Basic Statistics

About Me

- Name: Shiv
- BSc and MSc in Mathematics from IIT Kharagpur

- PhD in Analytics
- Experience: 19 years (9 Years in Analytics and 10 years in IT)

- Current Role: Chief Data Scientist

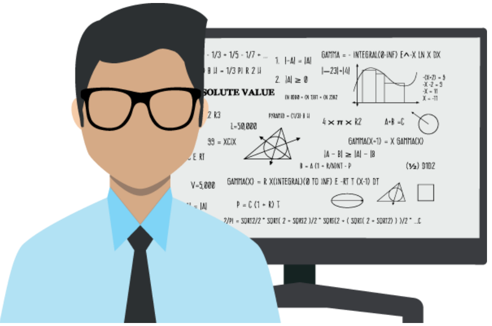


Image source link is at reference section

Prerequisites

- Interest in Mathematics and Statistics
- Expert in XL 2016

Training Content: Part 1 - Introduction and Basics

i. Introduction to statistics
   - What is Statistics
   - Population and Sample
   - Descriptive and Inferential Statistics
   - Parameters and Statistics
   - Type of Data/Variables

ii. Measures of Central Tendency: Mean, median, mode

iii. Measures of Dispersion: Range, quartile deviation, mean deviation , standard deviation

iv. Measure of Shape: Skewness, Kurtosis

v. Sampling Procedure: Probability & Non-Probability

vi. Probability & Distribution: Bernoulli, Binomial, Poisson, Normal, t, Chi Square etc

vii. Normal Distribution
   - The Vials Filling Simulation
   - The Histograms
   - Calculating Proportions
   - The Continuous Distribution
   - Calculating Area
   - Using Tables
   - The Normal Distribution Calculator

viii. Inferences from Samples
   - The Testing Process
   - Sample Averages
   - The Central Limit Theorem
   - Distribution of Strengths
   - Introducing Alpha
   - Standard Deviation Unknown
   - Deciding on the Sample Size

ix. Hypothesis Testing
   - The Testing Process
   - Sample Averages
   - Confidence Intervals
   - Hypothesis Tests
   - The Null Hypothesis
   - The p-Value
   - Interpreting the Test Results
   - One & Two sided Tests
   - Type I & II Errors and Power
   - Deciding on the Sample Size

Training Content: Part 2 – Graphs and Tests

i. The t Distribution

- Calculating Probability with the t Distribution
- Calculating the t statistic
- Using t Distribution Tables
- Comparing Two Samples
- Two sample t Test
- Paired and Unpaired Test

II. The Chi Square Distribution

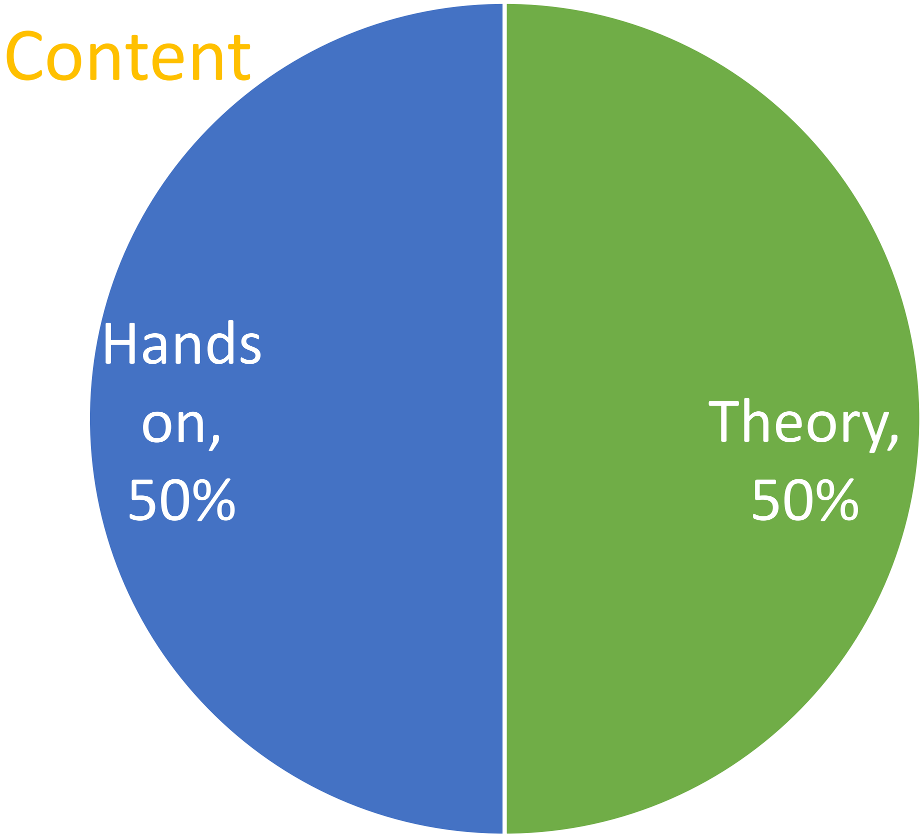
- Calculating the Chi Square statistic
- Using Chi Square Distribution Tables
- Goodness of fit
- Test of Independence
- Comparing Proportions
- Contingency Tables

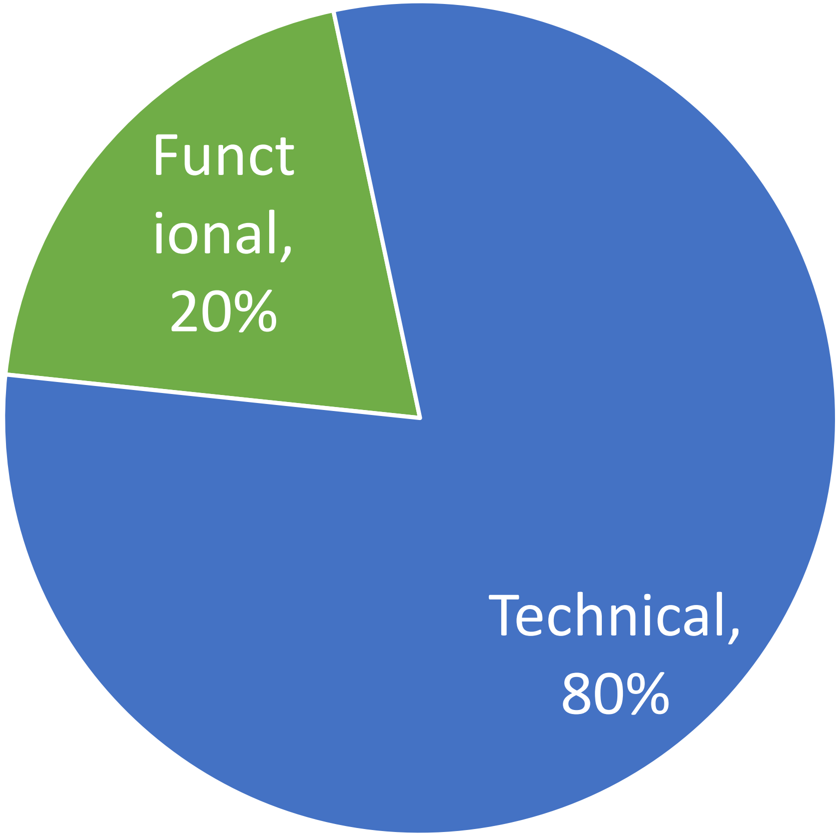
III. ANOVA

IV. Representing Data - Graphical /Tabular

- XY Graphs
- Scatter Graphs
- Correlation
- Box Plots
- Calculating the Quartiles
- Box Plots for Comparison
- Grouped Data
- Cumulative Frequency
- Percentiles
- Pareto Charts
- Stem and Leaf Plots
- Multi variance Charts

Methodology



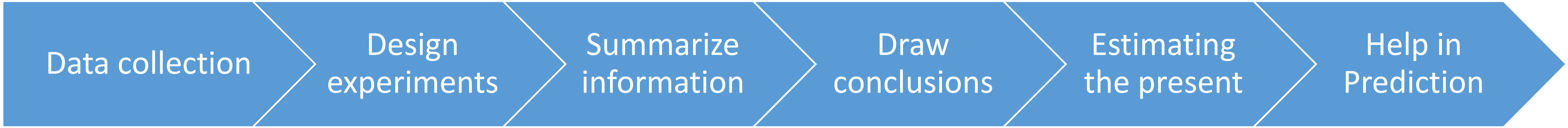


Why Statistics

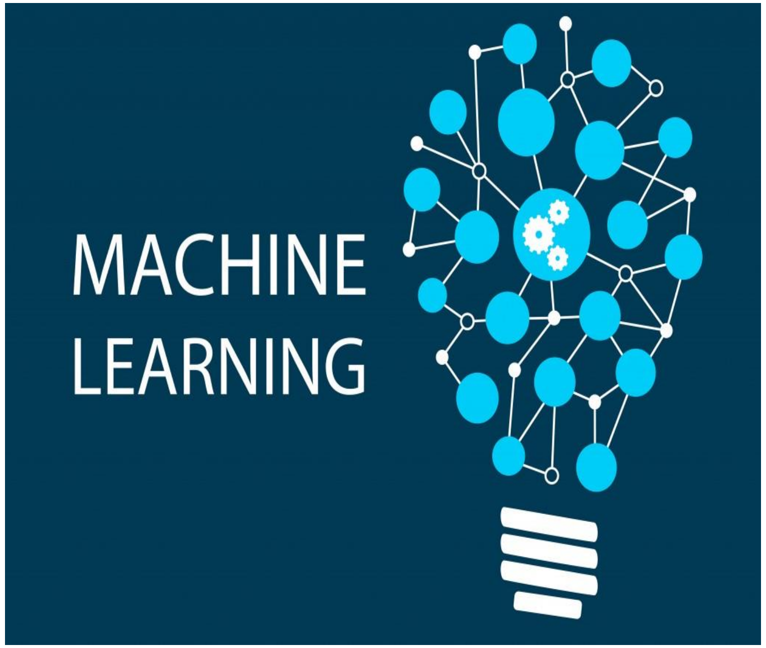
- It is back bone of data science or any analysis
- Helps an analyst to make sound business decisions
- Descriptive statistics helps us to understand the data and its properties by use of central tendency and variability
- Inferential statistics helps us to infer properties of the population from a given sample of data
- Knowledge of both descriptive and inferential statistics is essential for an aspiring data scientist or analyst.

What is Statistics

Statistics is a branch of mathematics dealing with the collection, analysis, interpretation, presentation, and organization of data



Data collection → Design experiments → Summarize information → Draw conclusions → Estimating the present → Help in Prediction



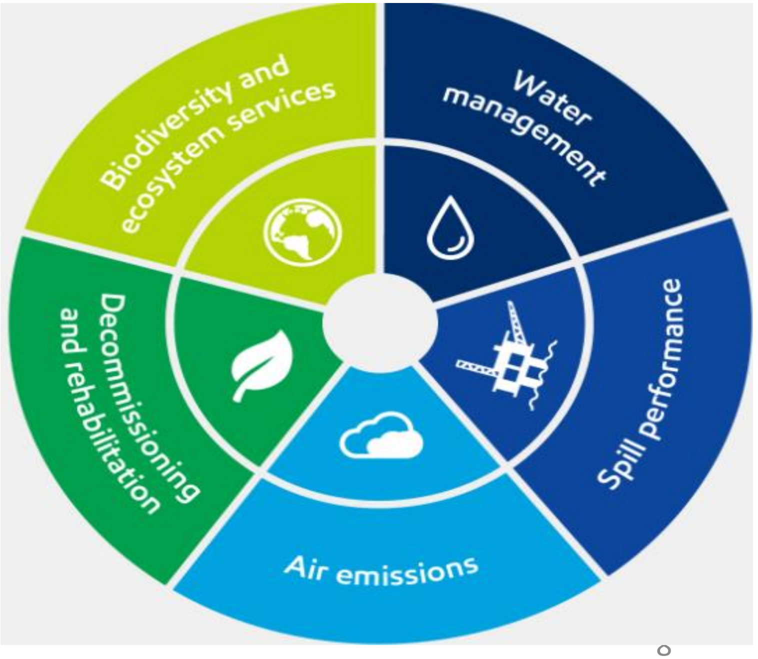
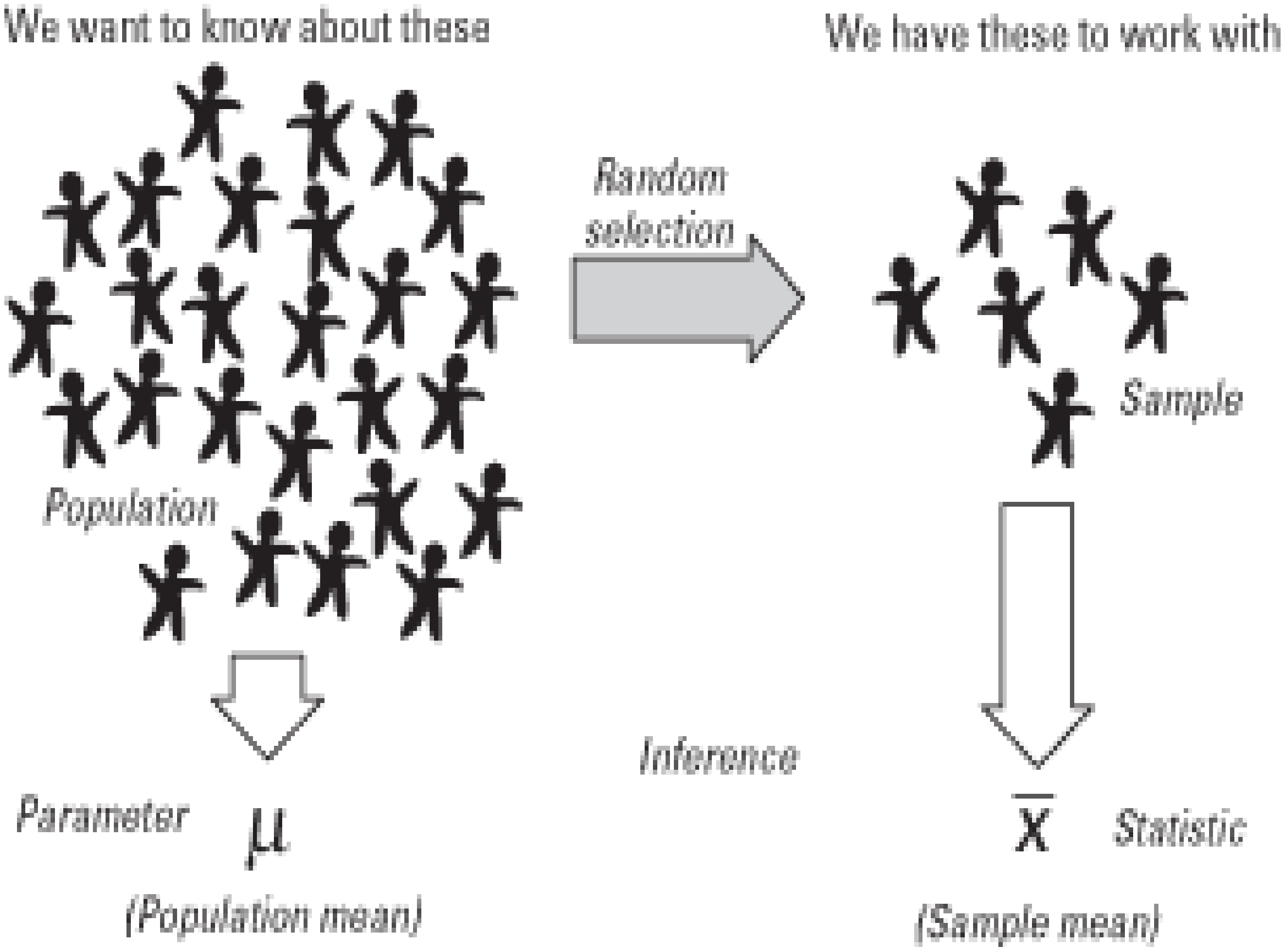
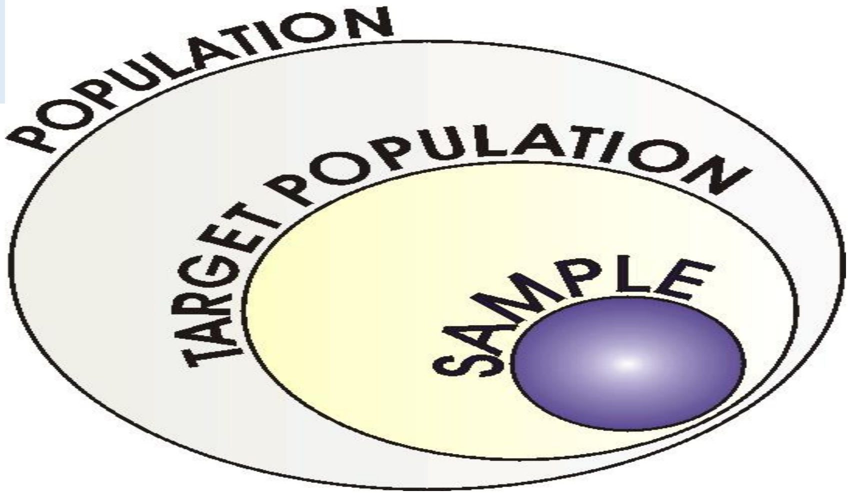

Biodiversity and ecosystem services
Water management
Spill performance
Air emissions
Decommissioning and rehabilitation

