WHAT STATISTICS ARE

- **This involves working with statistics generated during data processing.**
- **In the presentation and analysis of data, we always make a distinction between different understandings of what statistics are:**
- **In the plural sense, statistics are just numbers and figures collectively referred to as data.**

WHAT STATISTICS ARE

- **In the singular sense, statistics are techniques and methods of data analysis.**

- **There are two aspects to the methods and techniques, namely:**

- **Descriptive statistics**
- **Inferential statistics**

THE IMPORTANCE OF STATISTICS

- **Statistical literacy is necessary because it possible for us to read, understand , and evaluate research findings critically and intelligently.**

- **A grasp of statistics enables researchers to undertake research on their own account with minimum difficulty.**

- **With such an understanding, the researcher can conduct research more confidently, make recommendations with greater knowledge and certainty.**

COMMON USES OF STATISTICS

- **STATISTICS are used in solving most PRACTICAL PROBLEMS.**
- **They are therefore useful in the following areas:**
- **RESEARCH, ADMINISTRATION, and CONSULTANCY**

EXAMPLES OF PRACTICAL APPLICATIONS:

- **QUALITY CONTROL**
-
- **Manufacturing of good for export in the SADC region (ZAMBIA BUREAU OF STANDARDS)**
-
- **MARKETING RESEARCH**
-
- **Determination of the popularity of a new product (eg, soft drink)**
-
- **OPINION POLLING**
-
- **Prediction of the probable outcome of an election.**
-
- **MEDICAL RESEARCH**
-
- **Testing the efficacy of an AIDS vaccine**

LIMITATIONS OF STATISTICS

- **STATISTICS is not a PANACEA.**
- **STATISTICS is simply a TOOL that can be used to deal with problems that are amenable to QUANTIFICATION.**
- **STATISTICS cannot deal with innumerable situations that do lend themselves to QUANTIFICATION or STATISTICAL ANALYSIS.**

STAGES IN DATA ANALYSIS

1. **DATA COLLECTION**
2. **EDITING OF QUESTIONNAIRES**
3. **PROCESSING OF DATA**
4. **DATA CLEANING**
5. **CHOICE OF STATISTICAL METHOD**

    - **TYPES OF VARIABLES**
    - **MEASUREMENT SCALES**
    - **NUMBER OF VARIABLES**
    - **SAMPLING DESIGN**

6. **DATA ANALYSIS**

DESCRIPTIVE STATISTICS

- **Descriptive statistics involve organizing, summarizing, and describing the data.**

- **ORGANIZATION OR CLASSIFICATION OF DATA**

- **Once data has been collected, it can be organized in terms of a frequency distribution.**

- **Organization primarily involves converting raw data into sorted data in the form of an array.**

- **On the basis of this, graphs and simple statistics can be generated.**

FREQUENCY DISTRIBUTION

- **One of the basic things is to represent data through frequency distributions.**
- **This is a description of data presented in tabular form so the data will be more manageable.**
- **It gives the number of times a particular observation or measurement appears in a distribution**.

TYPES OF FREQUENCY DISTRIBUTIONS

- **FREQUENCIES**

- **Use these if all you want to know is how many individuals responded to the question on any issue.**

There are two types of frequency distributions:

UNGROUPED FREQUENCY DISTRIBUTION – This shows how many times a single observation or measurement appears on its own.

GROUPED FREQUENCY DISTRIBUTION – This shows how many times grouped observations or measurements appear together.

UNGROUPED FREQUENCY DISTRIBUTION

x	f
22	1
26	1
27	1
31	1
32	1
33	1
35	1
36	3
40	1
41	2
43	1
44	1
46	1
47	1
48	1
49	1
50	1
51	1
53	1
55	1
58	1
62	1
Total	25

GROUPED FREQUENCY DISTRIBUTION

Class Interval	Frequency
20-24	1
25-29	2
30-34	3
35-39	4
40-44	5
45-49	4
50-54	3
55-59	2
60-64	1
Total	25

THE RELATIVE FREQUENCY DISTRIBUTION

- **This shows the percentage of the total number observed at each score value (for ungrouped data) and four each class interval (for grouped data).**

- **It is useful in making comparisons because using absolute frequency distributions can be misleading.**

RELATIVE FREQUENCY DISTRIBUTION

	CBU		UNZA	
Class Interval	Frequency	%	Frequency	%
19.5 - 24.5	1	4	2	4
24.5 - 29.5	2	8	4	8
29.5 - 34.5	3	12	6	12
34.5 - 39.5	4	16	8	16
39.5 - 44.5	5	20	10	20
44.5 - 49.5	4	16	8	16
49.5 - 54.5	3	12	6	12
54.5 -59.5	2	8	4	8
59.5 – 64.5	1	4	2	4
Total	25	100	50	100

