

Machine Learning

Lecture 1 (Spring-2021)

EDA – Part I

Dr. Tariq Mahmood

Assistant Professor IBA and Head of BDA-LAB

tmahmood@iba.edu.pk

Profile

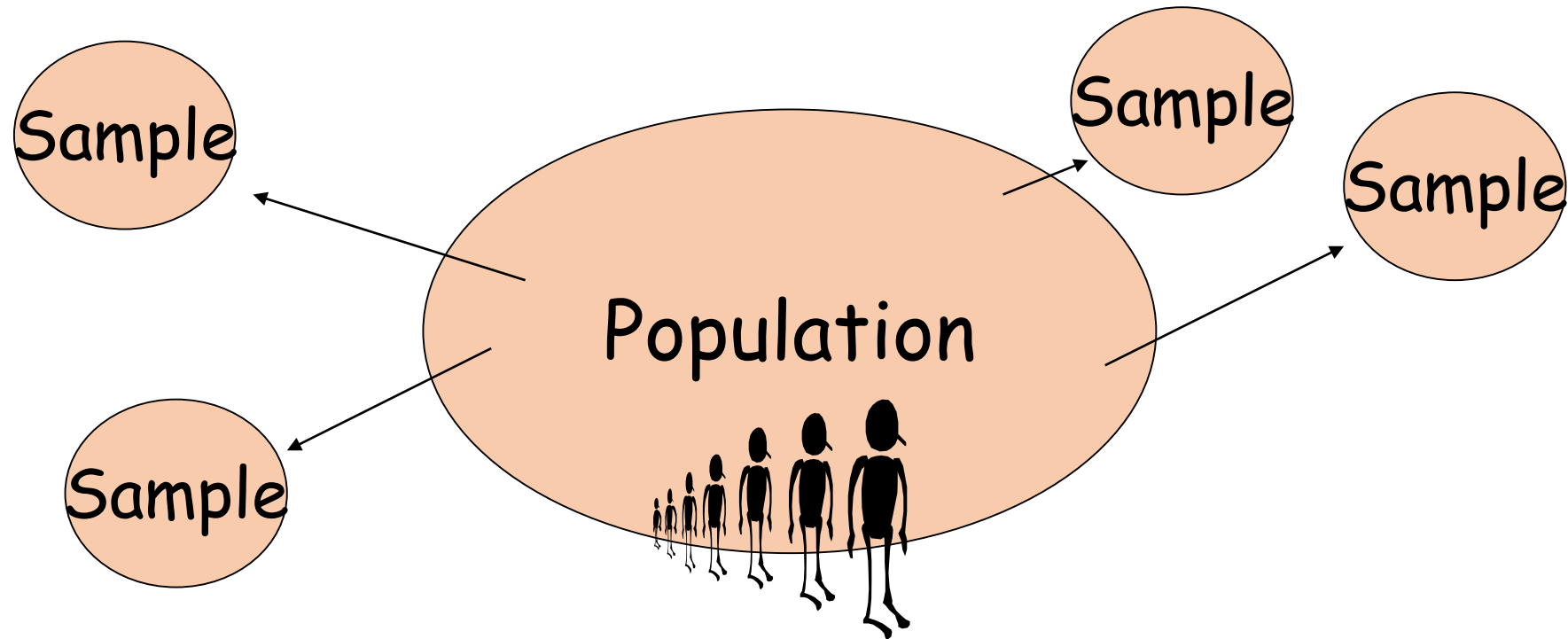
- MS (Statistical Machine Learning), France
- PhD (Machine Learning), Italy
- Research domains: Big Data Analytics, Deep Learning, Machine Learning
- Workshops Conducted: Ministry of Defense, Telecom, Finance, FMCG, Automobile, Pakistan Navy, Healthcare, Insurance, Analytics Service Providers
- Diploma for BDA (4 month course – 5th batch of pure industry professionals in progress)
(<http://cict.iba.edu.pk/BigDataAnalytics.php>)

Statistics and Probability

Basics of Statistics

- Science of collection, presentation, analysis, and reasonable interpretation of data.
- presents a rigorous scientific method for gaining insight into data.
- Weights of 500 patients – difficult to analyze each one separately but through statistics can give an instant overall picture of data based on graphical presentation or numerical summarization irrespective to the number of data points

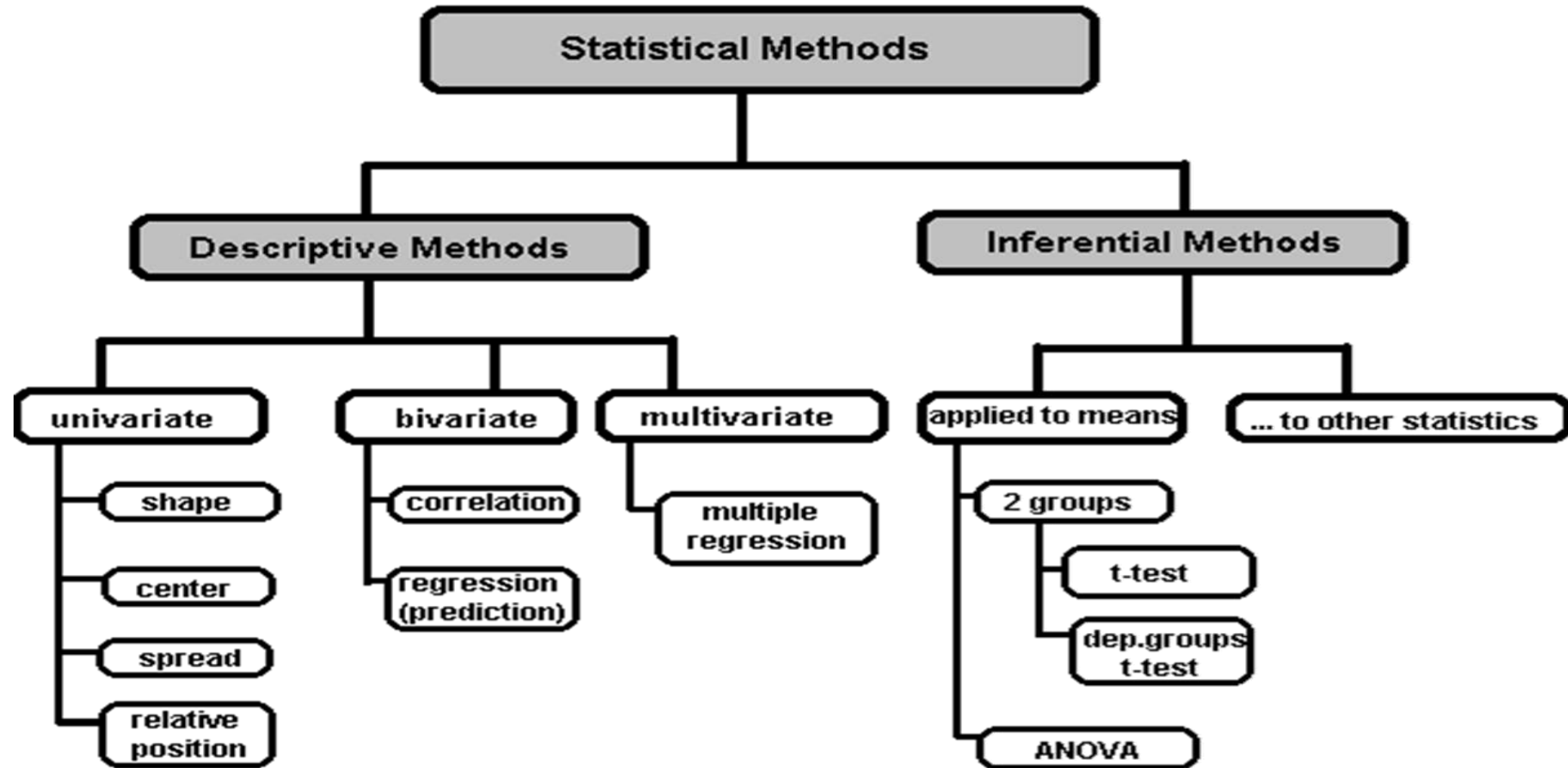
What is a Statistic?



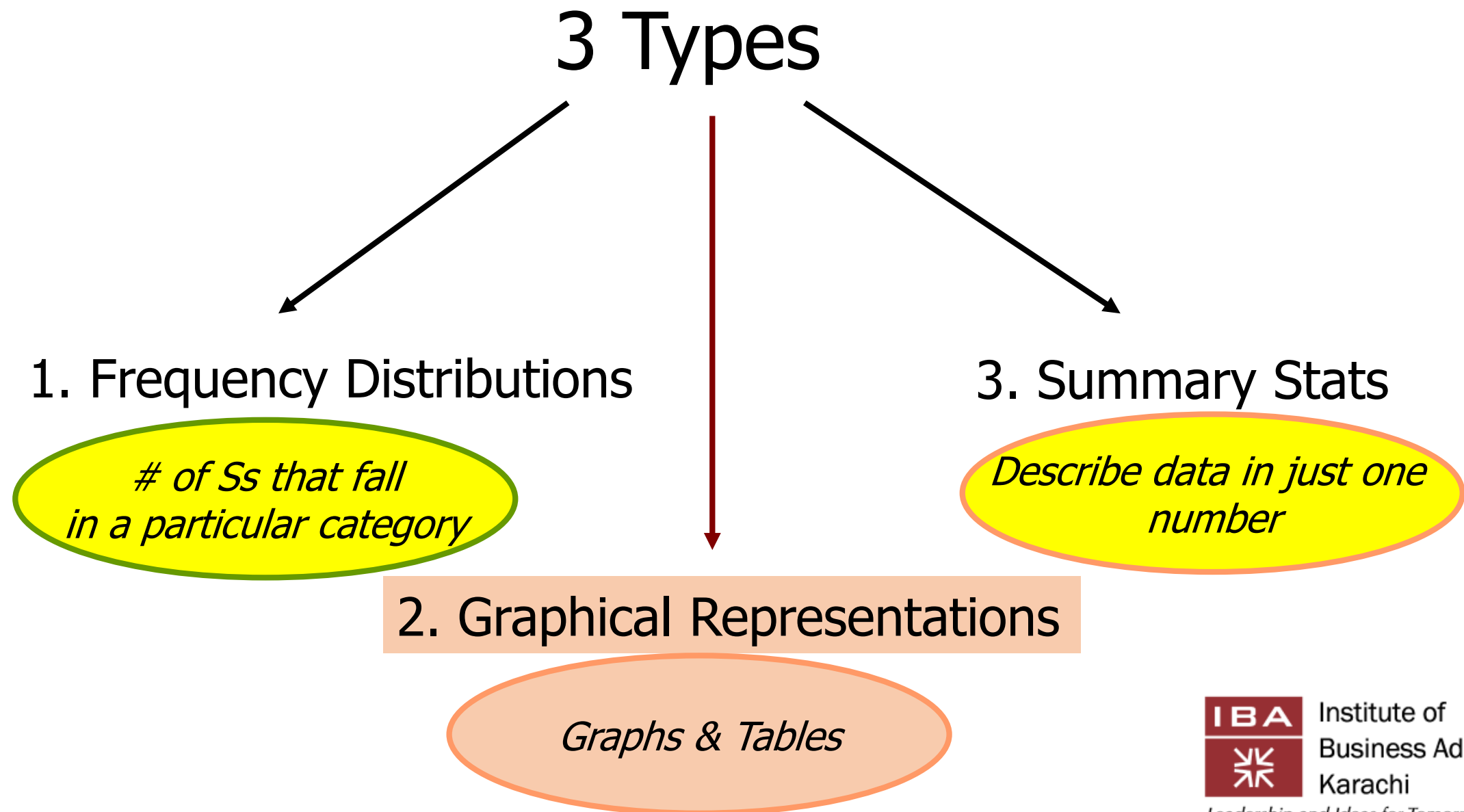
Parameter: value that describes a population

Statistic: a value that describes a sample

Taxonomy



Descriptive



Descriptive vs Inferential

- Organize
- Summarize
- Simplify
- Presentation of data

Describing data

- Generalize from samples to pops
- Hypothesis testing
- Relationships among variables

Make predictions



Institute of
Business Administration
Karachi

Leadership and Ideas for Tomorrow

Statistical Description of Data

- Statistics describes a numeric set of data by its
 - Center
 - Variability
 - Shape
- Statistics describes a categorical set of data by
 - Frequency, percentage or proportion of each category
- Date data types can also be analyzed using date functions

Variables

- **Nominal** - Categorical variables with no inherent order or ranking sequence such as names or classes (e.g., gender) – can be enumerated
- **Ordinal** - Variables with an inherent rank or order, e.g. mild, moderate, severe. Can be compared for equality, or greater or less, but not *how much* greater or less.
- **Interval** - Values of the variable are ordered as in Ordinal, and additionally, differences between values are meaningful, however, the scale is not absolutely anchored. Calendar dates and temperatures on the Fahrenheit scale are examples.
- **Ratio** - Variables with all properties of Interval plus an absolute, non-arbitrary zero point, e.g. age, weight, temperature (Kelvin).

