.846	.846	5
trunk	74	13.757	4.277	5	23
weight	74	3019.459	777.194	1760	4840
length	74	187.932	22.266	142	233
turn	74	39.649	4.399	31	51
displacement	74	197.297	91.837	79	425
<u>gear_ratio</u>	74	3.015	.456	2.19	3.89
foreign	74	.297	.46	0	1

# Descriptive Statistics in R for Data Frames

- `max()`: Returns the largest value in the entire data frame.
- `min()`: Returns the smallest value in the entire data frame.
- `sum()`: Returns the sum of the entire data frame.
- `fivenum()`: Returns the Tukey summary values for the entire data frame.
- `length()`: Returns the number of columns in the data frame.
- `summary()`: Returns the summary for each column.





# R Special Summary Commands

There are two types of special summary commands:

- Row Summary Commands – Applied to work with row data. Two commands here are *rowMeans()* and *rowSums()*.
- Column Summary Commands – Also, applied to work with row data but the two commands here are *colMeans()* and *colSums()*.

Q1	Q2	Q3	Q4	Q5
0	0	0	1	1
0	1	1	1	0
1	0	0	1	1
0	0	1	1	0
1	1	1	1	1

```
patientsurvey <- data.frame("q1" = c(0, 0, 1, 0, 1),
                             "q2" = c(0, 1, 0, 0, 1),
                             "q3" = c(0, 1, 0, 1, 1),
                             "q4" = c(1, 1, 1, 1, 1),
                             "q5" = c(1, 0, 1, 0, 1))
```

```
rowMeans(patientsurvey)
rowSums(patientsurvey)
colMeans(patientsurvey)
colSums(patientsurvey)
```



# Frequency tables: Used to describe categorical variables

- You can generate contingency (frequency) tables using:
  - the `table( )` function
  - tables of proportions using the `prop.table( )` function, and
  - marginal frequencies using `margin.table( )`
- Compute table margins and relative frequency
  - `table(x)`
  - `margin.table(x, margin = NULL)`
  - `prop.table(x, margin = NULL)`

# Frequency tables: Used to describe categorical variables

Weight (Kg)	Frequency	Cumulative Frequency
0 up to 20	2	2
20 up to 40	7	9
40 up to 60	12	21
60 up to 80	6	27
80 up to 100	3	30

- For demonstration, see R program FrequencyTables.R

# Hmsic package: Describe function

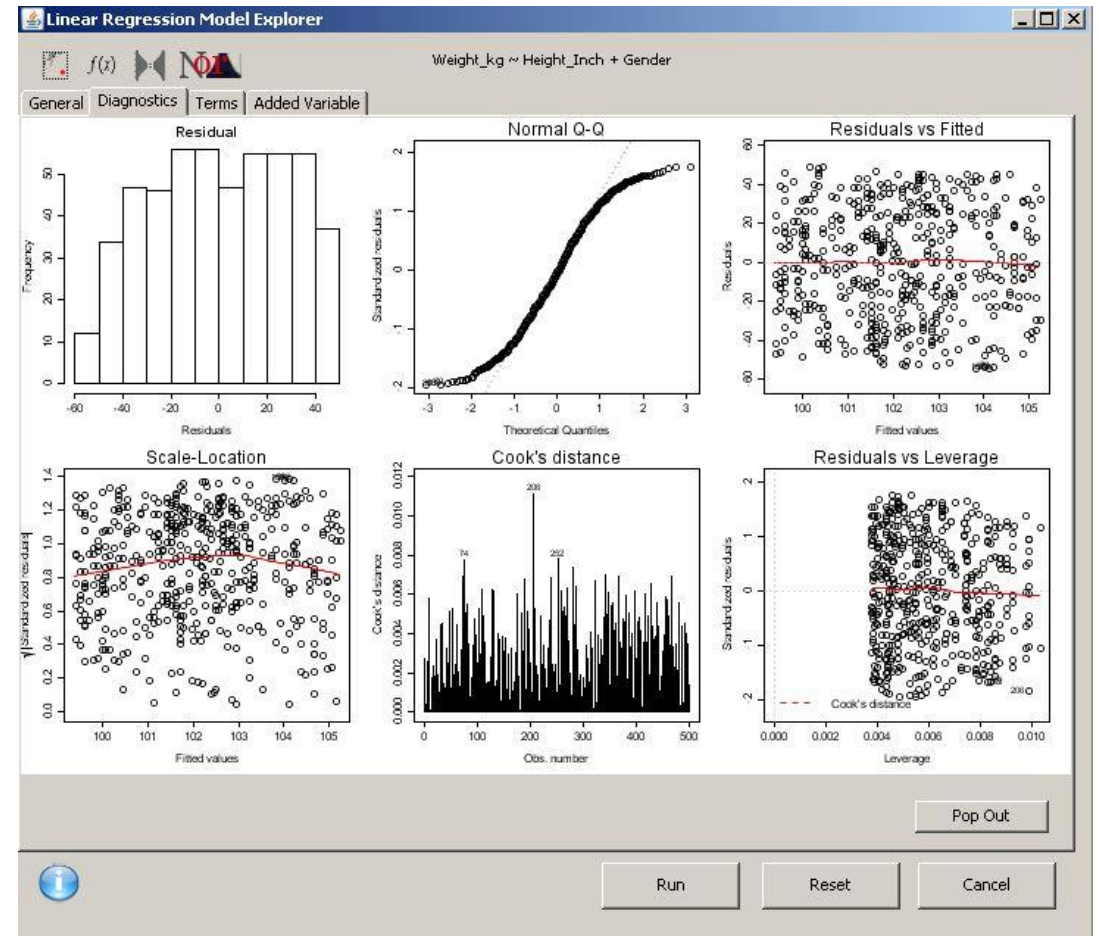
- `Install.packages("Hmisc")`
- `Library("Hmisc")`
- Similar to the `summary` function is the `describe` function.

Data Frame Summary  
data  
N: 64

No	variable	Stats / values	Freqs (% of valid)	Text Graph	valid	Missing
1	type [factor]	1. A 2. B 3. C 4. D 5. E	7 (11.7%) 15 (25.0%) 11 (18.3%) 8 (13.3%) 19 (31.7%)	IIIIII IIIIIIIIIIIIII IIIIIIIIII IIIIII IIIIIIIIIIIIIIIIII	60 (93.75%)	4 (6.25%)
2	score [integer]	mean (sd) : 48.29 (31.71) min < med < max : 0 < 45 < 124 IQR (cv) : 54 (0.66)	47 distinct val.	: : : . : .	58 (90.62%)	6 (9.38%)
3	category [factor]	1. X 2. Y 3. Z	17 (26.6%) 10 (15.6%) 37 (57.8%)	IIIIIIII IIII IIIIIIIIIIIIIIIIII	64 (100%)	0 (0%)
4	rating [integer]	mean (sd) : 1389025.28 (12598684.85) min < med < max : -11102847 < 5 < 1e+08 IQR (cv) : 7.25 (9.07)	22 distinct val.	: : : : : :	64 (100%)	0 (0%)

# Deducer package: frequencies

- Adds menu driven analysis and plotting
  - `install.packages("Deducer")`
  - `library("Deducer")`
- frequency functions



# Methods of Standardization and Normalization

S.NO.	Normalisation	Standardisation
1	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3	Scales values between $[0, 1]$ or $[-1, 1]$ .	It is not bounded to a certain range.
4	It is really affected by outliers.	It is much less affected by outliers.
5	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
6	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
7	It is an often called as Scaling Normalization	It is an often called as Z-Score Normalization.



# Methods of Standardization and Normalization

1. Z-score:  $z = \frac{x - \text{mean}}{\text{std.dev}}$

2. Min-Max Scaling:  $x - \min(x) / \max(x) - \min(x)$

3. Standard Deviation Method:  $x / \text{stdev}(x)$

4. Range:  $x / (\max(x) - \min(x))$

5. Centering: Subtracting a constant value from every value of a variable. The constant value can be average, min or max.

# # Creating a sample data

```
set.seed(272)
```

```
#create a dataframe with k1=1,000 values between 100 and 1,000 and k2= 1,000  
values between 10 and 100
```

```
X = data.frame(k1 = sample(100:1000, 1000, replace=TRUE),  
               k2 = sample(10:100, 1000, replace=TRUE))  
X.scaled = scale(X, center= TRUE, scale=TRUE)
```

```
#Check Mean and Variance of Standardized Variable
```

```
colMeans(X.scaled)
```

```
var(X.scaled)
```

## Permutations & Combinations

A **combination** is an arrangement of items in which **ORDER DOES NOT MATTER.**

A **permutation** is an arrangement of items in a particular order.  
Notice, **ORDER MATTERS!**

- A berry salad is a **combination** of blueberries, strawberries and raspberries, since it's the same fruit salad regardless of the order of fruits.
- To open the iPad, you need the right order of numbers, thus the code is a **permutation**.

Order Matters	Repetition Allowed	Formula
Yes (Permutation)	Yes	$P(n, r) = n^r$
Yes (Permutation)	No	$P(n, r) = \frac{n!}{(n - r)!}$
No (Combination)	No	$C(n, r) = \frac{n!}{r!(n - r)!}$
No (Combination)	Yes	$C(n + r - 1, r) = \frac{(n + r - 1)!}{r!(n - 1)!}$

- Combination Lock: permutations with repetitions

```
> pwr<- expand.grid(rep(list(0:9), 3))
```

```
> head(pwr)
```

```
#take a look at pwr
```



- Lottery choosing 6 out of 49 balls: combinations without repetitions. Use the `combn()` function for finding all possibilities.

```
> cwop <- combn(1:49, 6)
```

```
> cwop[, 1:10]
```



```
> #install and load library ("gtools")
```

```
> #urn with 3 balls
```

```
> x <- c('yellow', 'green', 'black')
```

```
[,1] [,2]
```

```
[1,] "black" "black"
```

```
[2,] "black" "green"
```

```
[3,] "black" "yellow"
```

```
[4,] "green" "black"
```

```
[5,] "green" "green"
```

```
[6,] "green" "yellow"
```

```
[7,] "yellow" "black"
```

```
[8,] "yellow" "green"
```

```
[9,] "yellow" "yellow"
```

```
> #pick 2 balls from the urn with replacement
```

```
> permutations(n=3,r=2,v=x, repeats.allowed=T) #get all permutations
```

```
> nrow(permutations(n=3,r=2,v=x, repeats.allowed=T)) #number of permutations
```

```
[1] 9
```

## Permutations with repetition



PRACTICE  
MAKES  
PERFECT

- We have 4 choices (A, C, G, and T) and we are choosing 3 nucleotides:  $n^r = 4^3 = 64$  (though some permutations code for the same amino acid).
- How about the number of 11-mers that are possible:  
$$n^r = 4^{\{11\}} = 4194304 .$$
- For miRNAs of size 21, there are  $4^{\{21\}} = 4398046511104$  possible miRNAs, but since the majority of these 21-mers are probably not biologically relevant or even possible, they most likely don't exist.

```
perm_without_replacement <- function(n, r){  
  return(factorial(n)/factorial(n - r))  
}
```

```
#sixteen choices, choose 16  
> perm_without_replacement(16,16)  
#[1] 2.092279e+13
```

```
#sixteen choices, choose 4  
> perm_without_replacement(16,4)  
#[1] 43680
```

## Permutations without repetition

#calculate the number of combinations without replacement/repetition

```
> choose(n=24,k=4)
```

```
[1] 10626
```

$$\frac{n!}{r!(n-r)!} = \frac{24!}{4!(24-4)!} = \frac{24!}{4!20!} = \frac{24 \times 23 \times 22 \times 21}{24} = 10626.$$

#calculate the number of combinations with replacement/repetition

```
> comb_with_replacement <- function(n, r){  
  return( factorial(n + r - 1) / (factorial(r) * factorial(n - 1)) )  
}
```

#have 3 elements, choosing 3

```
> comb_with_replacement(3,3)
```

```
[1] 10
```

$$\frac{(3+3-1)!}{3!(3-1)!} = \frac{5!}{3!2!} = \frac{120}{12} = 10.$$

The area of *descriptive statistics* is concerned with meaningful and efficient ways of presenting data.

When it comes to *inferential statistics*, though, our goal is to make some statement about a characteristic of a population based on what we know about a sample drawn from that population.

S. No	Descriptive Statistics	Inferential Statistics
1	Concerned with the describing the target population	Make inferences from the sample and generalize them to the population.
2	Organize, analyze and present the data in a meaningful manner	Compares, test and predicts future outcomes.
3	Final results are shown in form of charts, tables and Graphs	Final result is the probability scores.
4	Describes the data which is already known	Tries to make conclusions about the population that is beyond the data available.
5	Tools- Measures of central tendency (mean/median/ mode), Spread of data (range, standard deviation etc.)	Tools- hypothesis tests, Analysis of variance etc.



# Theory of Probability

## Deterministic vs. Random Processes

- In **deterministic** processes, the outcome can be predicted exactly in advance
  - Eg. Force = mass x acceleration. If we are given values for mass and acceleration, we exactly know the value of force
- In **random** processes, the outcome is not known exactly, but we can still describe the *probability distribution* of possible outcomes
  - Eg. 10 coin tosses: we don't know exactly how many heads we will get, but we can calculate the probability of getting a certain number of heads

# Probability



**Experiments** are the uncertain situations, which could have multiple outcomes. Whether it rains on a daily basis is an experiment.



**Outcome** is the result of a single trial. So, if it rains today, the outcome of today's trial from the experiment is "It rained"



**Event** is one or more outcome from an experiment. "It rained" is one of the possible event for this experiment.



**Probability** is a measure of how likely an event is. So, if it is 60% chance that it will rain tomorrow, the probability of Outcome "it rained" for tomorrow is 0.6


# Basic common terminologies

An **event** is simply the outcome of a random experiment.

- Getting a heads when we toss a coin is an event.
- Getting a 6 when we roll a fair die is an event.

We associate probabilities to these events by defining the event and the sample space.

A **sample space** is nothing but the collection of all possible outcomes of an experiment. This means that if we perform a particular task again and again, all the possible results of the task are listed in the sample space.