Image source link is at reference section

Population and Sample

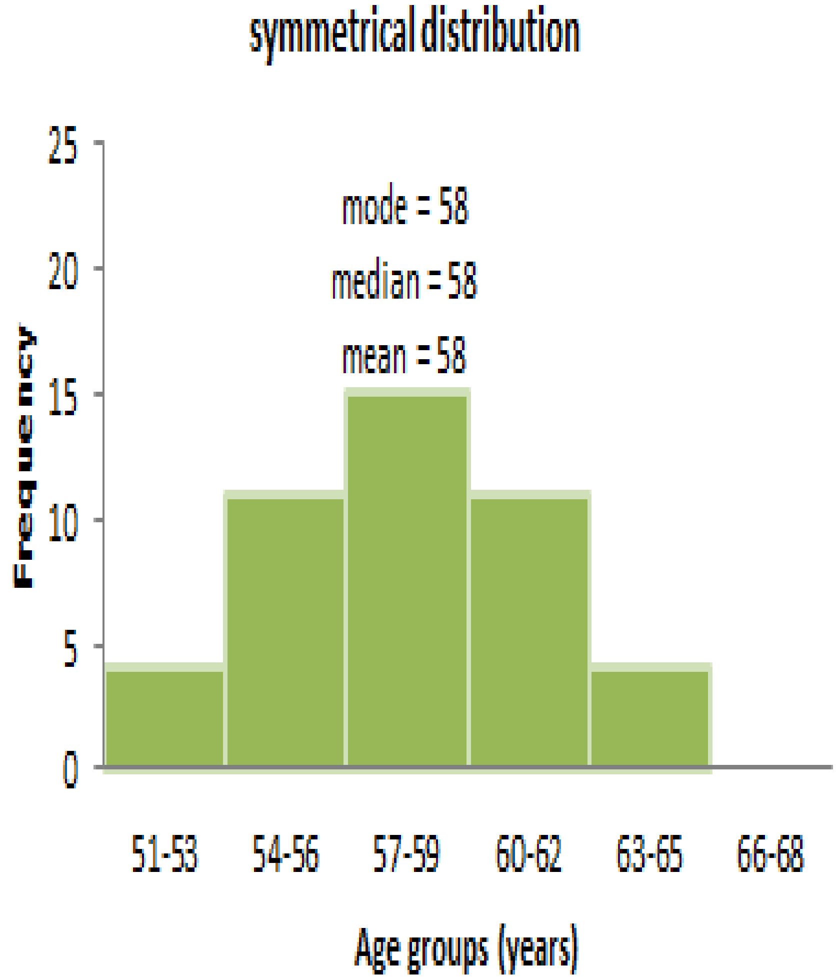


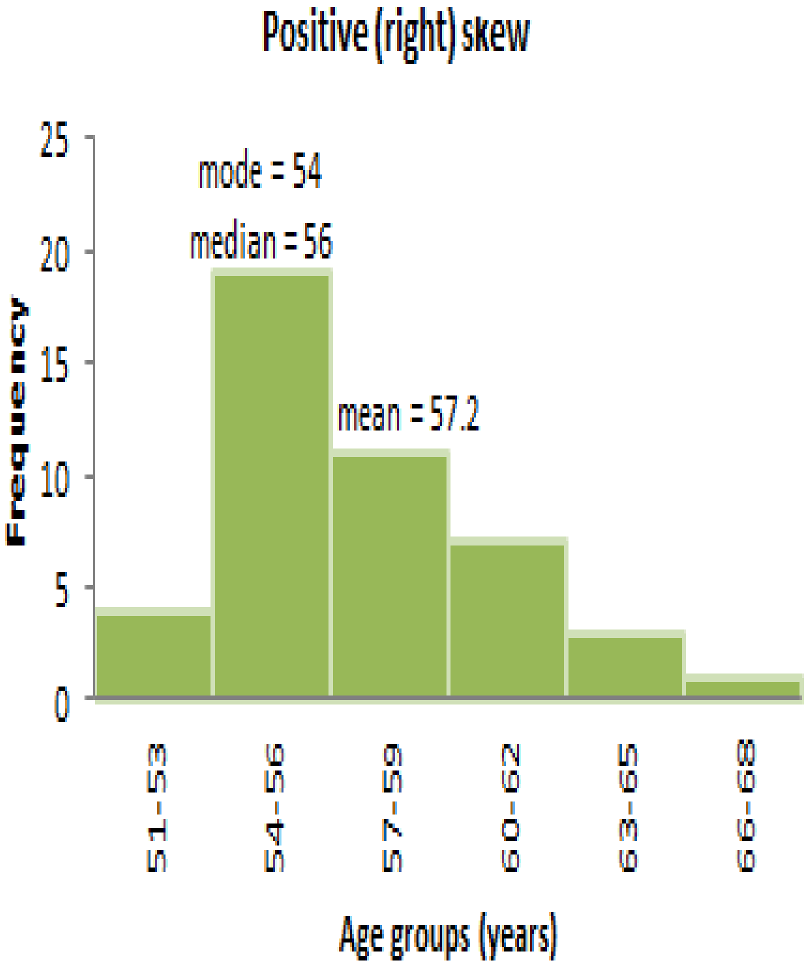


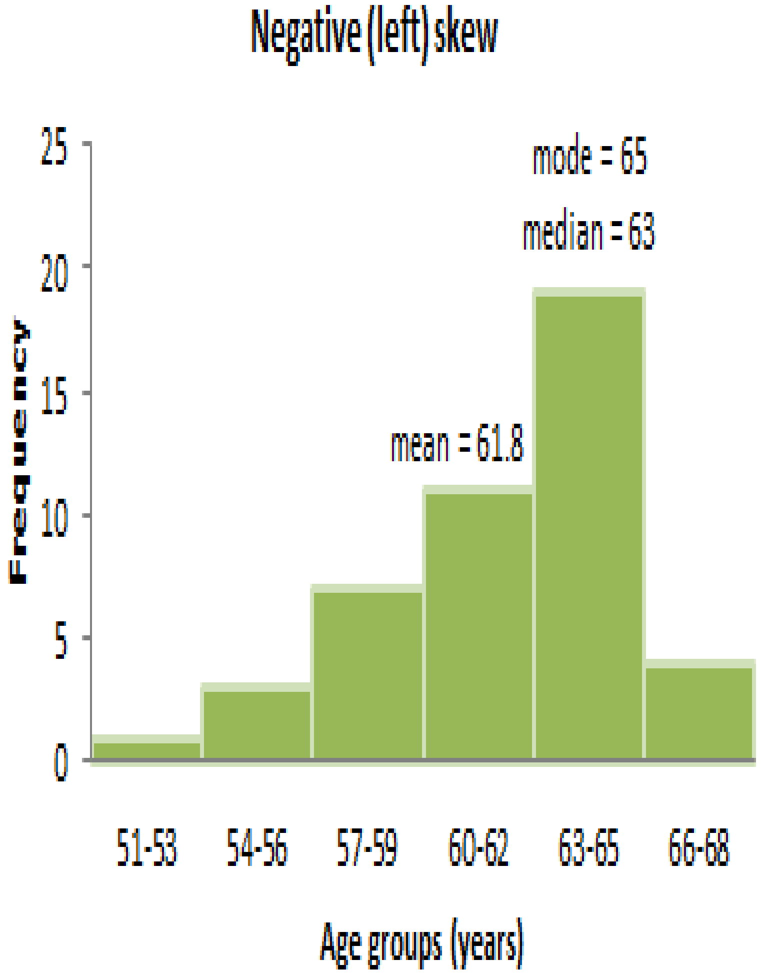
	Population **Parameters**	Sample **Statistics**
Mean	μ	$\overline{X}$
Standard deviation	σ	s
Variance	σ^2	s^2

https://brownmath.com/swt/symbol.htm

Measures of Central Tendency (Mean, Median, Mode)


symmetrical distribution

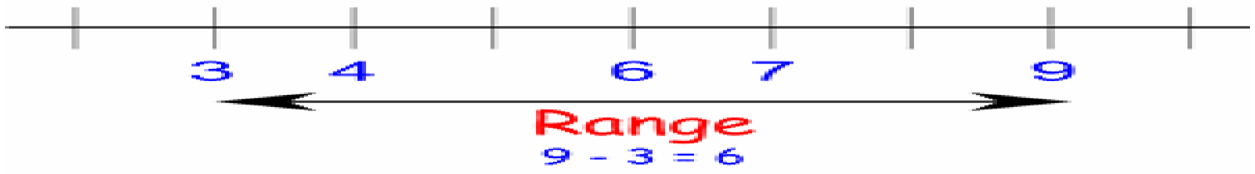

Positive (right) skew

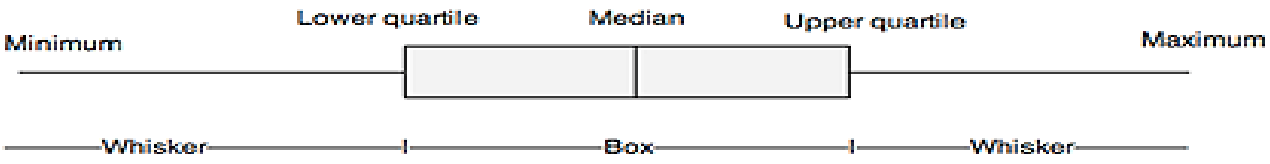

Negative (left) skew

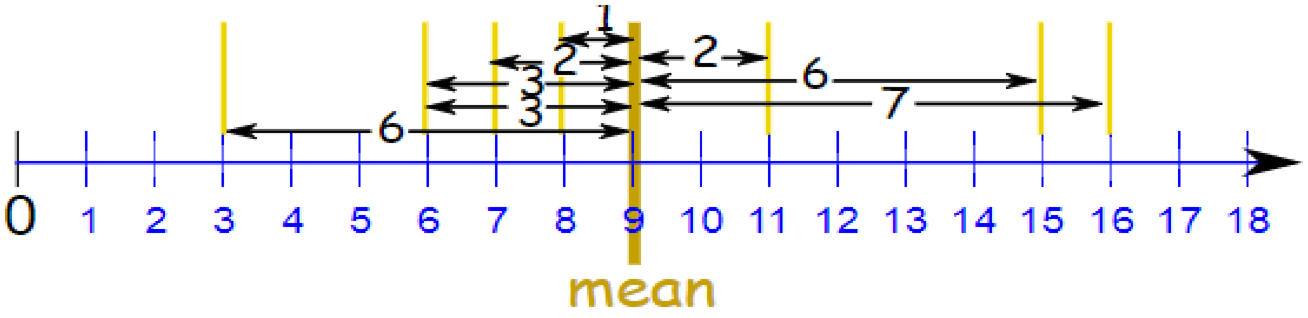
Important to detect outliers

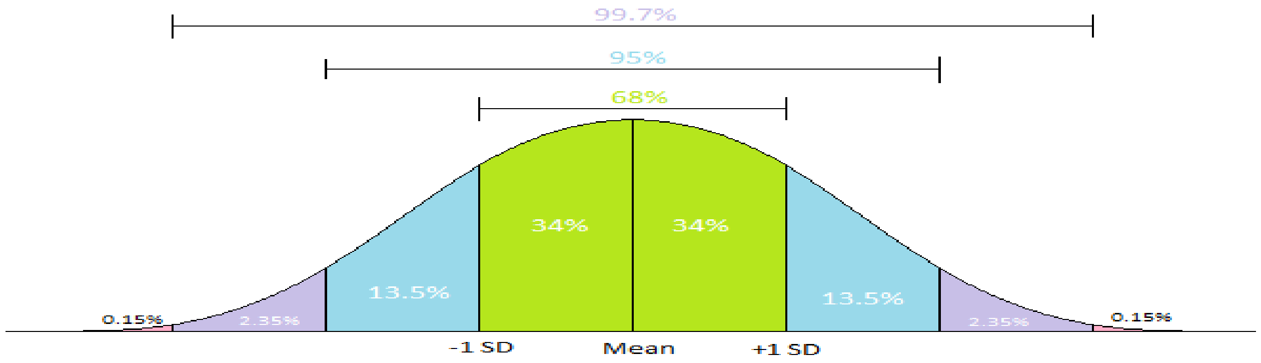
Measures of Dispersion

- **Range**: The difference between the smallest value and the largest value of a series.
- **Quartile deviation**: Take into account 'Upper quartile (Q3)' and the 'Lower quartile' (Q1). Also called 'inter-quartile range'.
- **Mean deviation**: The average of the deviations of various items from a measure of central tendency Mean or Median (default) or Mode, ignoring negative signs.
- **Standard deviation**: The square root of average of squared deviations taken from actual mean.









Practical example of 1 sigma to 6 sigma (https://en.wikipedia.org/wiki/List_of_Six_Sigma_companies)

Measure of Shape (Skewness and Kurtosis)

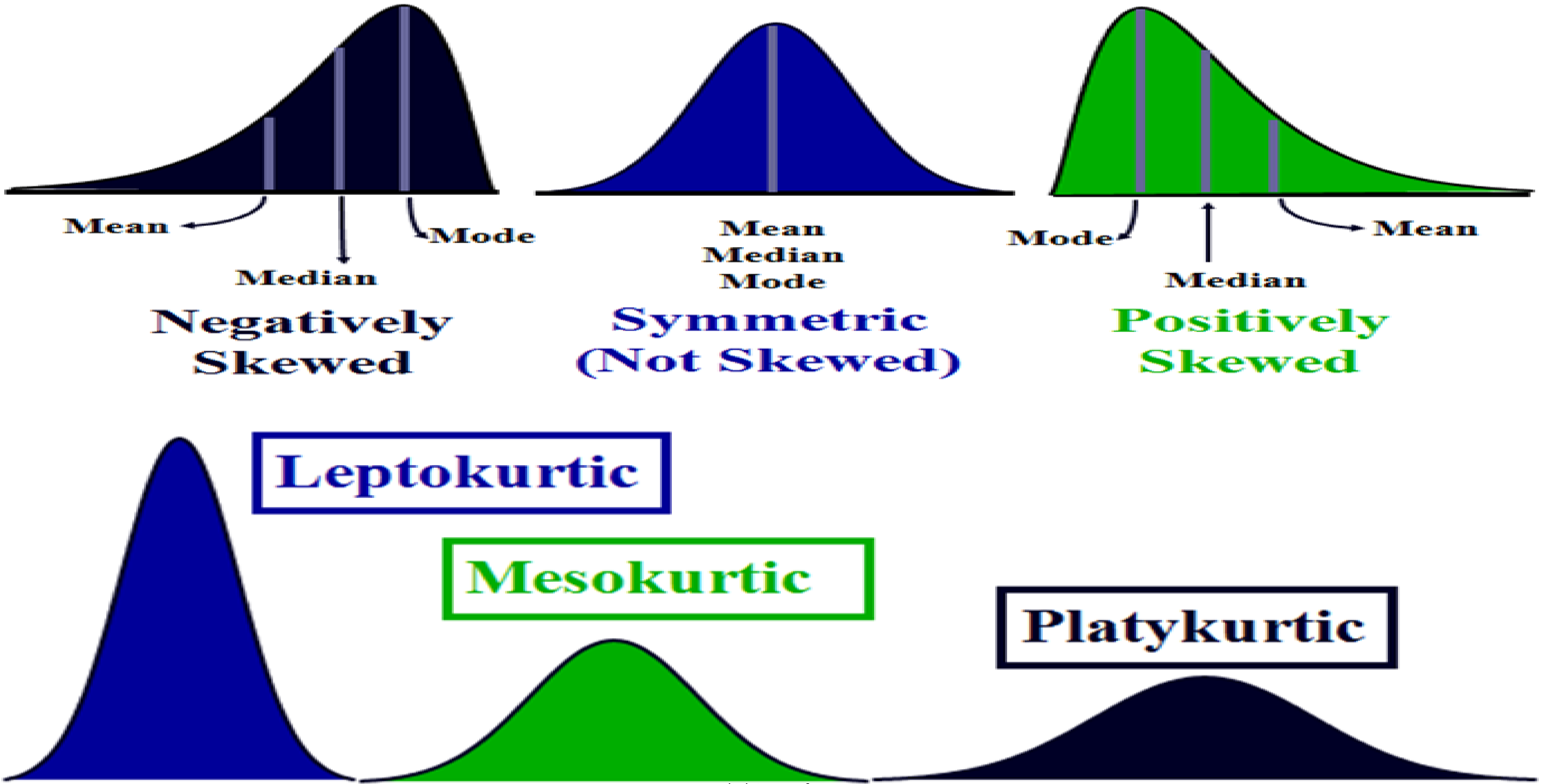


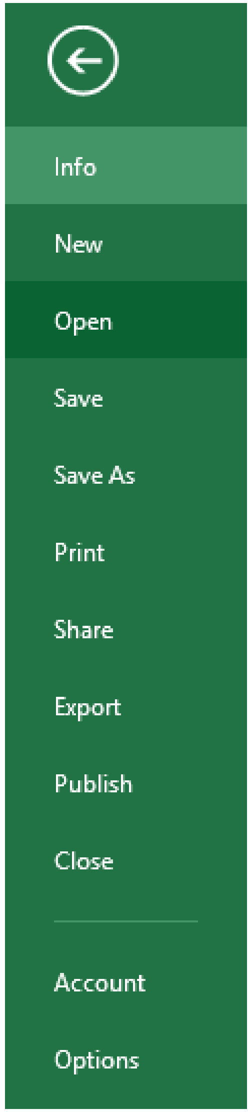
Image source link is at reference section

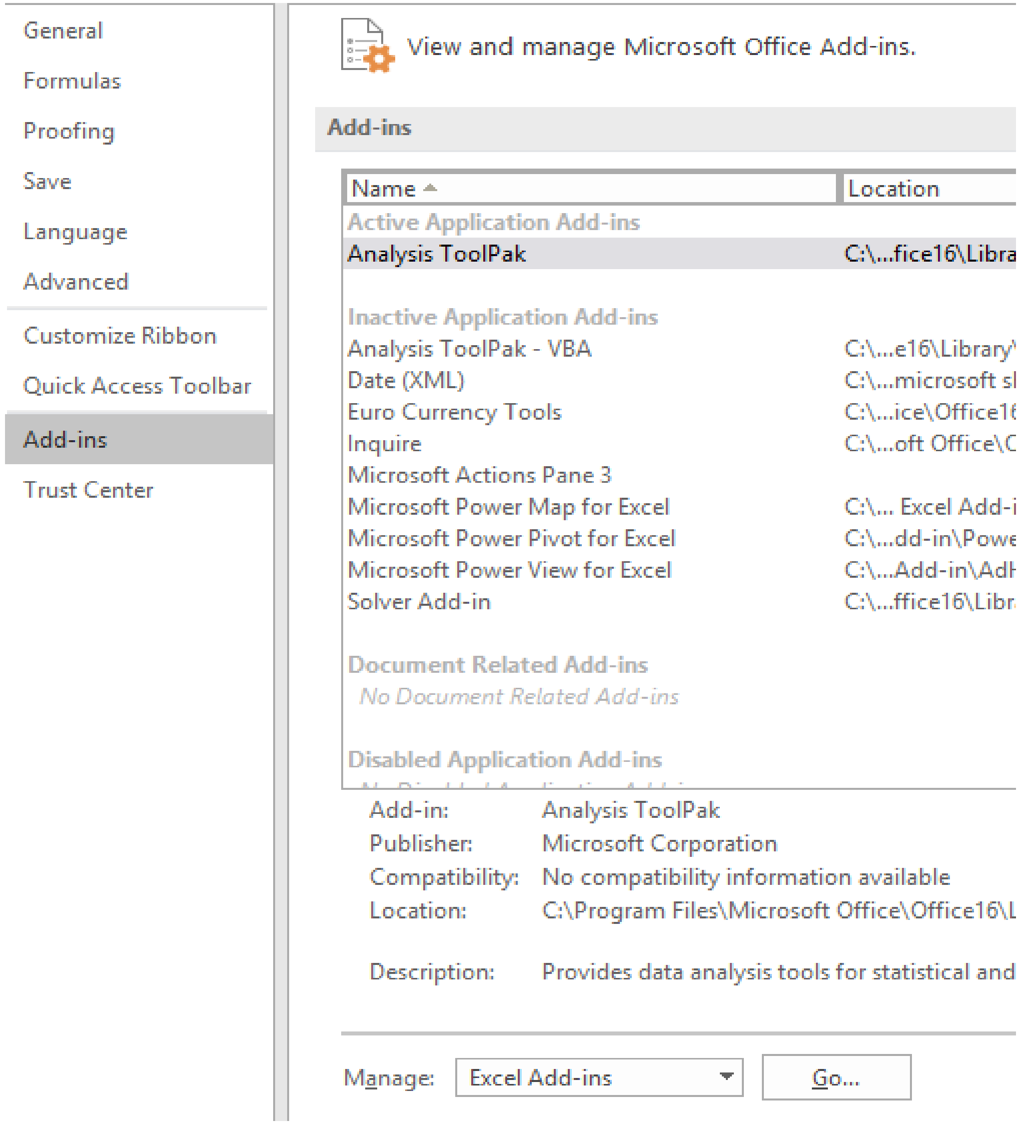
Hands On

How to add Data Analysis tab -> Next Slide

How to add Data Analysis tab

File -> Options -> Add-Ins -> Manage Excel Add Ins -> Go -> Select Analysis tool Pak





View and manage Microsoft Office Add-ins.