Distribution

- Distribution - (of a variable) tells us what values the variable takes and how often it takes these values.
 - Unimodal - having a single peak
 - Bimodal - having two distinct peaks
 - Symmetric - left and right half are mirror images.

Frequency Distribution

- Consider a data set of 26 children of ages 1-6 years. Then the frequency distribution of variable 'age' can be tabulated as follows:

Frequency Distribution of Age

Age	1	2	3	4	5	6
Frequency	5	3	7	5	4	2

Grouped Frequency Distribution of Age:

Age Group	1-2	3-4	5-6
Frequency	8	12	6

Cumulative Frequency

Age	1	2	3	4	5	6
Frequency	5	3	7	5	4	2
Cumulative Frequency	5	8	15	20	24	26

Age Group	1-2	3-4	5-6
Frequency	8	12	6
Cumulative Frequency	8	20	26



Institute of
Business Administration
Karachi

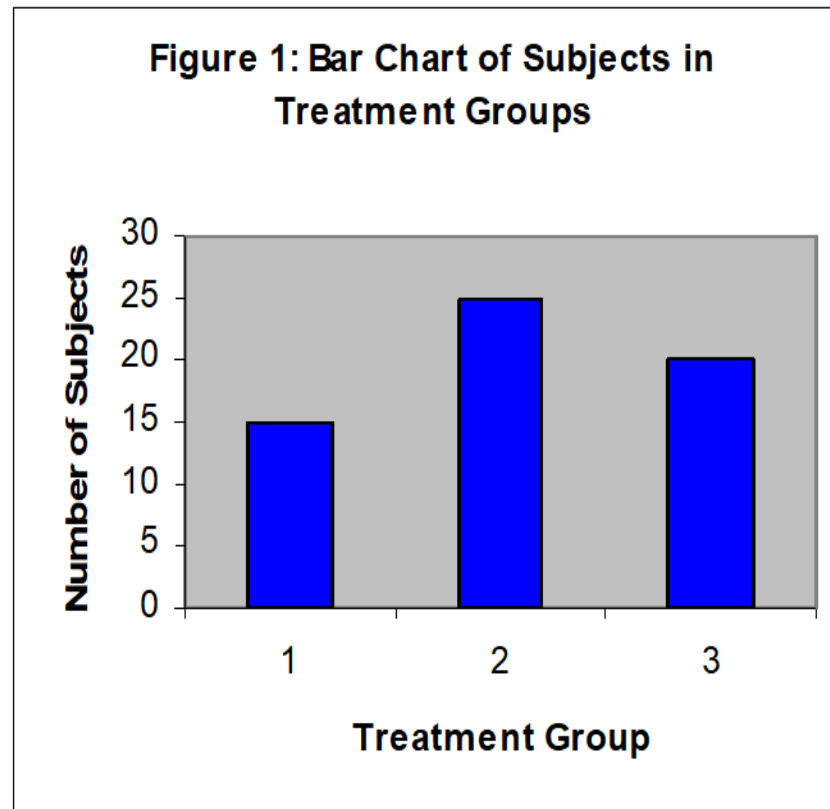
Leadership and Ideas for Tomorrow

Data Presentation

- Graphical and Numerical
- Graphical Presentation: We look for the overall pattern and for striking deviations from that pattern. Overall pattern usually described by shape, center, and spread of the data. An individual value that falls outside the overall pattern is called an outlier.
- Bar diagram and Pie charts are used for categorical variables (many others are now available in BI tools eg. Heatmaps/treemaps)
- Histogram, stem and leaf and Box-plot are used for numerical variable (density plots super-imposed on histograms)

Categorical Data Presentation

Bar Diagram: Lists the categories and presents the percent or count of individuals who fall in each category.

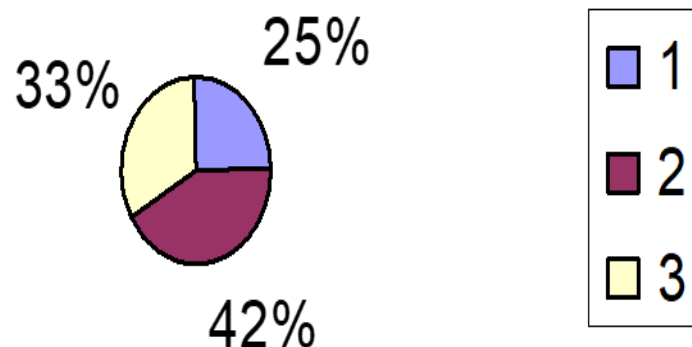


Treatment Group	Frequency	Proportion	Percent (%)
1	15	$(15/60)=0.25$	25.0
2	25	$(25/60)=0.333$	41.7
3	20	$(20/60)=0.417$	33.3
Total	60	1.00	100

Categorical Data Presentation

Pie Chart: Lists the categories and presents the percent or count of individuals who fall in each category.

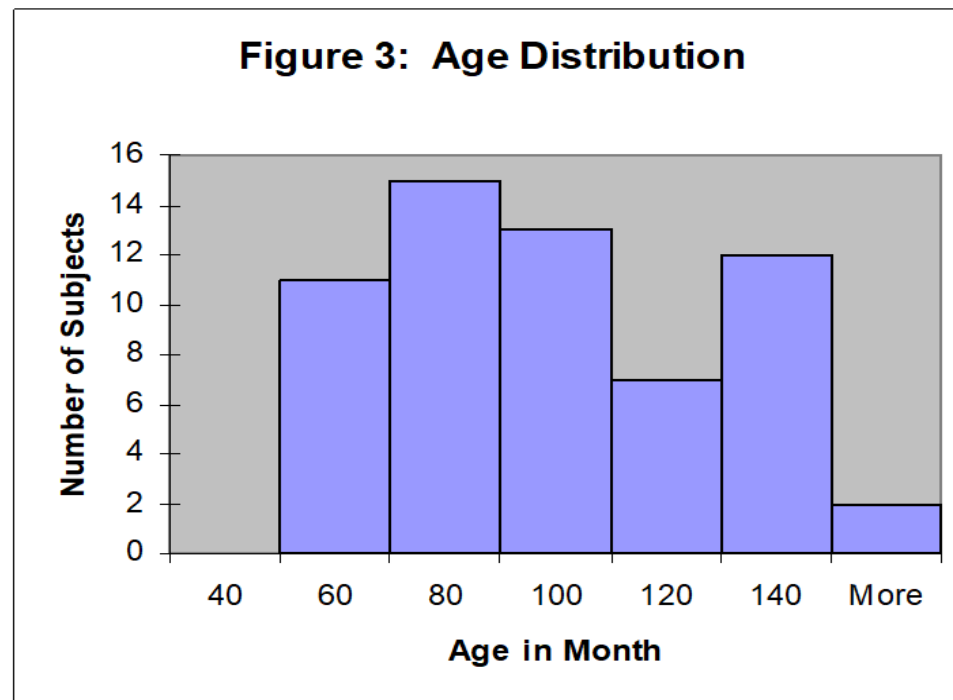
Figure 2: Pie Chart of Subjects in Treatment Groups



Treatment Group	Frequency	Proportion	Percent (%)
1	15	$(15/60)=0.25$	25.0
2	25	$(25/60)=0.333$	41.7
3	20	$(20/60)=0.417$	33.3
Total	60	1.00	100

Numerical Data Presentation

Histogram: Overall pattern can be described by its **shape**, **center**, and **spread**. The following age distribution is **right skewed**. The **center** lies between **80 to 100**. **No outliers**.



Mean	90.41666667
Standard Error	3.902649518
Median	84
Mode	84
Standard Deviation	30.22979318
Sample Variance	913.8403955
Kurtosis	-1.183899591
Skewness	0.389872725
Range	95
Minimum	48
Maximum	143
Sum	5425
Count	60

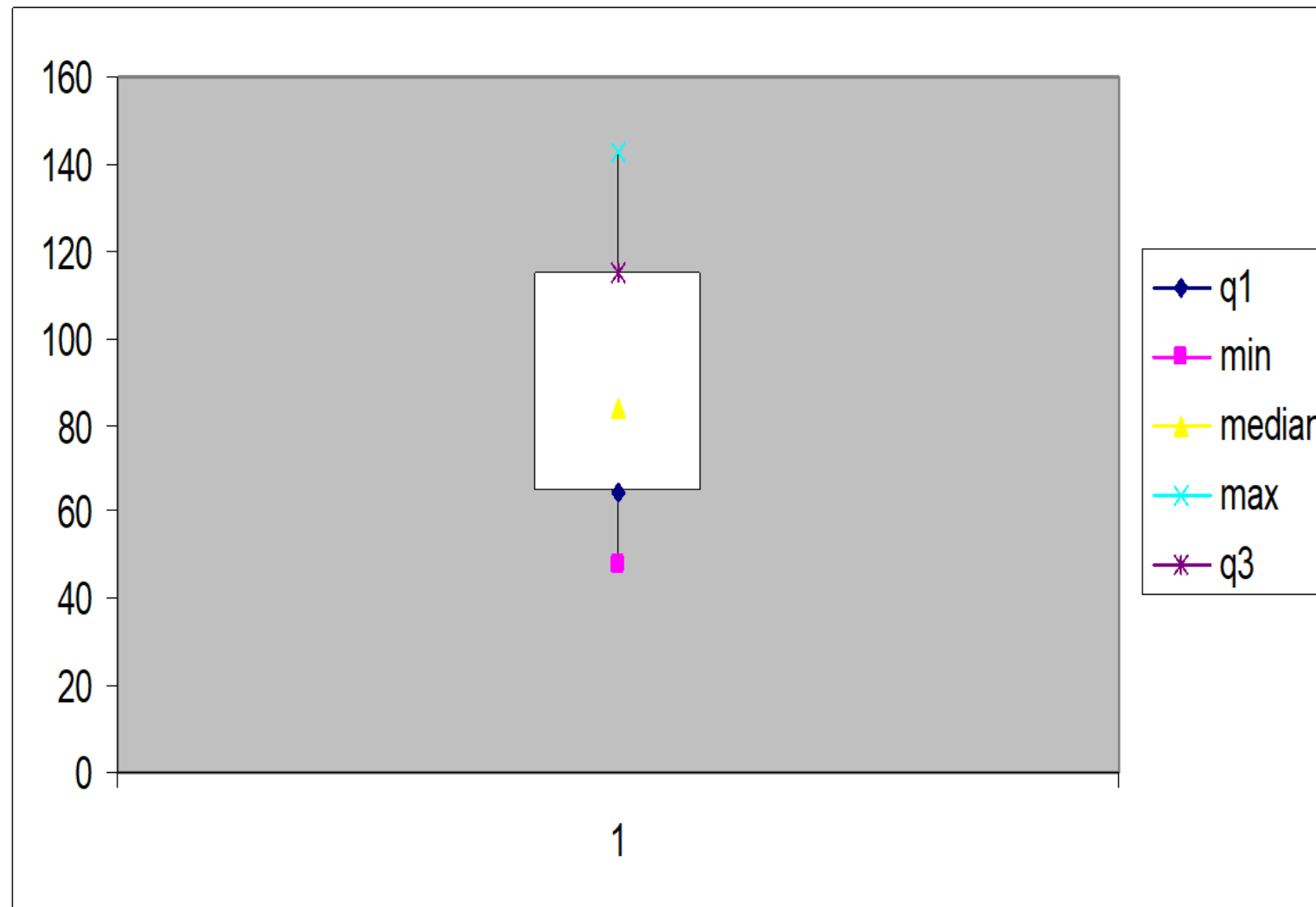


Institute of
Business Administration
Karachi

Leadership and Ideas for Tomorrow

Numerical Data Presentation

Box-Plot: Describes the five-number summary



Numerical Data Presentation

- Central Value: extent to which the central value characterizes the whole set of data.
- Measures of central value such as the mean or median must be coupled with measures of data dispersion to indicate how well the central value characterizes the data as a whole.
- A: 30, 50, 70
- B: 40, 50, 60
 - Mean = 50
 - But spread of B is less than A

Numerical Data Presentation

- Mean: Notation : Let $x_1, x_2, \dots, x_n$ are n observations of a variable x . Then the mean of this variable,

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n}$$

- Median: The middle value in an ordered sequence of observations.
- Mode: Value observed most frequently

Outlier Effect

- The median is less sensitive to outliers (extreme scores) than the mean and thus a better measure than the mean for highly skewed distributions, e.g. family income.
- For example mean of 20, 30, 40, and 990 is $(20+30+40+990)/4 = 270$.
 - The median of these four observations is $(30+40)/2 = 35$.
 - Here 3 observations out of 4 lie between 20-40.
 - So, the mean 270 really fails to give a realistic picture of the major part of the data. It is influenced by extreme value 990.
- Can only be determined after thorough analysis

Variability

- Variability (or dispersion) measures the amount of scatter
- Commonly used methods: range, variance, standard deviation, interquartile range, coefficient of variation etc.
- Range: The difference between the largest and the smallest observations. The range of 10, 5, 2, 100 is (100-2)=98 (crude)
- Variance: Average of the squares of the deviations of the observations from their mean. In symbols, the variance of the n observations $x_1, x_2, \dots, x_n$ is

$$S^2 = \frac{(x_1 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n-1}$$

- Standard Deviation: Square root of the variance.