A close-up, shallow depth-of-field photograph of a computer keyboard. The central focus is on a single black key with white markings for multiplication (x), division (/), and a remainder symbol (a vertical line followed by a horizontal line). To the left, a portion of a key with blue markings is visible. To the right, a key with a white arrow pointing up is visible. Below the central key, a key with a white plus sign is visible. The background is blurred, showing other keys in the keyboard.





# Five Theorems of Probability

- $P(A) = 1 - P(A')$ .
- $P(\emptyset) = 0$ .
- If events  $A$  and  $B$  are such that  $A \subseteq B$ , then  $P(A) \leq P(B)$ .
- $P(A) \leq 1$ .
- For any two events  $A$  and  $B$ ,  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ .

# Definitions and Notation

- Two events are **mutually exclusive** or **disjoint** if they cannot occur at the same time.
- The **probability** probability that Event A occurs, given that Event B has occurred, is called a **conditional probability**. The conditional probability of Event A, given Event B, is denoted by the symbol  $P(A|B)$ .
- The **complement** of an event is the event not occurring. The probability that Event A will not occur is denoted by  $P(A')$ .
- The probability that Events A **and** B *both* occur is the probability of the **intersection** of A and B. The probability of the intersection of Events A and B is denoted by  $P(A \cap B)$ . If Events A and B are mutually exclusive,  $P(A \cap B) = 0$ .
- The probability that Events A **or** B occur is the probability of the **union** of A and B. The probability of the union of Events A and B is denoted by  $P(A \cup B)$ .
- If the occurrence of Event A changes the probability of Event B, then Events A and B are **dependent**. On the other hand, if the occurrence of Event A does not change the probability of Event B, then Events A and B are **independent**.

## Four Types of Probability

Marginal	Union	Joint	Conditional
$P(X)$ The probability of an event <b>X</b> occurring 	$P(X \cup Y)$ The probability of <b>X or Y or both</b> occurring 	$P(X \cap Y)$ The probability of <b>both X and Y</b> occurring 	$P(X Y)$ The probability of <b>X occurring given that Y</b> has occurred 



# Conditional Probability

$$P(A | B) = \frac{P(A \cap B)}{P(B)}$$

Probability of event A  
given B has occurred

Probability of event A occurred  
and event B occurred

Probability of event B

- Conditional probabilities arise naturally in the investigation of experiments where an outcome of a trial may affect the outcomes of the subsequent trials.
- We try to calculate the probability of the second event (event B) given that the first event (event A) has already happened. If the probability of the event changes when we take the first event into consideration, we can safely say that the probability of event B is dependent of the occurrence of event A.
- Probability relates to all past experiments, or occurrences, of that sequence. However, it is often useful to narrow down the situation to include only those with specific criteria.

# Conditional Probability Equation

$$P\left(\frac{A}{B}\right) = \frac{P(A \text{ and } B)}{P(B)} = \frac{\text{Probability of the occurrence of both A and B}}{\text{Probability of B}}$$

- **Joint probability** is the probability of two events occurring simultaneously.
- **Marginal probability** is the probability of an event irrespective of the outcome of another variable.
- **Conditional probability** is the probability of one event occurring in the presence of a second event.

A predictive model can easily be understood as a statement of conditional probability.

For example, the probability of a customer from segment A buying a product of category Z in next 10 days is 0.80.

In other words, the probability of a customer buying product from Category Z, given that the customer is from Segment A is 0.80.

A research group collected the yearly data of road accidents with respect to the conditions of following and not following the traffic rules of an accident prone area. They are interested in calculating the probability of accident given that a person followed the traffic rules. The table of the data is given as follows:

Condition	Follow Traffic Rule	Does not follow Traffic Rule
Accident	50	500
No Accident	2000	5000

- $P(\text{Accident} \mid \text{A person follow Traffic Rule}) = P(\text{Accident and follow Traffic Rule}) / P(\text{Follow Traffic Rule})$
- $P_{\text{Accident\_who\_follow\_Traffic\_Rule}} < 50$
- $P_{\text{who\_follow\_Traffic\_Rule}} = 50 + 2000$
- $\text{Conditional\_Probability} = (P_{\text{Accident\_who\_follow\_Traffic\_Rule}} / P_{\text{who\_follow\_Traffic\_Rule}})$
- Conditional\_Probability

Suppose we have a test for the flu that is positive 90% of the time when tested on a flu patient ( $P(\text{test} + \mid \text{flu}) = 0.9$ ) and is negative 95% of the time when tested on a healthy person ( $P(\text{test} - \mid \text{no flu}) = 0.95$ ). We also know that the flu is affecting about 1% of the population ( $P(\text{flu})=0.01$ ). You go to the doctor and test positive. What is the chance that you truly have the flu?

```
flu <- sample(c('No','Yes'), size=100000, replace=TRUE, prob=c(0.99,0.01))
```

```
test <- rep(NA, 100000) #create a dummy variable first
```

```
test[flu=='No'] <- sample(c('Neg','Pos'), size=sum(flu=='No'), replace=TRUE,  
prob=c(0.95,0.05))
```

```
test[flu=='Yes'] <- sample(c('Neg','Pos'), size=sum(flu=='Yes'), replace=TRUE,  
prob=c(0.1, 0.9))
```

# Bayesian

THE PROBABILITY OF "B"  
BEING TRUE GIVEN THAT  
"A" IS TRUE

↓

THE PROBABILITY  
OF "A" BEING  
TRUE

↙

$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$

↑

THE PROBABILITY  
OF "A" BEING TRUE  
GIVEN THAT "B" IS  
TRUE

↖

THE PROBABILITY  
OF "B" BEING  
TRUE

# Bayes Theorem

- **Bayes' Theorem** is simply *an alternate* way of calculating conditional probability. It describes the probability of occurrence of an event related to any condition.
- Here's the conditional probability for outcome 4 using a joint probability:  
P(G) = 'Probability that first child is a girl' (1/2)  
P(B) = 'Probability that **both (B)** children are girls' (1/4)  
P(B | G) = P(B,G) / P(G)  
P(B | G) = (1/4) / (1/2) = **1/2** or roughly **50%**
- Technically, we *can't* use joint probability because the two events are *not independent*.

	Boy	Girl
1	1	0
2	1	1
3	0	1
4	0	0



- To clarify, the probability of the older child being a certain gender and the probability of the younger child being a certain gender is independent, but  $P(B|G)$  the 'probability of **both** child being a girl' and 'the probability of the older child being a girl' are not independent; and hence we express it as a conditional probability.
- So, the joint probability of  $P(B,G)$  is just event  $B, P(B)$ .

A man is known to speak truth 2 out of 3 times. He throws a die and reports that the number obtained is a four. Find the probability that the number obtained is a four.

Let A be the event that the man reports that number four is obtained.

Let E1 be the event that four is obtained and E2 be its complementary event.

Then,  $P(E1) = \text{Probability that four occurs} = 1/6$

$P(E2) = \text{Probability that four does not occurs} = 1 - P(E1) = 1 - 1/6 = 5/6$

Also,  $P(A|E1) = \text{Probability that man reports four and it is actually a four} = 2/3$

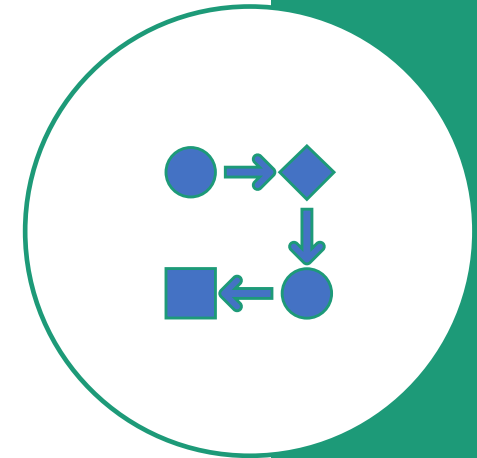
$P(A|E2) = \text{Probability that man reports four and it is not a four} = 1/3$

By using Bayes' theorem, probability that number obtained is a four,

$$P(E1|A) = P(E1)P(A|E1)/[P(E1)P(A|E1) + P(E2)P(A|E2)] = (1/6 \times 2/3) / (1/6 \times 2/3 + 5/6 \times 1/3) = 2/7$$

# Bayes' Theorem in Machine Learning

- Bayes' theorem tells use how to gradually update our knowledge on something as we get more evidence about that something.
- Bayes' theorem is used in machine learning; both in regression and classification, to incorporate previous knowledge into our models and improve them.



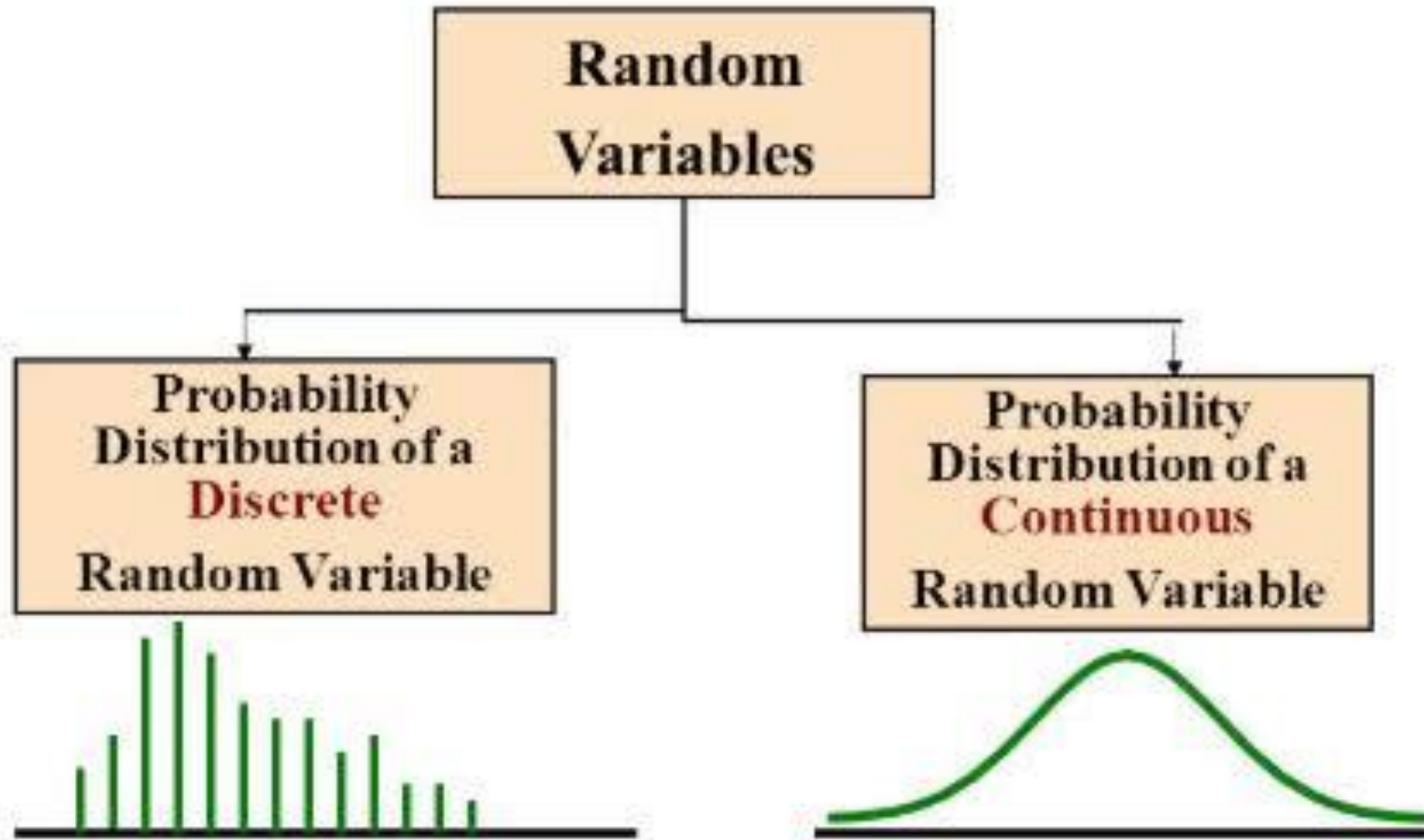


# Random Experiment

---

“.....a random experiment is an action or process that leads to one of several possible outcomes/events. For example

<u>Experiment</u>	<u>Outcome</u>
Patient Wait Time	$t > 0$ seconds
Blood Types	O, A-, B+, AB
Exam Marks	Numbers 0, 1, 2,...100



# Random Variables

1. Generate random numbers from uniform distribution

**runif**(n, min=a, max=b)

2. Generate random numbers from normal distribution

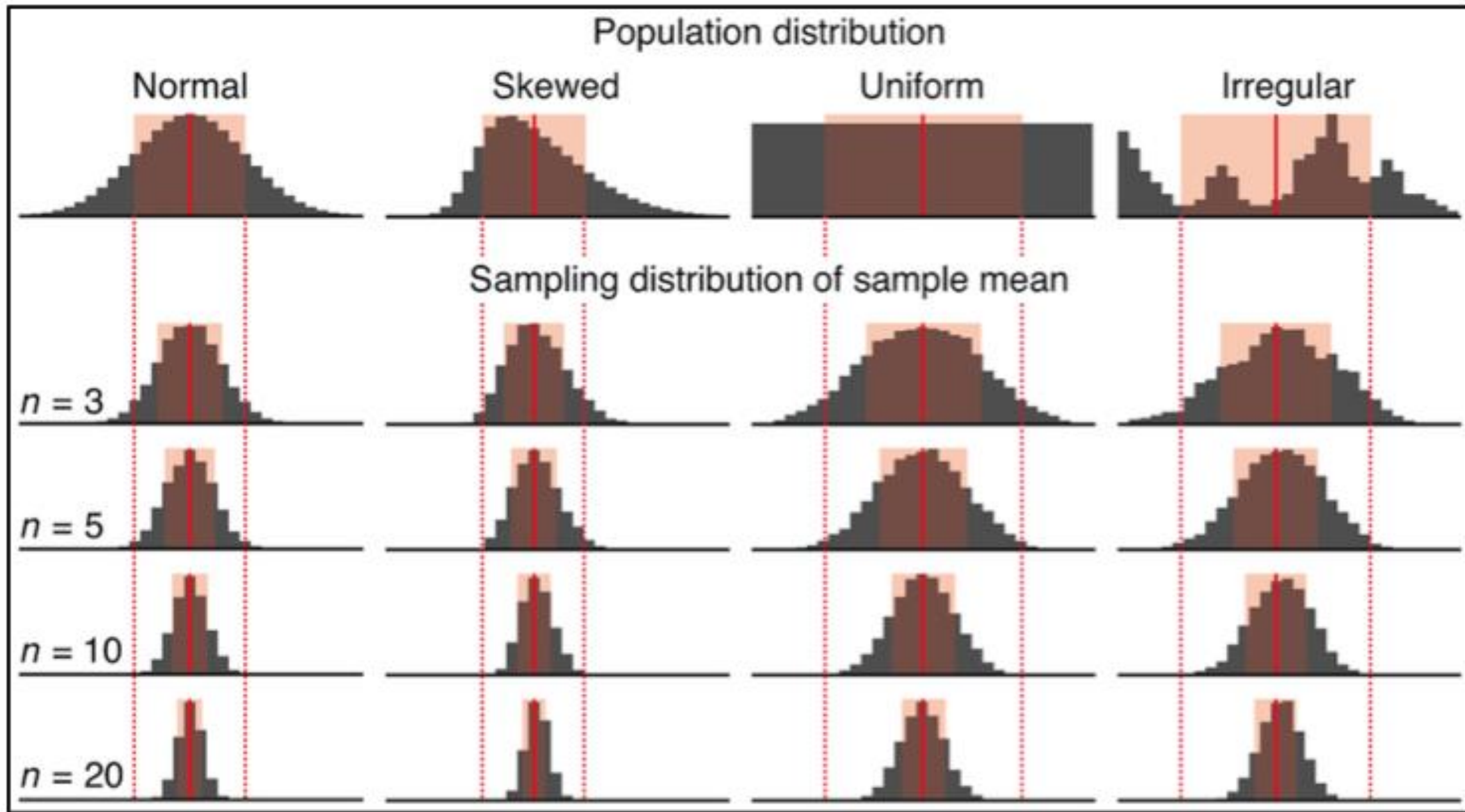
**rnorm**(n, mean=a, sd=b)