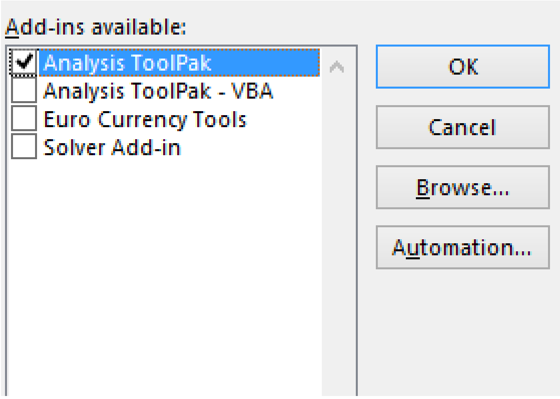
Add-ins

Name ▲	Location
Active Application Add-ins	
Analysis ToolPak	C:\...fice16\Libra
Inactive Application Add-ins	
Analysis ToolPak - VBA	C:\...e16\Library\
Date (XML)	C:\...microsoft sl
Euro Currency Tools	C:\...ice\Office16
Inquire	C:\...oft Office\C
Microsoft Actions Pane 3	
Microsoft Power Map for Excel	C:\... Excel Add-i
Microsoft Power Pivot for Excel	C:\...dd-in\Powe
Microsoft Power View for Excel	C:\...Add-in\Adh
Solver Add-in	C:\...ffice16\Libr
Document Related Add-ins	
No Document Related Add-ins	
Disabled Application Add-ins	

Add-in: Analysis ToolPak
Publisher: Microsoft Corporation
Compatibility: No compatibility information available
Location: C:\Program Files\Microsoft Office\Office16\L

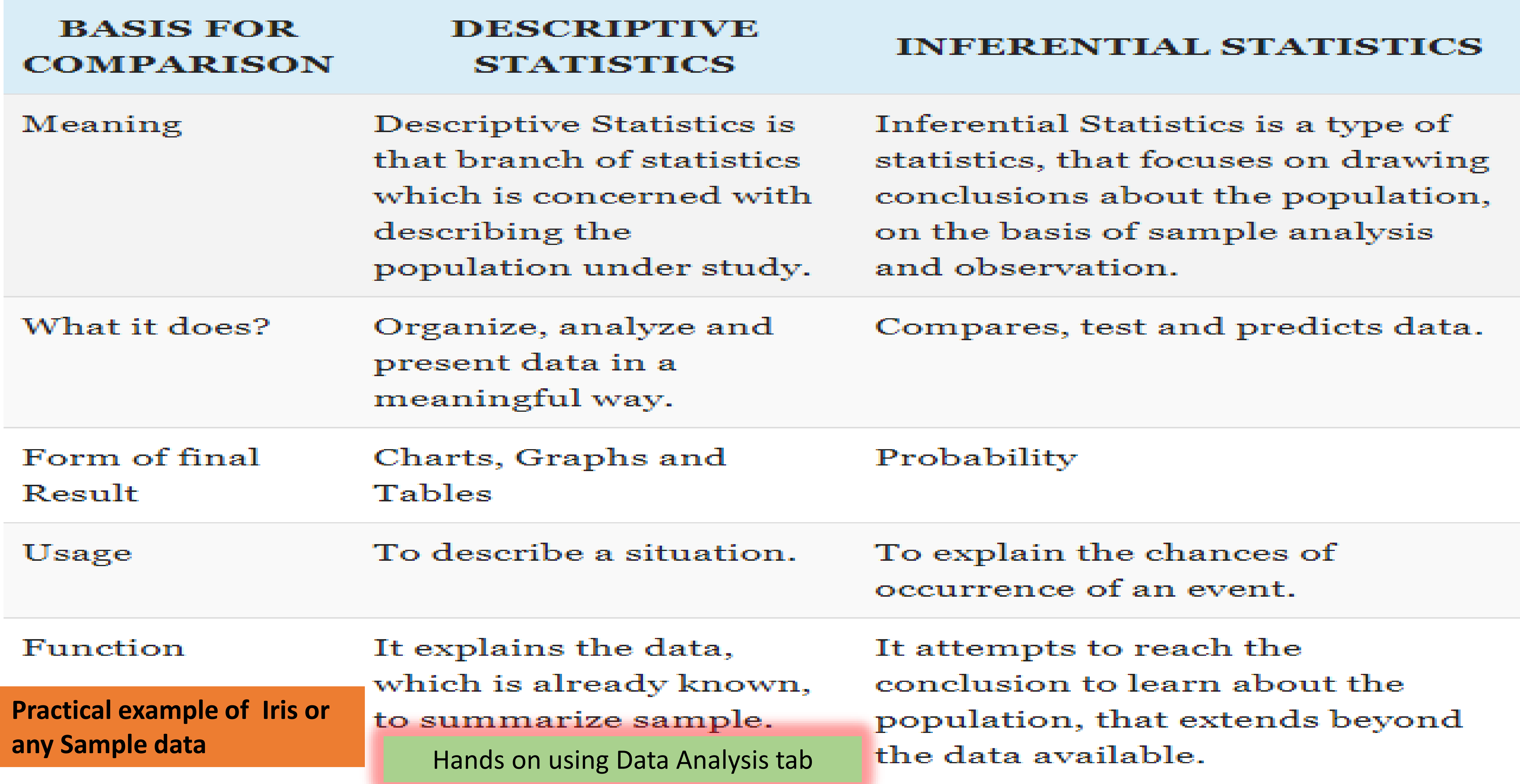
Description: Provides data analysis tools for statistical and

Manage: Excel Add-ins


Add-ins available:

- ☑ Analysis ToolPak
- □ Analysis ToolPak - VBA
- □ Euro Currency Tools
- □ Solver Add-in

Descriptive and Inferential Statistics



BASIS FOR COMPARISON	DESCRIPTIVE STATISTICS	INFERENTIAL STATISTICS
Meaning	Descriptive Statistics is that branch of statistics which is concerned with describing the population under study.	Inferential Statistics is a type of statistics, that focuses on drawing conclusions about the population, on the basis of sample analysis and observation.
What it does?	Organize, analyze and present data in a meaningful way.	Compares, test and predicts data.
Form of final Result	Charts, Graphs and Tables	Probability
Usage	To describe a situation.	To explain the chances of occurrence of an event.
Function	It explains the data, which is already known, to summarize sample.	It attempts to reach the conclusion to learn about the population, that extends beyond the data available.

Practical example of Iris or any Sample data

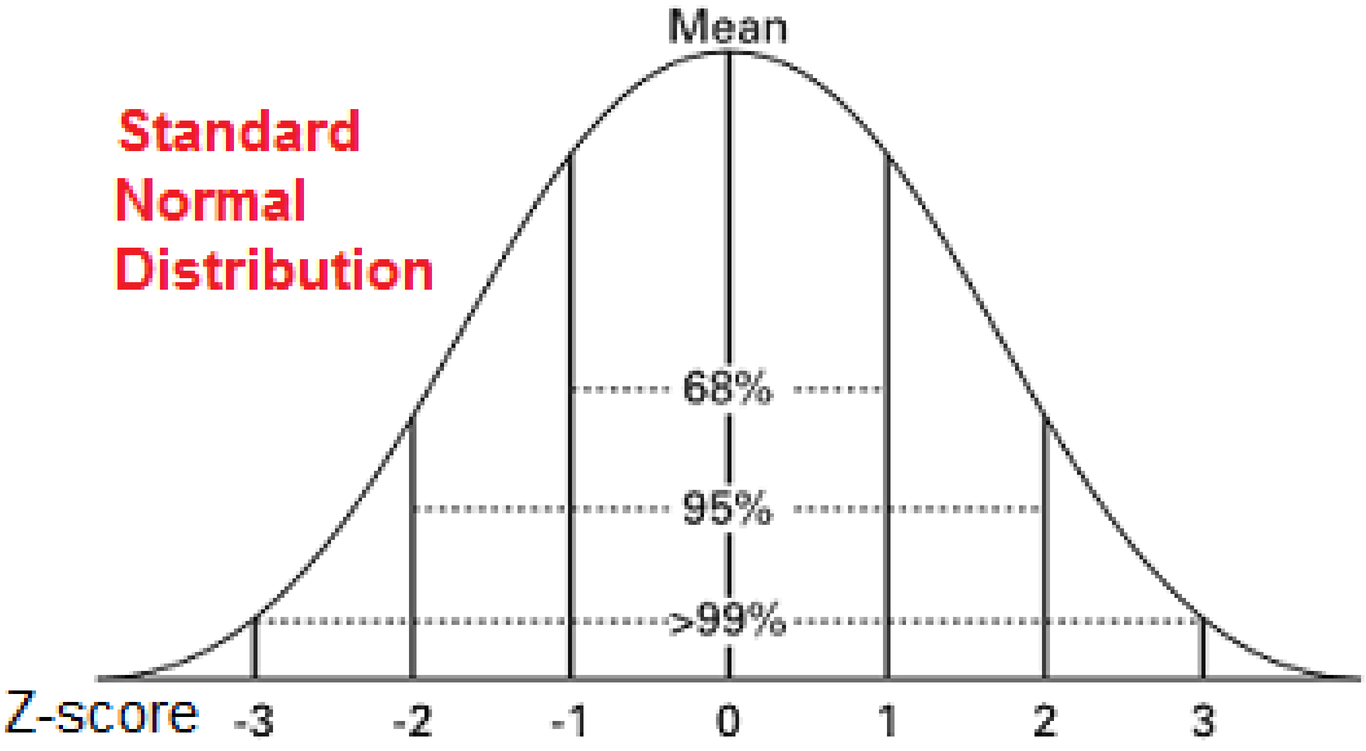
Hands on using Data Analysis tab

Type of Data/Variables

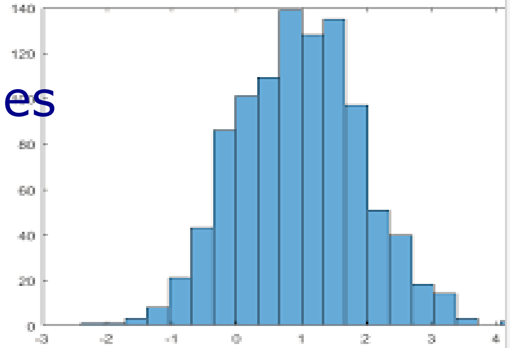
- Numeric data
  - Continuous (measurements)
    - Ratio (interval + clear definition of 0)
    - Interval (difference between two values is meaningful)
  - Discrete (counts)
- Categorical
  - Nominal
  - Ordinal (Likert scale)
  - Dichotomous
- Independent Variables (experimental or predictor)
- Dependent (outcome)

	Nominal	Ordinal	Interval	Ratio
Frequency distribution	Yes	Yes	Yes	Yes
Median and percentiles	No	Yes	Yes	Yes
Add or subtract	No	No	Yes	Yes
mean, std	No	No	Yes	Yes
Ratio	No	No	No	Yes

Normal



- Heights of people
- Size of things produced by machines
- Errors in measurements
- Blood pressure
- Marks on a test

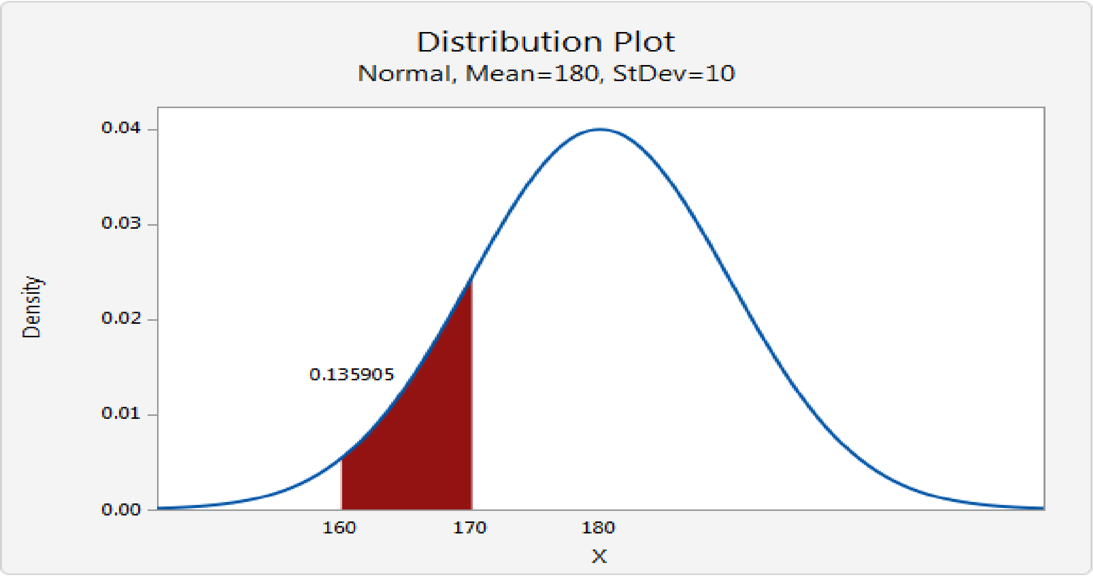


Live

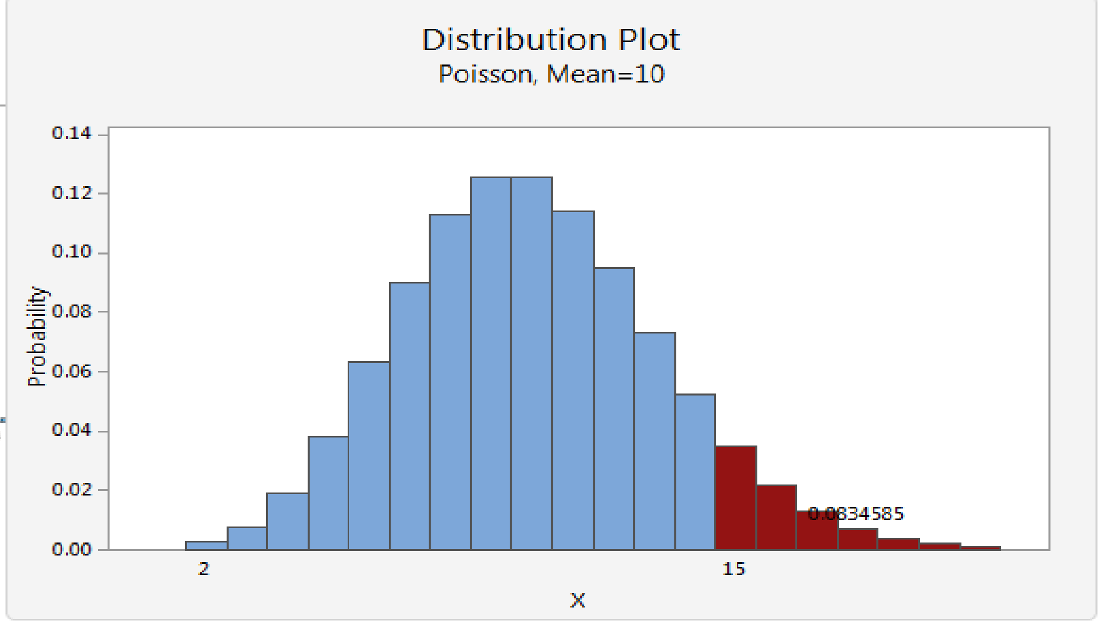
https://www.mathsisfun.com/data/quincunx.html

http://onlinestatbook.com/2/calculators/normal_dist.html

Continuous distribution



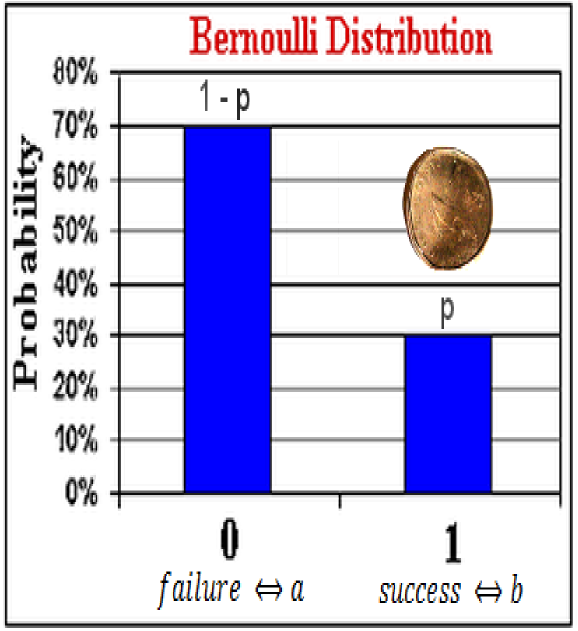
Discrete distribution



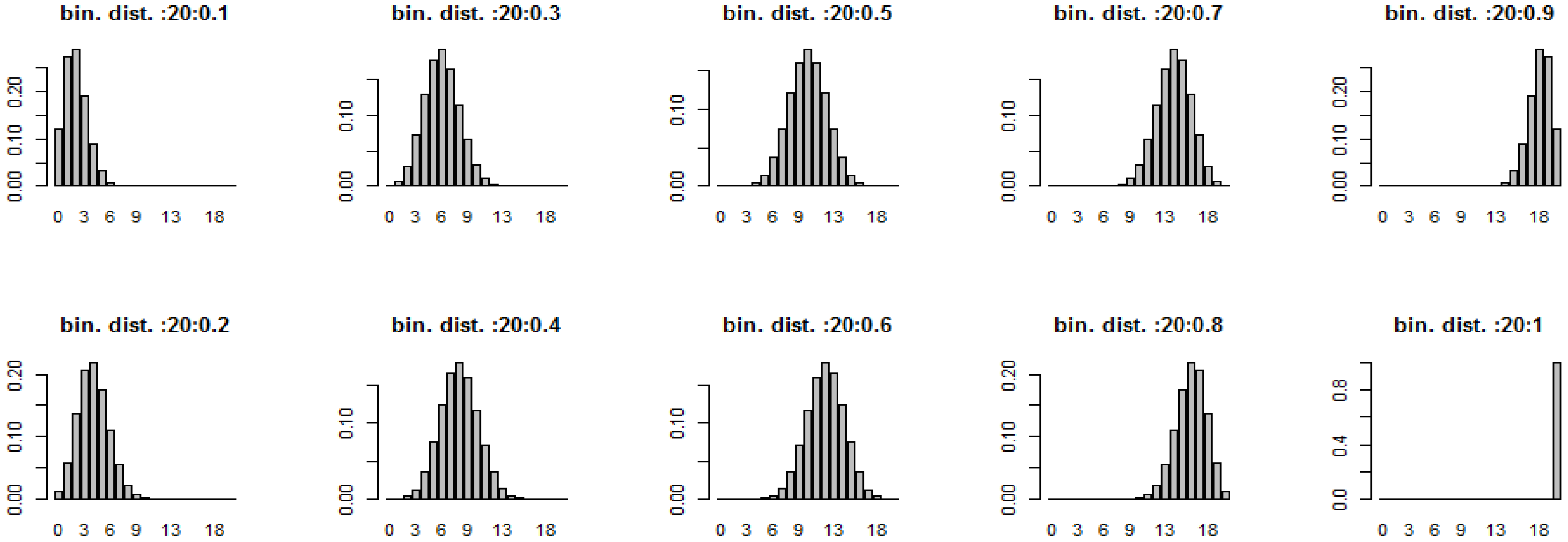
Examples of Standard Normal Distribution will be taken after Hypothesis Testing slide

Various Distributions cont

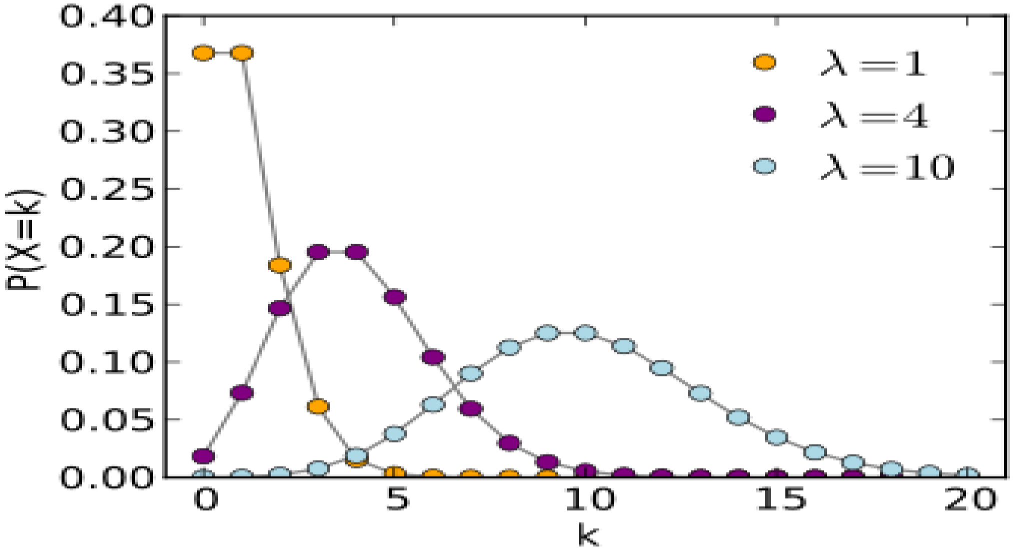
Bernoulli



Binomial

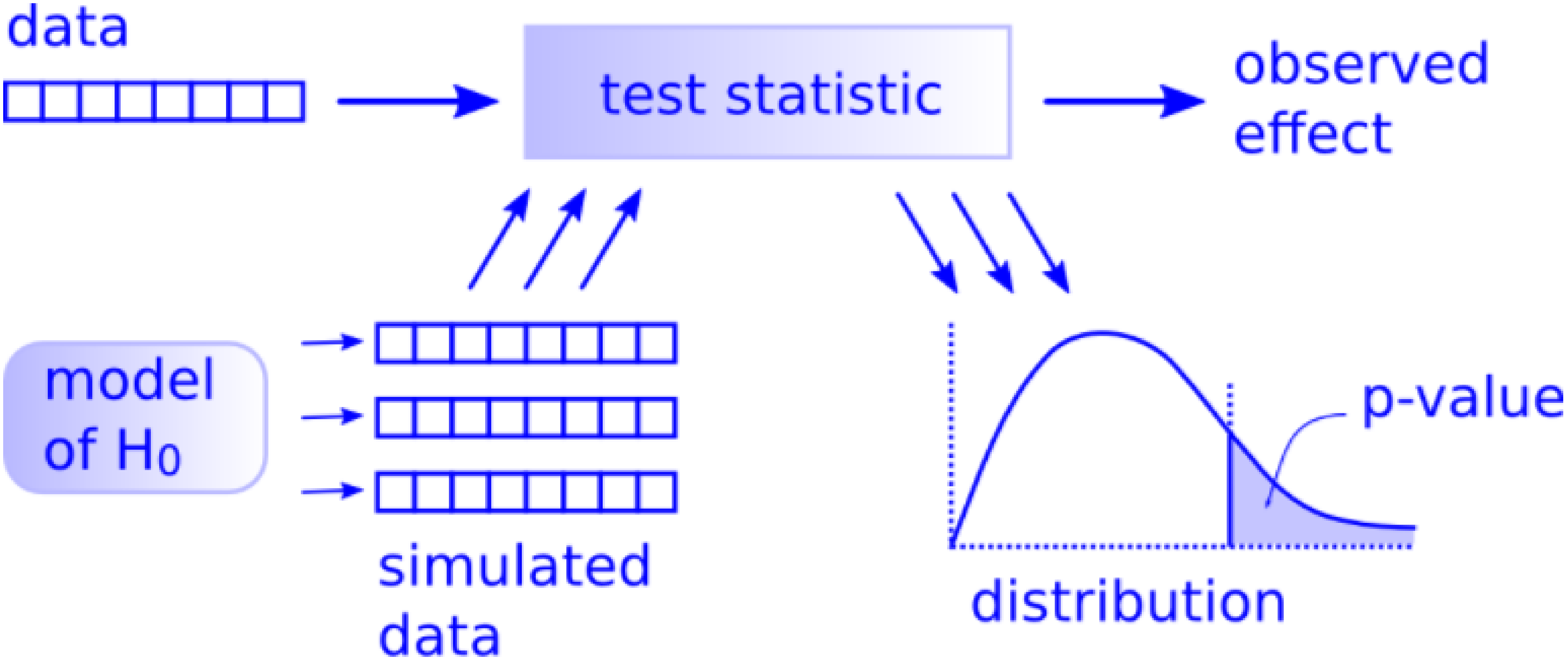


Poisson: A discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time if these events occur with a known average rate and independently of the time since the last event