Variability

- Quartiles: Data can be divided into four regions that cover the total range of observed values. Cut points for these regions are known as quartiles.
- Quartiles of a data is the $((n+1)/4)q$ th observation of the data, where q is the desired quartile and n is the number of observations of data.
- The first quartile (Q1) is the first 25% of the data. The second quartile (Q2) is between the 25th and 50th percentage points in the data. The upper bound of Q2 is the median. The third quartile (Q3) is the 25% of the data lying between the median and the 75% cut point in the data.
- Q1 is the median of the first half of the ordered observations and Q3 is the median of the second half of the ordered observations.

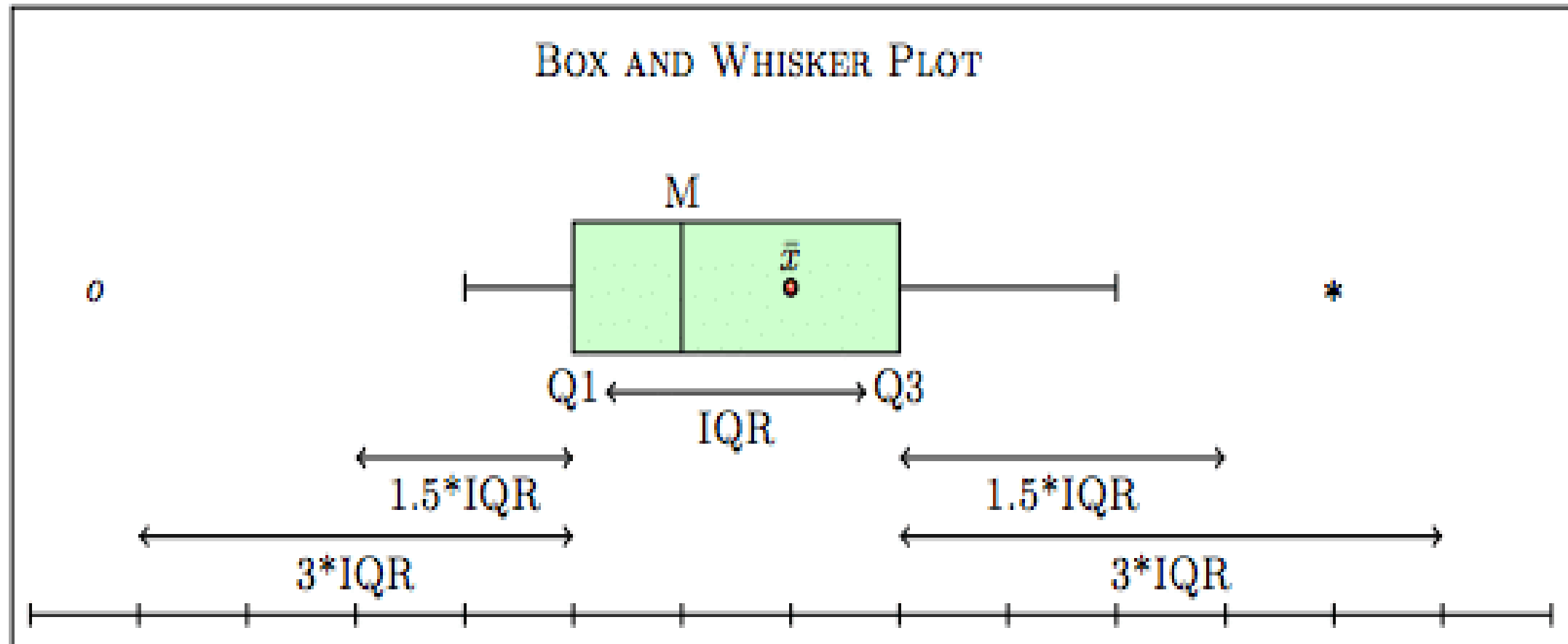
Variability

- An example with 15 numbers

3 6 7 11 13 22 30 40 44 50 52 61 68 80 94
 Q1 Q2 Q3

- The first quartile is $Q1=11$. The second quartile is $Q2=40$ (This is also the Median.) The third quartile is $Q3=61$.
- **Inter-quartile Range:** Difference between $Q3$ and $Q1$. Inter-quartile range of the previous example is $61 - 40 = 21$. The middle half of the ordered data lie between 40 and 61.

BOX AND WHISKER PLOT



Variability

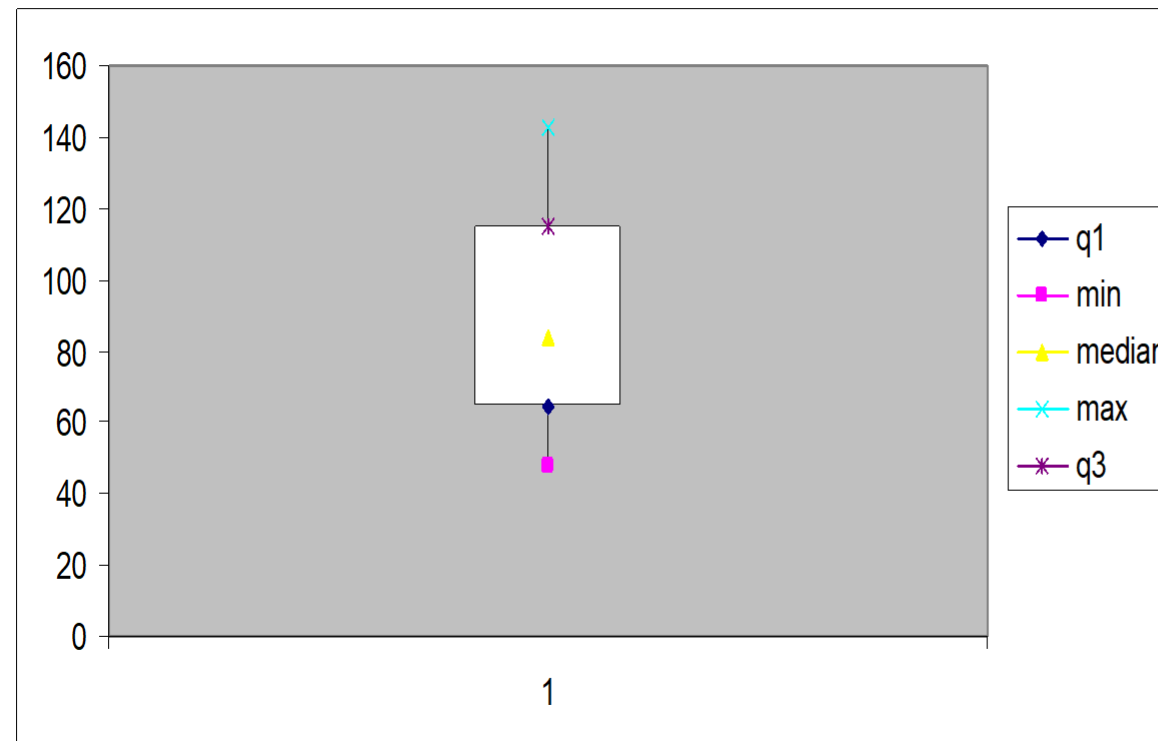
- Deciles: If data is ordered and divided into 10 parts, then cut points are called Deciles
- Percentiles: If data is ordered and divided into 100 parts, then cut points are called Percentiles. 25th percentile is the Q1, 50th percentile is the Median (Q2) and the 75th percentile of the data is Q3.
 - Percentile of a data is the $((n+1)/100)p$ -th observation of the data, where p is the desired percentile and n is the number of observations of data.
- Coefficient of Variation: The standard deviation of data divided by its mean. It is usually expressed in percent.

$$\text{Coefficient of Variation} = \frac{\sigma}{\bar{x}} \times 100$$

Five Number Summary

- The five number summary of a distribution consists of the smallest (Minimum) observation, the first quartile (Q1), the median (Q2), the third quartile, and the largest (Maximum) observation written in order from smallest to largest.

Distribution of Age in Month



Summary Selection

- The five number summary is usually better than the mean and standard deviation for describing a skewed distribution or a distribution with extreme outliers. The mean and standard deviation are reasonable for symmetric distributions that are free of outliers.
- In real life we can't always expect symmetry of the data. It's a common practice to include number of observations (n), mean, median, standard deviation, and range as common for data summarization purpose. We can include other summary statistics like Q_1 , Q_3 , Coefficient of variation if it is considered to be important for describing data.

Shape of Data

- Measured by Kurtosis and Skewness
- Skewness: Measures asymmetry of data
 - Positive or right skewed: Longer right tail
 - Negative or left skewed: Longer left tail

Let $x_1, x_2, \dots, x_n$ be n observations. Then,

$$\text{Skewness} = \frac{\sqrt{n} \sum_{i=1}^n (x_i - \bar{x})^3}{\left(\sum_{i=1}^n (x_i - \bar{x})^2 \right)^{3/2}}$$

Shape of Data

- Measures peakedness of the distribution of data. The kurtosis of normal distribution is 0.

Let $x_1, x_2, \dots, x_n$ be n observations. Then,

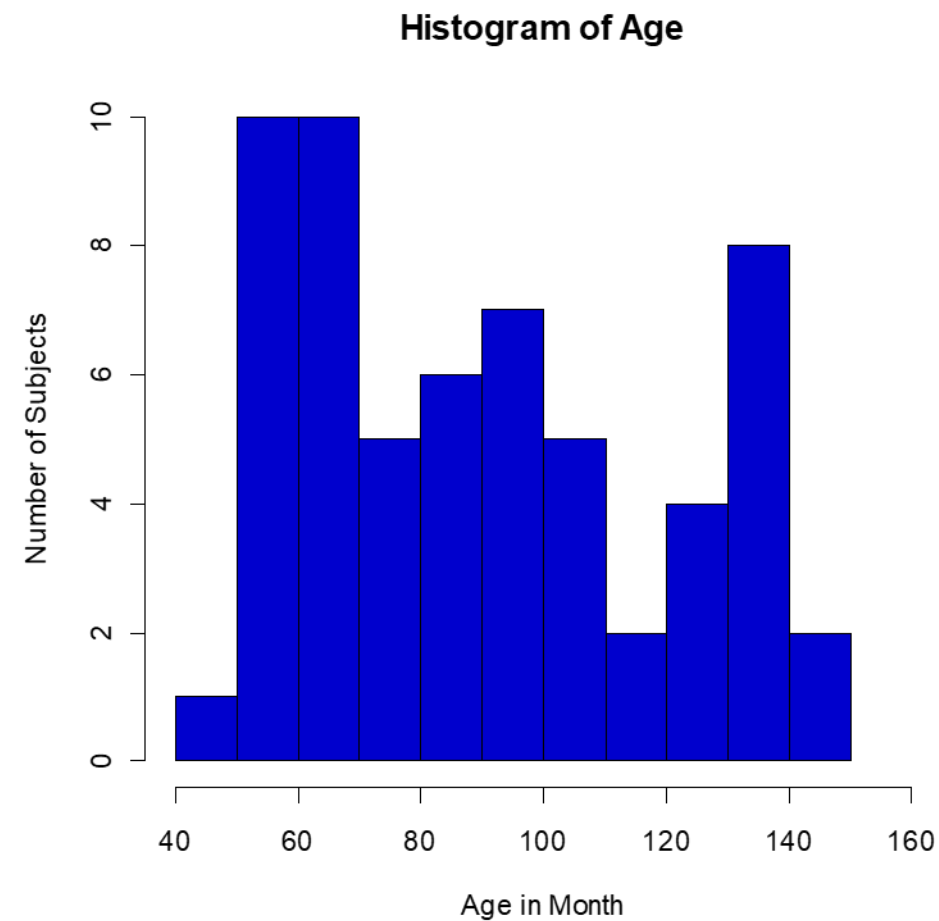
$$\text{Kurtosis} = \frac{n \sum_{i=1}^n (x_i - \bar{x})^4}{\left(\sum_{i=1}^n (x_i - \bar{x})^2 \right)^2} - 3$$

Ideal Values?

- The values for asymmetry and **kurtosis** between -2 and +2 are considered **acceptable** in order to prove normal univariate distribution (George & Mallery, 2010). Hair et al. (2010) and Bryne (2010) argued that data is considered to be normal if **skewness** is between -2 to +2 and **kurtosis** is between -7 to +7.
- The rule of thumb seems to be: If the **skewness** is between -0.5 and 0.5, the data are fairly symmetrical. If the **skewness** is between -1 and -0.5 or between 0.5 and 1, the data are moderately **skewed**. If the **skewness** is less than -1 or greater than 1, the data are highly **skewed**.