3. Generate random numbers from binomial distribution

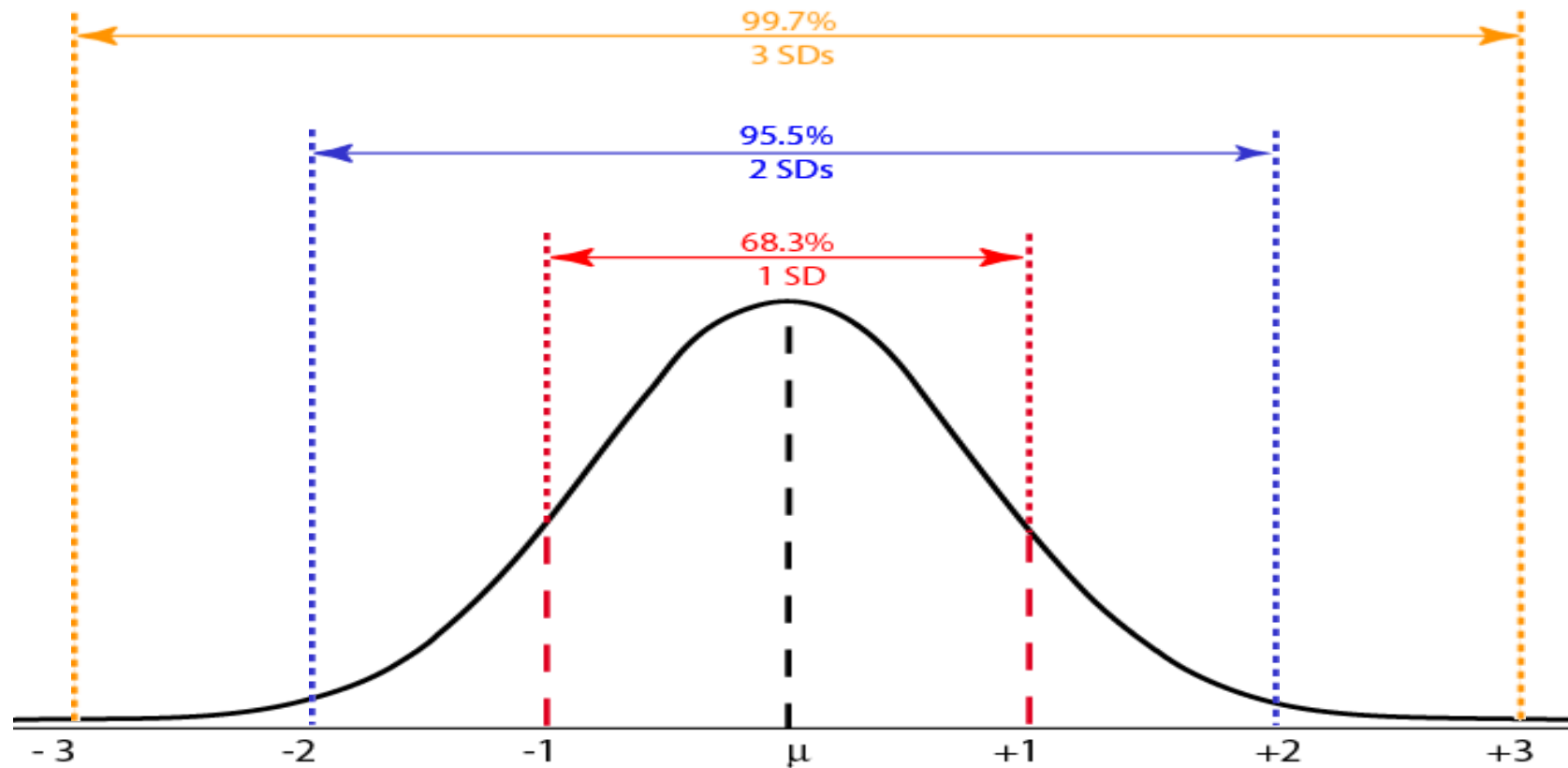
**rbinom** (# observations, # trials/observation, probability of success )

4. Generate random numbers from bernoulli distribution

**rbinom**(10, 1,.5)







## Empirical Rule

- 68% of data lie within  $1\sigma$  of the mean  $\mu$
- 95% of data lie within  $2\sigma$  of the mean  $\mu$
- 99.7% of data lie within  $3\sigma$  of the mean  $\mu$

# Example: Dimension of Matrix or Data Frame

```
> set.seed(8212)
```

```
> N <- 500
```

```
# Set Seed for reproducibility
```

```
# Sample size
```

```
> x1 <- round(rnorm(N, 1, 20))
```

```
> x2 <- round(runif(N, 5, 10))
```

```
> x3 <- round(runif(N, 1, 4), 1)
```

```
> x4 <- round(runif(N, 5, 50))
```

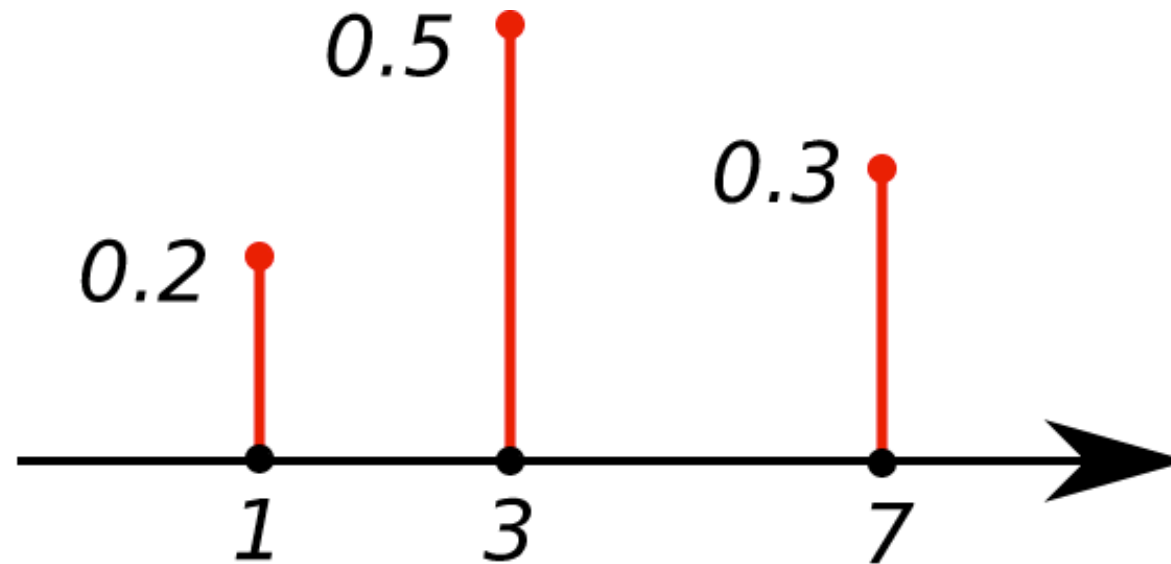
```
> x5 <- rpois(N, 5)
```

```
# Create 5 random variables
```

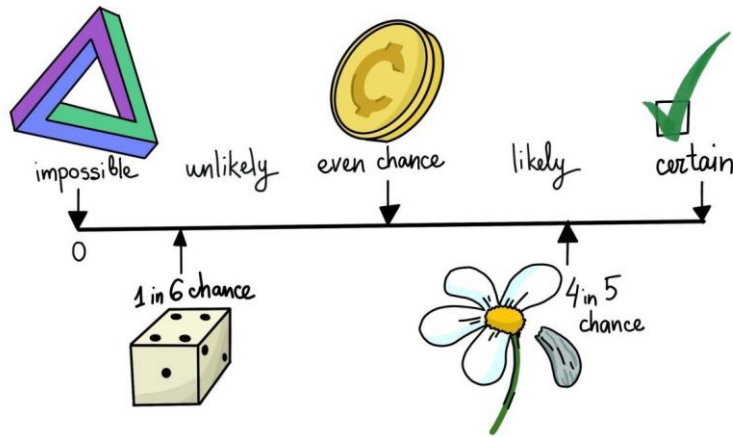
	x1	x2	x3	x4	x5
1	29	6	2.5	6	8
2	26	8	1.1	47	6
3	-7	6	3.9	38	5
4	-11	6	1.1	33	7
5	17	7	1.7	27	4
6	-11	7	3.0	41	6

# Distributions of Random Variables

A **probability mass function** (pmf) assigns a probability to each possible value of a **discrete** random variable



# What is a Probability Distribution?



## Discrete Probability Distributions

- The probability distribution of a [discrete](#) random variable can always be represented by a table.

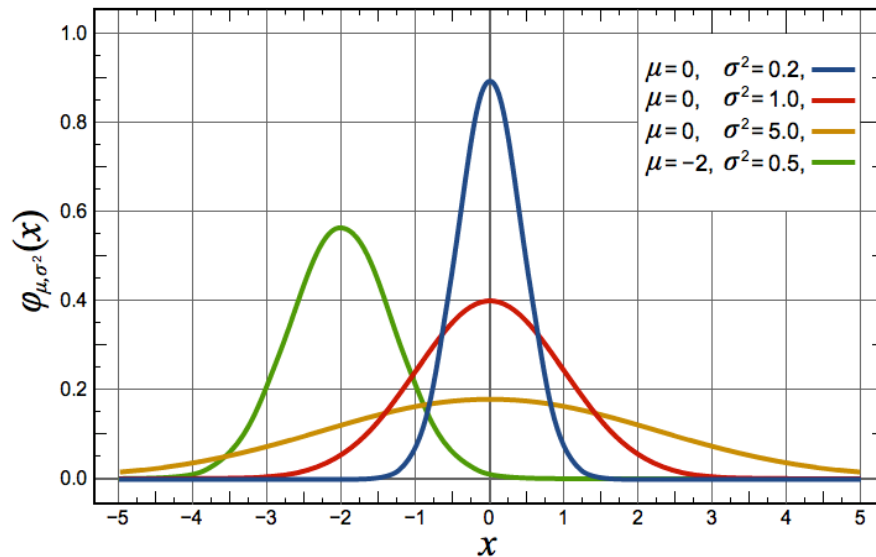
## Continuous Probability Distributions

- The probability distribution of a [continuous](#) random variable is represented by an equation, called the **probability density function** (pdf). All probability density functions satisfy the following conditions:
- The random variable  $Y$  is a function of  $X$ ; that is,  $y = f(x)$ .
- The value of  $y$  is greater than or equal to zero for all values of  $x$ .
- The total area under the curve of the function is equal to one.

	Discrete	Continuous
Type	Countable	Not countable
Definition	Probability mass function, Cumulative distribution function	Probability density function, Cumulative distribution function
Examples	Binomial distribution, Poisson distribution	Gaussian distribution, Exponential distribution

# Distribution of Random Variables

- A **continuous** random variable  $X$  is described through the **probability density function** (pdf)
- PDF describes the *relative* likelihood for  $X$  to take a given value

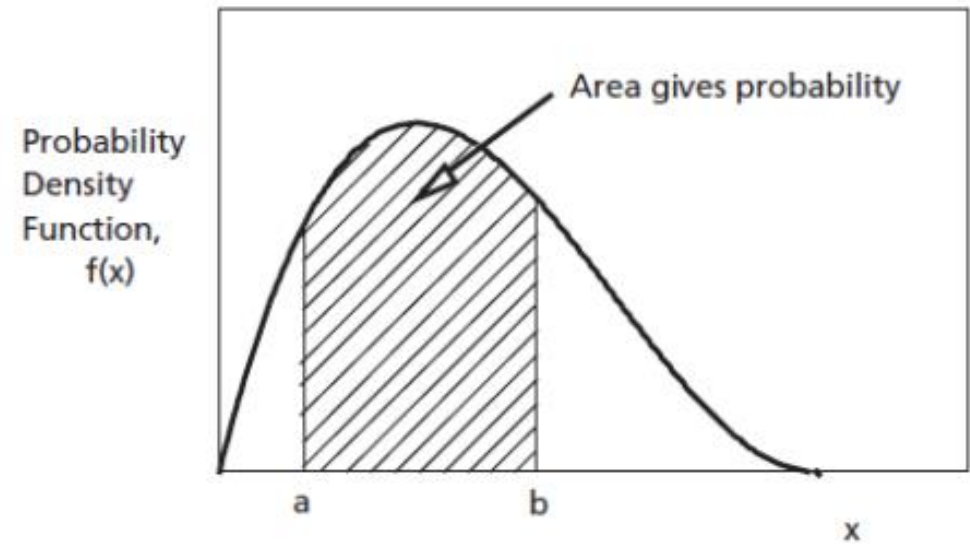


$$P(a \leq X \leq b) = \int_a^b f(x) dx$$

**What is the probability that  $X = b$ ?**

# Normal Distribution

- ***dnorm()***: returns the height of the probability distribution at each point. If you only give the points it assumes you want to use a mean of zero and standard deviation of one.
- ***pnorm()***: Given a number or a list it computes the probability that a normally distributed random number will be less than that number. Goes by Cumulative Distribution Function
- ***qnorm()***: inverse of *pnorm*. The idea behind *qnorm* is that you give it a probability, and it returns the number whose cumulative distribution matches the probability (quantiles)
- ***rnorm()***: generates random numbers whose distribution is normal``







# The t Distribution

- Values are normalized to mean zero and standard deviation one
  - Specify the number of degrees of freedom
    - *dt*: Distribution function
    - *pt*: Cumulative probability function
    - *qt*: Inverse cumulative probability distribution
    - *rt*: Random
-

# Binomial Distribution

The binomial distribution requires two extra parameters, the number of trials and the probability of success for a single trial.