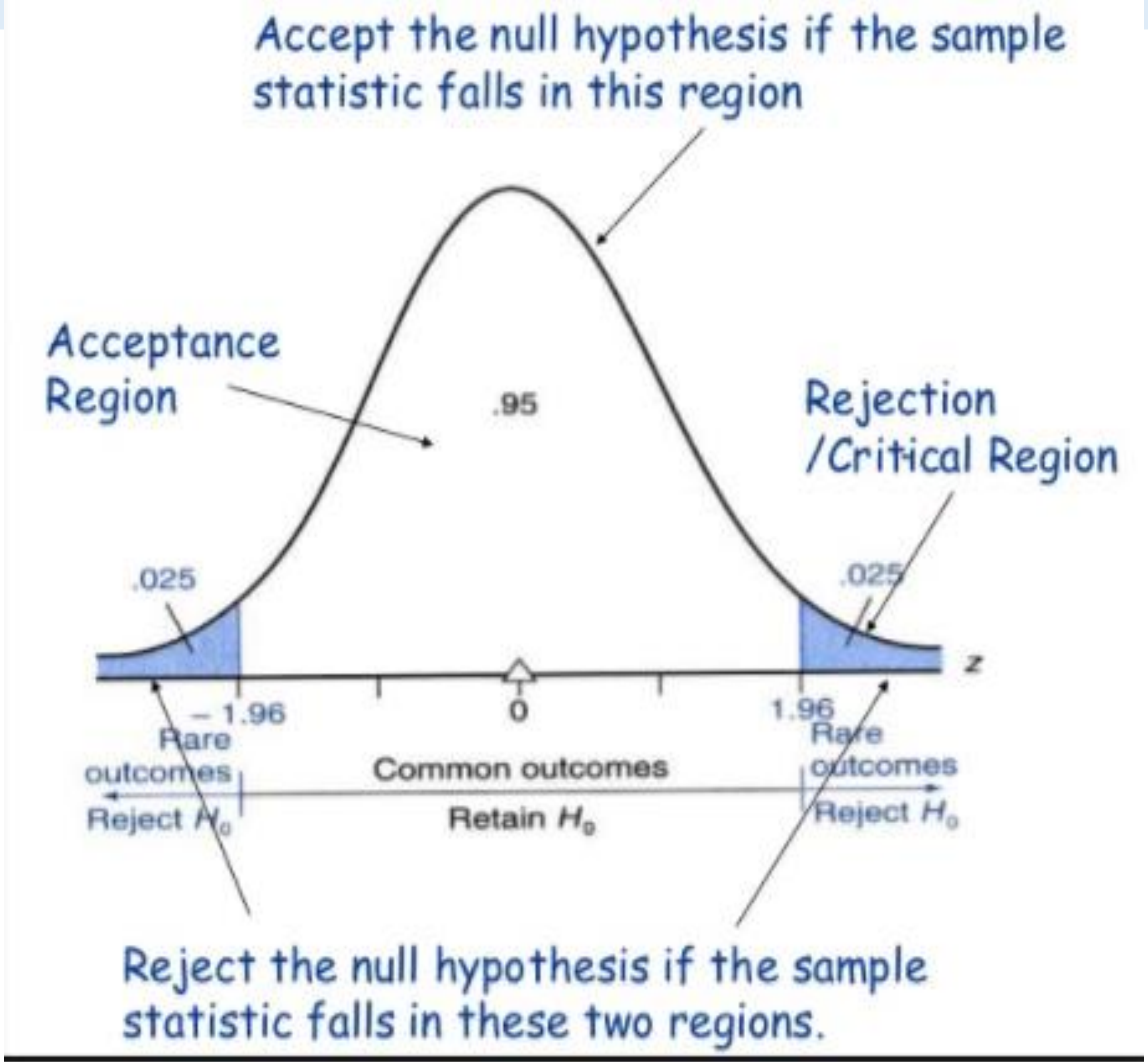
The horizontal axis is the index *k*, the number of occurrences. *λ* is the expected number of occurrences. The vertical axis is the probability of *k* occurrences given *λ*.

Hypothesis testing

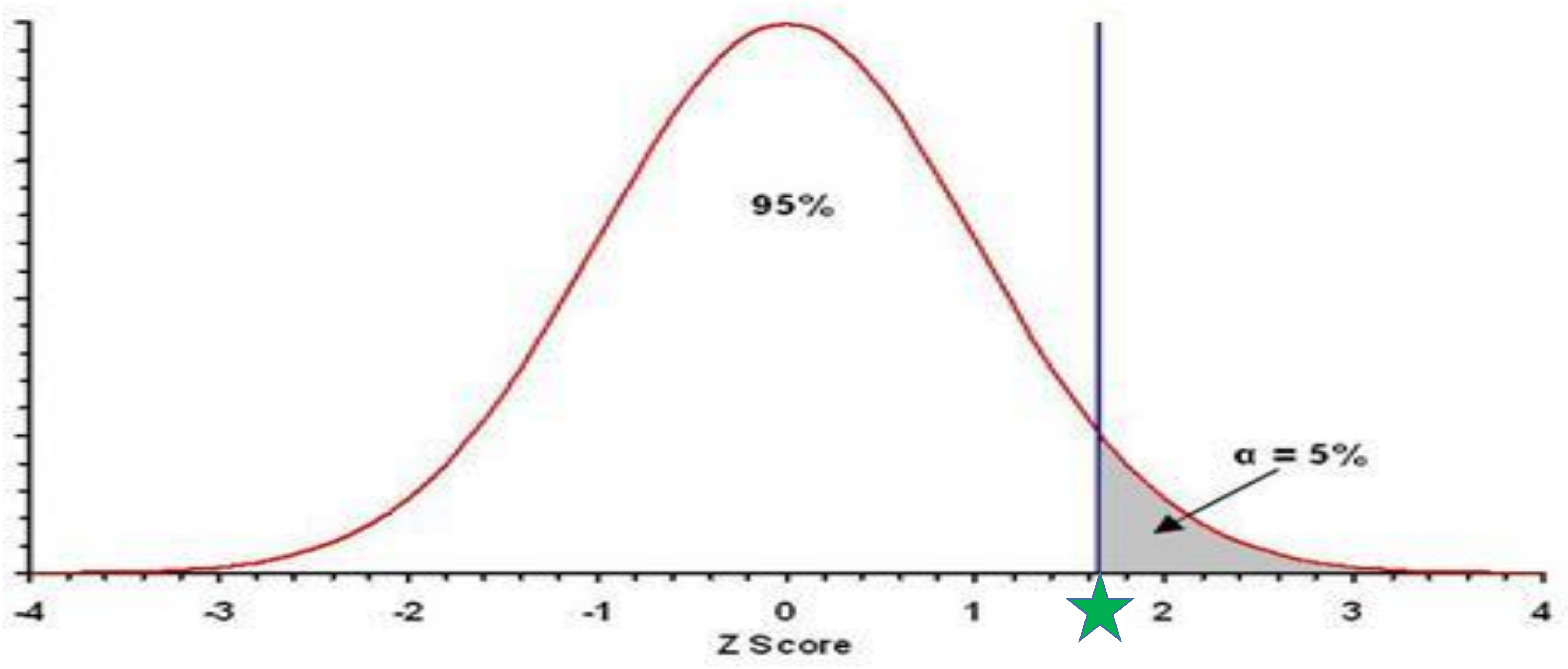


Critical Values

- What is a Critical Value?
- Critical Value of Z
- Critical Value in Any tail
- Critical value for a confidence level
- Common confidence levels and their critical values.
- Critical Value: Two-Tailed Test.
- Critical Value: Right-Tailed Test.
- Critical Value: Left-Tailed Test.
- What does Significance Testing Tell


Accept the null hypothesis if the sample statistic falls in this region
Acceptance Region
.95
Rejection /Critical Region
.025
.025
z
− 1.96
0
1.96
Rare outcomes
Reject H_0
Common outcomes
Retain H_0
Rare outcomes
Reject H_0
Reject the null hypothesis if the sample statistic falls in these two regions.

CW: Find Value of z for one tail at 95% confidence



Confusion matrix & Types of Error: Alpha and Beta

		Predicted	
		good	bad
Actual	good	TP	FN
	bad	FP	TN

2, β

1, α

Power of Test: 1-β

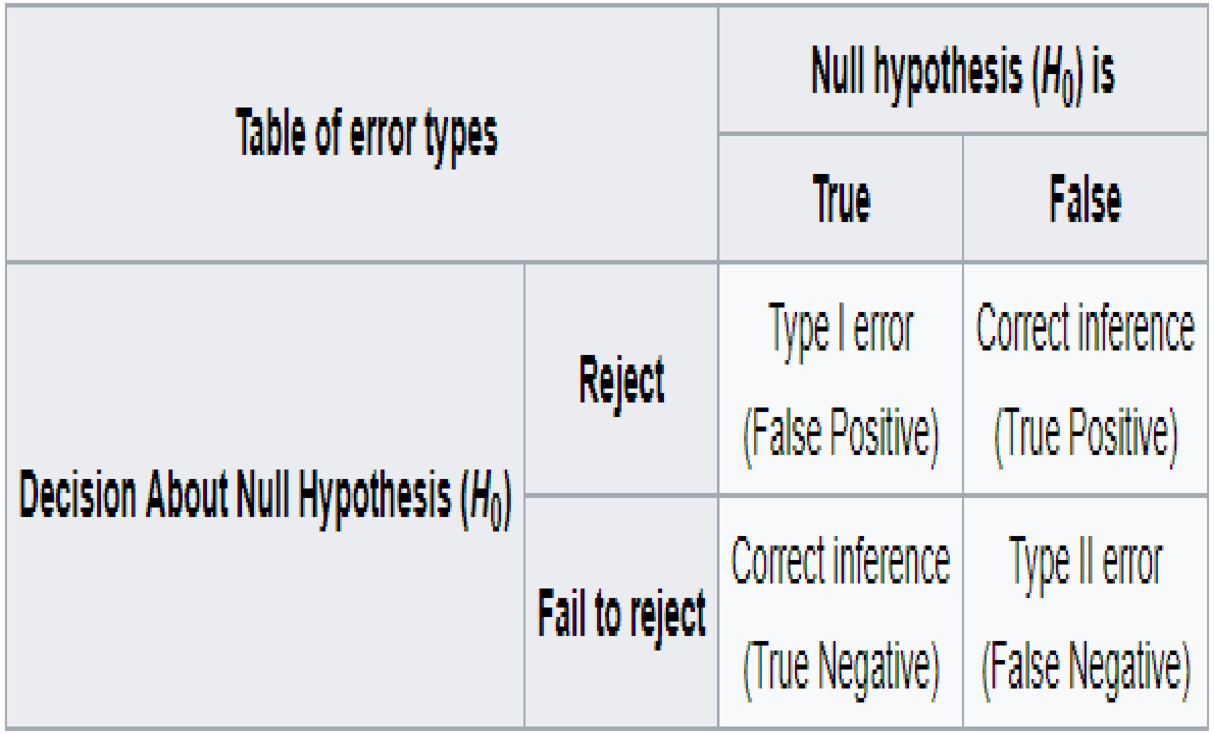


Table of error types		Null hypothesis (H_0) is	
		True	False
Decision About Null Hypothesis (H_0)	Reject	Type I error (False Positive)	Correct inference (True Positive)
	Fail to reject	Correct inference (True Negative)	Type II error (False Negative)

Examples

Example: Standard Normal Distribution $z = (x - \mu) / (\sigma / \sqrt{n})$

- A medical doctor wants to reduce blood sugar level of all his patients by altering their diet. He finds that the mean sugar level of all patients is 180 with a standard deviation of 18. Nine of his patients start dieting and the mean of the sample is observed to 175. Now, he is considering to recommend all his patients to go on a diet.
- What is the probability of getting a mean of 175 or less after all the patients start dieting?

CW

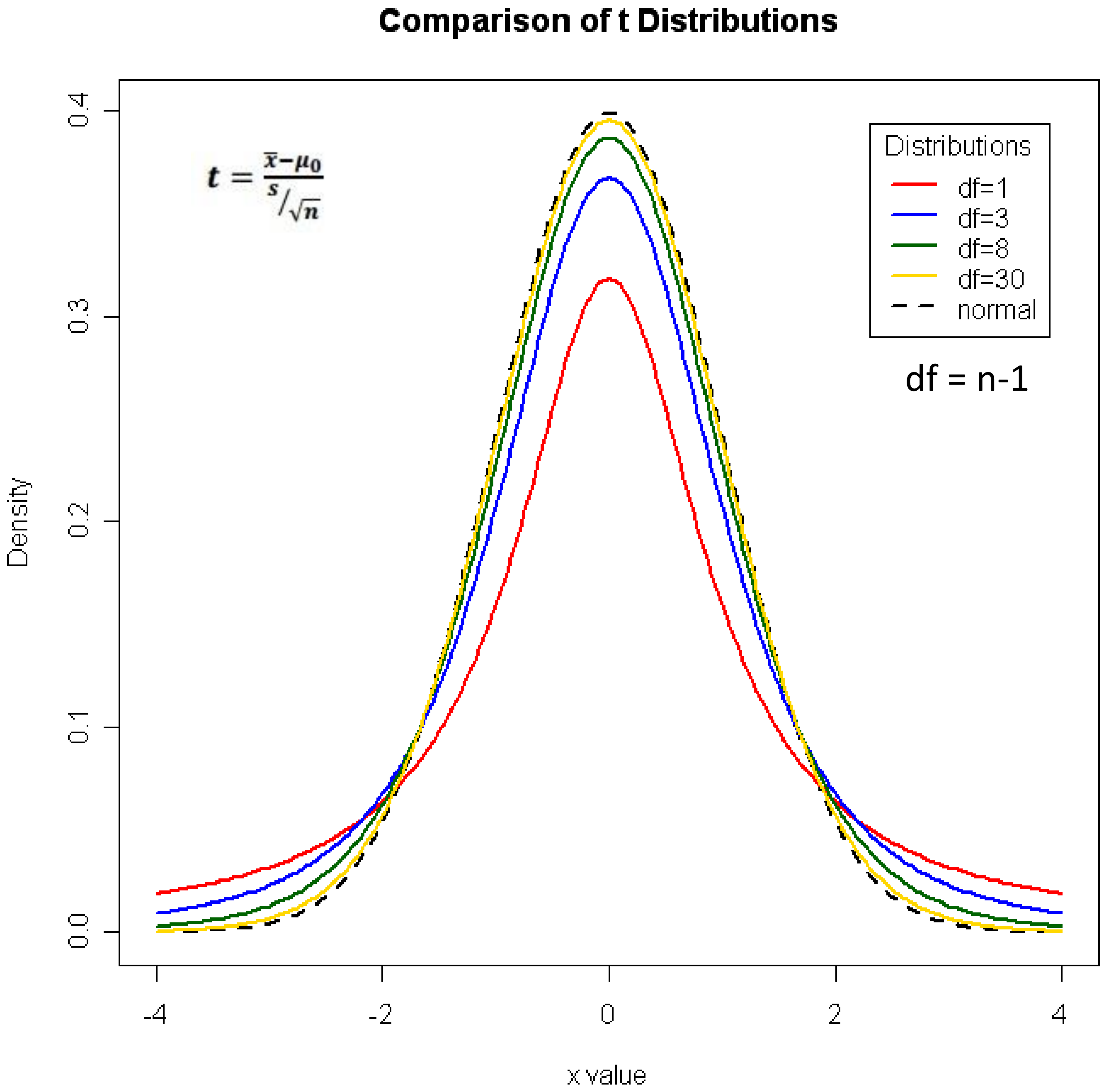
- Studies show that listening to music while studying can improve your memory. To demonstrate this, a researcher obtains a sample of 36 college students and gives them a standard memory test while they listen to some background music. Under normal circumstances (without music), the mean score obtained was 25 and standard deviation is 6. The mean score for the sample after the experiment (i.e With music) is 28.
- After performing the Z-test, what can we conclude

source link is at reference section

Student's t Distribution

It is used to estimate population parameters when the sample size is small and/or when the population variance is unknown.

- Paired Two Sample for means
- Two Sample t test assuming Equal Variances
- Two-sample t test assuming Unequal variances



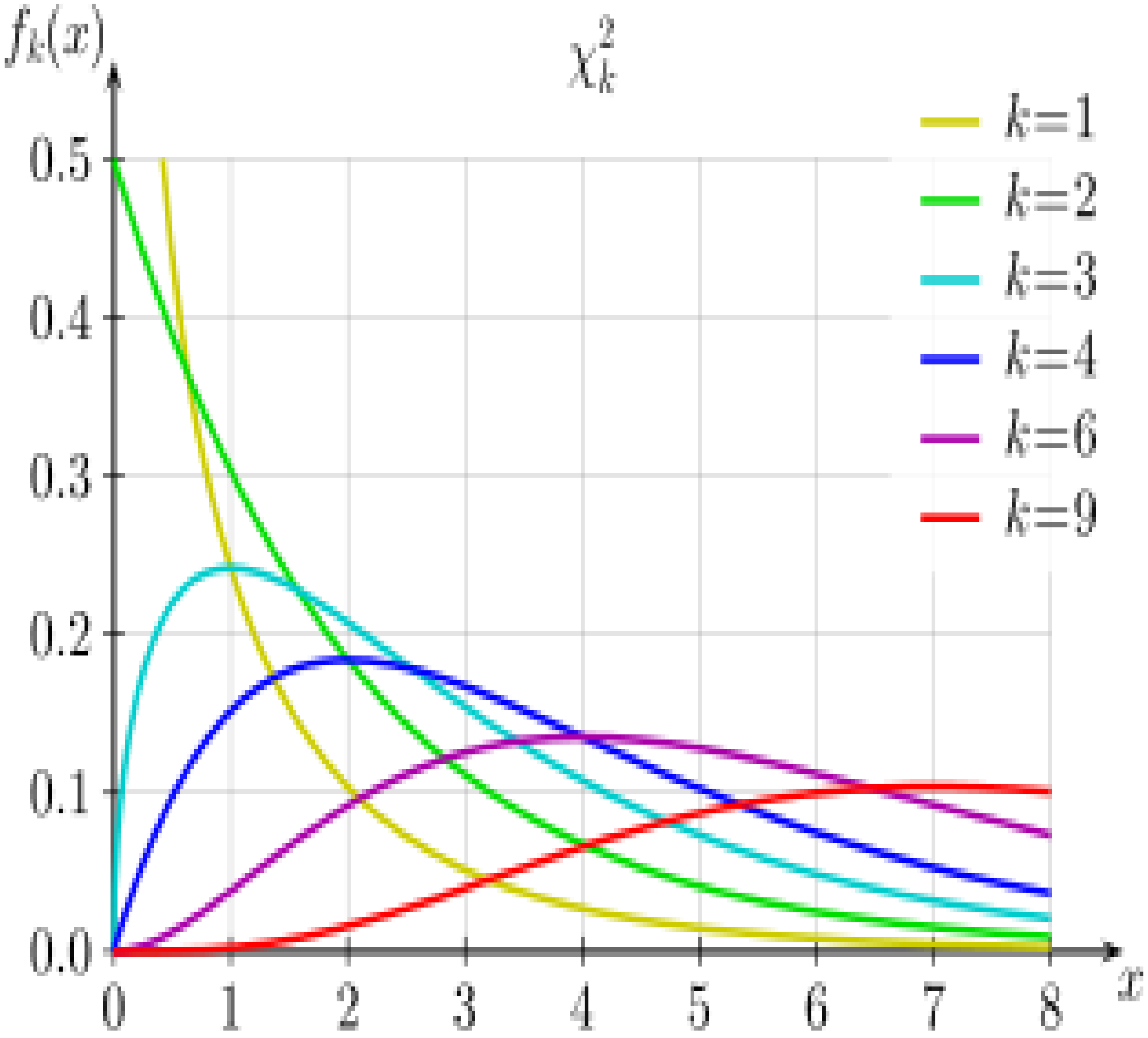
Hands On

Chi-square goodness of fit

A Chi-square goodness of fit test determines if a sample data matches a population.