Summary of Age

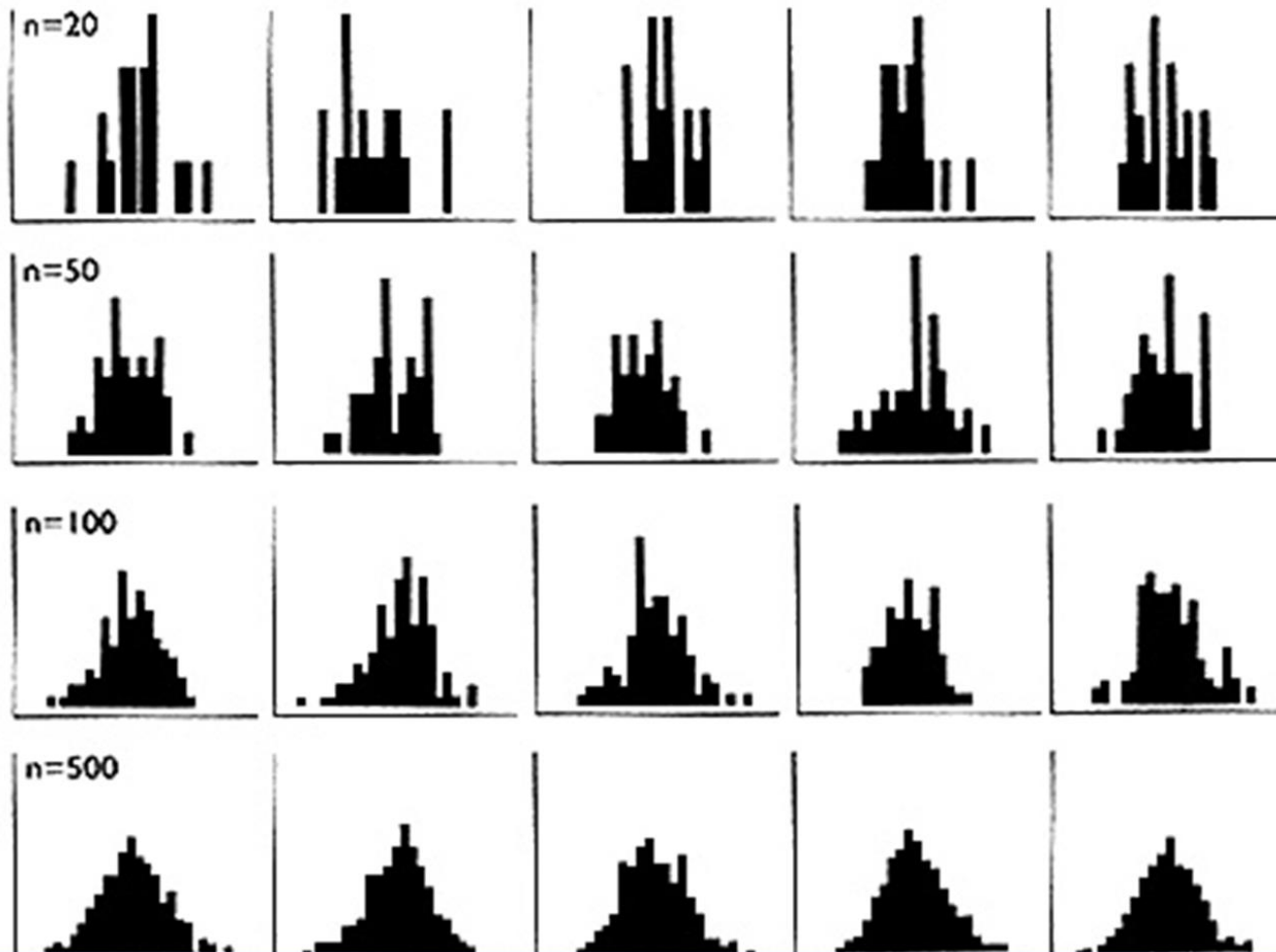
Mean	90.41666667
Standard Error	3.902649518
Median	84
Mode	84
Standard Deviation	30.22979318
Sample Variance	913.8403955
Kurtosis	-1.183899591
Skewness	0.389872725
Range	95
Minimum	48
Maximum	143
Sum	5425
Count	60



Institute of
Business Administration
Karachi

Leadership and Ideas for Tomorrow

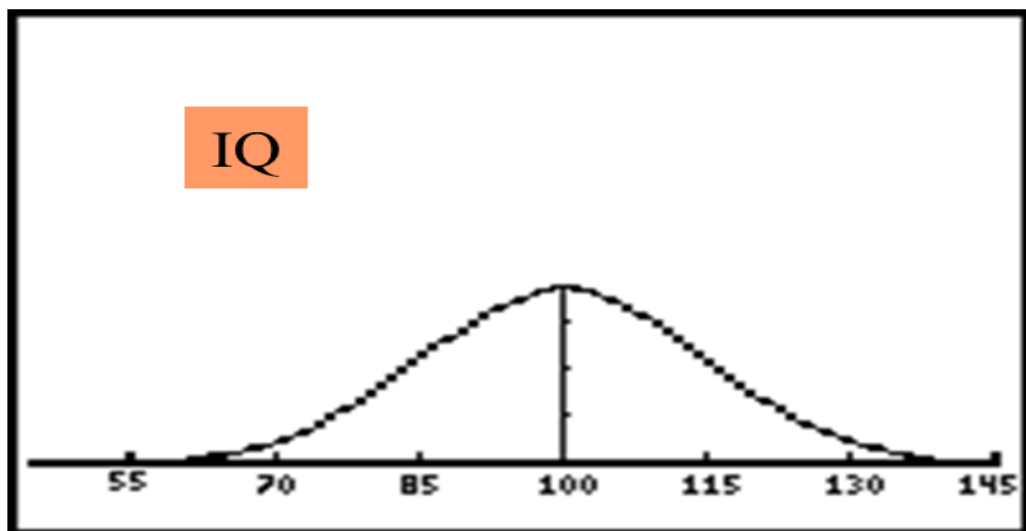
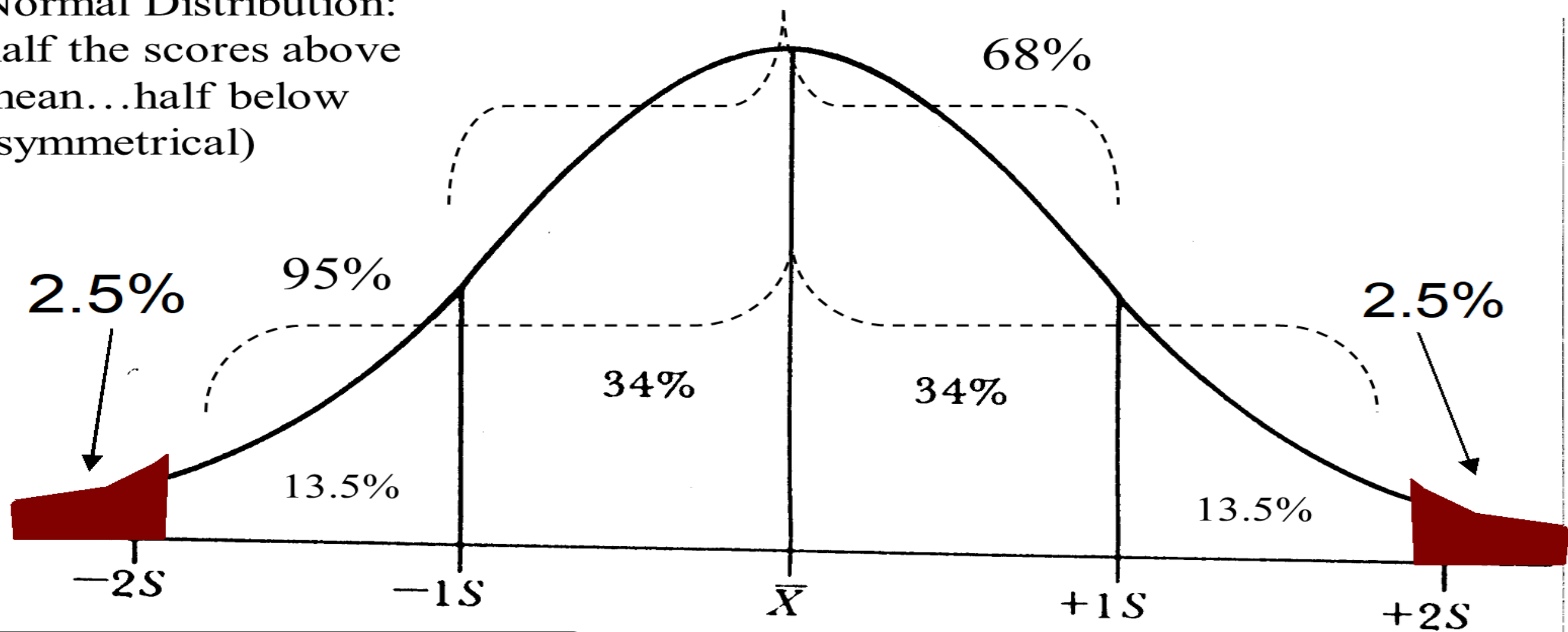
Central Limit Theorem



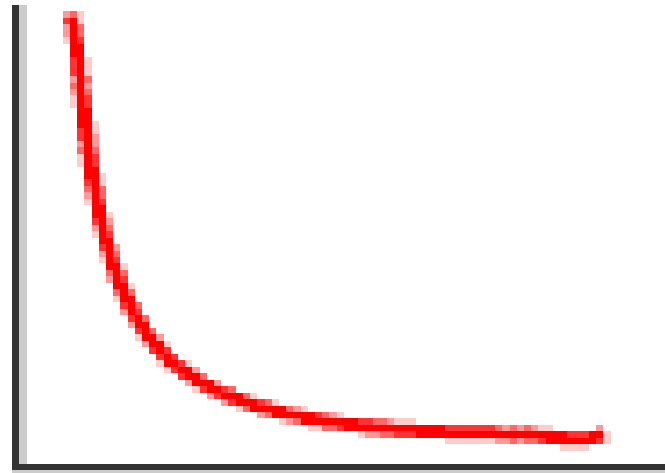
Central Limit Theorem: the larger the sample size, the closer a distribution will approximate the normal distribution or

A distribution of scores taken at random from any distribution will tend to form a normal curve

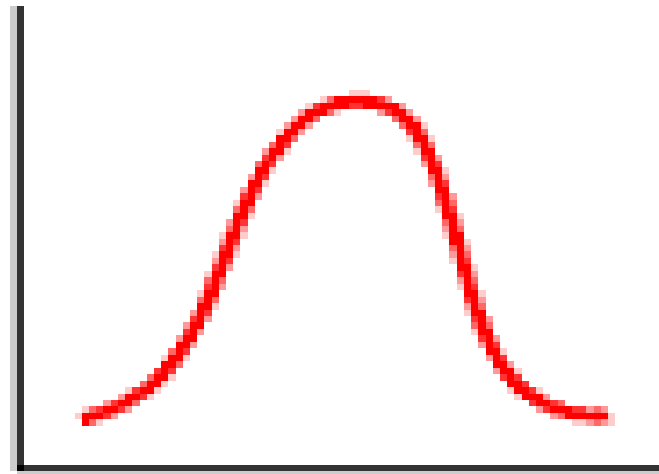
Normal Distribution:
half the scores above
mean...half below
(symmetrical)



body temperature, shoe sizes, diameters of trees,
Wt, height etc...