- *dbinom*: Value of the probability density function (pdf)
- *pbinom*: Cumulative probability distribution function (cdf)
- *qbinom*: Inverse cumulative probability distribution
- *rbinom*: random numbers

# Standard probability density function for the binomial distribution:

#If we flip a fair coin 10 times, what is the probability of getting exactly 5 heads? (a fair coin ( $p(\text{head})=.5$ ))

```
> dbinom(5, size=10, prob=0.5) #calculate binomial probability
```

# cumulative probability of getting X successes

# If we flip a fair coin 10 times, what is the probability of getting 5 or less heads?

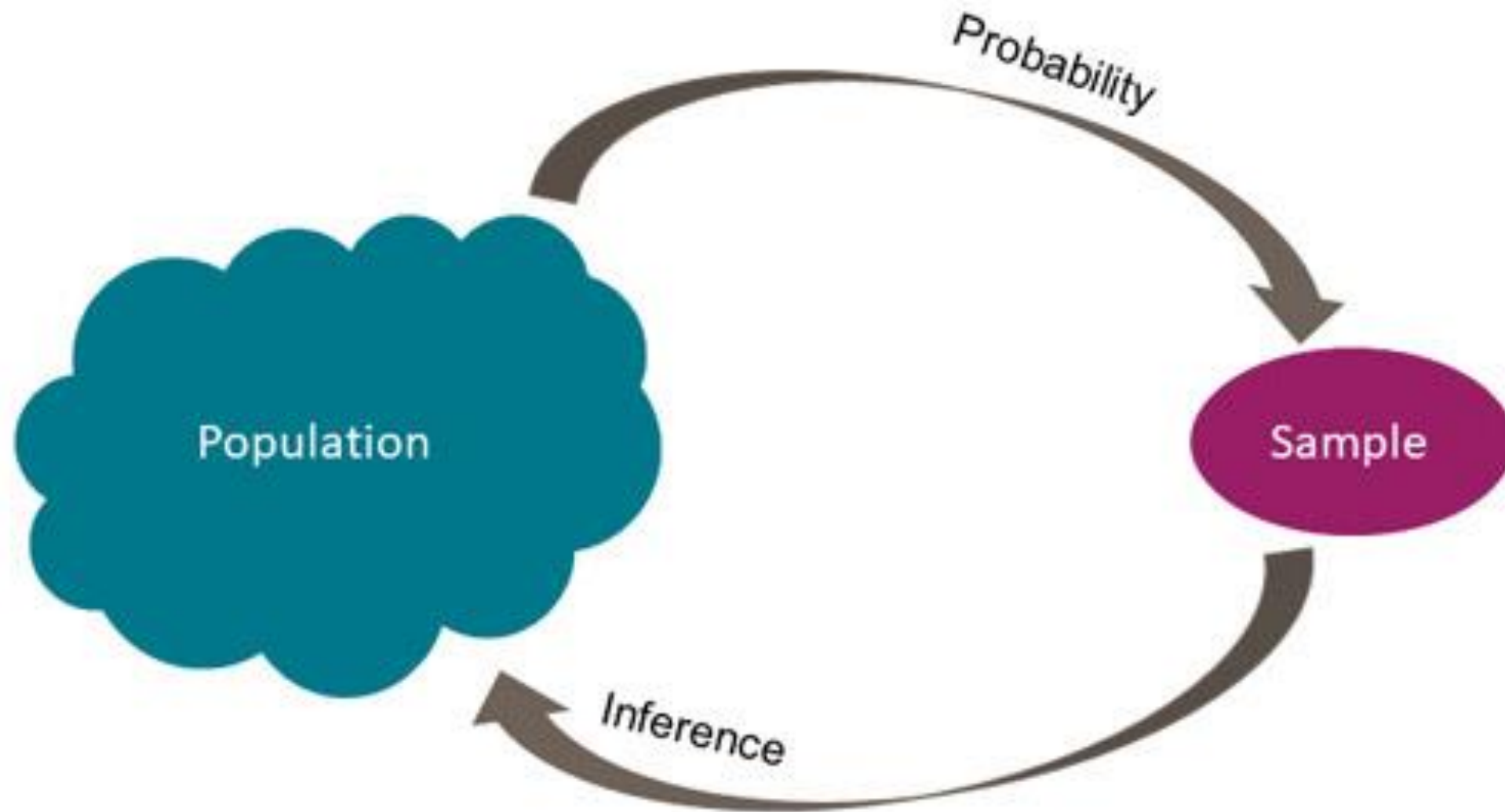
```
> pbinom(5,10,0.5)
```

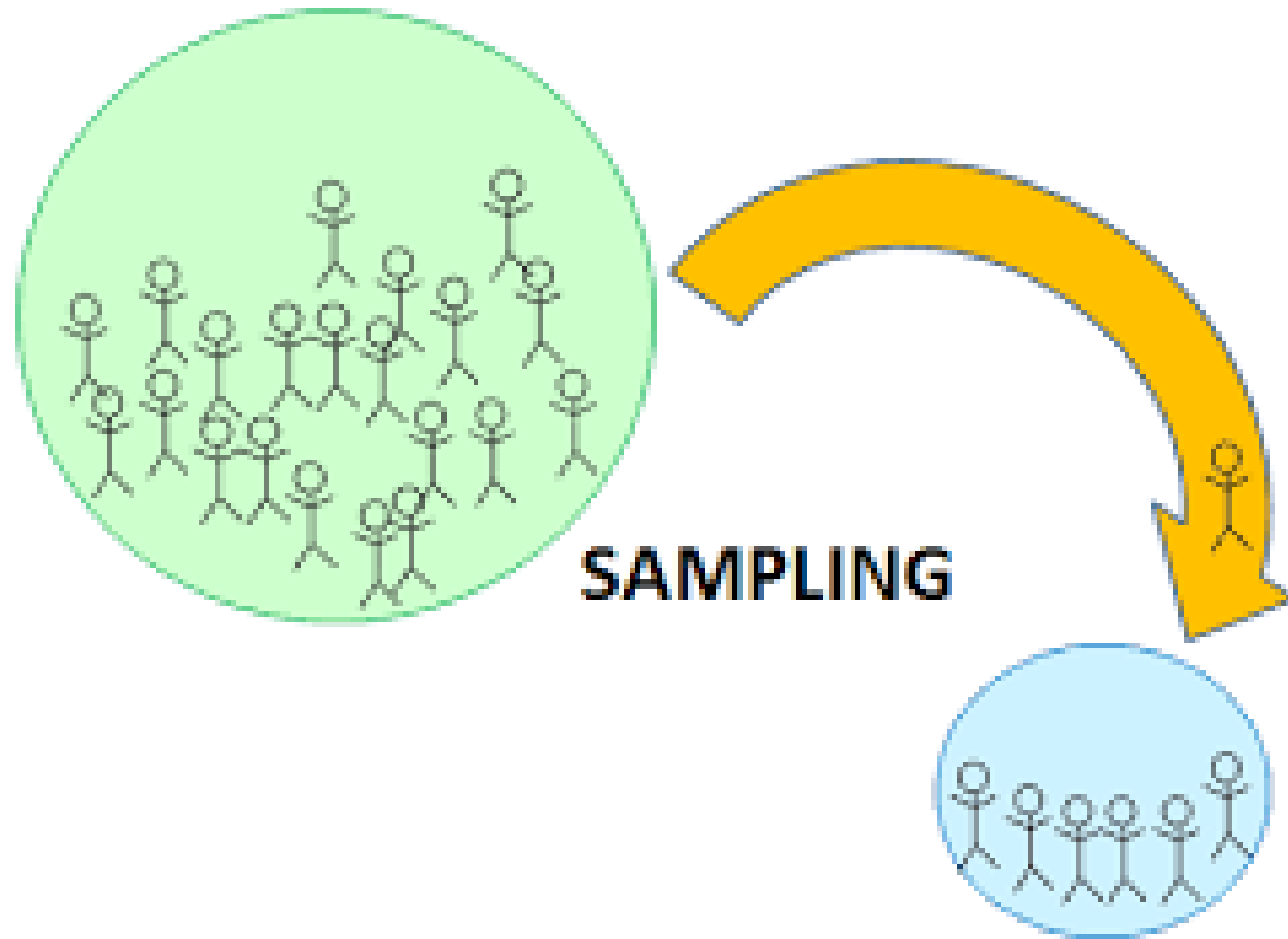
There is a difference between pbinom and dbinom!!

# Exercises

1. In some group of people the average body weight was 65 Kilograms (SD=7). It was assumed that body weight distribution is normal. When carrying out screening examination among adults, the probability was calculated, that a person who will participate in this study will have a body weight  $< 66$  Kilograms. How was it done?
2. In a certain population aged 40 - 50 where a blood pressure is equal to 135 -155 mmHg it was concluded that sport activity prevents from heart attack (ha) episodes. It turned out that probability of preventing ha episodes is equal to 60 %.  
What is a probability that:
  - a. sport activity would protect 8 out of 10 persons from ha episodes ?
  - b. sport activity would protect at least 4 persons from this population ?
3. See the Binomial Distribution.R for more exercises

# Inferential statistics





# Data Collection



First problem a statistician faces: how to obtain the data.



It is important to obtain *good*, or *representative*, data.



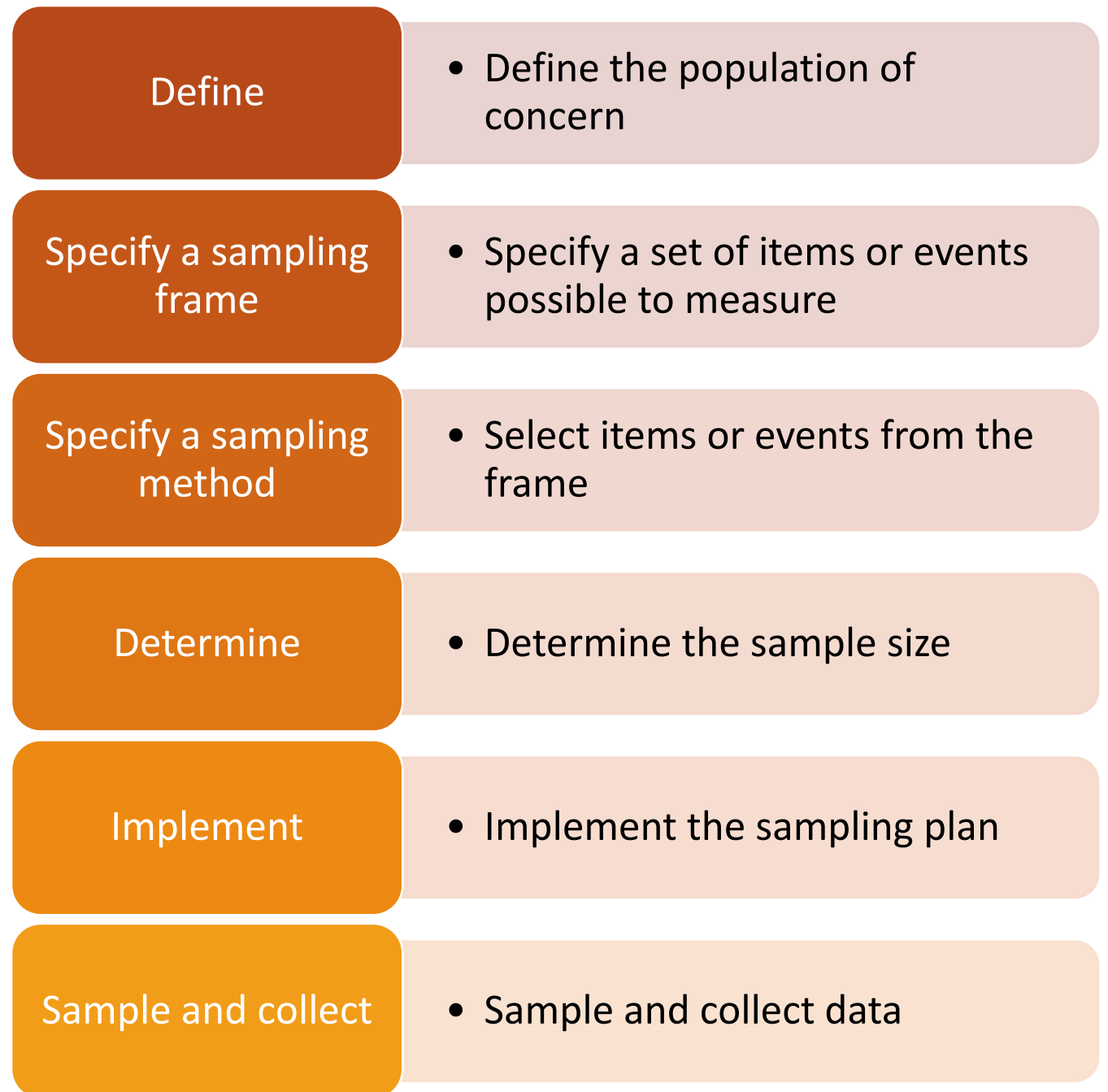
Inferences are made based on statistics obtained from the data.