THE CUMULATIVE AND DECUMULATIVE FREQUENCY DISTRIBUTION

- **This shows the percentage on number of observations located below or above a certain limit.**
- **To know the percentage or number of observations below a certain limit - use a cumulative frequency or “less than” distribution**
- **To do this, cumulate the values down.**
- **To know the percentage or number of observations above a certain limit - use a decumulative frequency or “greater than” distribution.**
- **To do this, cumulate the values up.**

CUMULATIVE AND DECUMULATIVE FRQUENCY DISTRIBUTION

Class interval	Frequency	%	Less than		Greater than	
20-24	1	4	1	4	25	100
25-29	2	8	3	12	24	96
30-34	3	12	6	24	22	88
35-39	4	16	10	40	19	76
40-44	5	20	15	60	15	60
45-49	4	16	19	76	10	40
50-54	3	12	22	88	6	24
55-59	2	8	24	96	3	12
60-64	1	4	25	100	1	4
Total	25	100				

THE CUMULATIVE AND DECUMULATIVE FREQUENCY DISTRIBUTION

- **INTERPRETATION**
- **Cumulative Distribution:**
- **Use the true upper limit as the point of comparison thus:**
- **15 or 60% of the UNZA employees were aged below 44.5 years.**

- **<u>Decumulative distribution</u>**

- **Use the true lower limit as the point of comparison thus:**

- **22 or 88% of the UNZA employees were aged above 29.5 years.**

GRAPHICAL PRESENTATION OF DATA

- **Once data is organized in the form of frequency distributions, it can then be summarized and then presented graphically.**

This can be done using graphical descriptive methods such as:

- **Pie charts and bar charts for categorical data.**
- **Categorical data refers to data measured on a nominal or ordinal scale.**
- **Frequency histograms and polygons for continuous data.**
- **Continuous data refers to data measured on an interval or ratio scale.**

THE PIE CHART

- **One of the simplest ways to represent the data if it is qualitative data is in the form of a pie chart.**

- **A pie chart is used to display the percentage of the total number of observations (or measurements) falling into each of the categories of variable by partitioning a circle must as one would slice a pie.**

THE PIE CHART

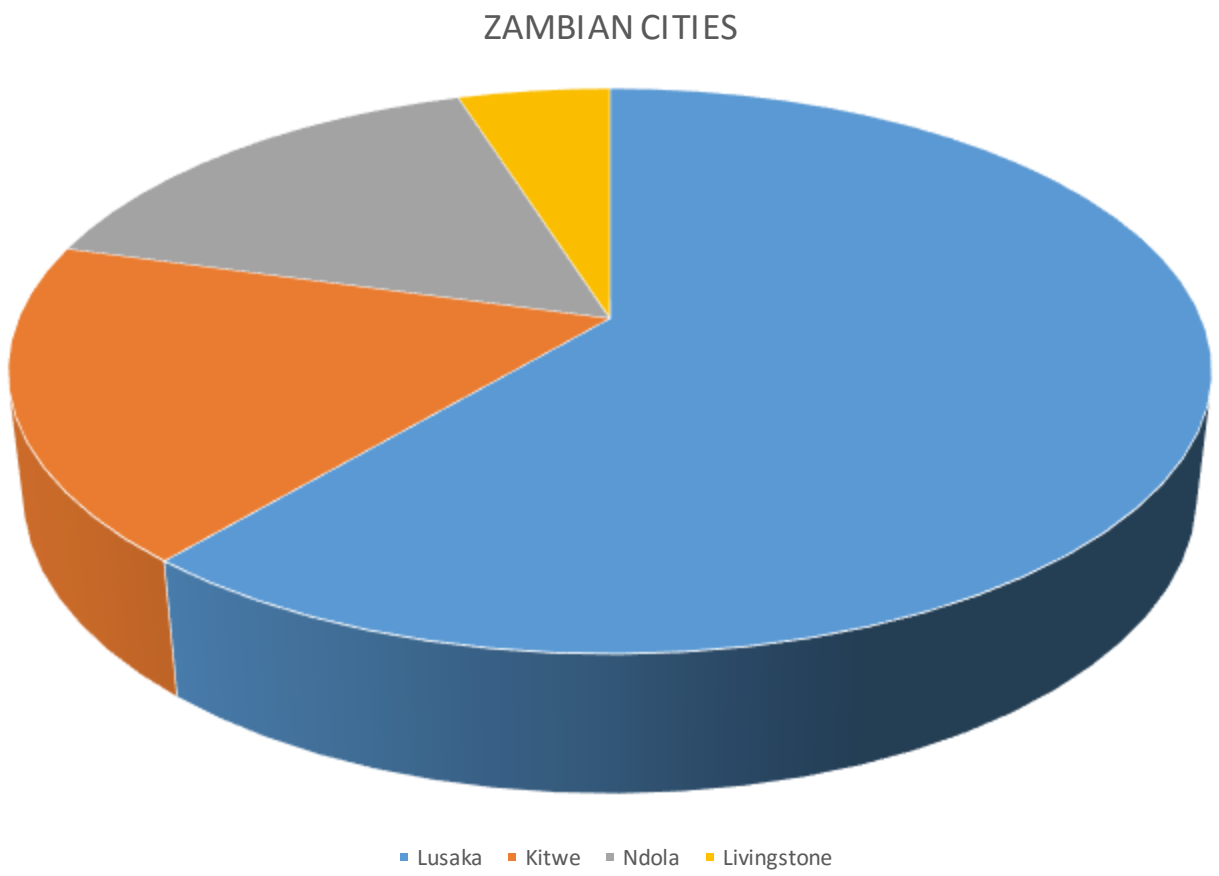
- **In constructing a pie chart, the following guidelines should be followed:-**