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

- **Null hypothesis:** There is no significant difference between the observed and the expected value.
- **Alternative hypothesis:** There is a significant difference between the observed and the expected value.



Chi-square test for independence

- It compares two variables in a contingency table to see if they are related. (distributions of categorical variables differ from each another)
  - H0: The two categorical variables are independent (No association).
  - H1: The two categorical variables are dependent (there is an association).
- Calculations
  - Formula: $\chi^2 = \sum (O - E)^2 / E$
  - Where E= (row total × column total)/ sample size
  - Compare the value of the test statistic to the critical value of χ2α with degree of freedom = (r - 1) (c - 1), and reject the null hypothesis if χ2>χ2α.

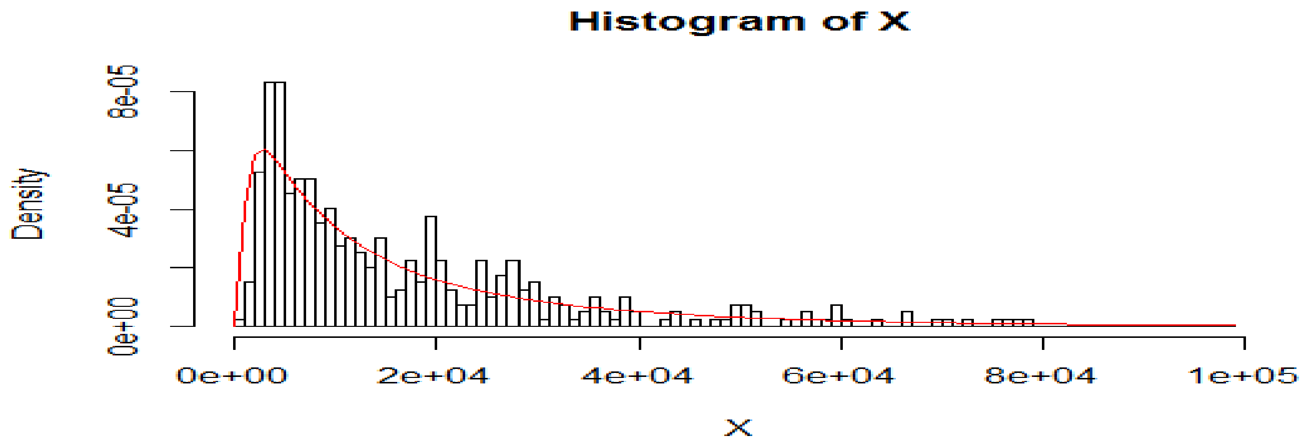
Chi-square Uses

- Independence of two criteria of classification of qualitative variables.
- Relationships between categorical variables (contingency tables).
- Tests of deviations of differences between expected and observed frequencies (one-way tables).
- The chi-square test (a goodness of fit test).
- Confidence interval estimation for a population by standard deviation of a normal distribution from a sample standard deviation.
- Sample variance study when the underlying distribution is normal.

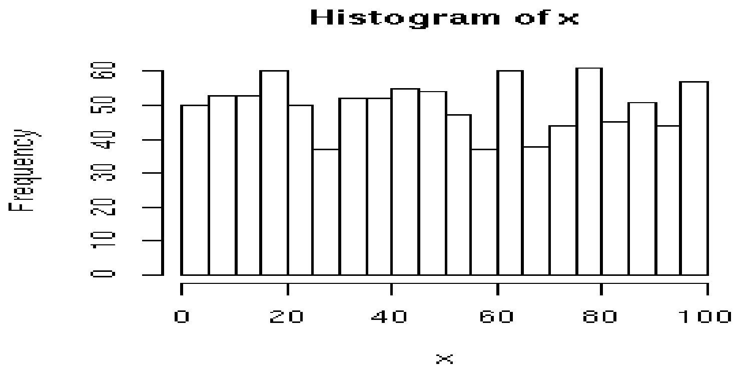
Hands On

Various Distributions – Some more

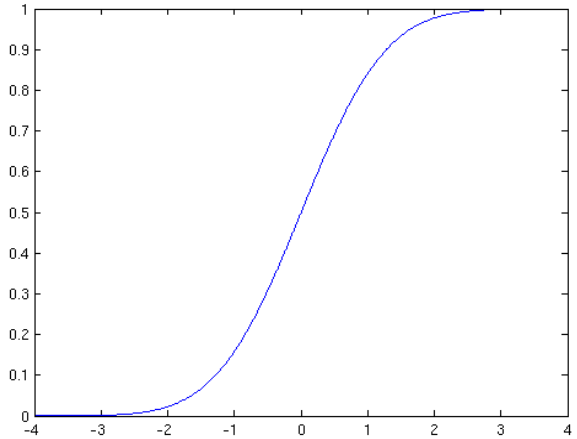
Tweedie



Uniform



Sigma



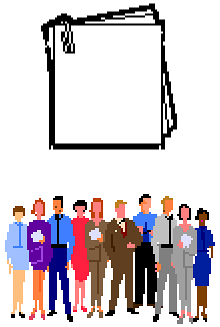
Sampling Procedure: **Probability Sampling**

A method of sampling that utilizes some form of *random selection*

Simple Random Sampling

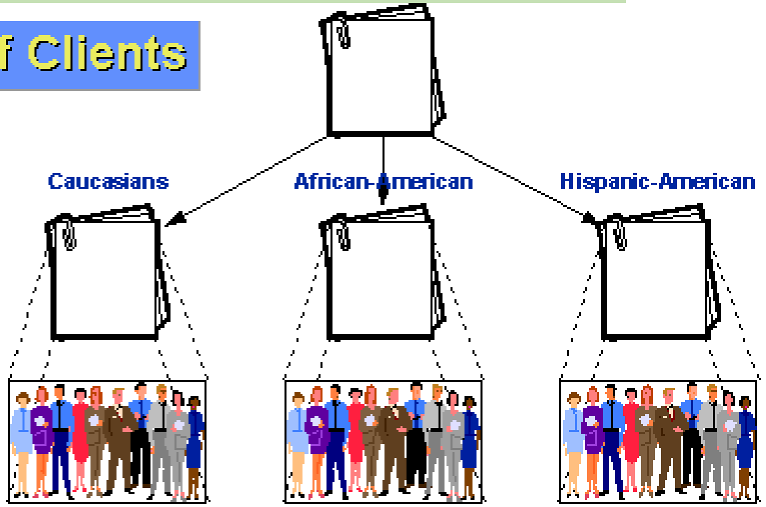
List of Clients

Random Subsample



Stratified Random Sampling

List of Clients



Strata

Random Subsamples of n/N

Systematic Random Sampling

N = 100

want n = 20

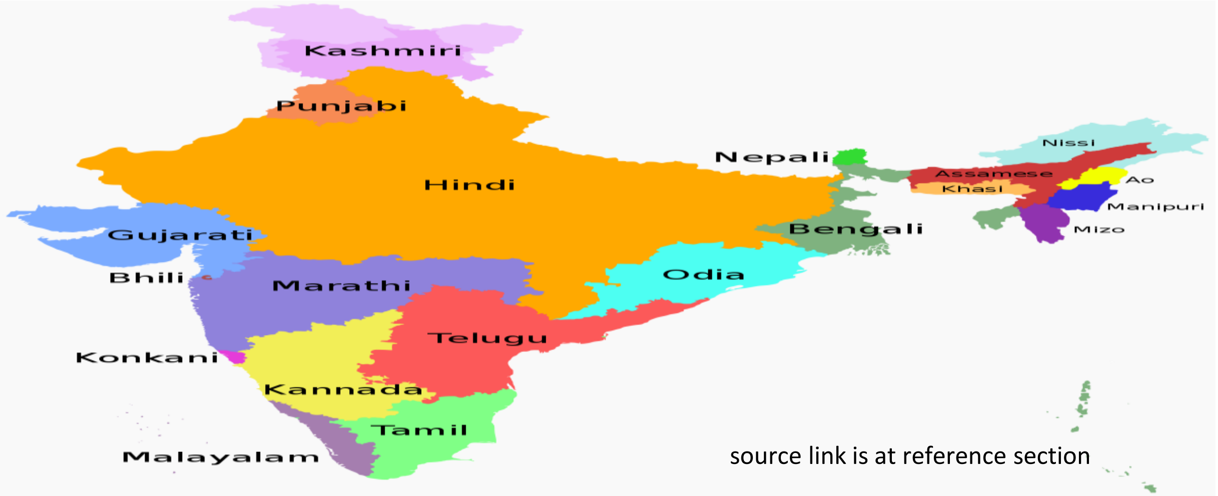
N/n = 5

select a random number from 1-5: chose 4

start with #4 and take every 5th unit

1	26	51	76
2	27	52	77
3	28	53	78
4	29	54	79
5	30	55	80
6	31	56	81
7	32	57	82
8	33	58	83
9	34	59	84
10	35	60	85
11	36	61	86
12	37	62	87
13	38	63	88
14	39	64	89
15	40	65	90
16	41	66	91
17	42	67	92
18	43	68	93
19	44	69	94
20	45	70	95
21	46	71	96
22	47	72	97
23	48	73	98
24	49	74	99
25	50	75	100

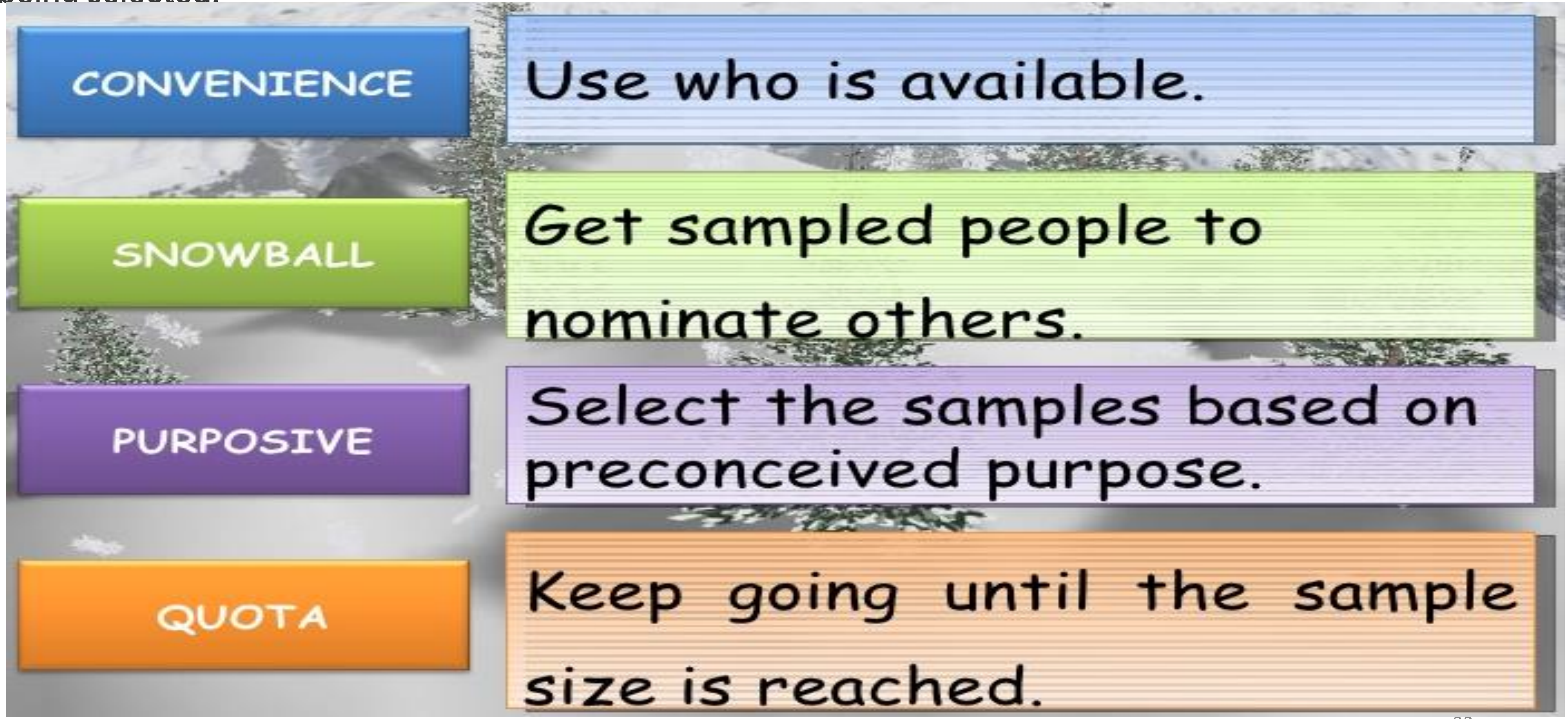
Cluster (Area) Random Sampling



source link is at reference section

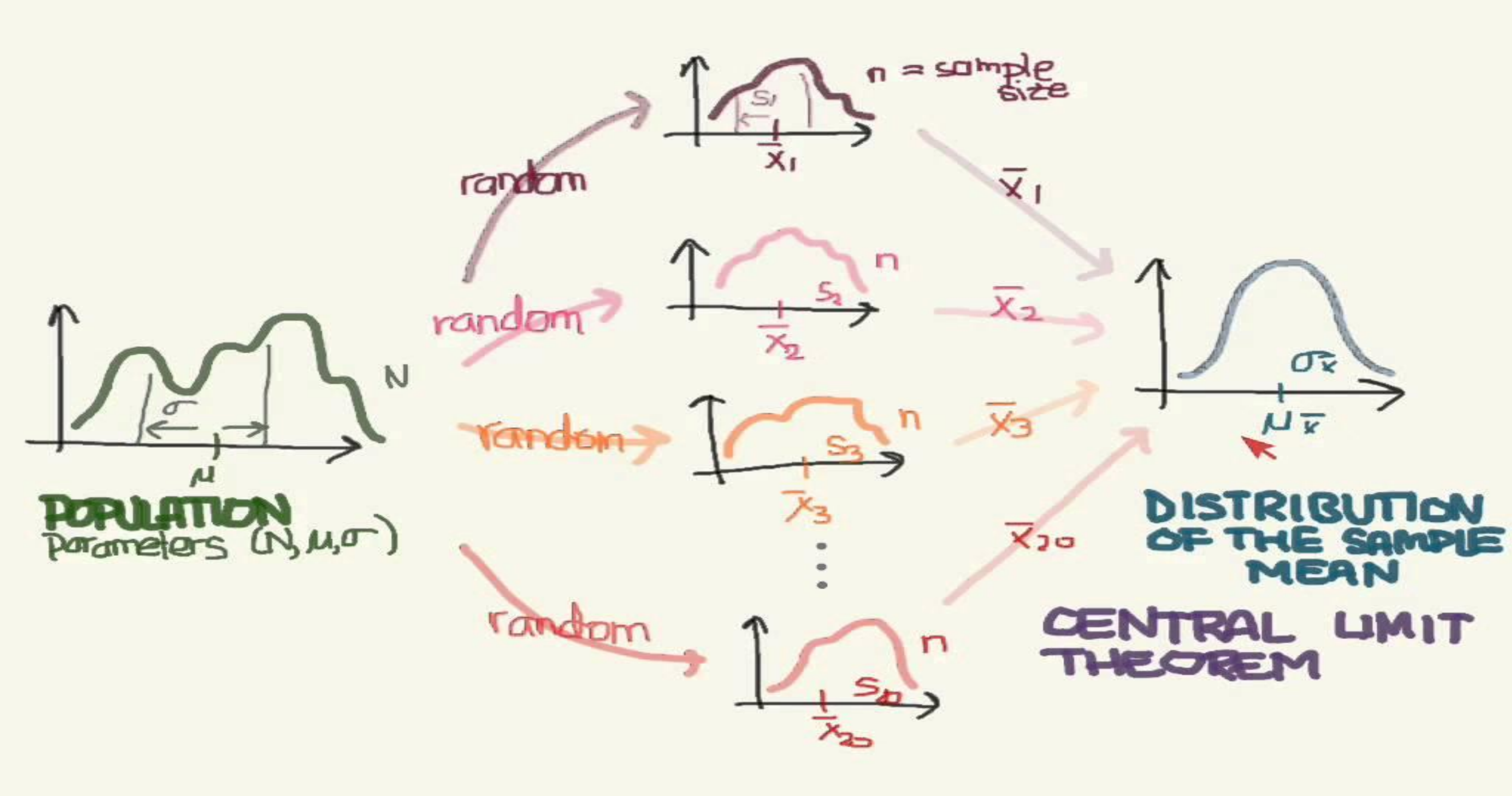
Sampling Procedure: Non Probability Sampling

Samples are gathered in a process that does not give all the individuals in the population equal chances of being selected.



CONVENIENCE	Use who is available.
SNOWBALL	Get sampled people to nominate others.
PURPOSIVE	Select the samples based on preconceived purpose.
QUOTA	Keep going until the sample size is reached.

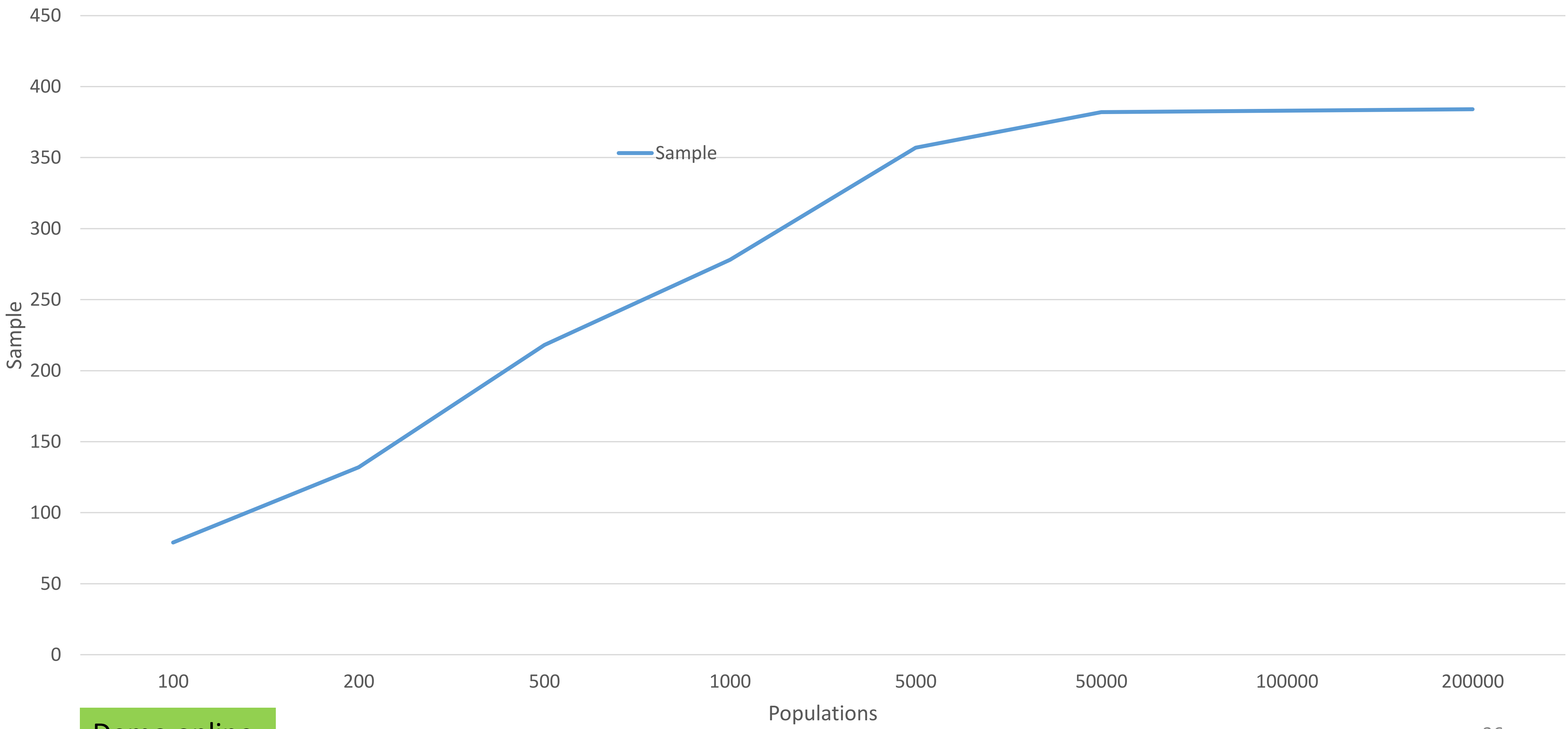
Central Limit Theorem - CLT



Sample size: Quiz by Guessing

Population	Sample
100	
200	
500	
1000	
5000	
50000	
100000	
200000	

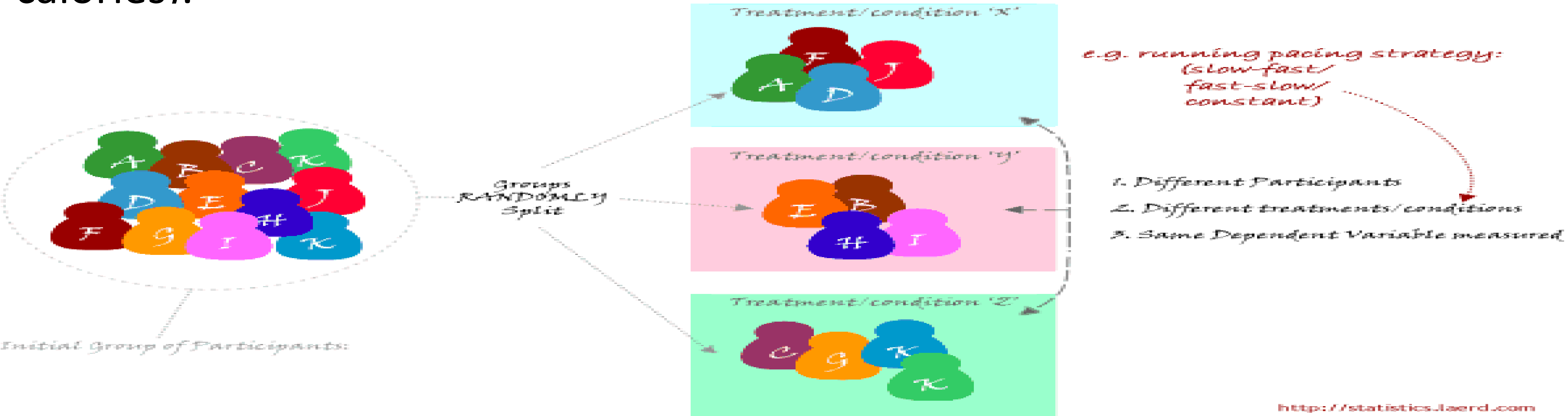
Sample size



Demo online

ANOVA - Analysis of variance

- It is a test to find out if survey or experiment results are significant. Basically, groups are getting tested to see if there's a difference between them.
- One-way or two-way refers to the number of independent variables (IVs) in your Analysis of Variance test. One-way has one independent variable (with 2 levels) and two-way has two independent variables (can have multiple levels). For example, a one-way Analysis of Variance could have one IV (brand of cereal) and a two-way Analysis of Variance has two IVs (brand of cereal, calories).