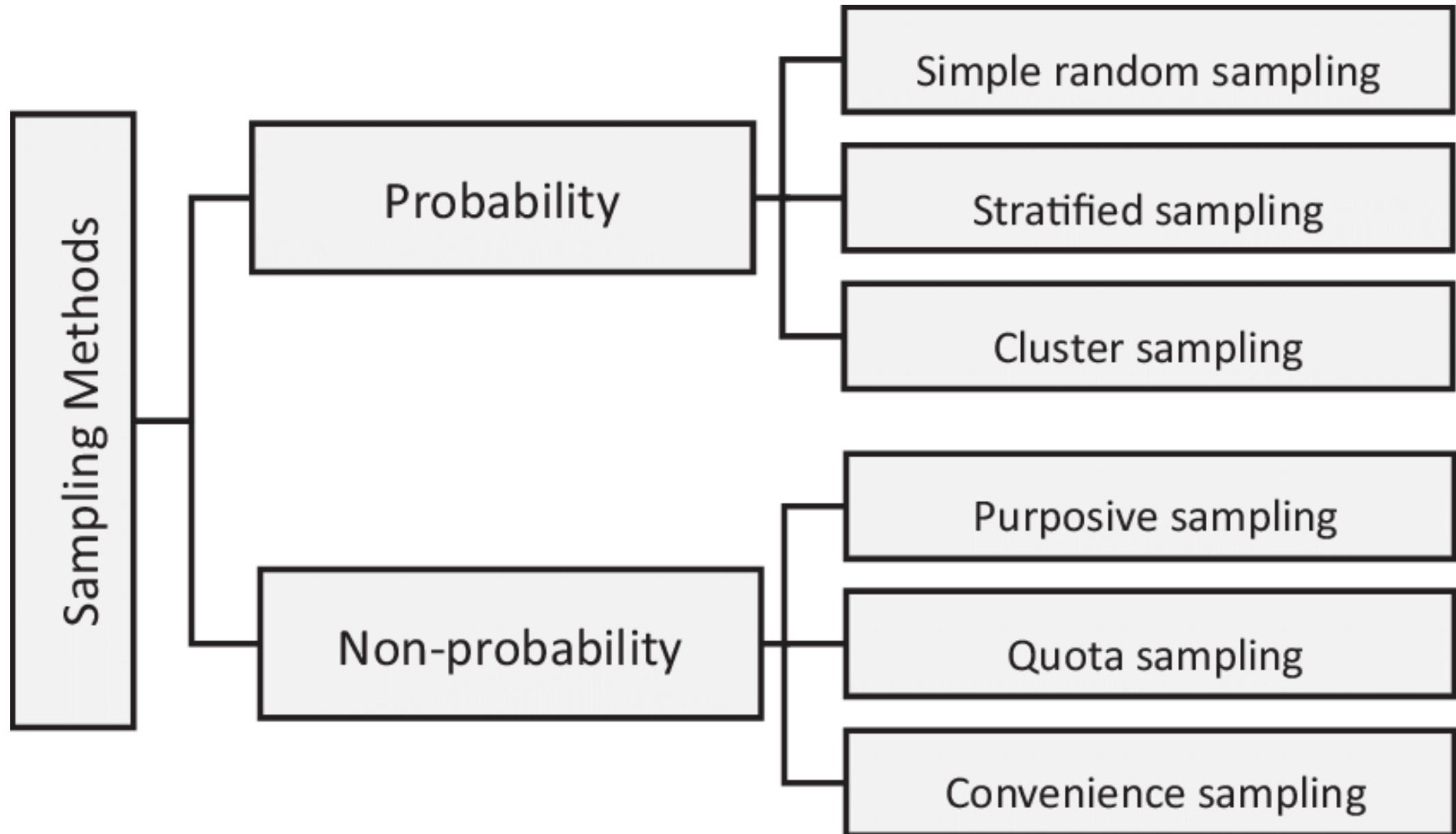


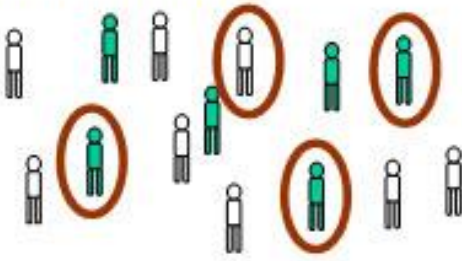
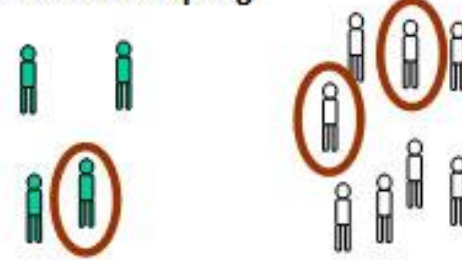
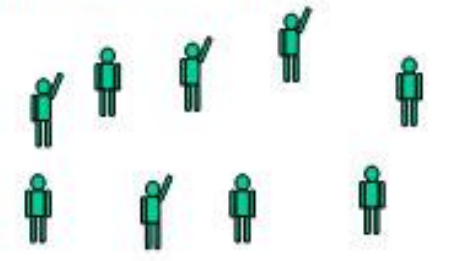
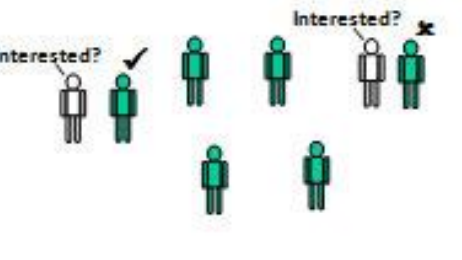
Inferences can only be as good as the data.



# The sampling process comprises several stages:





<b>Random sampling</b> 	<p>Every member of a population has an equal chance of being selected</p> <p>E.g. Pulling names out of a hat</p>	<p>For very large samples it provides the best chance of an unbiased representative sample</p>	<p>For large populations it is time-consuming to create a list of every individual.</p>
<b>Stratified sampling</b> 	<p>Dividing the target population into important subcategories</p> <p>Selecting members in proportion that they occur in the population</p> <p>E.g. 2.5% of British are of Indian origin, so 2.5% of your sample should be of Indian origin... and so on</p>	<p>A deliberate effort is made to make the sample representative of the target population</p>	<p>It can be time consuming as the subcategories have to be identified and proportions calculated</p>
<b>Volunteer sampling</b> 	<p>Individuals who have chosen to be involved in a study. Also called self-selecting</p> <p>E.g. people who responded to an advert for participants</p>	<p>Relatively convenient and ethical if it leads to informed consent</p>	<p>Unrepresentative as it leads to bias on the part of the participant. E.g. a daytime TV advert would not attract full-time workers.</p>
<b>Opportunity sampling</b> 	<p>Simply selecting those people that are available at the time.</p> <p>E.g. going up to people in cafés and asking them to be interviewed</p>	<p>Quick, convenient and economical. A most common type of sampling in practice</p>	<p>Very unrepresentative samples and often biased by the researcher who will likely choose people who are 'helpful'</p>

# (Simple) Random Sampling

- A sample selected in such a way that every element in the population has a equal probability of being chosen.
- Equivalently, all samples of size  $n$  have an equal chance of being selected.
- Obtained either by sampling with replacement from a finite population or by sampling without replacement from an infinite population.
- Inherent in the concept of randomness: the next result (or occurrence) is not predictable.
- Proper procedure for selecting a random sample: use a random number generator or a table of random numbers
- Assumed when performing conventional statistical analyses
- No guarantee of a representative sample
- May not be feasible (e.g., costly, impractical)

# Stratified Sampling (Proportional Sample)

- Population gets partitioned into groups based on a factor that may influence the variable that is being measured.
- These groups called strata.
- An individual group is called a stratum.
- To perform **stratified sampling**:
  - Partition the population into groups (strata)
  - Obtain a simple random sample from each group (stratum)
  - Collect data on each sampling unit that was randomly sampled from each group (stratum)
  - Works best when a heterogeneous population is split into fairly homogeneous groups.
- Generally, produces more precise estimates of the population percent than estimates that would be found from a simple random sample. More control over representativeness. Allows for intentional oversampling which permits greater statistical precision (i.e., decreases standard errors).
- Must have data on the characteristics of the population in order to select the sample.

# Cluster Sampling

- Stratifying the sampling frame and then selecting some or all of the items from some of, but not all, the strata.
  - Divide the population into groups (clusters).
  - Obtain a simple random sample of so many clusters from all possible clusters.
  - Obtain data on every sampling unit in each of the randomly selected clusters.
- Note that, unlike with the strata in stratified sampling, the clusters should be microcosms, rather than subsections, of the population.
- Each cluster should be heterogeneous.
- Statistical analysis using cluster sampling often more complicated than stratified sampling.
- Decreases statistical precision (individuals within groups tend to be more similar so we have less unique information)

# Syntax for Sample Function in R:

```
sample(x, size, replace = FALSE, prob = NULL)
```

# A single roll of a die is a number between one and six

```
> set.seed(1)
```

```
> sample(1:6, 10, replace=TRUE) #Sample with replacement
```

```
> sample(1:6, 10, replace=TRUE)
```

#Using the Parkinson dataset

```
> set.seed(123) #Setting a seed
```

```
> index <- sample(1:nrow(pd), 5)
```

```
> index
```

```
> pd[index, ] #all rows chosen with column information
```



#See script “Stratified Sampling.R”. Script features functions:

- group\_by
- sample\_n
- sample\_frac

#Stratified sampling

```
> set.seed(1) #set a seed
```

```
> d1 <- data.frame(ID = 1:100, A = sample(c("AA", "BB", "CC", "DD", "EE"), 100,  
    replace = TRUE),  
    B = rnorm(100),  
    C = abs(round(rnorm(100), digits = 1)),  
    D = sample(c("CA", "NY", "TX"), 100, replace = TRUE),  
    E = sample(c("M", "F"), 100, replace = TRUE))
```

# Non-probability Sampling

Sampling types that should be avoided:

- Convenience (accidental) – selected based on availability
- Quota – selected based on availability with “quotas” being selected to represent the distribution in the population.
- Judgmental – researcher selects units he/she thinks are more representative of the population. Every unit is not eligible for inclusion in the sample, personal biases
- Snowball – unit with a desired characteristic is identified. This unit then identifies other units with desired characteristics and so on.... (i.e., social networks)

These are referred to as "sampling disasters".

- Biased samples
- Based on human choice rather than random selection
- Statistical theory cannot explain how they might behave, and potential sources of bias are rampant.

# Discussion: Variable Selection

How do these limitations impact your selection of data?

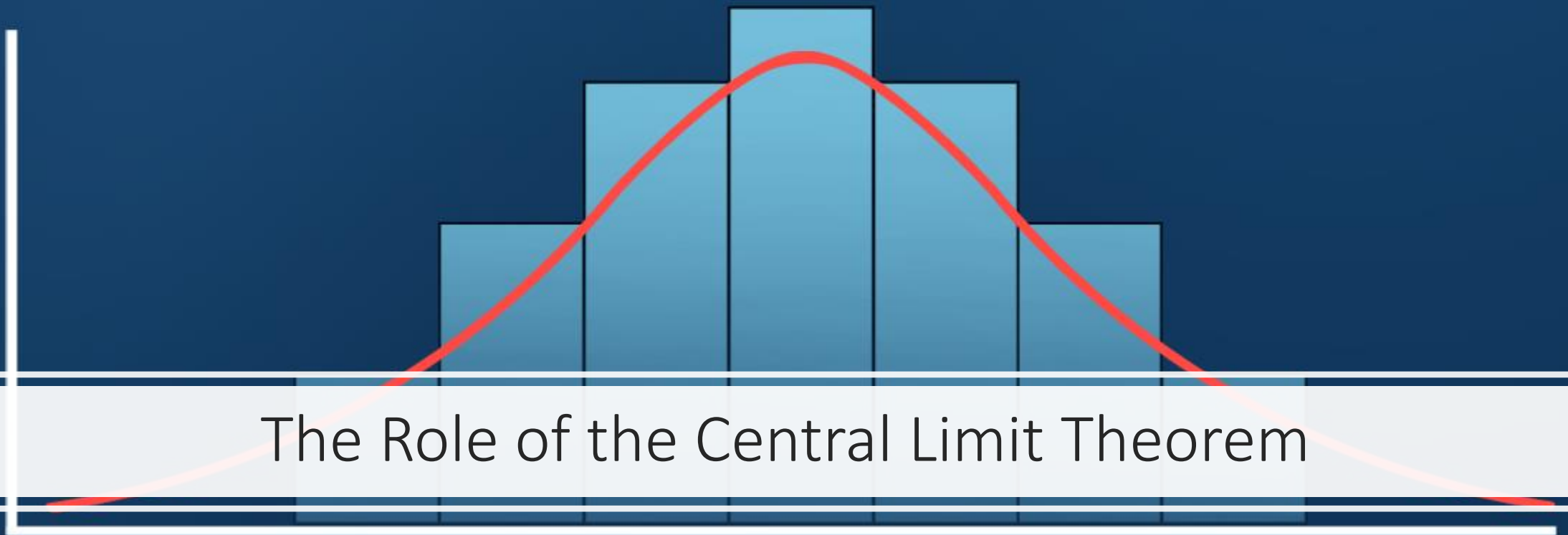
- Highly predictive variables — the use of which is prohibited by legal, ethical or regulatory rules.
- Some variables might not be available or might be of poor-quality during modeling or production stages.
- The business will always have the last word and might insist that only business-sound variables are included or request monotonically increasing or decreasing effects.

# Sampling Challenges



- *Sampling error* - discrepancies between the sample and the population on a certain parameter that are due to random differences; no fault of the researcher.
- *Systematic error* - difference between the sample and the population that is due to a systematic difference between the two rather than random chance alone.
- *Response rate* - sample can become self-selecting, and that there may be something about people who choose to participate in the study that affects one of the variables of interest.
- *Coverage error* - refers to the fact that sometimes researchers mistakenly restrict their sampling frame to a subset of the population of interest
- \*The more participants a study has, the less likely the study is to suffer from sampling error.