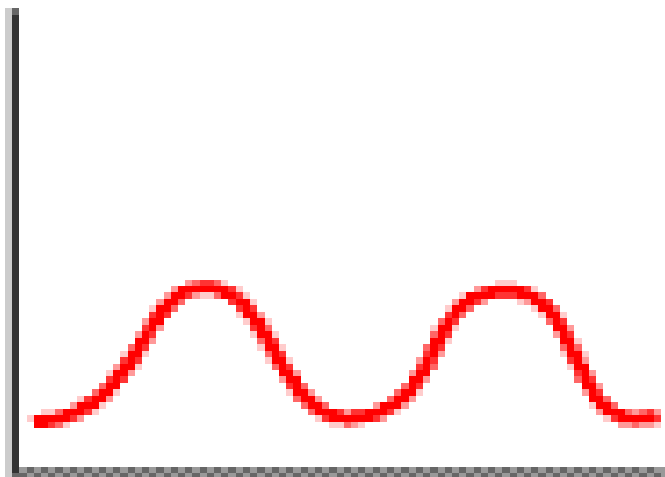
J-shaped



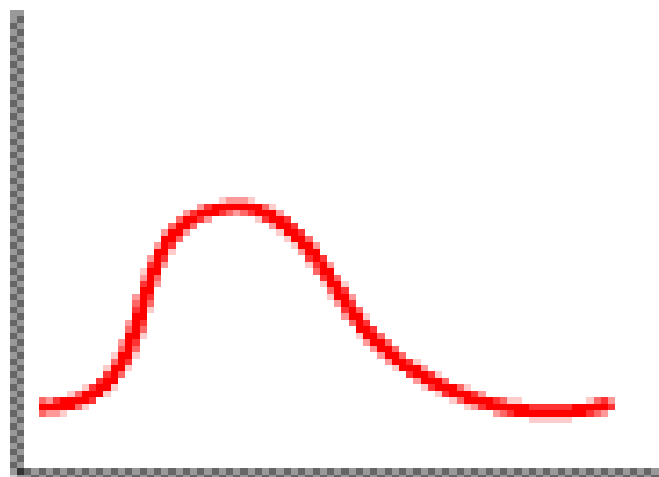
Normal



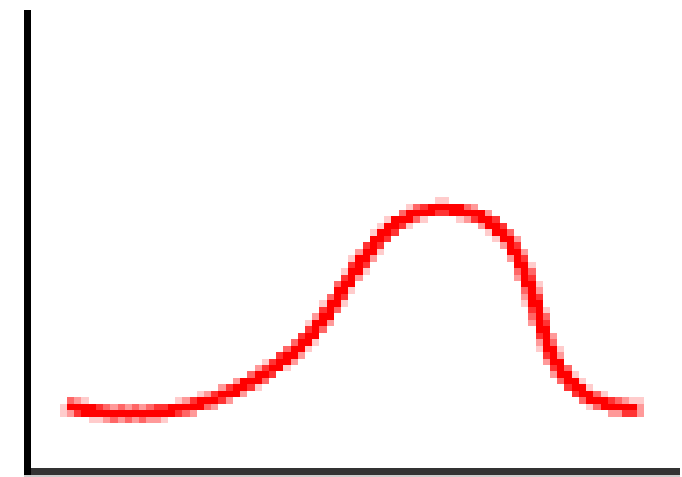
Rectangular



Bimodal

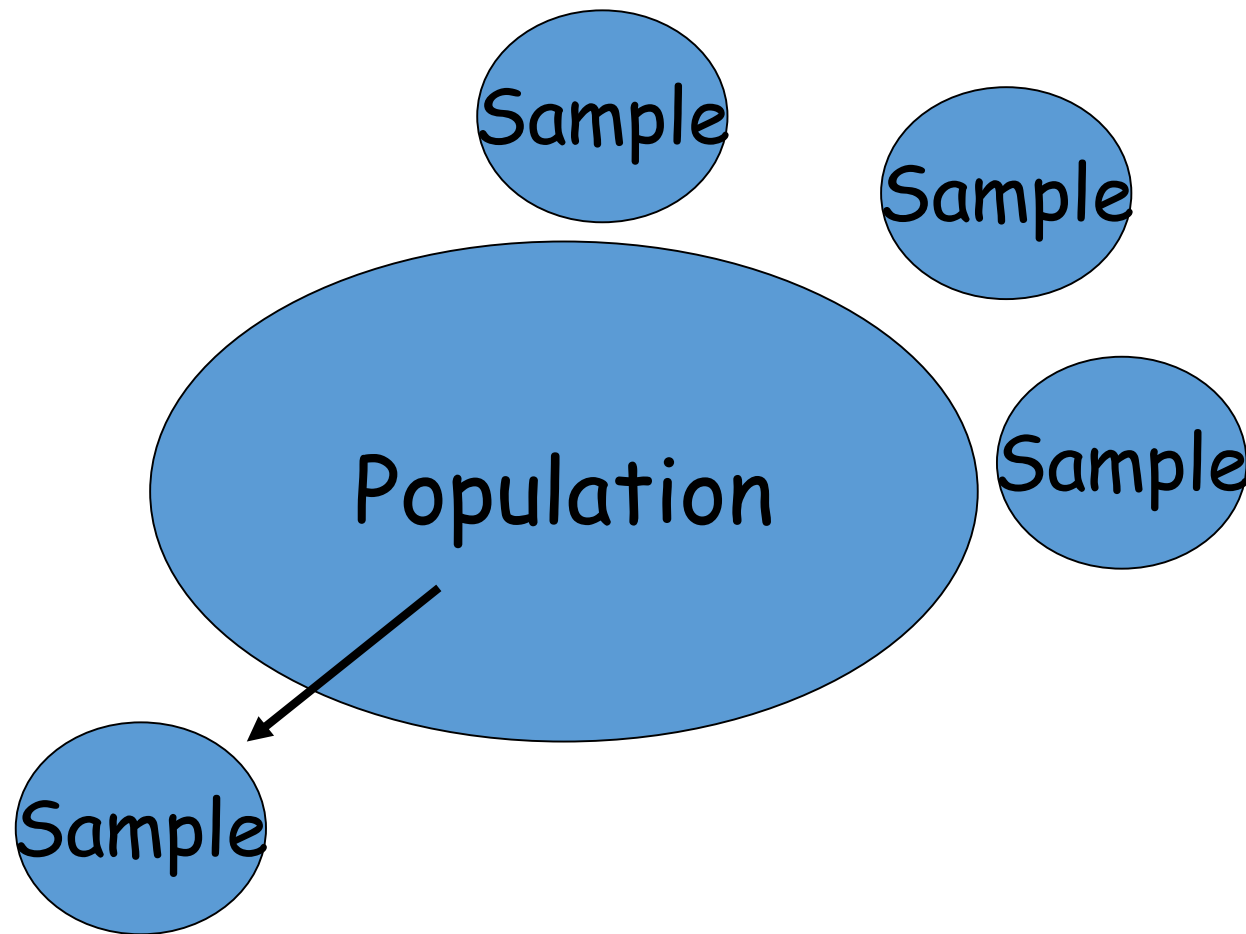


Positive (right) skew



Negative (left) skew

Inferential Statistics



Draw inferences about the
larger group

Sampling Error

Sampling Error: variability among samples due to chance vs population



Or true differences? Are just due to sampling error?
Probability.....

Probability

- Numerical indication of how likely it is that a given event will occur (General Definition) “*hum... what’s the probability it will rain?*”

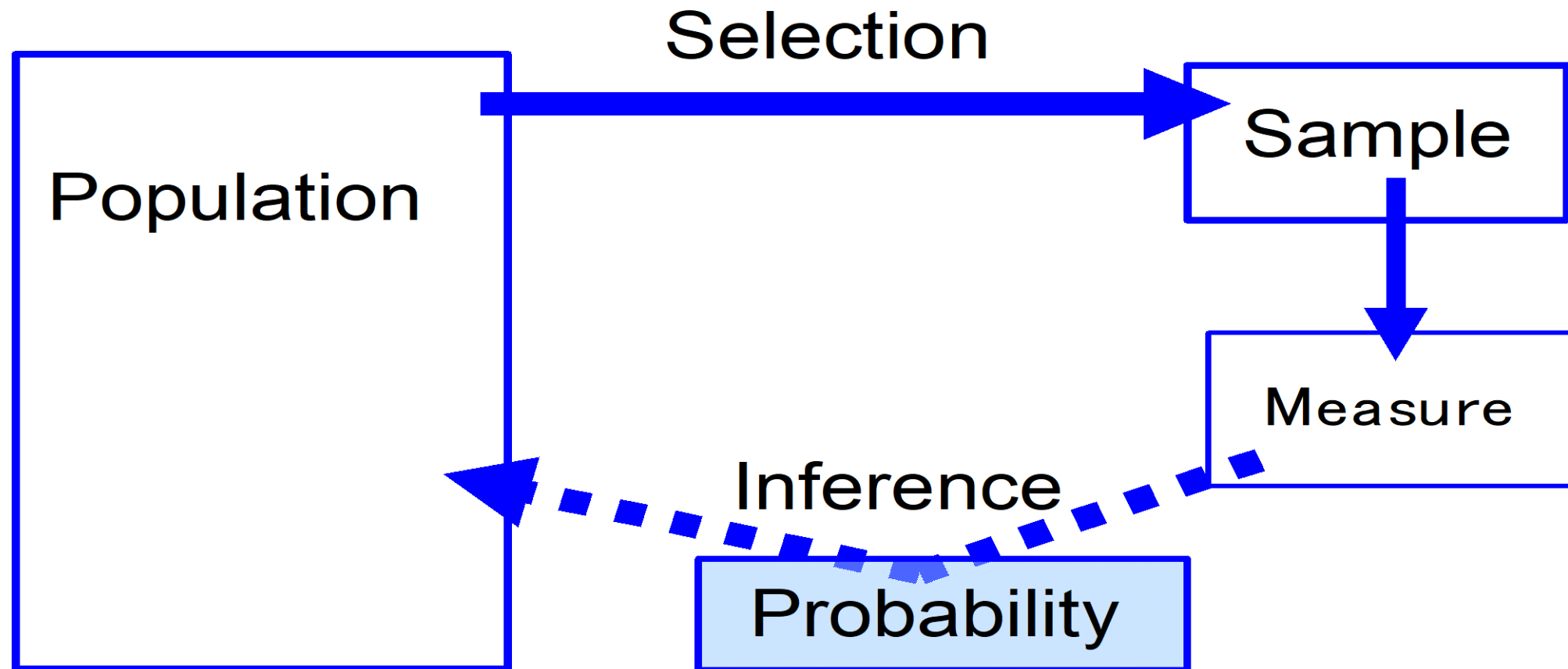


- Statistical probability:

the odds that what we observed in the sample did not occur because of error (random and/or systematic) “*hum... what’s the probability that my results are not just due to chance*”

- In other words, the probability associated with a statistic is the level of confidence we have that the sample group that we measured actually represents the total population

Chain of Reasoning for Inferential Statistics



Are our inferences valid? ...Best we can do is to calculate probability about inferences

Inferential Statistics: uses sample data to evaluate the credibility of a hypothesis about a population



NULL Hypothesis:

NULL (*nullus* - latin): "not any" → no differences between means

$$H_0 : \mu_1 = \mu_2$$

Always testing the null hypothesis

"H- Naught"



Institute of
Business Administration
Karachi

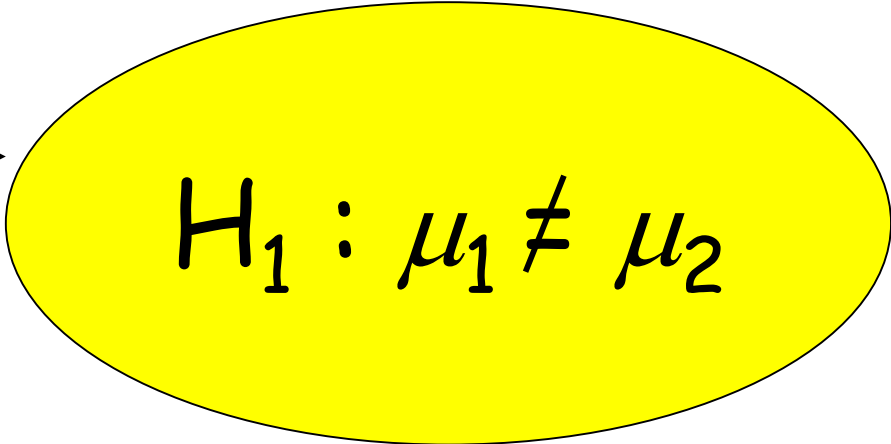
Leadership and Ideas for Tomorrow

Inferential statistics: uses sample data to evaluate the credibility of a hypothesis about a population