A
B
C
K
D
E
J
F
G
H
I
K
Initial Group of Participants:
Groups
RANDOMLY
Split
Treatment/condition 'X'
F
A
D
J
Treatment/condition 'Y'
E
B
H
I
Treatment/condition 'Z'
C
G
K
K
e.g. running pacing strategy:
(slow-fast/
fast-slow/
constant)
1. Different Participants
2. Different treatments/conditions
3. Same Dependent Variable measured

http://statistics.laerd.com

ANOVA - Examples

- A group of psychiatric patients are trying three different therapies: counseling, medication and biofeedback. You want to see if one therapy is better than the others.
- A manufacturer has two different processes to make light bulbs. They want to know if one process is better than the other.
- Students from different colleges take the same exam. You want to see if one college outperforms the other.
- To find out if there is an interaction between income and gender for anxiety level at job interviews. The anxiety level is the outcome, or the variable that can be measured. Gender and Income are the two categorical variables. These categorical variables are also the independent variables, which are called factors in a Two Way ANOVA.

ANOVA Hypothesis

One Way ANOVA

A one way ANOVA is used to compare two means from two independent (unrelated) groups using the F-distribution.

H0: The two means are equal.

H1: The two means are unequal.

Two Way ANOVA (**Two null hypotheses are tested)**

H01: All the income groups have equal mean stress.

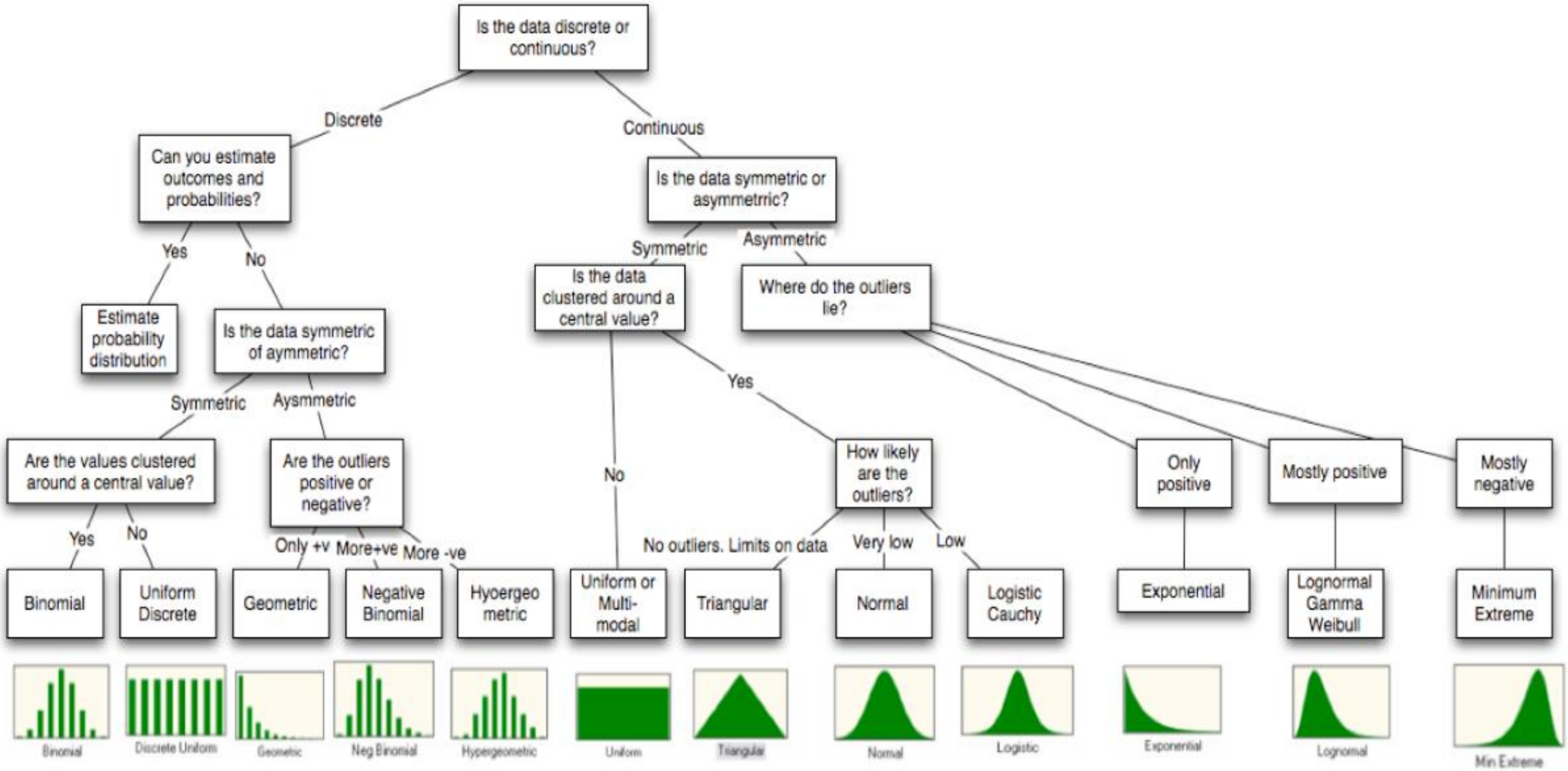
H02: All the gender groups have equal mean stress.

For multiple observations in cells, Testing a third hypothesis may also required.

H03: The factors are independent or the interaction effect does not exist.

Hands On

Various Distributions – Choices

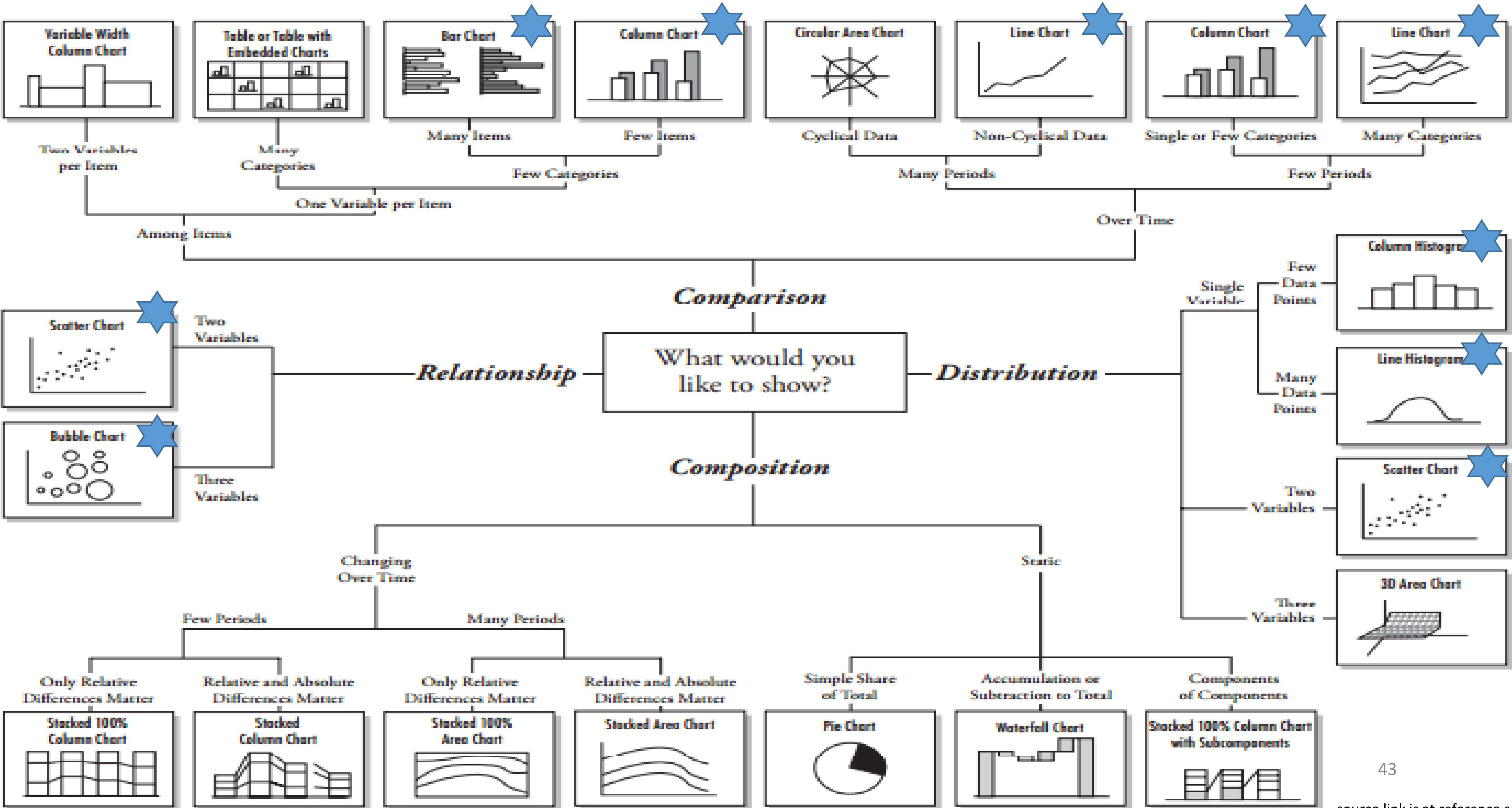


source link is at reference section

Graphs

i. XY Graphs
ii. Scatter Graphs
iii. Correlation
iv. Box Plots
v. Calculating the Quartiles
vi. Box Plots for Comparison
vii. Grouped Data
viii. Cumulative Frequency
ix. Percentiles
x. Pareto Charts
xi. Stem and Leaf Plots
xii. Multi-Variable Charts

Chart Suggestions—A Thought-Starter



What would you like to show?

Comparison

Among Items

Two Variables per Item

Variable Width Column Chart

One Variable per Item

Many Categories

Table or Table with Embedded Charts

Few Categories

Many Items

Bar Chart

Few Items

Column Chart

Over Time

Many Periods

Cyclical Data

Circular Area Chart

Non-Cyclical Data

Line Chart

Few Periods

Single or Few Categories

Column Chart

Many Categories

Line Chart

Relationship

Two Variables

Scatter Chart

Three Variables

Bubble Chart

Distribution

Single Variable

Few Data Points

Column Histogram

Many Data Points

Line Histogram

Two Variables

Scatter Chart

Three Variables

3D Area Chart

Composition

Changing Over Time

Few Periods

Only Relative Differences Matter

Stacked 100% Column Chart

Relative and Absolute Differences Matter

Stacked Column Chart

Many Periods

Only Relative Differences Matter

Stacked 100% Area Chart

Relative and Absolute Differences Matter

Stacked Area Chart

Static

Simple Share of Total

Pie Chart

Accumulation or Subtraction to Total

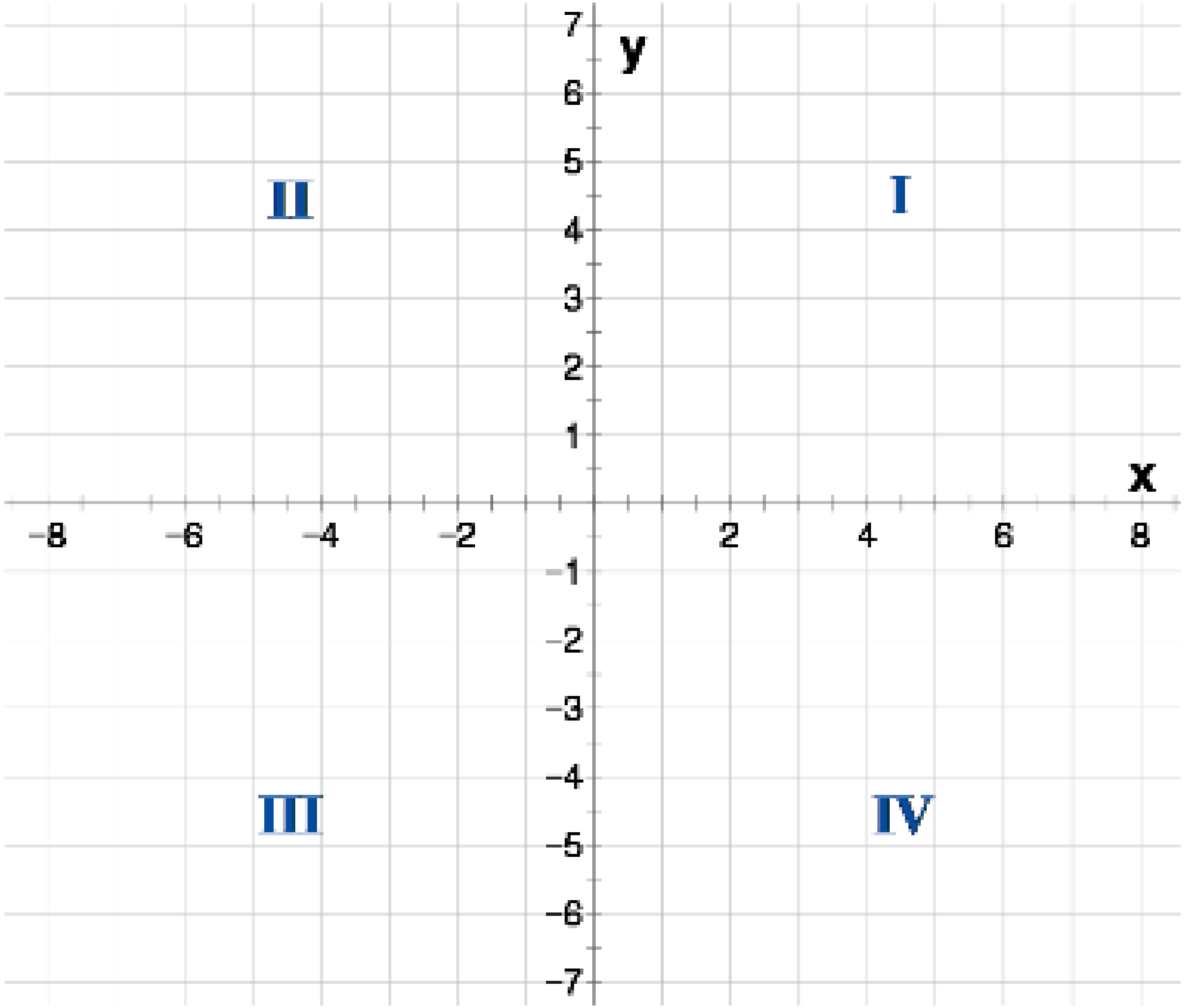
Waterfall Chart

Components of Components

Stacked 100% Column Chart with Subcomponents

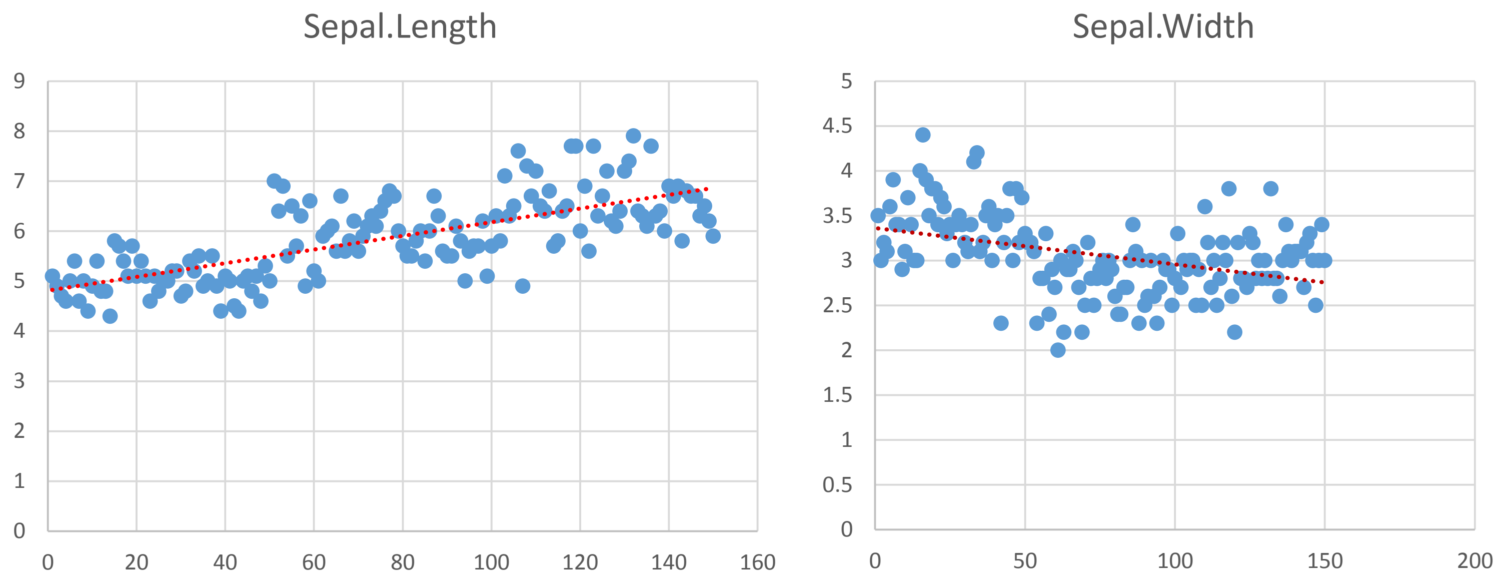
source link is at reference

XY Graphs



A. (5, 3), Quadrant I
B. (-3, 1), Quadrant II
C. (-6, -4), Quadrant III
D. (0, -3). It lies on an axis so it's not in a quadrant.
E. (0, 0). No quadrant because it's at the origin.
F. (4, -5), Quadrant IV
G. (6.5, 0). It lies on an axis so it's not in a quadrant.