# CENTRAL LIMIT THEOREM



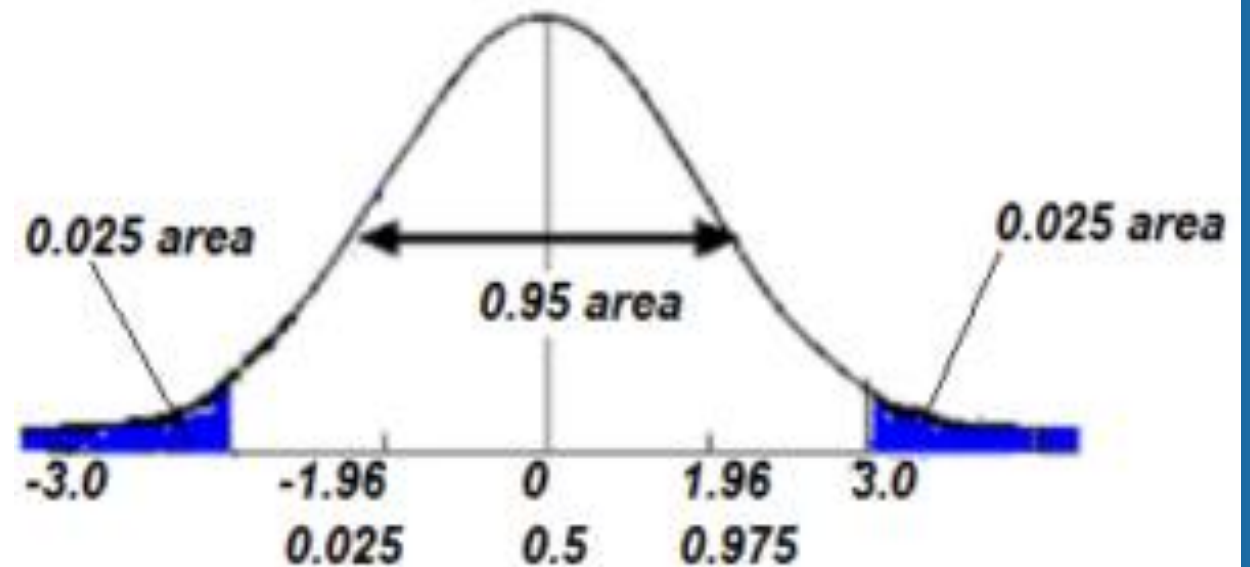
The Role of the Central Limit Theorem

# Confidence Level and Intervals

- A confidence interval is a particular kind of interval estimate of a population parameter. Instead of estimating the parameter by a single value, an interval likely to include the parameter is given. e.g.  $40 \pm 2$  or  $40 \pm 5\%$ .
- **Population Size:** population is the entire entities concerning which statistical inferences are to be drawn. The population size is the total number of the entire entities.

[Sample Size Calculator](#)

*An illustration of 95% confidence interval for the mean*



# Confidence Intervals

```
> a <- 5
```

**<- Sample Mean**

```
> s <- 2
```

**<- The Standard Deviation**

```
> n <- 20
```

**<- Sample Size**

```
> error <- qnorm(0.975)*s/sqrt(n)
```

**<- Calculate the error term**

```
> left <- a-error
```

**<- Left (Lower Bound)**

```
> right <- a+error
```

**<- Right (Upper Bound)**

```
> left
```

```
[1] 4.123477
```

```
> right
```

```
[1] 5.876523
```



# The Process of Statistical Analysis

When we have resource constraints, Statistical Analysis enables us to make quantitative inferences based on an amount of information we can analyze (a sample).

## Form Hypotheses

- Null: Nothing special
- Alternative: Something unique, an actionable finding, etc.

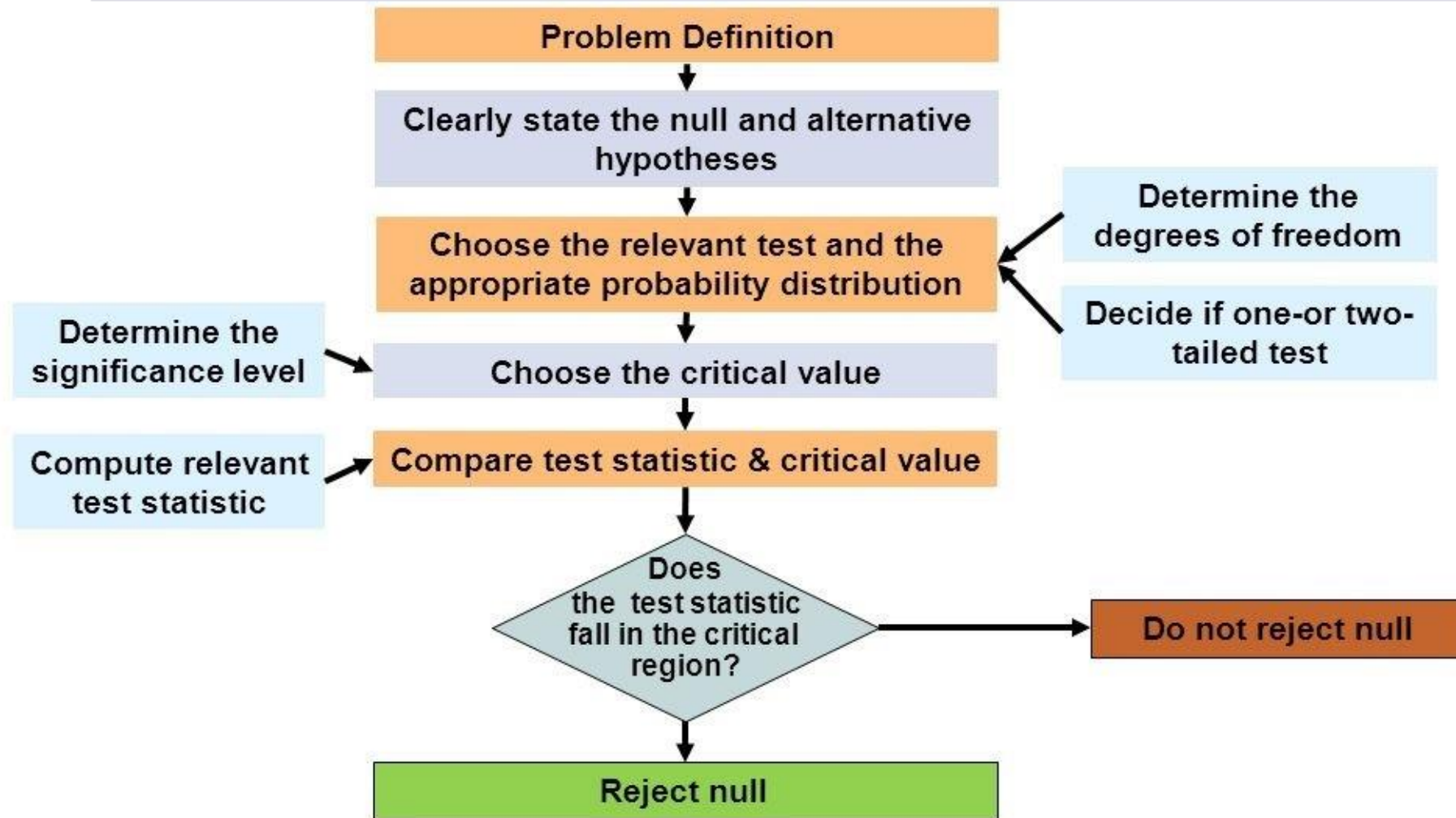
## Identify Data Source

- Don't go overboard!
- Collect your own, OR
- Use secondary data

## Prove/Disprove Hypothesis

- Is Type I or Type II error worse?
- Choose confidence level
- Reject/not reject null

# Hypothesis Testing Process



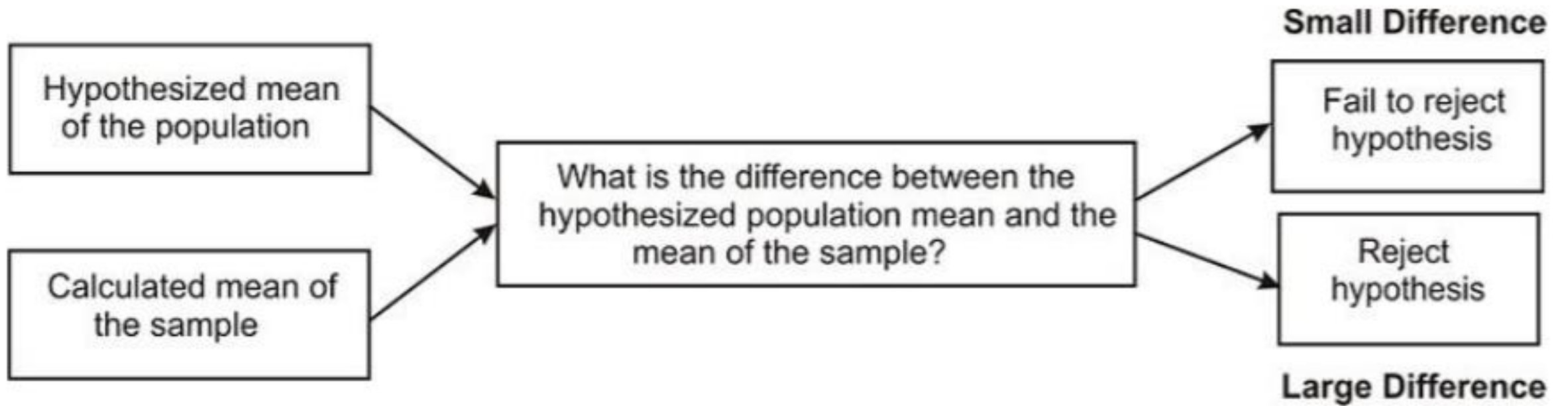
# Example

Consider a person on trial for a “criminal” offense in the United States. Under the US system a jury (or sometimes just the judge) must decide if the person is innocent or guilty while in fact the person may be innocent or guilty.

**Person Is:**

		Innocent	Guilty
		No Error	Error
Jury Says:	Innocent	No Error	Error
	Guilty	Error	No Error

There is a favored assumption, an initial bias. The jury is instructed to assume the person is innocent.



# Hypothesis Testing

- You have sample data, and you are asked to assess the credibility of a statement about population using sample data.
- Suppose we have a website that has a white background and the average engagement on the website by any user is around 20 minutes. Now we are proposing a new website with a yellow background that might increase the average engagement by any user to more than 20 minutes. So, we state the null and alternate hypothesis as:
  1.  $H_0: \mu = 20 \text{ min after the change} \mid H_a: \mu > 20 \text{ min after the change}$
  2. Significance Level :  $\alpha = 0.05$

Recall the **Central Limit Theorem**:

- Using this, we determine if our assumption for the null hypothesis (**H0**) is reasonable or not. If it is unlikely, by the **Rare Event rule**, our hypothesis is probably incorrect (i.e., reject **H0**).

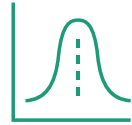
In general, we use the following:

- If the test sample yields an unlikely result, it is probably incorrect (reject **H0** (**large difference**))
- If the test sample yields a likely result, it is probably correct (fail to reject **H0** (**small difference**))

# P – values and Significance tests



There is an extremely close relationship between confidence intervals and hypothesis testing.



When a 95% confidence interval is constructed, all values in the interval are considered plausible values for the parameter being estimated. Values outside the interval are rejected as relatively implausible.



The confidence interval tells you **more than just the possible range around the estimate**. It also tells you about **how stable the estimate is**.



If exact p-value is reported, then the relationship **between confidence intervals and hypothesis testing** is very close.



However, the objective of the two methods is different:  
**Hypothesis testing** relates to a single conclusion of statistical significance vs. no statistical significance.





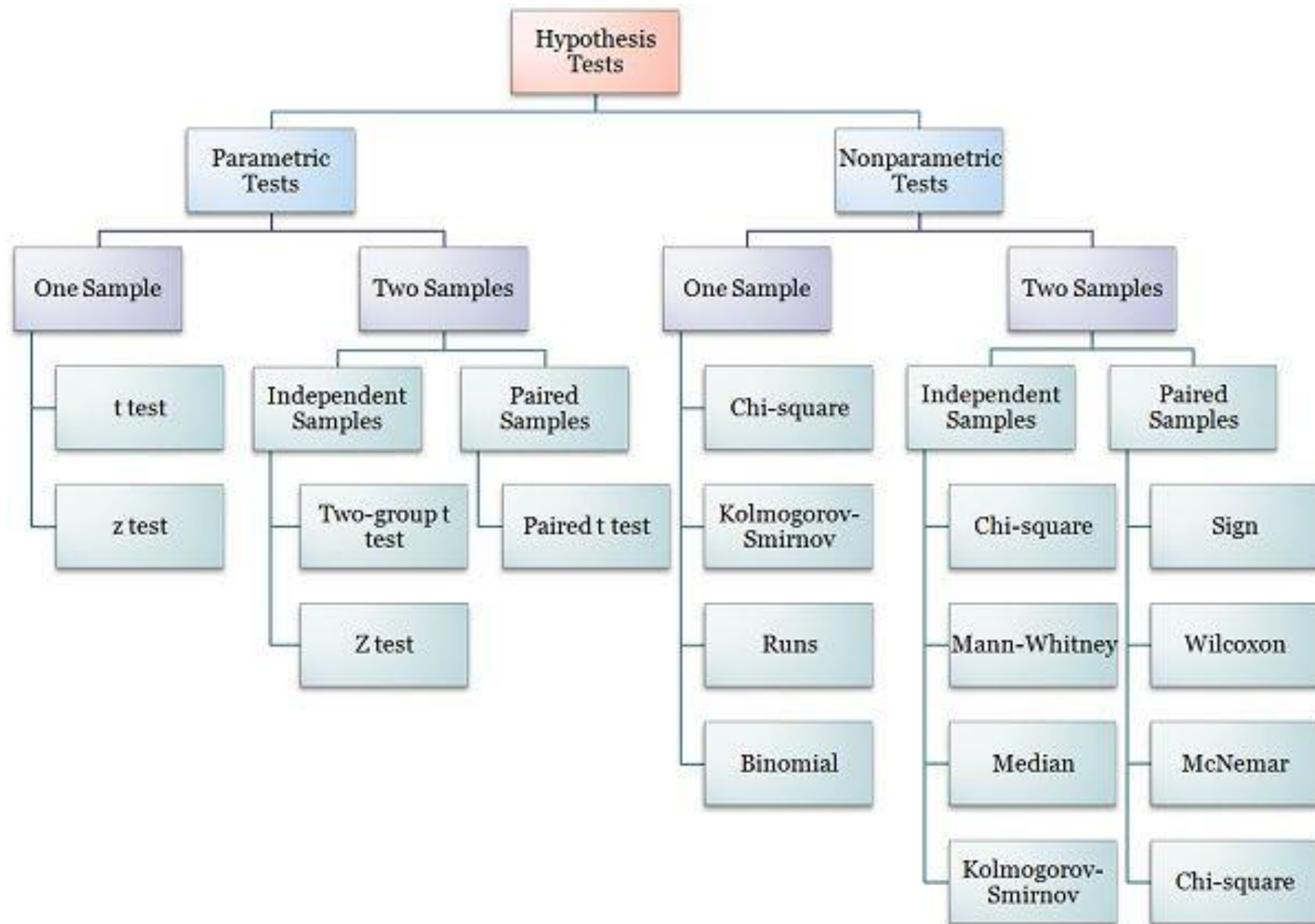
Probability theory: Allows us to calculate the exact *probability* that chance was the real reason for the relationship.