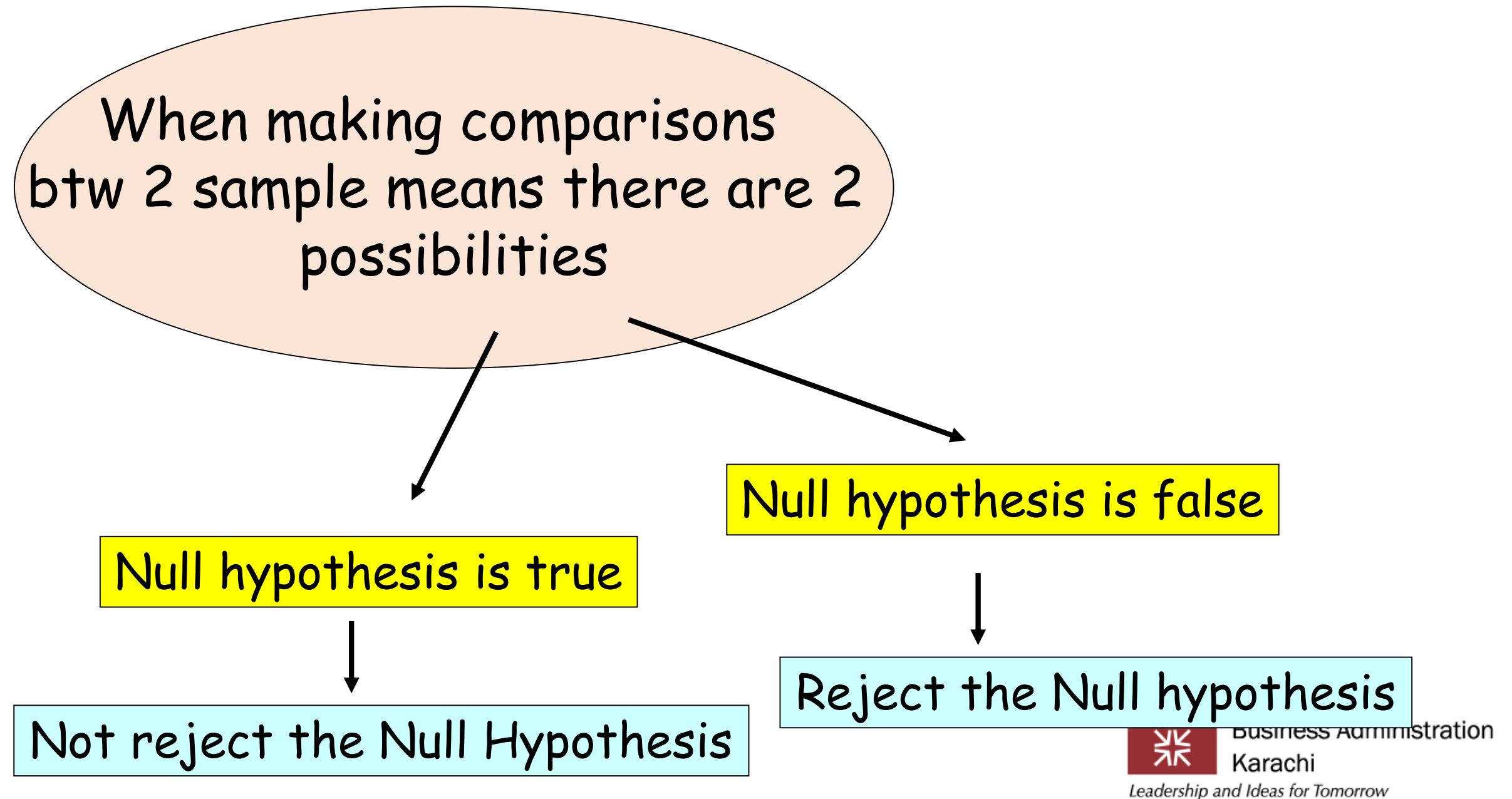


Hypothesis: Scientific or alternative hypothesis

Predicts that there are differences between the groups


$$H_1 : \mu_1 \neq \mu_2$$

Inferential Statistics



Outcomes

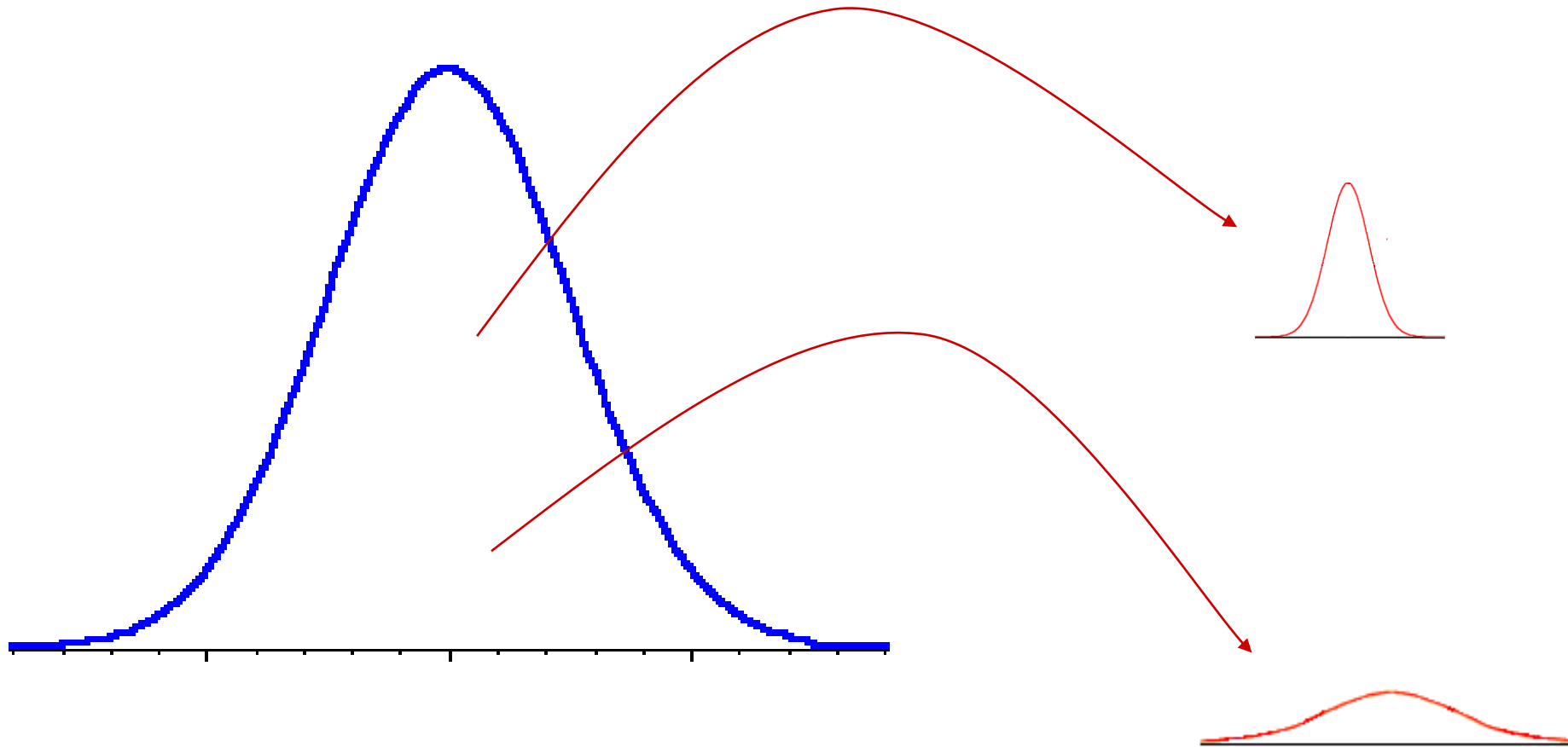
Possible Outcomes in Hypothesis Testing (Decision)

	Null is True	Null is False
Accept	Correct Decision	Error Type II Error
Reject	Error Type I Error	Correct Decision

Type I Error: Rejecting a True Hypothesis

Type II Error: Accepting a False Hypothesis

Issues when making Inferences



P-Values

- So how do we know whether the effect observed in our sample was genuine?
 - We don't
- Instead we use **p values** to indicate our level of certainty that our results represent a genuine effect present in the whole population

the probability of making a type I error → depends on the criterion you use to accept or reject the null hypothesis = **significance level** (smaller you make alpha, the less likely you are to commit error) **0.05** (5 chances in 100 that the difference observed was really due to sampling error - 5% of the time a type I error will occur)

Possible Outcomes in Hypothesis Testing

Alpha (α)

Difference observed is really
just sampling error
The prob. of type one error

Accept

Reject

	Null is True	Null is False
Accept	Correct Decision	Error Type II Error
Reject	Error Type I Error	Correct Decision

Criteria

When we do statistical analysis... if alpha
(p value- significance level) greater than 0.05



WE ACCEPT THE NULL HYPOTHESIS

is equal to or less than 0.05 we



REJECT THE NULL (difference btw means)



Institute of
Business Administration
Karachi

Leadership and Ideas for Tomorrow

Inferential statistics

Used for Testing for Mean Differences

T-test: when experiments include only 2 groups

- a. Independent
- b. Correlated
 - i. Within-subjects
 - ii. Matched

Based on the t statistic (critical values) based on
df & alpha level

Inferential statistics

Used for Testing for Mean Differences

Analysis of Variance (ANOVA): used when comparing more than 2 groups

1. Between Subjects
2. Within Subjects - repeated measures

Based on the f statistic (critical values) based on df & alpha level

More than one IV = factorial (iv=factors)
Only one IV=one-way anova

Hypothetical experiment



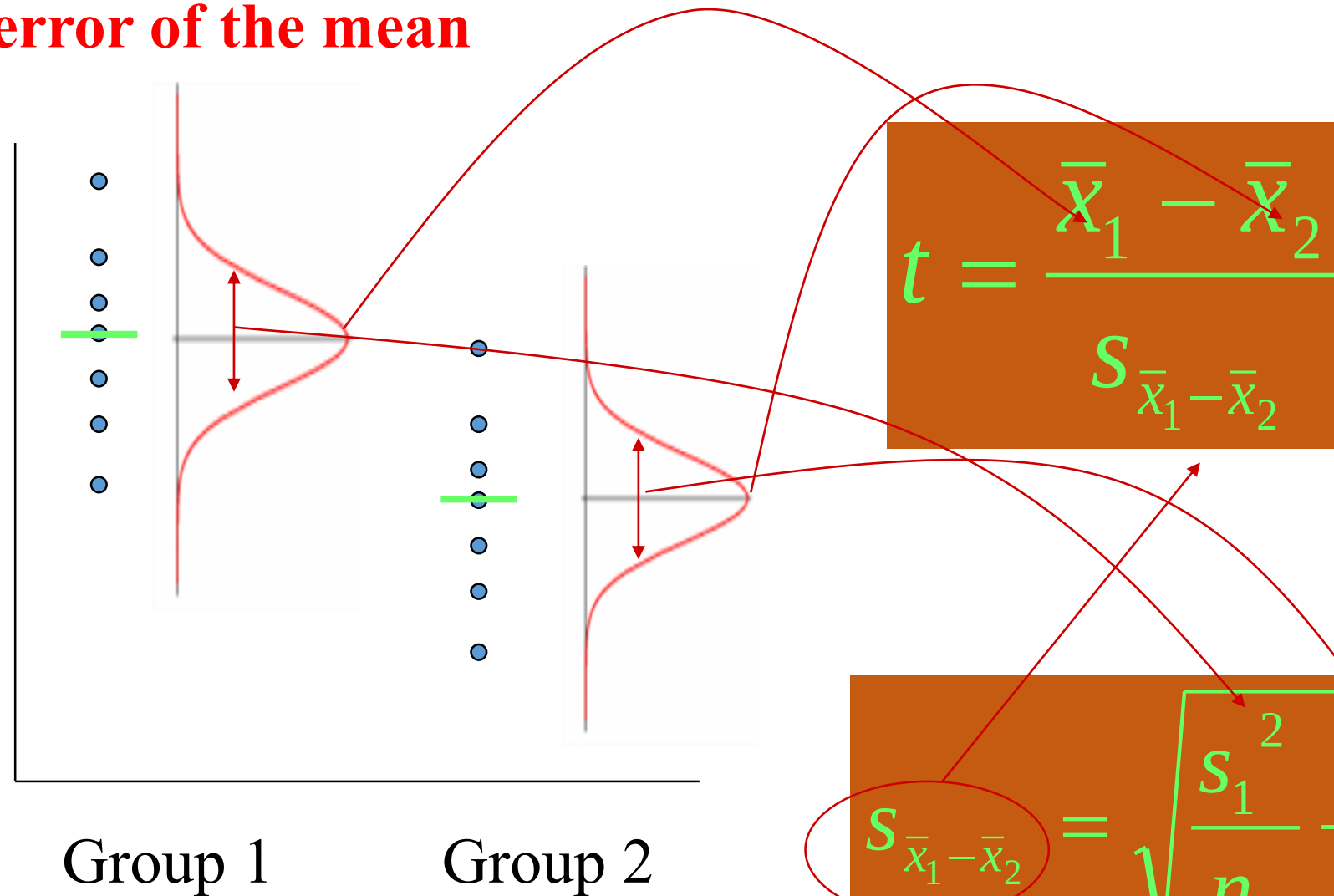
Q – does viewing pictures of the Simpson and the Griffin family activate the same brain regions?

Condition 1 = Simpson family faces

Condition 2 = Griffin family faces

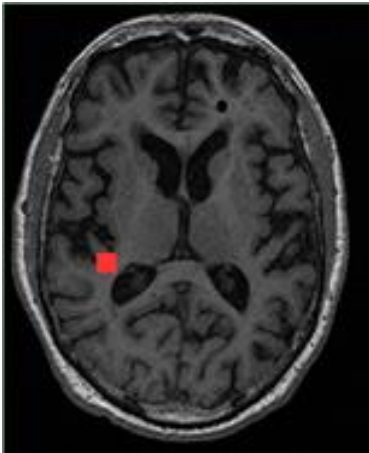
Calculating T

Difference between the means divided by the pooled
standard error of the mean



$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

How do we apply this to fMRI
data analysis?