Scatter



When to use it

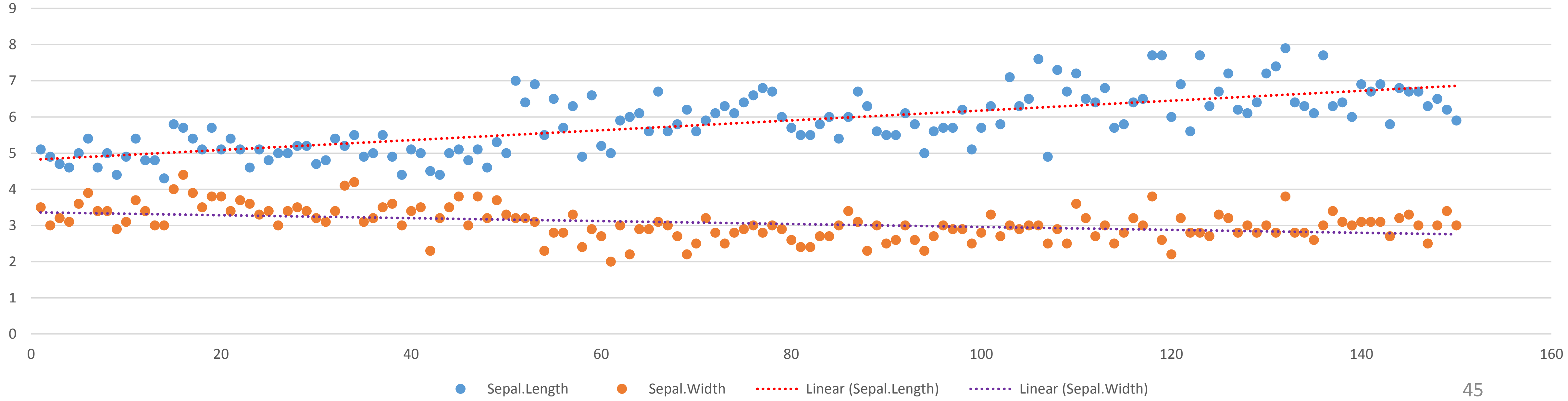
- Potential relationships between values
- Find outliers in data sets
- Each instances has at least two metrics

Advantages

- Visualize the correlation of two or more measures at the same time.

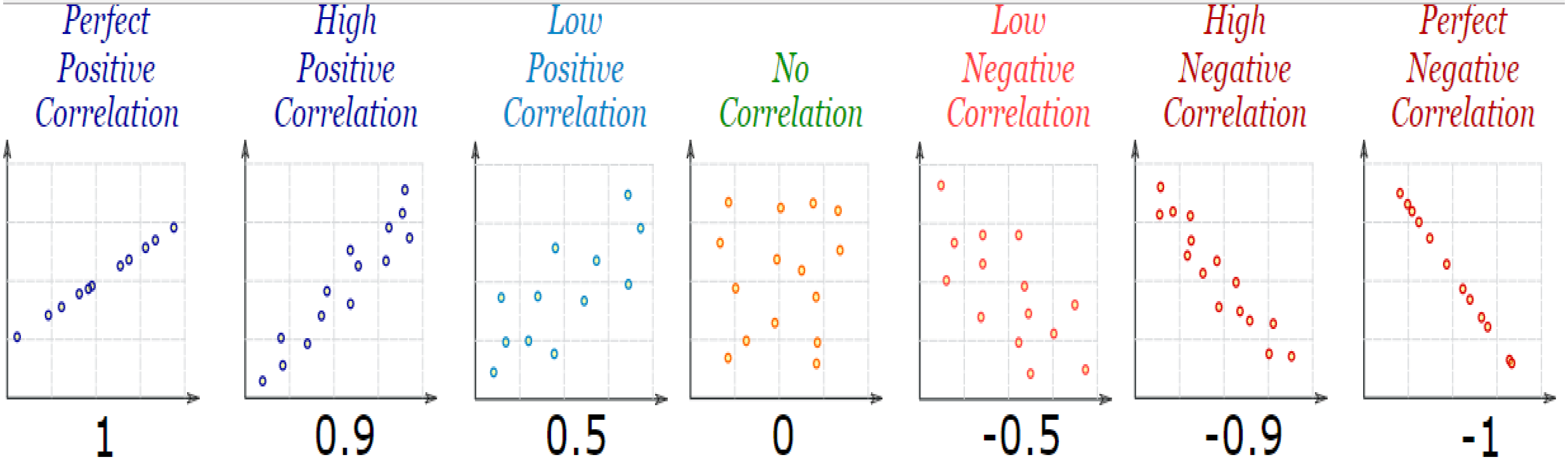
Disadvantages

- Difficult to understand for an inexperienced user

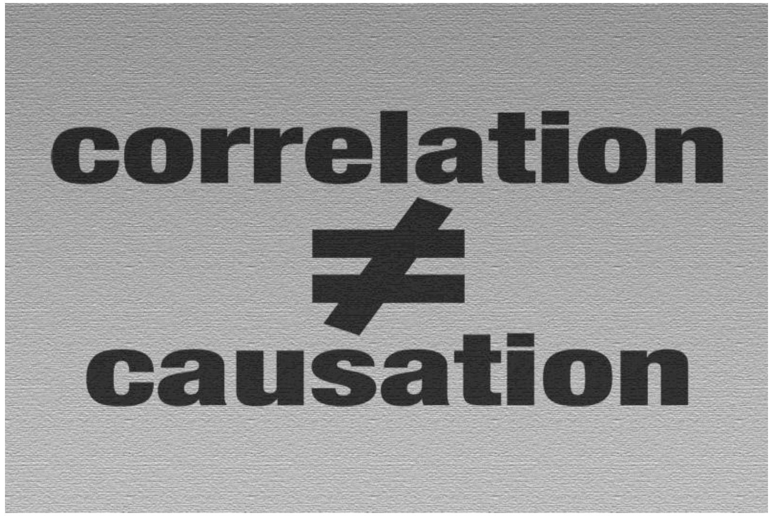


Hands On

Correlation



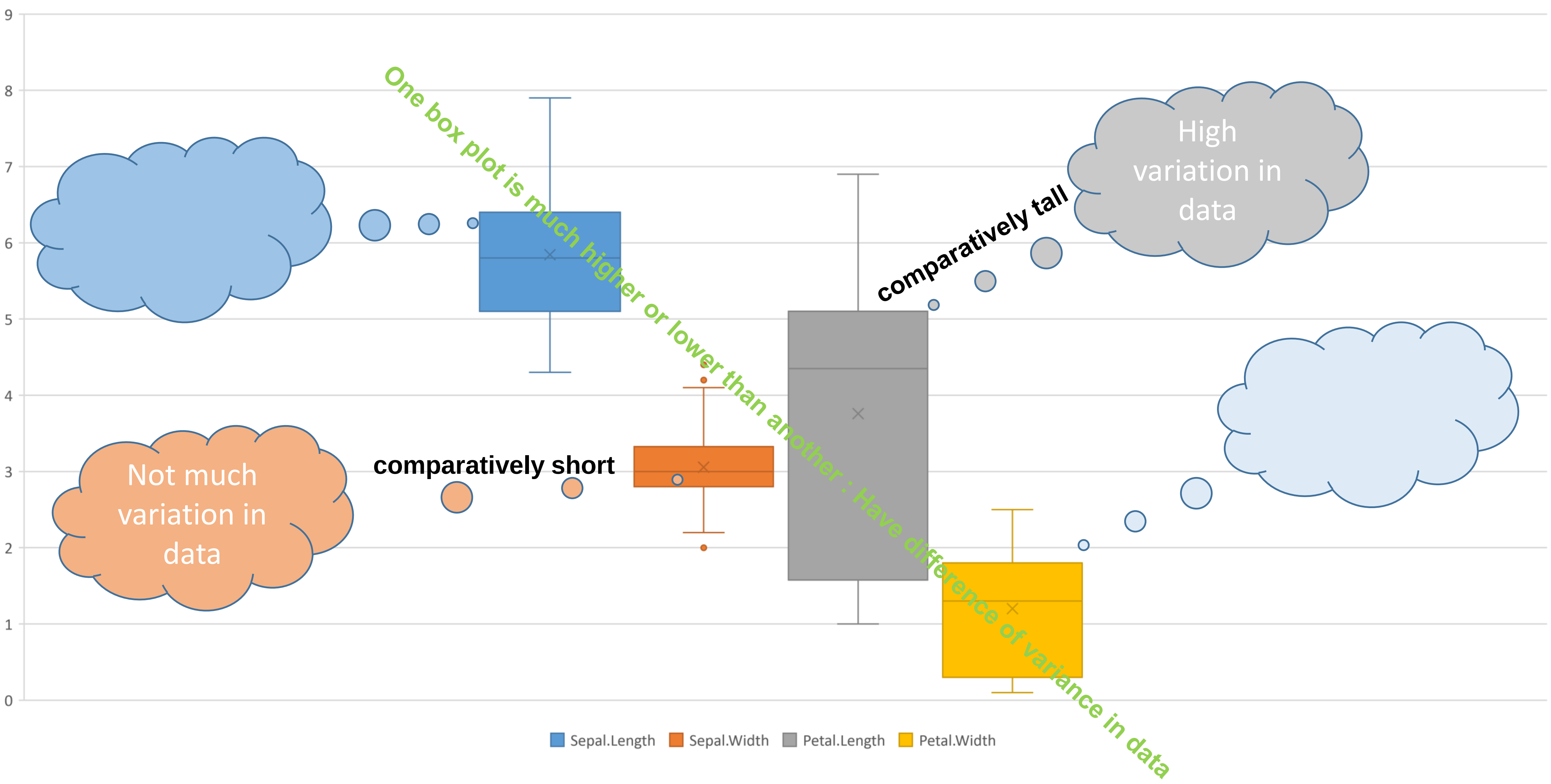
	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Sepal.Length	1			
Sepal.Width	W -0.11757	1		
Petal.Length	0.871754	M -0.42844	1	
Petal.Width	H 0.817941	-0.36613	0.962865	1



correlation ≠ causation

Hands On

Box Plots



9
8
7
6
5
4
3
2
1
0
One box plot is much higher or lower than another : Have difference of variance in data
comparatively tall
High variation in data
comparatively short
Not much variation in data
Sepal.Length
Sepal.Width
Petal.Length
Petal.Width

Hands On

Grouped Data

Grouped data is raw data that has been sorted into groups called classes. A class-interval is the range from the lowest value to the highest value in each class.

Number of Library Visits ... ungrouped (raw data).

1, 2, 2, 2, 3, 3, 3, 4, 4, 5, 5, 5, 5, 6, 6, 6, 6, 7, 7, 7, 7, 7, 7, 7, 8, 8, 9, 9, 10, 11

Number of Library Visits ... grouped (into classes).

1, 2, 2, 2,| 3, 3, 3, 4, 4,| 5, 5, 5, 5, 6, 6, 6, 6,| 7, 7, 7, 7, 7, 7, 7, 8, 8,| 9, 9, 10,| 11

4 | 5 | 8 | 9 | 3 | 1

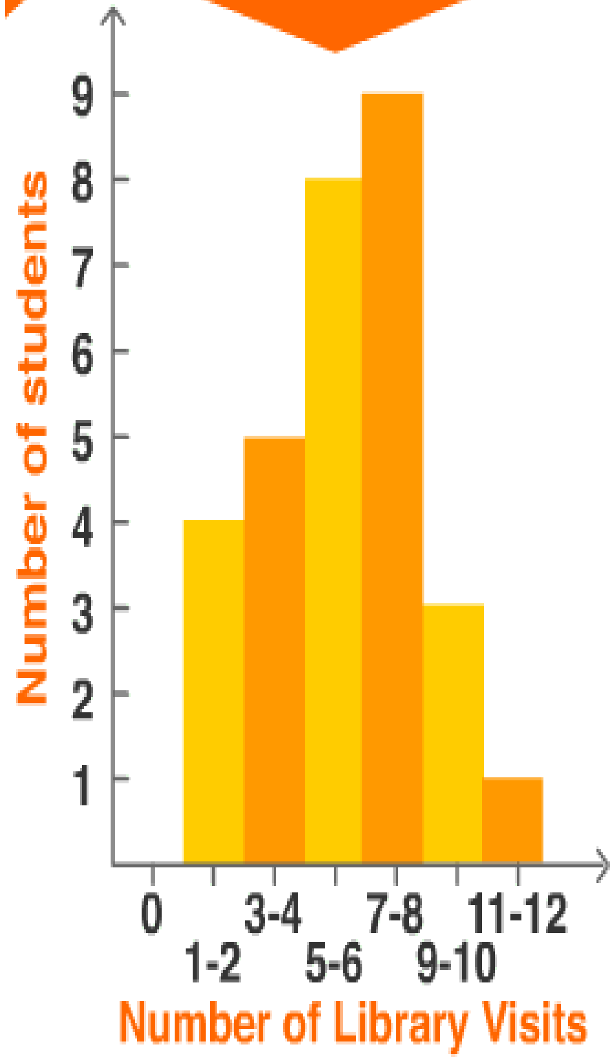
The grouped data may be displayed as a frequency disribution table.

The graph of grouped data is called a histogram.

frequency distribution table

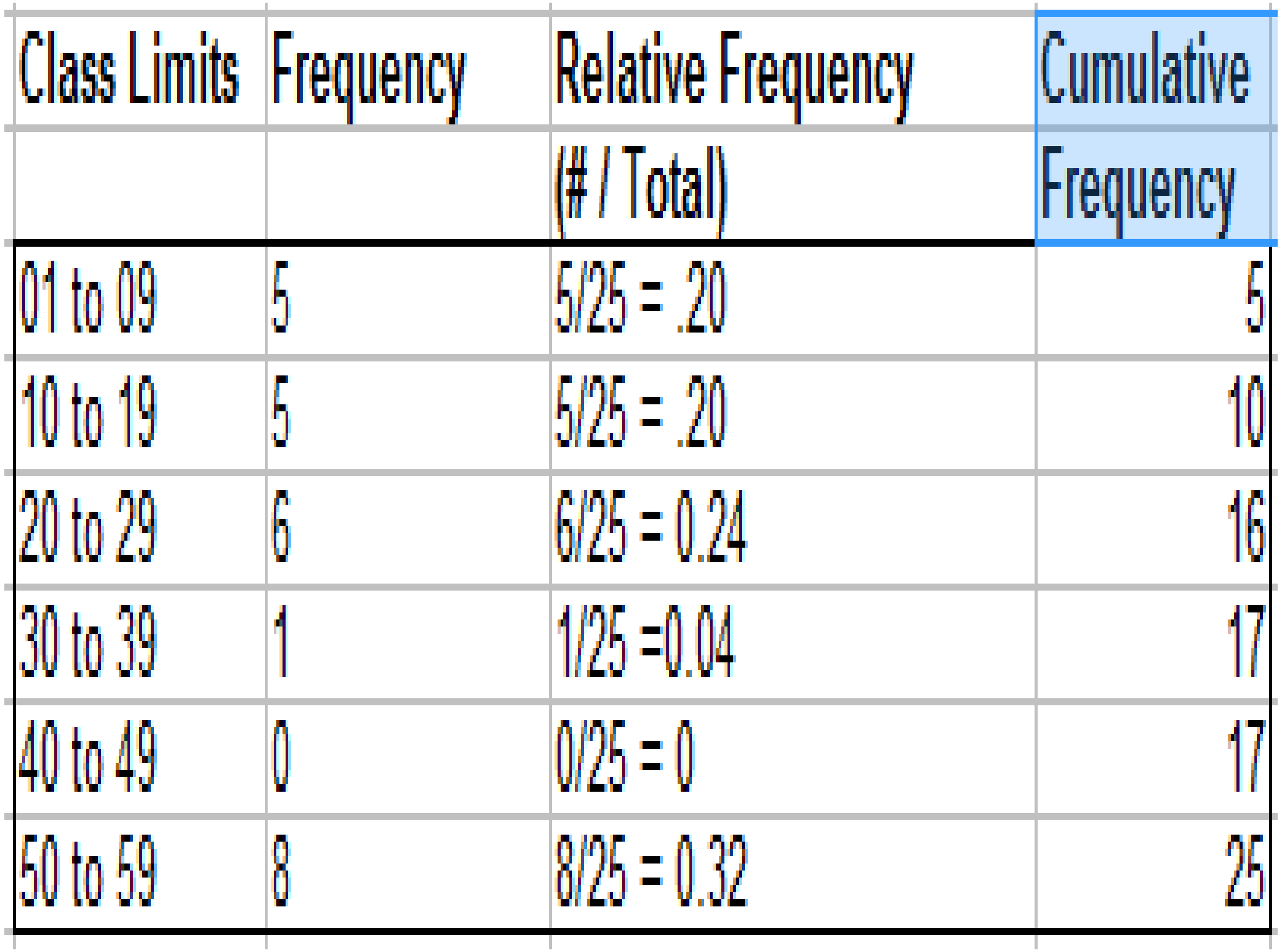
Classes - Number of Visits to the Library	Frequency - Number of Students
1 - 2	4
3 - 4	5
5 - 6	8
7 - 8	9
9 - 10	3
11 - 12	1

histogram

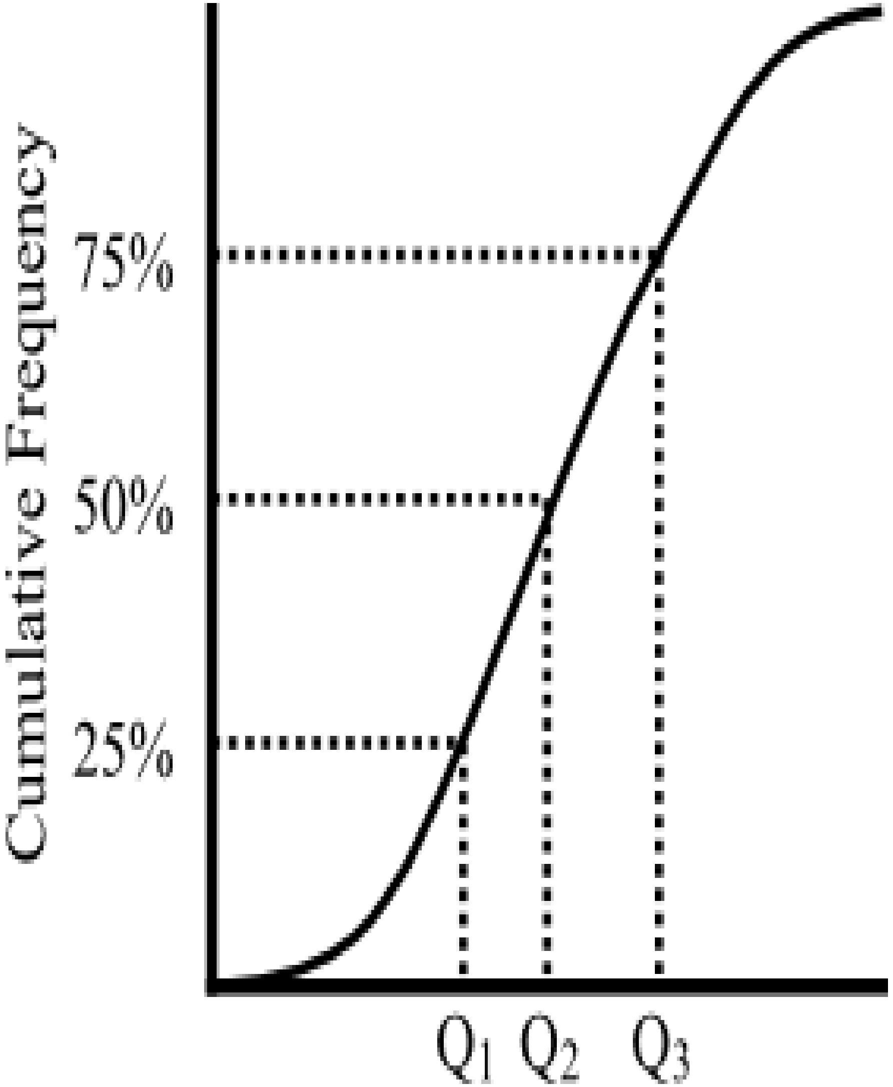


HW: Draw the diagram

Cumulative Frequency

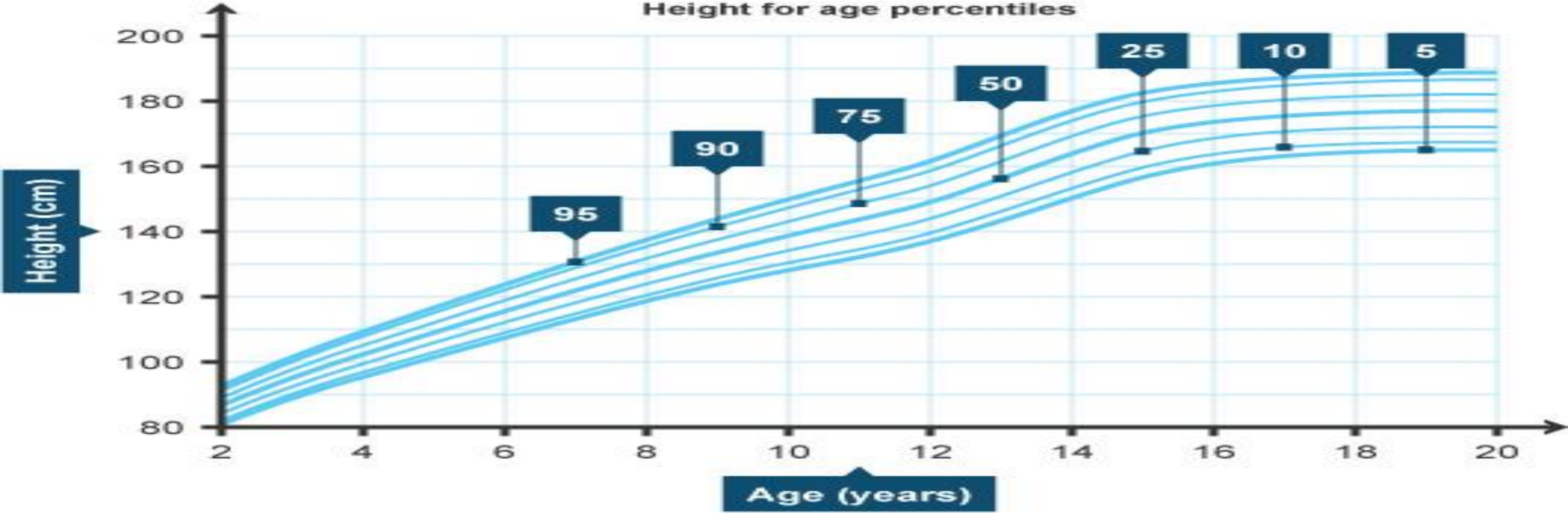


Class Limits	Frequency	Relative Frequency	Cumulative
		(# / Total)	Frequency
01 to 09	5	5/25 = .20	5
10 to 19	5	5/25 = .20	10
20 to 29	6	6/25 = 0.24	16
30 to 39	1	1/25 =0.04	17
40 to 49	0	0/25 = 0	17
50 to 59	8	8/25 = 0.32	25