Probability theory allows us to produce test statistics (using mathematical formulas)



A test statistic is a number that is used to decide whether to accept or reject the null hypothesis.



# Applications of Hypothesis Testing

1. When you want to compare the sample mean with the population mean. For example – You would like to determine if the average life of a bulb from brand X is 12 years or not.  
In this case, when you want to check if the sample mean represents the population mean, then you should run **One Sample t-test**.
2. When you want to compare the means of two independent variables. One of which can be a categorical variable. In this case, we run **Two sample t-test**.
3. When you want to compare the before and after-effects of an experiment or a treatment. Then, in that case, we run **Paired t-test**.
4. When you want to compare more than two independent variables; in that case, we run **ANOVA test**.
5. In all the above applications, we assumed that variables are numeric.  
However, When you want to compare two categorical variables, we run **Chi-square test**.

# $t$ -test statistic

- The  $t$  statistic allows researchers to use sample data to test hypotheses about an unknown population mean.
- The  $t$  statistic is mostly used when a researcher wants to determine whether or not a treatment intervention causes a significant change from a population or untreated mean.
- The goal for a hypothesis test is to evaluate the *significance* of the observed discrepancy between a sample mean and the population mean.
- Therefore, the  $t$  statistic requires that you use the sample data to compute an estimated standard error of  $M$ .
- A large value for  $t$  (a large ratio) indicates that the obtained difference between the data and the hypothesis is greater than would be expected if the treatment has no effect.

# Example

Does caffeine improve our reaction time?

We can find 40 people and give random assignment:

20 get a caffeine pill (experimental group)

20 get a sugar pill (control group)

We give the people time for any reaction and record the results.

experimental group (caffeine)

Mean = 501.78ms

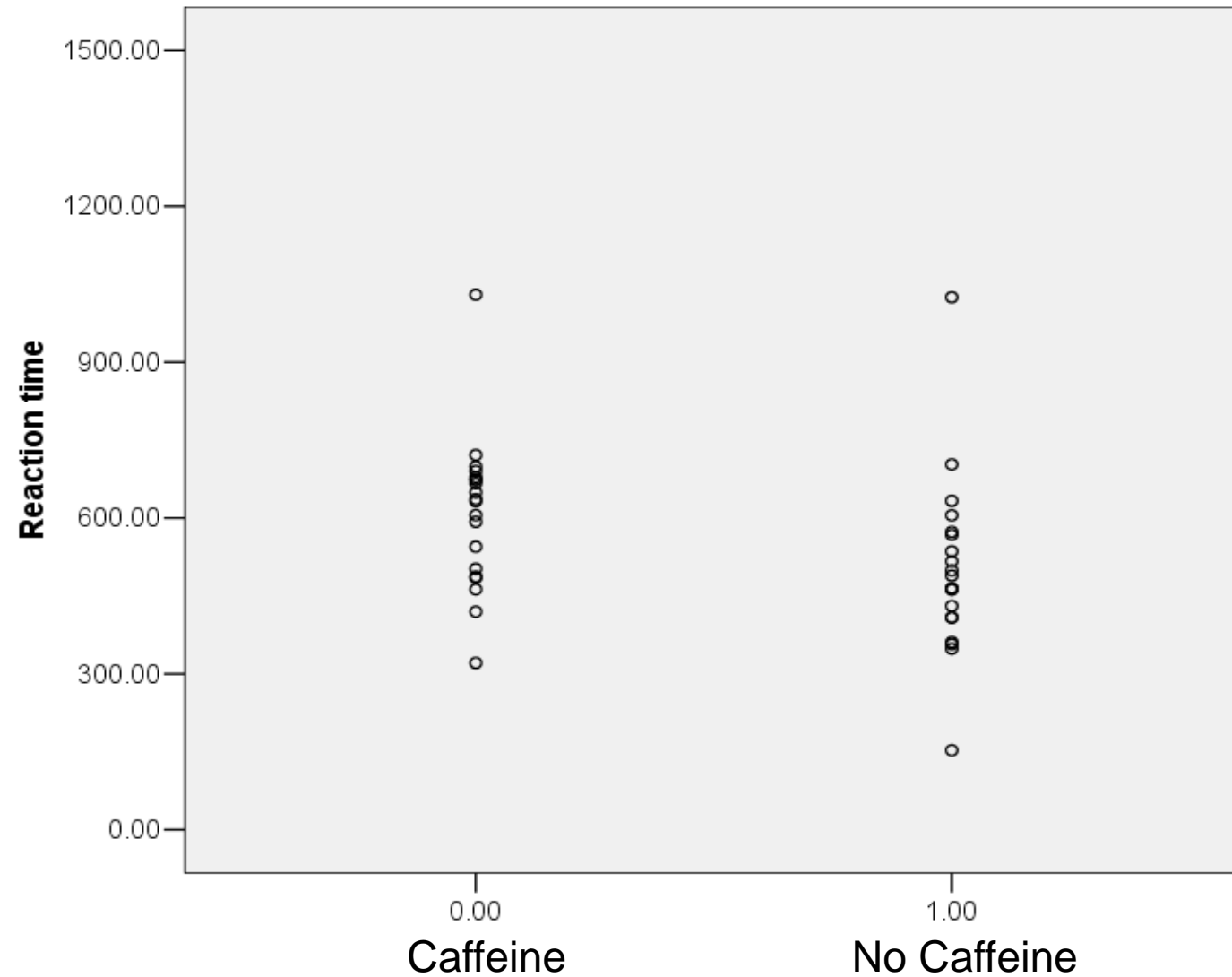
standard deviation = 171.4ms

control group results (placebo)

Mean = 603.47ms

standard deviation = 141.9

# Example of when to do a t-test





- A t test yields a p-value generally.
- The t test gives us a p-value that allows a measurement of confidence in the observed difference.
- It allows us to say that the difference is real and not just by chance.
- A p-value of  $<0.05$  is a common criteria for significance. This is statistically significant.

# Perform One-Sample t-test

The basic method of applying a t-test to compare two vectors of numeric data:

```
t.test(data.1, data.2)
```

```
> set.seed(100)    # Generating random data with normal distribution
```

```
> x <- round(rnorm(30, mean = 12, sd = 1), 0)
```

```
> t.test(x, mu=12) # Checking if mean is really 12 years:
```

p-value = 0.813, this value is greater than alpha value, and thus we must accept the null hypothesis. Here the null hypothesis was that the average life of the bulb is 12. And the alternative hypothesis was that it is not equal to 12.

95 percent confidence interval:

(11.74772, 12.31895) – The 95% CI also includes the twelve, and thus it is fine to state that the mean value is 12.



## #Create sample data

#Define Sample 1

```
smp2014 <- c(222, 823, 1092, 400, 948, 836)
```

#Define Sample 2

```
smp2019 <- c(910, 650, 700, 892, 229, 1051)
```

#Two sample T-test

```
t.test(smp2014, smp2019, var.equal=FALSE)
```

*#What is the p-value?*

#Run Welch's T-test of Equal Variance

```
t.test(smp2014, smp2019, var.equal=TRUE)
```

# Conditions for a One-Sample z-test

A one sample z test is one of the most basic types of hypothesis test.

1. Normality -normal population. Data roughly fits a [bell curve](#) shape.  
Central Limit Theorem ( $n \geq 30$ )

Graphing - Qplots/box plots/normal probability

2. Independence - population is greater than 10 times the sample size ( $N \geq 10n$ )

Null hypothesis for a 1 sample z-test

- The mean of a population is equal to the sample mean,  $\mu = \mu_0$

Alternative hypotheses for a 1 sample z-test

- The mean of a population is (not equal to/less than/greater than) the sample mean,  
 $\mu \neq \mu_0$  OR  $\mu < \mu_0$  OR  $\mu > \mu_0$

# Conditions for a Two-Sample z-test

Both (two) populations are normally distributed

Null hypothesis for a 2-sample z-test

- The difference between the means of 2 populations is zero (equal means)  
 $\mu_1 - \mu_2 = 0$  OR  $\mu_1 = \mu_2$

Alternative hypotheses for a 2-sample z-test

- Population 1 mean is (not equal to/greater than/less than) the population 2 mean

```
z.test = function(x,mu,popvar),
```

the first argument is the vector of data, the second is the population mean, and the third is the population variance. The left curly bracket signifies that the remainder of the code is what happens inside the function.

1. Create a vector that will hold the one-tailed probability of the z-score.

```
one.tail.p <- NULL
```

2. Then calculate the z-score and round it to three decimal places:

```
z.score <- round((mean(x)-mu)/(popvar/sqrt(length(x))),3)
```

3. Calculate the one-tailed probability (the proportion of area beyond the calculated z-score), and again round to three decimal places:

```
one.tail.p <- round(pnorm(abs(z.score),lower.tail = FALSE),3)
```

# z Testing: Calculating the score and probabilities

```
z.test = function(x,mu,popvar){  
  one.tail.p <- NULL  
  z.score <- round((mean(x)-mu)/(popvar/sqrt(length(x))),3)  
  one.tail.p <- round(pnorm(abs(z.score),lower.tail = FALSE),3)  
  cat(" z =",z.score,"\n",  
    "one-tailed probability =", one.tail.p,"\n",  
    "two-tailed probability =", 2*one.tail.p )}
```

**Conclusions** are sentence answers which include whether there is enough evidence or not (based on the decision), the level of significance, and whether the original claim is supported or rejected.

**Conclusions** are based on the original claim, which may be the null or alternative hypotheses. The decisions are always based on the null hypothesis

Decision	Original Claim	
	$H_0$ "REJECT"	$H_1$ "SUPPORT"
Reject $H_0$ "SUFFICIENT"	There is <b>sufficient</b> evidence at the alpha level of significance to <b>reject</b> the claim that (insert original claim here)	There is <b>sufficient</b> evidence at the alpha level of significance to <b>support</b> the claim that (insert original claim here)
Fail to reject $H_0$ "INSUFFICIENT"	There is <b>insufficient</b> evidence at the alpha level of significance to <b>reject</b> the claim that (insert original claim here)	There is <b>insufficient</b> evidence at the alpha level of significance to <b>support</b> the claim that (insert original claim here)

# Were you right ? ... Confusion matrix

		Predicted Class		
		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) <b>Type II Error</b>	<b>Sensitivity</b> $\frac{TP}{(TP + FN)}$
	Negative	False Positive (FP) <b>Type I Error</b>	True Negative (TN)	<b>Specificity</b> $\frac{TN}{(TN + FP)}$
		<b>Precision</b> $\frac{TP}{(TP + FP)}$	<b>Negative Predictive Value</b> $\frac{TN}{(TN + FN)}$	<b>Accuracy</b> $\frac{TP + TN}{(TP + TN + FP + FN)}$

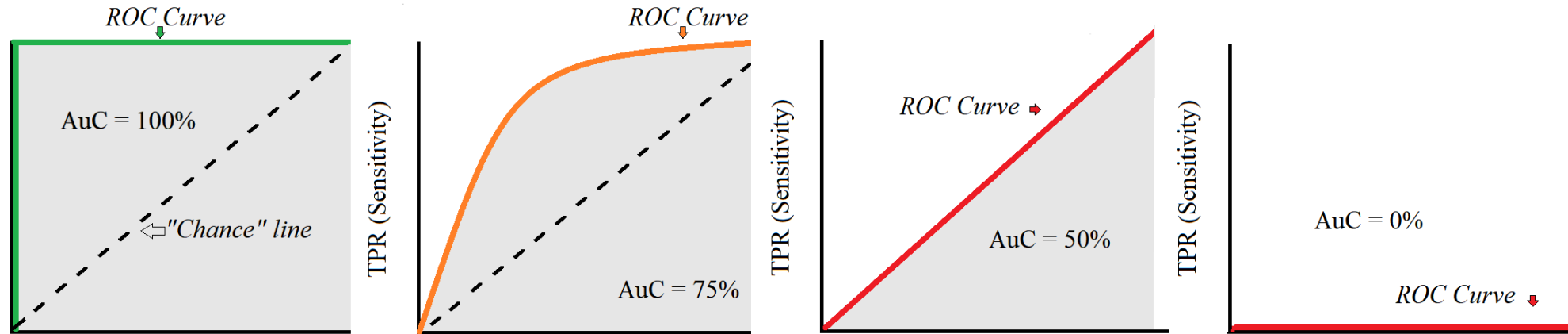
## Which of the two errors is more serious? Type I or Type II ?

- Since *Type I is the more serious error* (usually), rationale is stick to the status quo or default assumption, at least you're not making things *worse*. But it depends...
- Note: alpha is not a Type I error. Alpha, ' $\alpha$ ', is the *probability of committing* a Type I error. (i.e.  $\alpha = .05$ ; 5% is the level of reasonable doubt that you are willing to accept when statistical tests are used to analyze the data after the study is completed.). Likewise beta, ' $\beta$ ', is the *probability of committing* a Type II error. A Type II error relates to the concept of "power,"; having enough power depends on whether the sample size is sufficiently large to detect a difference when it exists.
- Although type I and type II errors can never be avoided entirely, the you can reduce the likelihood of occurrence by increasing the sample size (the larger the sample, the lesser is the likelihood that it will differ substantially from the population).
- False-positive and false-negative results can also occur because of bias (observer, instrument, recall, etc.). (Errors due to bias, however, are not referred to as type I and type II errors.) Such errors are troublesome, since they may be difficult to detect and cannot usually be quantified.