Simpsons
Family Faces

Griffin Family
Faces

Time

Degrees of freedom

- = number of unconstrained data points
- Which in this case = number of data points – 1.
- Can use t value and df to find the associated p value
- Then compare to the α level

Different types of t-test

- 2 sample t tests
 - Related = two samples related, i.e. same people in both conditions
 - Independent = two independent samples, i.e. diff people in 2 conditions
- One sample t tests
 - compare the mean of one sample to a given value

Another approach to group differences

- Analysis Of VAriance (ANOVA)
 - Variances not means
- Multiple groups
 - e.g. Different facial expressions
- H_0 = no differences between groups
- H_1 = differences between groups

Calculating F

- F = the between group variance divided by the within group variance
 - the model variance/error variance
- for F to be significant the between group variance should be considerably larger than the within group variance

What can be concluded from a significant ANOVA?

- There is a significant difference between the groups
- **NOT** where this difference lies
- Finding exactly where the differences lie requires further statistical analyses

	fertilizer	weight	
1	None	55	
2	None	45	
3	None	46	
4	Biological	64	
5	Biological	52	
6	Biological	42	
7	Chemical	65	
8	Chemical	51	
9	Chemical	66	
10	Chemical	55	

ONE-WAY ANOVA

Population Means Equal?

1 metric outcome variable
3(+) groups of cases

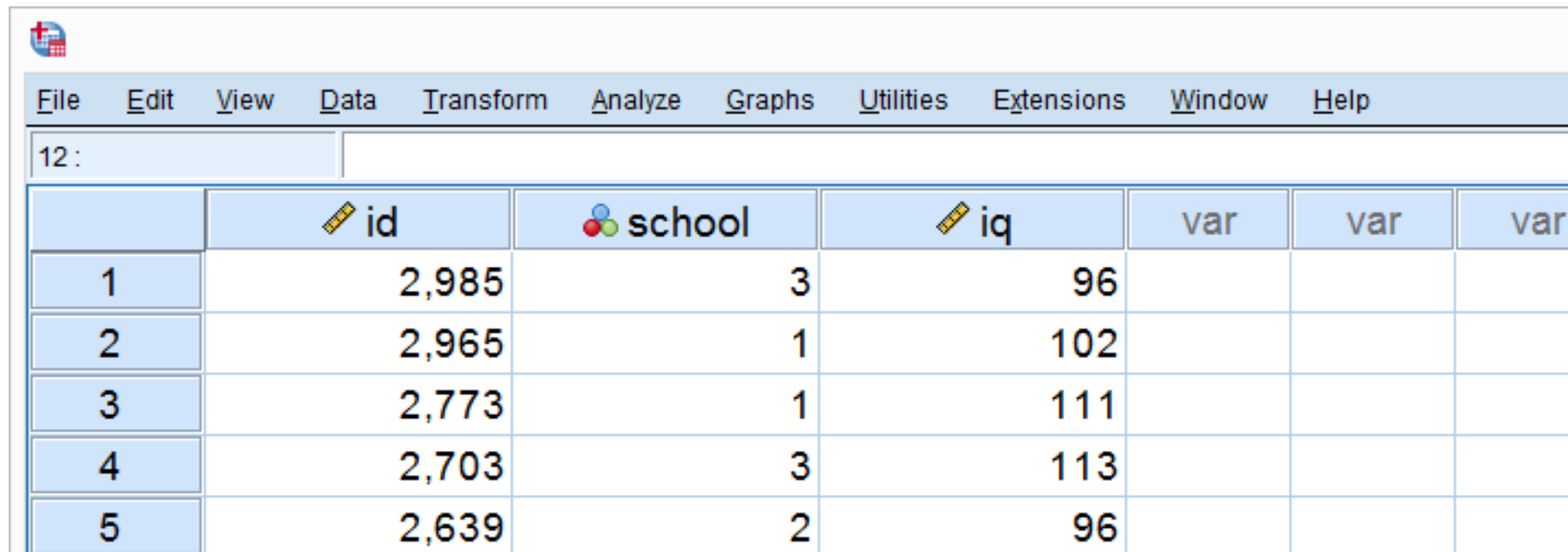


Institute of
Business Administration
Karachi

Leadership and Ideas for Tomorrow

Example

- A scientist wants to know if all children from schools A, B and C have equal mean IQ scores. Each school has 1,000 children. It takes too much time and money to test all 3,000 children. So a simple random sample of $n = 10$ children from each school is tested.



The screenshot shows a software interface with a menu bar (File, Edit, View, Data, Transform, Analyze, Graphs, Utilities, Extensions, Window, Help) and a toolbar. Below the toolbar is a table with 7 columns: an unlabeled index column, 'id', 'school', 'iq', and three 'var' columns. The table contains 5 rows of data. The first row is highlighted in blue.

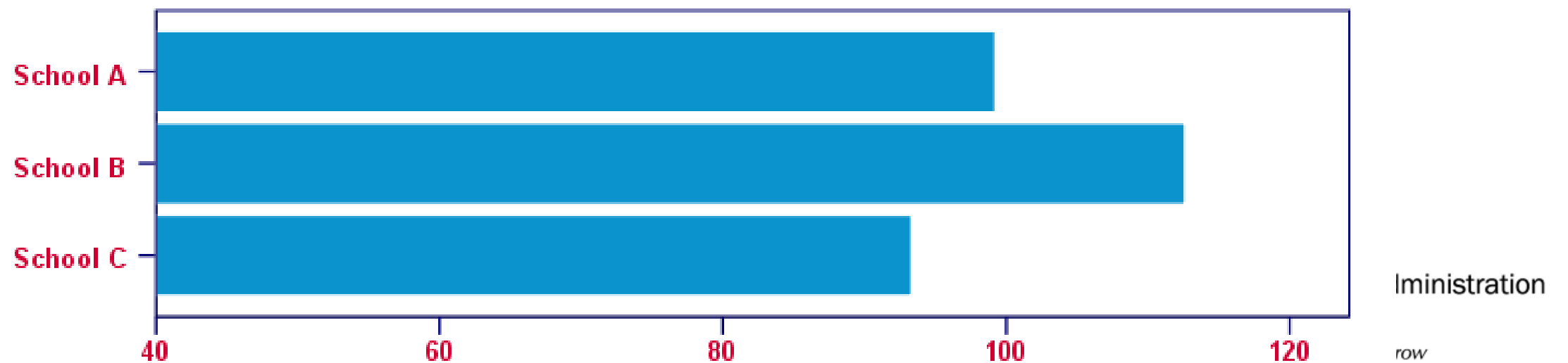
	id	school	iq	var	var	var
1	2,985	3	96			
2	2,965	1	102			
3	2,773	1	111			
4	2,703	3	113			
5	2,639	2	96			

Example

Outcome of IQ test

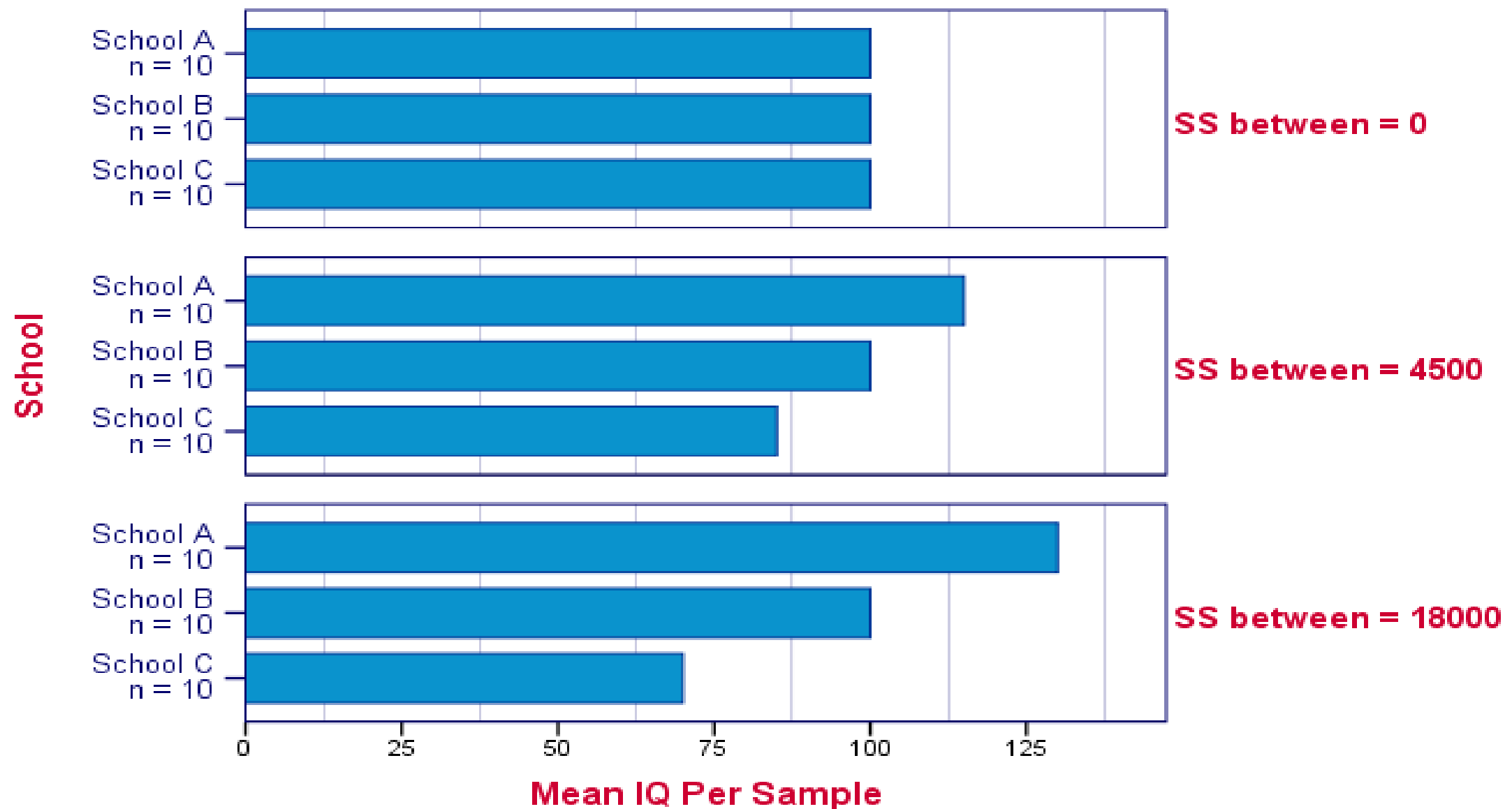
Respondent's school	N	Mean	Variance
School A	10	99.2	111.1
School B	10	112.6	205.6
School C	10	93.3	160.5
Total	30	101.7	215.5

Mean IQ Scores by School (n = 10 per School)



Example

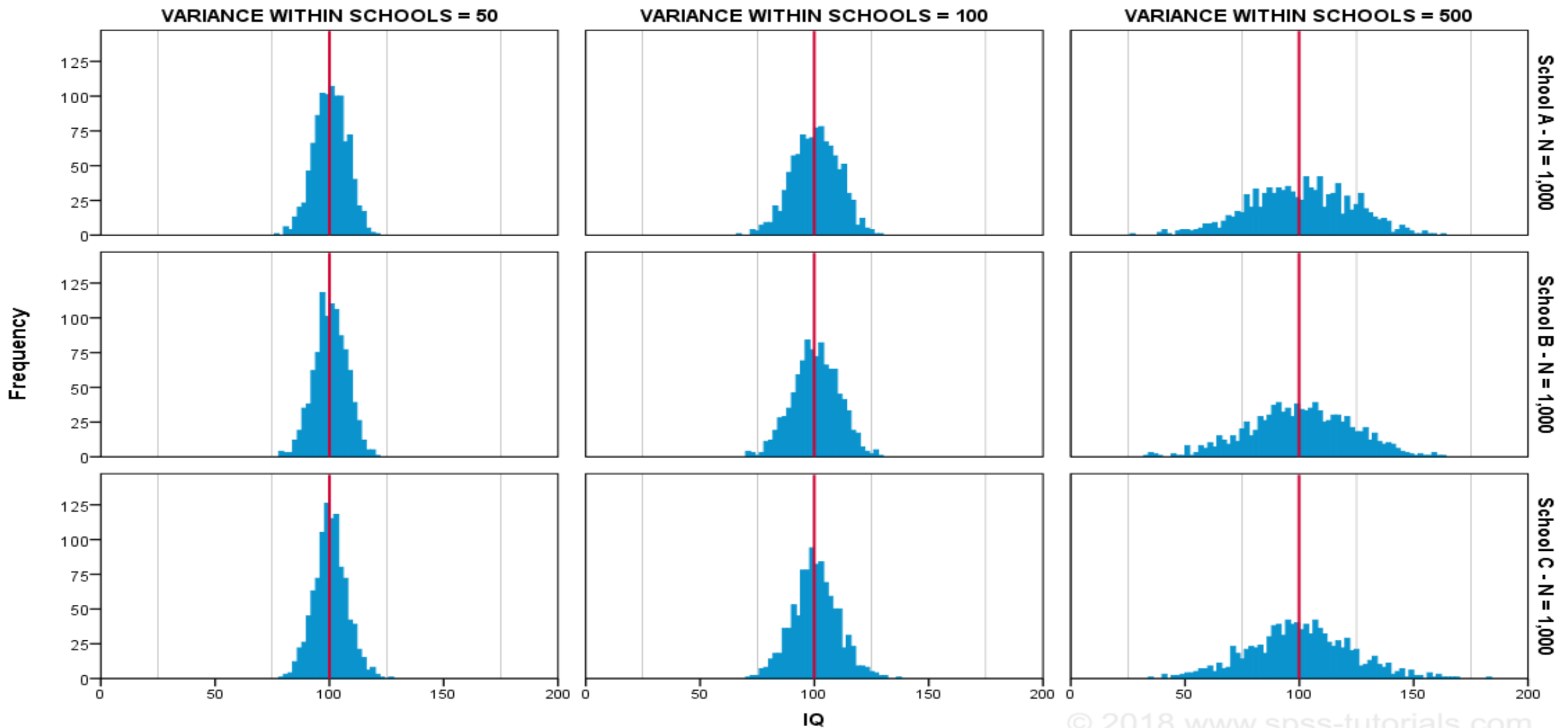
SS BETWEEN INDICATES HOW FAR SAMPLE MEANS LIE APART