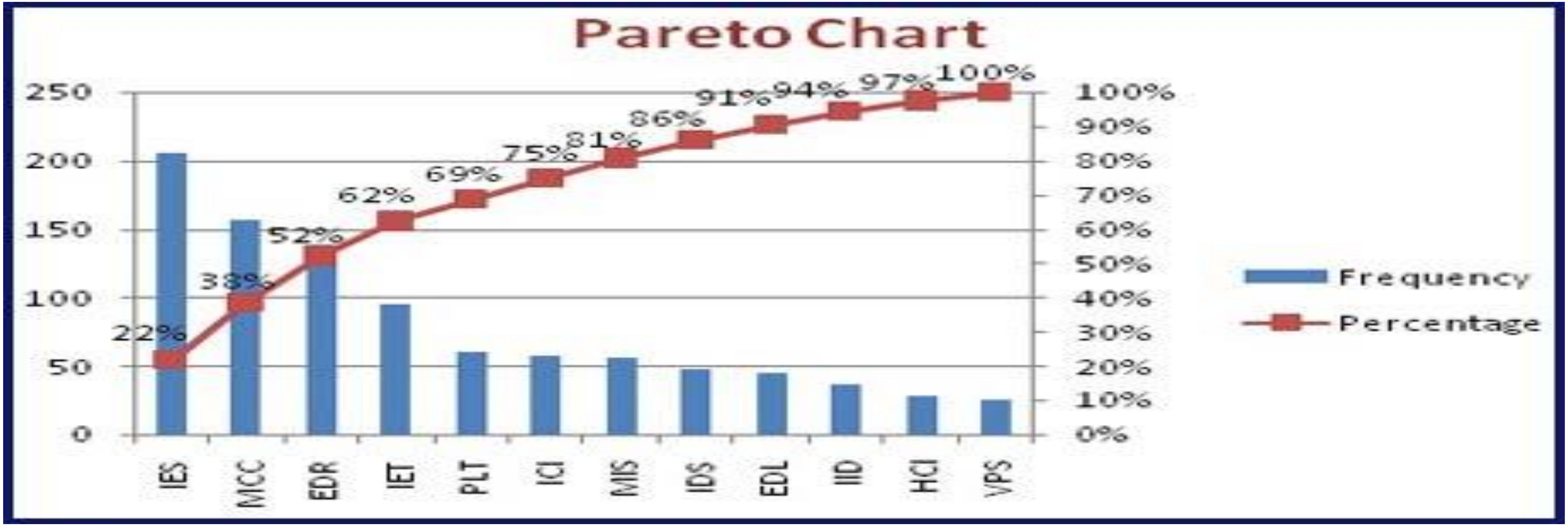
Percentiles

A percentile is a measure used in statistics indicating the value below which a given percentage of observations in a group of observations fall. For example, the 20th percentile is the value (or score) below which 20% of the observations may be found.



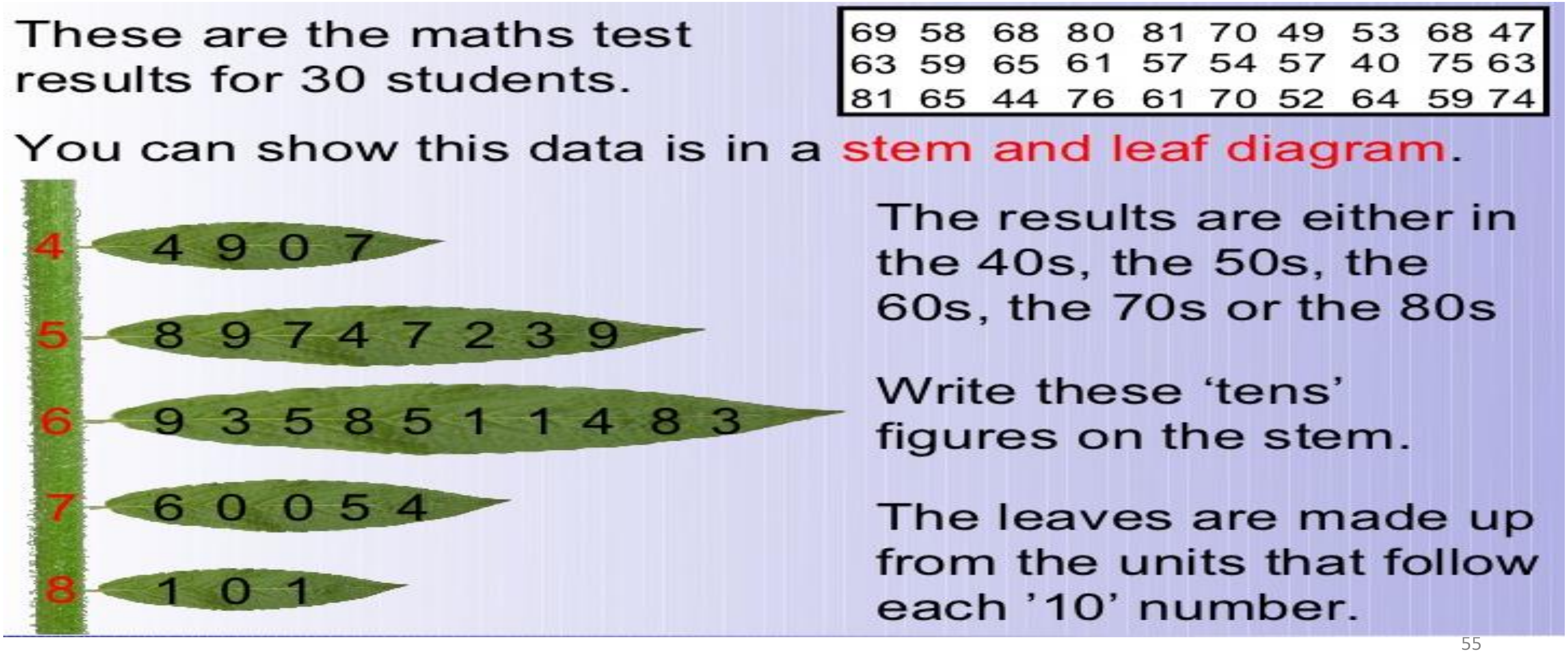
Pareto Charts

A Pareto chart, named after Vilfredo Pareto, is a type of chart that contains both bars and a line graph, where individual values are represented in descending order by bars, and the cumulative total is represented by the line.



Stem and Leaf Plots

It is a special table where each data value is split into a "stem" (the first digit or digits) and a "leaf" (usually the last digit).



These are the maths test results for 30 students.

69	58	68	80	81	70	49	53	68	47
63	59	65	61	57	54	57	40	75	63
81	65	44	76	61	70	52	64	59	74

You can show this data is in a stem and leaf diagram.

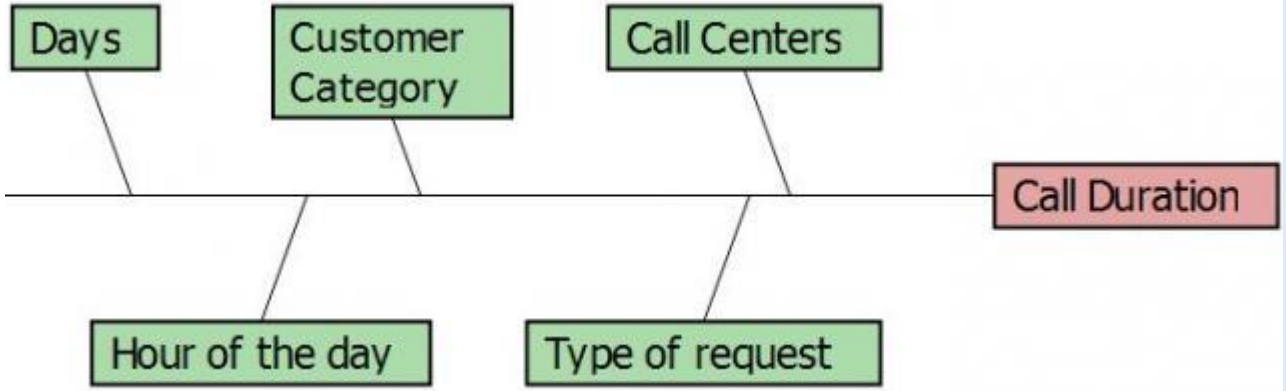
4	4 9 0 7
5	8 9 7 4 7 2 3 9
6	9 3 5 8 5 1 1 4 8 3
7	6 0 0 5 4
8	1 0 1

The results are either in the 40s, the 50s, the 60s, the 70s or the 80s

Write these 'tens' figures on the stem.

The leaves are made up from the units that follow each '10' number.

Multi variance Charts



Days

Customer Category

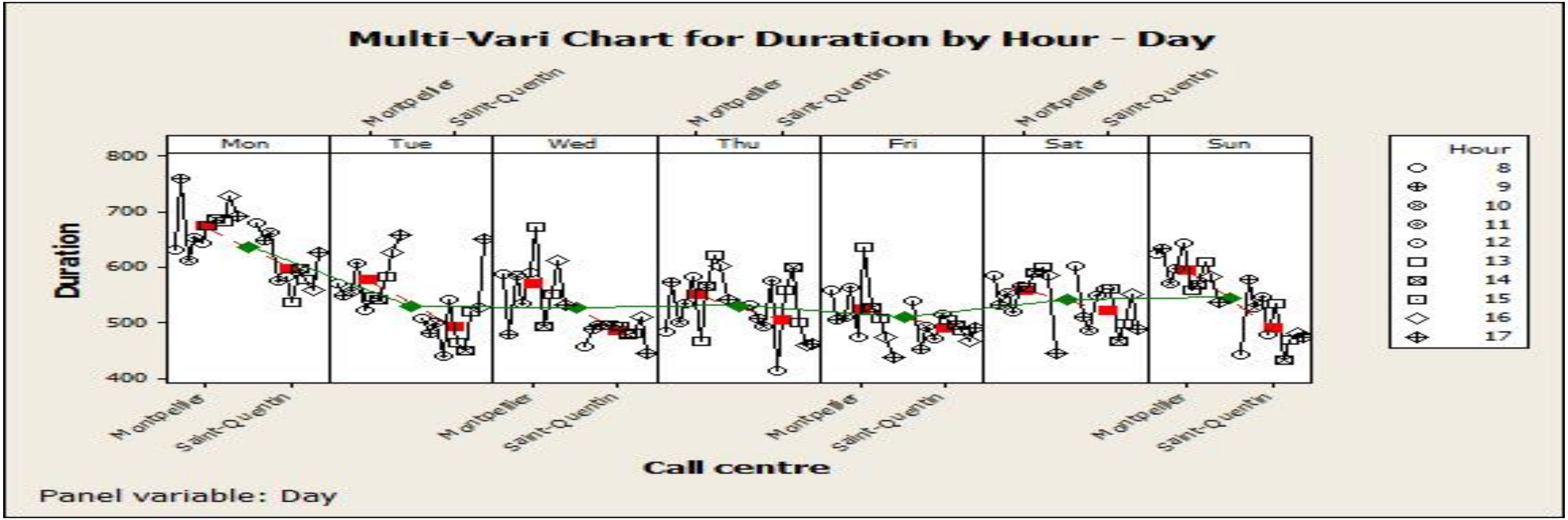
Call Centers

Hour of the day

Type of request

Call Duration

The multi variable chart below displays differences between the two call centers (Montpellier and Saint-Quentin: red points on the graph), the weekdays (green points on the graph) and the day hours (several black and white symbols). It suggests that waiting times are longer on Mondays



Future

- Advance Statistics

References

- https://www.google.co.in/imgres?imgurl=https://www.cliffsnotes.com/assets/267169.png&imgrefurl=https://www.cliffsnotes.com/study-guides/statistics/sampling/populations-samples-parameters-and-statistics&h=198&w=303&tbnid=vNA6IRElQFnWjM:&tbnh=137&tbnw=211&usg=__tnZczRZSn8ev-V3ygolcUuIxez4=&vet=1&docid=DSGpAawnRkZPQM&sa=X&ved=0ahUKEwjEtomzpPTVAhUDTI8KHUpWCq8Q9QEIKzAA
- https://www.cliffsnotes.com/study-guides/statistics/sampling/populations-samples-parameters-and-statistics
- http://keydifferences.com/difference-between-descriptive-and-inferential-statistics.html
- https://image.slidesharecdn.com/dansfinalnonprobabilityreport97-2003-100913002650-phpapp02/95/nonprobability-sampling-21-728.jpg?cb=1284337677
- https://en.wikipedia.org/wiki/Poisson_distribution
- https://sol.du.ac.in/mod/book/view.php?id=1317&chapterid=1066
- http://idolosol.com/images/range-3.jpg
- http://www.mathsisfun.com/data/images/range.gif
- http://topdrawer.aamt.edu.au/var/aamt/storage/images/media/tdt/statistics/s_cm_m4_ta3_fig1/283459-1-eng-AU/S_CM_M4_TA3_fig1.png
- https://www.mathsisfun.com/data/images/mean-deviation.svg
- http://www.biologyforlife.com/uploads/2/2/3/9/22392738/sd2_orig.png
- http://3.bp.blogspot.com/-mups5RZsDPE/Vn_JUutKo5I/AAAAAAAAA9Q/C6ks_UwNTmA/s1600/Types%2Bof%2Bskewness%2Bmean%2Bmdeian%2Bmode.PNG
- http://1.bp.blogspot.com/-lOeUCjVN9VE/Vn_Lmc5P0MI/AAAAAAAAA9c/CD799lTsrAw/s1600/types%2Bof%2Bkurtosis%252C%2Bleptokurtic%2Bmesokurtic%2Bplatykurtic.PNG
- https://www.socialresearchmethods.net/kb/sampprob.php
- https://upload.wikimedia.org/wikipedia/commons/thumb/3/3c/Language_region_maps_of_India.svg/1200px-Language_region_maps_of_India.svg.png
- https://explorable.com/non-probability-sampling
- http://blog.minitab.com/blog/applying-statistics-in-quality-projects/using-multi-vari-charts-to-analyze-families-of-variations
- https://www.google.com/imgres?imgurl=https%3A%2F%2Fcdn.edureka.co%2Fblog%2Fwp-content%2Fuploads%2F2013%2F06%2FData-Scientist.jpg&imgrefurl=https%3A%2F%2Fwww.edureka.co%2Fblog%2Fwho-is-a-data-scientist%2F&docid=q3ij004thhLBlM&tbnid=6_eIIKVKkQtWlM%3A&vet=10ahUKEwi74J_L7sbUAhWHRY8KHYcUBbUQMwi8AShFMEU..i&w=601&h=351&bih=632&biw=1366&q=Data%20Science&ved=0ahUKEwi74J_L7sbUAhWHRY8KHYcUBbUQMwi8AShFMEU&iact=mrc&uact=8
- https://en.wikipedia.org/wiki/R_(programming_language)
- http://www.chioka.in/differences-between-roc-auc-and-pr-auc/
- https://www.google.com/imgres?imgurl=http%3A%2F%2Fstanford.edu%2F~cpiech%2Fcs221%2Fimg%2FkmeansViz.png&imgrefurl=http%3A%2F%2Fstanford.edu%2F~cpiech%2Fcs221%2Fhandouts%2Fkmeans.html&docid=xBG90gMlnM_nKM&tbnid=0dV6dbzMcO1mgM%3A&vet=10ahUKEwiv8PWX49HUAhXKtI8KHVk3BigQMwhSKAUwBQ..i&w=501&h=338&bih=632&biw=1366&q=K-means&ved=0ahUKEwiv8PWX49HUAhXKtI8KHVk3BigQMwhSKAUwBQ&iact=mrc&uact=8
- https://www.otexts.org/

References

- http://www.moderndive.com
- http://www.statisticshowto.com/wp-content/uploads/2014/11/chi-square-distribution.png
- https://www.mathsisfun.com/data/quincunx.html
- http://onlinestatbook.com/2/calculators/normal_dist.html
- https://www.mathsisfun.com/data/standard-normal-distribution.html
- http://support.minitab.com/en-us/minitab-express/1/help-and-how-to/basic-statistics/probability-distributions/supporting-topics/basics/continuous-and-discrete-probability-distributions/
- https://i.ytimg.com/vi/3SKwerKHbRk/maxresdefault.jpg
- http://www.investopedia.com/terms/c/central_limit_theorem.asp
- https://www.quanterion.com/interference-stressstrength-analysis/
- http://www.theanalysisfactor.com/confusing-statistical-terms-1-alpha-and-beta/
- http://www.nss.gov.au/nss/home.nsf/pages/Sample+size+calculator
- http://www.statisticshowto.com/anova/
- https://www.google.com/imgres?imgurl=https%3A%2F%2Fwww.nervanasys.com%2Fwp-content%2Fuploads%2F2016%2F04%2FScreen-Shot-2016-04-27-at-10.59.50-AM.png&imgrefurl=https%3A%2F%2Fwww.nervanasys.com%2Fopenai%2F&docid=GggjYZ3JzfRoVM&tbnid=teq-_Mt_keBQiM%3A&vet=10ahUKEwiDqYXb-7_UAhWLs48KHd2dD00QMwh2KC8wLw..i&w=758&h=423&bih=632&biw=1366&q=Reinforced%20learning&ved=0ahUKEwiDqYXb-7_UAhWLs48KHd2dD00QMwh2KC8wLw&iact=mrc&uact=8
- https://images.google.com/
- robjhyndman.com/hyndsight/forecasting/
- OTexts.org/fpp

The American Statistical Association (ASA) Board statement on "*The Role of Statistics in Data Science*".

- The rise of data science, including big data and data analytics, has recently attracted enormous attention in the popular press for its spectacular contributions in a wide range of scholarly disciplines and commercial endeavors. These successes are largely the fruit of the innovative and entrepreneurial spirit that characterize this burgeoning field. Nonetheless, its interdisciplinary nature means that a substantial collaborative effort is needed for it to realize its full potential for productivity and innovation. While there is not yet a consensus on what precisely constitutes data science, three professional communities, all within computer science and/or statistics, are emerging as foundational to data science:
- Database Management enables transformation, conglomeration, and organization of data resources;
- Statistics and Machine Learning convert data into knowledge; and
- Distributed and Parallel Systems provide the computational infrastructure to carry out data analysis.

- Certainly, data science intersects with numerous other disciplines and areas of research. Indeed it is difficult to think of an area of science, industry, commerce, or government that is not in some way involved in the data revolution. But it is databases, statistics, and distributed systems that provide the core pipeline. At its most fundamental level, we view data science as a mutually beneficial collaboration among these three professional communities, complemented with significant interactions with numerous related disciplines. For data science to fully realize its potential requires maximum and multifaceted collaboration among these groups.

- Statistics and machine learning play a central role in data science. Framing questions statistically allows us to leverage data resources to extract knowledge and obtain better answers. The central dogma of statistical inference, that there is a component of randomness in data, enables researchers to formulate questions in terms of underlying processes and to quantify uncertainty in their answers. A statistical framework allows researchers to distinguish between causation and correlation and thus to identify interventions that will cause changes in outcomes. It also allows them to establish methods for prediction and estimation, to quantify their degree of certainty, and to do all of this using algorithms that exhibit predictable and reproducible behavior. In this way, statistical methods aim to focus attention on findings that can be reproduced by other researchers with different data resources. Simply put, statistical methods allow researchers to accumulate knowledge.