# ROC AUC



AUC means Area Under Curve, which is calculated for the ROC.

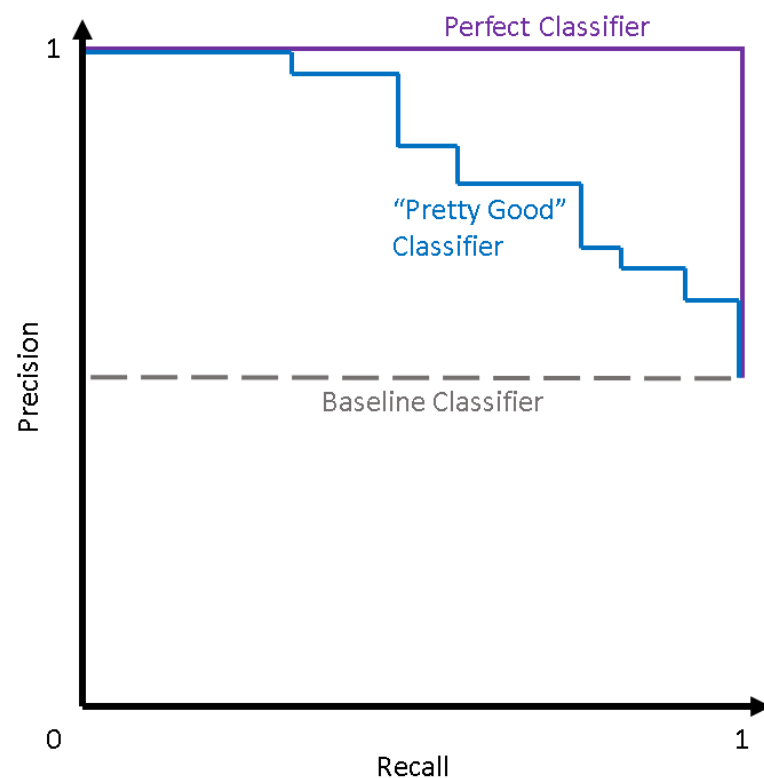
An ROC curve is a graph plotted between Sensitivity and False positive rate.

AUC - ROC curve is a performance measurement for the classification problems at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes.

Higher the AUC, the better the model is at predicting 0s as 0s and 1s as 1s.

An excellent model has AUC near to the 1 which means it has a good measure of separability. A poor model has AUC near to the 0 which means it has the worst measure of separability.

# Precision Recall Curve



AUC-PR stands for area under the (precision-recall) curve. Much like the ROC curve, the precision-recall curve is used for evaluating the performance of binary classification algorithms.

It is often used in situations where classes are heavily imbalanced.

Also, like ROC curves, precision-recall curves provide a graphical representation of a classifier's performance across many thresholds, rather than a single value.

Generally, the higher the AUC-PR score, the better a classifier performs for the given task.

One way to calculate AUC-PR is to find the AP, or average precision.

# Precision Recall Curve: popular model performance metrics to evaluate binary classification model

Cutoff	0.9	0.75	0.6	0.5	0.4
Recall (X)	0.14	0.41	0.69	0.88	0.90
Precision (Y)	0.90	0.85	0.82	0.81	0.50

```
#Area under the precision recall area
recall=c(0.14, 0.41, 0.69, 0.88, 0.90)
precision=c(0.90, 0.85, 0.82, 0.81, 0.50)
i = 2:length(recall)
recall = recall[i] - recall[i-1]
precision = precision[i] + precision[i-1]
(AUPRC = sum(recall * precision)/2)
```

# Chi-square

The function used for performing chi-Square test is `chisq.test()`. The syntax is `chisq.test(data)`.

```
> install.library("MASS") #install package
```

```
> library ("MASS")      #load library
```

```
# Create a data frame from the main data set.
```

```
> car.data <- data.frame(Cars93$AirBags, Cars93$Type)
```

```
# Create a table with the needed variables.
```

```
> car.data = table(Cars93$AirBags, Cars93$Type)
```

```
> print(car.data)
```

```
# Perform the Chi-Square test.
```

```
> print(chisq.test(car.data))
```

# Analysis of Variance (ANOVA)

#One Way ANOVA (Completely Randomized Design)

```
> fit <- aov(y ~ A, data=mydataframe)
```

#Randomized Block Design (B is the blocking factor)

```
> fit <- aov(y ~ A + B, data=mydataframe)
```

#Two Way Factorial Design

```
> fit <- aov(y ~ A + B + A:B, data=mydataframe)
```

```
> fit <- aov(y ~ A*B, data=mydataframe) # same thing
```

#Analysis of Covariance

```
> fit <- aov(y ~ A + x, data=mydataframe)
```

#Diagnostic **plots** provide checks for heteroscedasticity, normality, and influential observations.

# Analysis of Variance (ANOVA)

Here, Null Hypothesis:  $\mu_1 = \mu_2 = \mu_3$

and, Alternative:  $\mu_1 \neq \mu_2 \neq \mu_3$  or  $\mu_1 = \mu_2 \neq \mu_3$  or  $\mu_1 \neq \mu_2 = \mu_3$

```
> result <- aov(Sepal.Length ~ Species, data = iris) #Running anova
```

```
> summary(result) # Checking the result
```

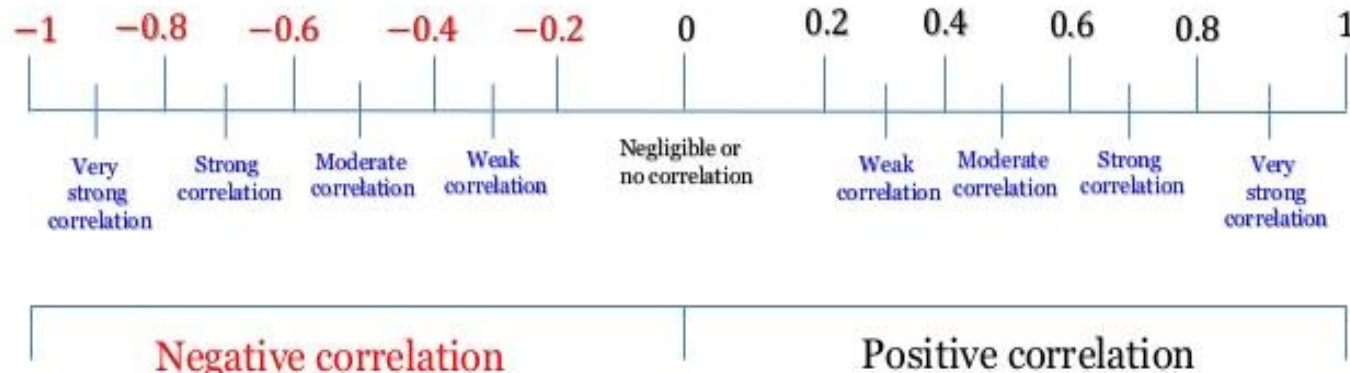
```
> TukeyHSD(result) # pass the model output
```

```
Df Sum Sq Mean Sq F value Pr(>F)
Species    2  63.21  31.606  119.3 <0.00000000000000002 ***
Residuals 147   38.96   0.265
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

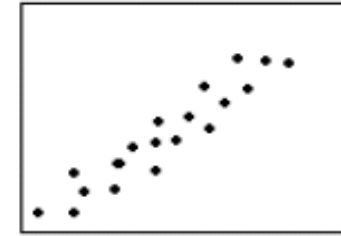
	Outcome variable						
Input Variable		Nominal	Categorical (>2 Categories)	Ordinal	Quantitative Discrete	Quantitative Non-Normal	Quantitative Normal
	Nominal	$\chi^2$ or Fisher's	$\chi^2$	$\chi^2$ -trend or Mann-Whitney	Mann-Whitney	Mann-Whitney or log-rank <sup>a</sup>	Student's <i>t</i> test
	Categorical (2>categories)	$\chi^2$	$\chi^2$	Kruskal-Wallis <sup>b</sup>	Kruskal-Wallis <sup>b</sup>	Kruskal-Wallis <sup>b</sup>	Analysis of variance <sup>c</sup>
	Ordinal (Ordered categories)	$\chi^2$ -trend or Mann-Whitney	e	Spearman rank	Spearman rank	Spearman rank	Spearman rank or linear regression <sup>d</sup>
	Quantitative Discrete	Logistic regression	e	e	Spearman rank	Spearman rank	Spearman rank or linear regression <sup>d</sup>
	Quantitative non-Normal	Logistic regression	e	e	e	Plot data and Pearson or Spearman rank	Plot data and Pearson or Spearman rank and linear regression
	Quantitative Normal	Logistic regression	e	e	e	Linear regression <sup>d</sup>	Pearson and linear regression

# Correlation Coefficient Interpretation Guideline

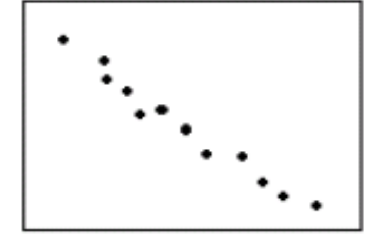
The correlation coefficient ( $r$ ) ranges from -1 (a perfect negative correlation) to 1 (a perfect positive correlation). In short,  $-1 \leq r \leq 1$ .



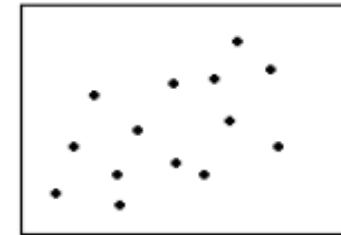
Degree of Correlation



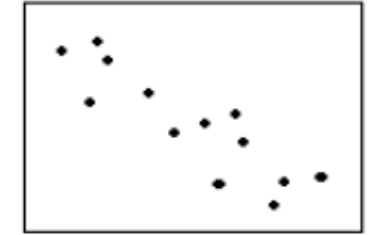
Strong Positive



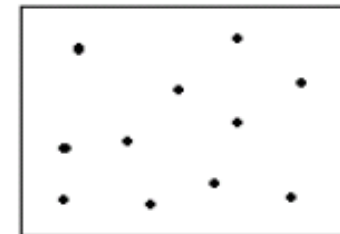
Strong Negative



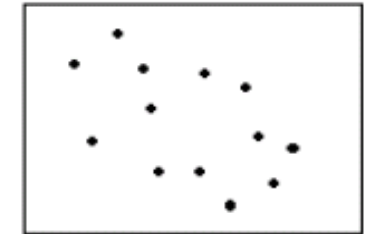
Weak Positive



Moderate Negative



None



Weak Negative



# Correlations

A simplified format is `cor(x, use=, method= )`

Pearson correlation – Pearson correlation is used when we want to assess the degree of association between two quantitative variables.

- `cor(x, method = "pearson")`

Spearman correlation – Use spearman correlation when you want to assess the degree of association between rank-ordered variables.

- `cor(x, method = "spearman")`

Kendall's correlation – Kendall's correlation can also be used to assess the degree of association between rank-ordered variables. However, it is a non-parametric measure.

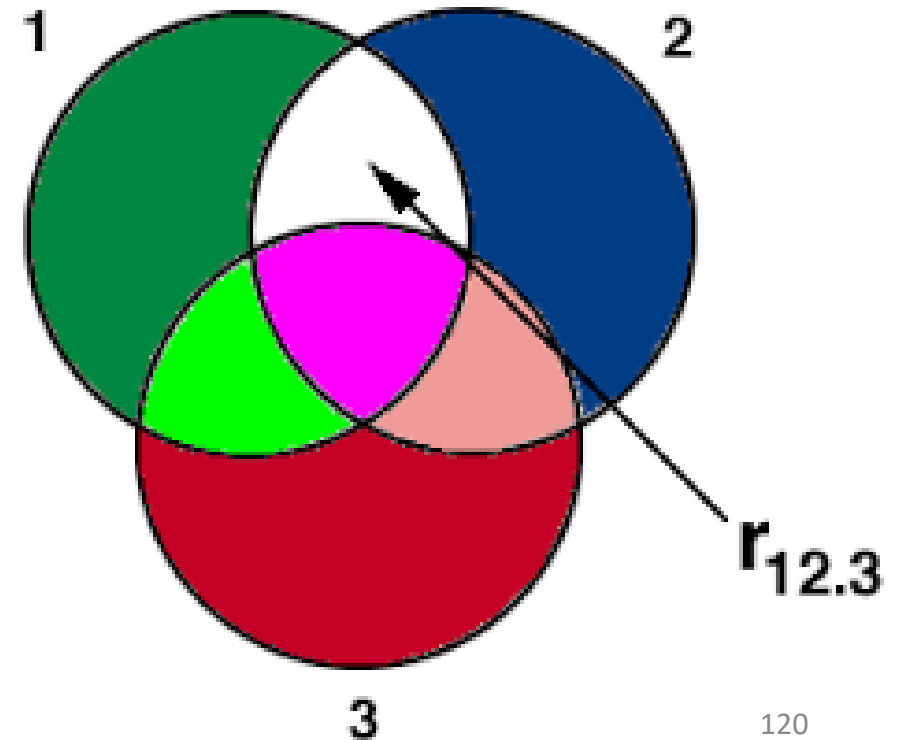
- `cor(x, method = "kendall")`

# Partial Correlation

- Partial Correlation measures relationship between two variables (X,Y) while eliminating influence of a third variable (Z).
- Called “correlation” but it is actually regression. It requires estimation of variances.

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2} \sqrt{1 - r_{23}^2}}$$

```
# Partial correlation  
library(ppcor)  
with(mydata, pcor.test(Height,Weight,Age))
```



# Semi-partial Correlation

- Semi-partial correlation measures the strength of linear relationship between variables X1 and X2 holding X3 constant for just X1 or just X2. It is also called part correlation.

$$r_{1(2.3)} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{23}^2}} \text{ and } r_{2(1.3)} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2}}$$

```
#Semipartial Correlation Coefficient
```

```
with(mydata, spcor.test(Height,Weight,Age))
```

```
#Semi partial correlation - Age constant for Weight only
```

## Data analysis flowchart