Example

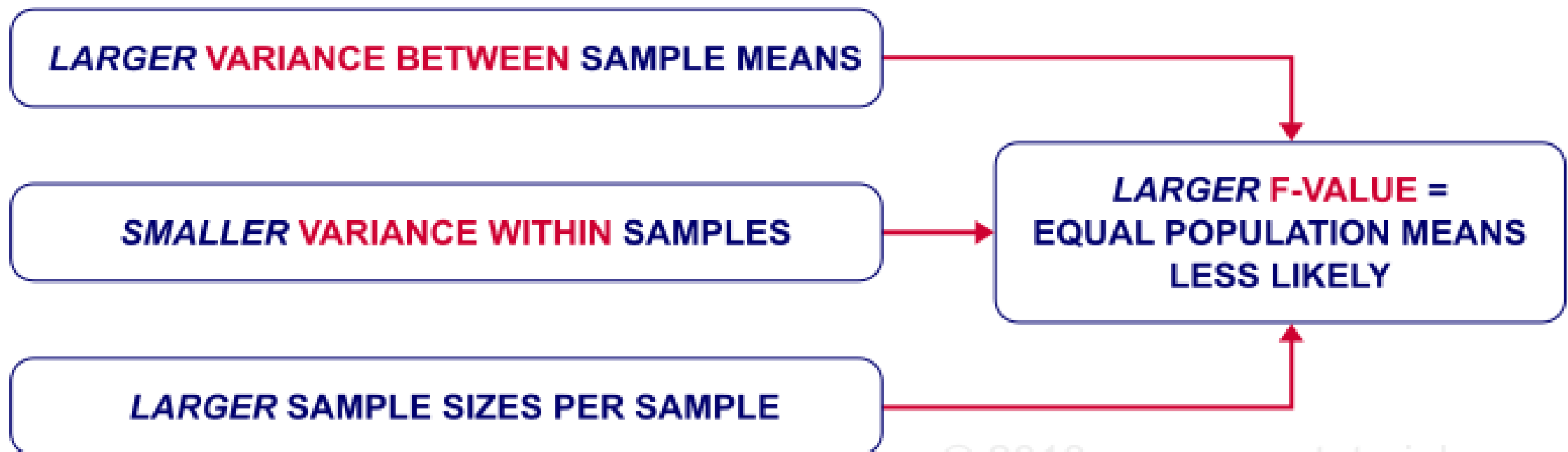
ONE-WAY ANOVA - SUMS OF SQUARES WITHIN GROUPS

MEAN IQ = 100 FOR EACH HISTOGRAM



Rule of Thumb

COMBINING VARIANCES INTO TEST STATISTIC F



© 2018 www.spss-tutorials.com



Institute of
Business Administration
Karachi

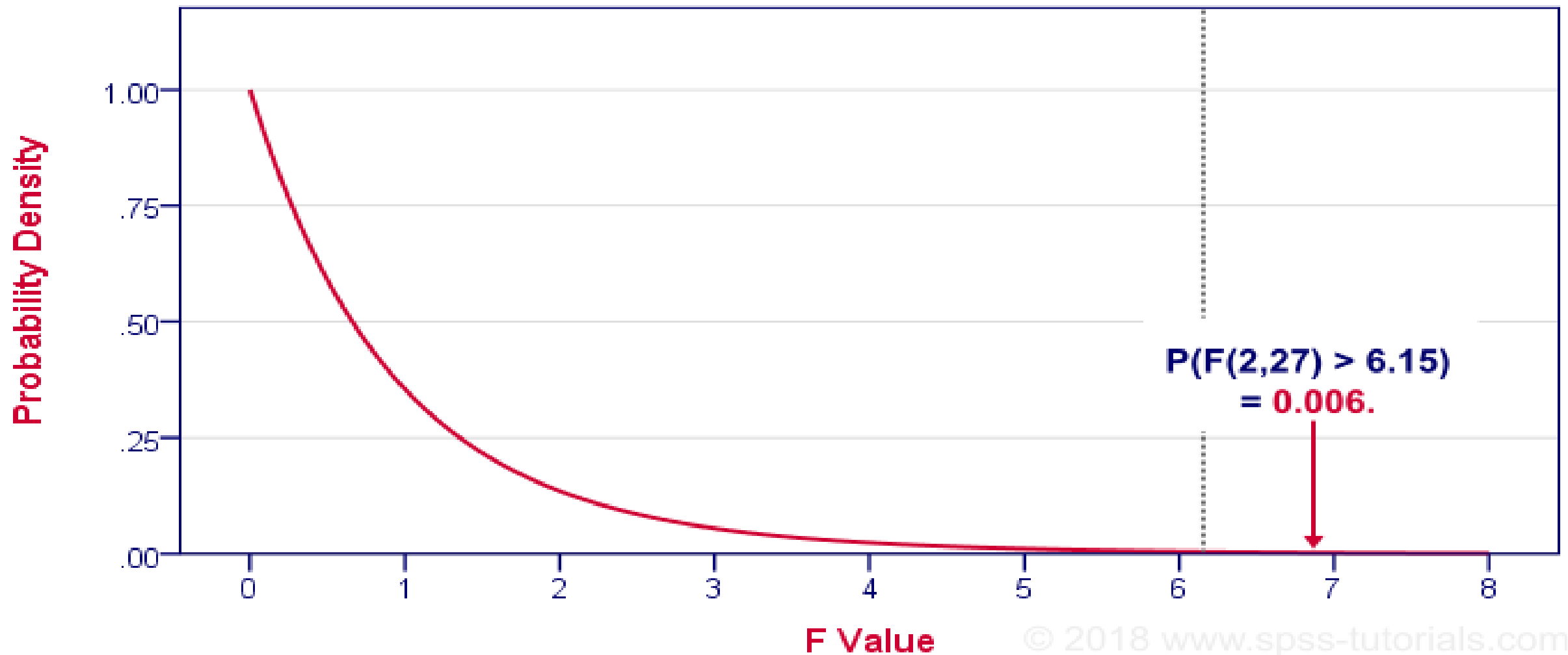
Leadership and Ideas for Tomorrow

Anova Assumptions

- Independent observations;
- **normality**: the outcome variable must follow a normal distribution in each subpopulation. Normality is really only needed for small sample sizes, say $n < 20$ per group.
- **homogeneity**: the variances within all subpopulations must be equal. Homogeneity is only needed if sample sizes are very unequal. In this case, Levene's test indicates if it's met.
- If these assumptions hold, then F follows an F-distribution

F-Distribution

F DISTRIBUTION (DF1 = 2, DF2 = 27)



Different types of ANOVA

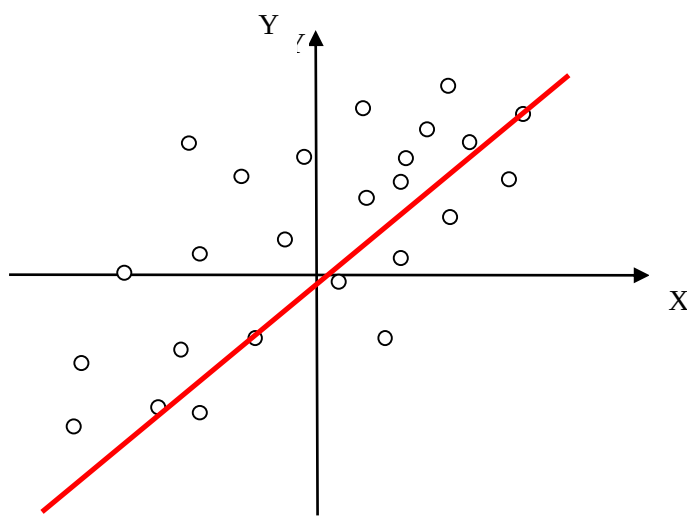
- One-way ANOVA
 - One factor with more than 2 levels
- Factorial ANOVAs (Two-Way)
 - More than 1 factor
- Mixed design ANOVAs
 - Some factors independent, others related

Conclusions

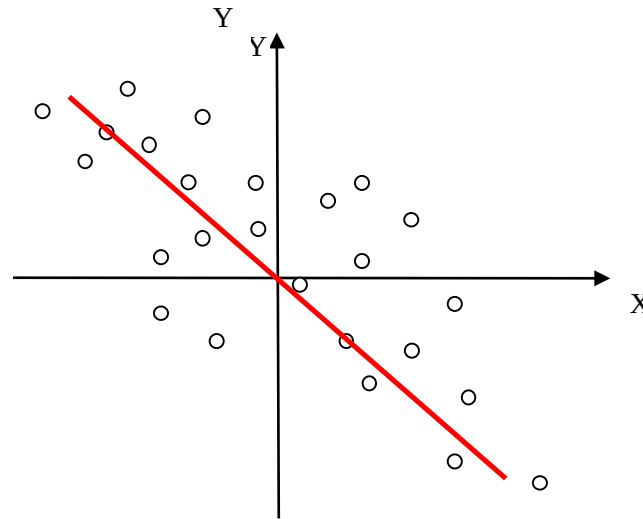
- T-tests assess if two group means differ significantly
- Can compare two samples or one sample to a given value
- ANOVAs compare more than two groups or more complicated scenarios
- They use variances instead of means

Correlation

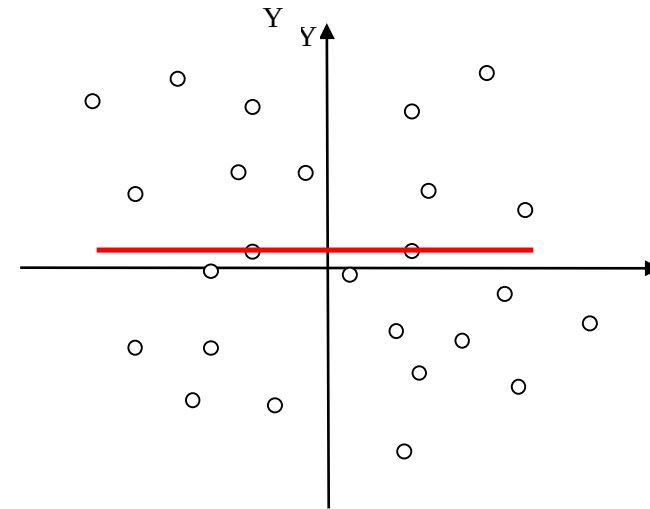
- Strength and direction of the relationship between variables
- Scattergrams



Positive correlation



Negative correlation



No correlation

Describe correlation: covariance

- A statistic representing the degree to which 2 variables vary together

- Covariance formula

$$\text{cov}(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n}$$

- cf. variance formula

$$S_x^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$$

but...

- the absolute value of $\text{cov}(x, y)$ is also a function of the standard deviations of x and y .

Describe correlation: Pearson correlation coefficient (r)

- Equation

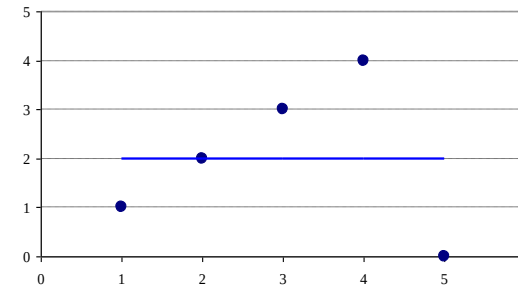
$$r_{xy} = \frac{\text{cov}(x, y)}{s_x s_y}$$

s = st dev of sample

- r = -1 (max. negative correlation); r = 0 (no constant relationship); r = 1 (max. positive correlation)

- Limitations:

- Sensitive to extreme values, e.g.



- r is an estimate from the sample, but does it represent the population parameter?
- Relationship not a prediction.