- **Choose a small number of categories for the variable, say, around 5 or 6.**

- **Too many categories make a pie char difficult to interpret.**

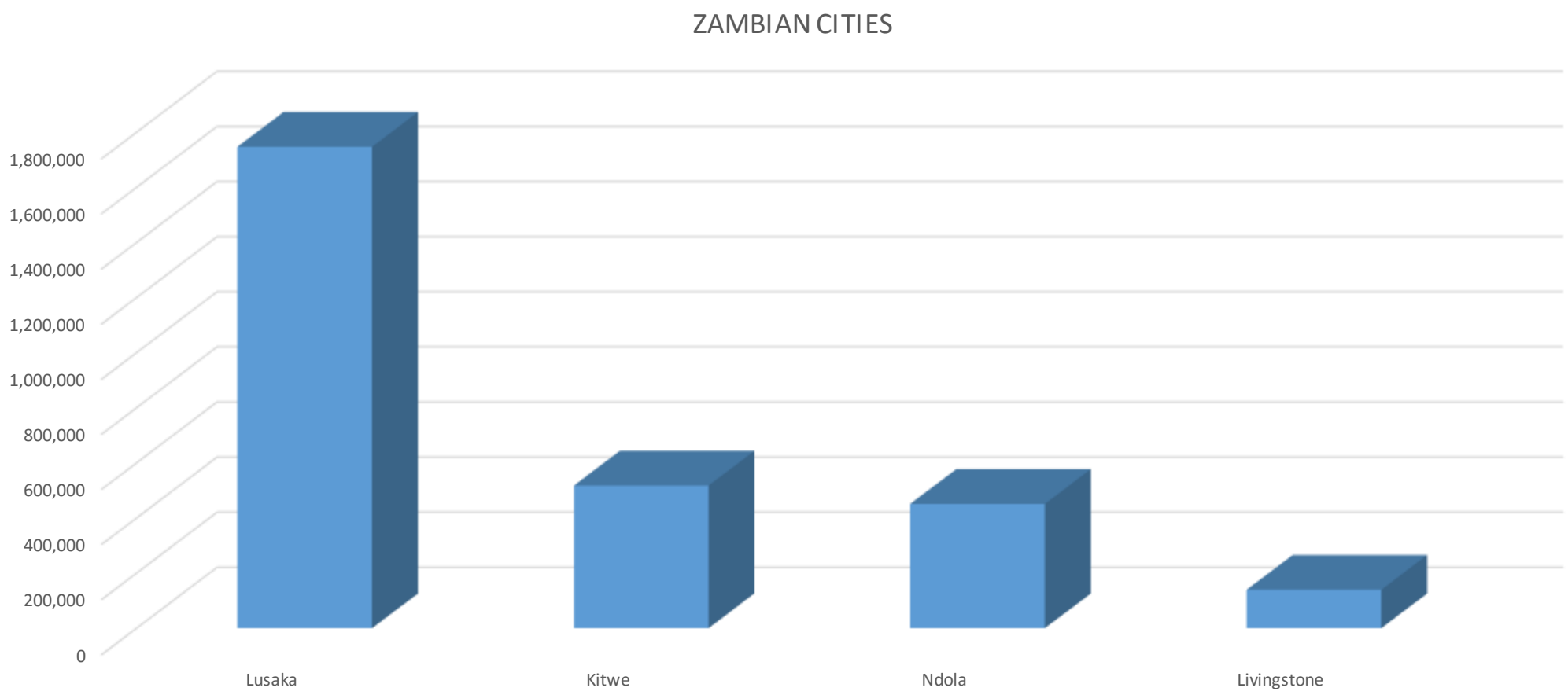
- **Whenever possible, construct the pie chart using percentages or counts either descending or ascending order.**

Lusaka	1,747,152
Kitwe	517,543
Ndola	451,246
Livingstone	139,509


ZAMBIAN CITIES
Lusaka
Kitwe
Ndola
Livingstone

THE BAR CHART

- **The second graphical technique is the bar chart.**
- **In this case, the following guidelines should be followed:-**
- **Label the frequencies along the vertical axis and the categories of the qualitative variable along the horizontal axis.**
- **Construct a rectangle over each category of the variable with a height equal to the frequency (number of observations) in the category.**
- **Leave a space between each category on the horizontal axis for clarity of presentation.**


ZAMBIAN CITIES
1,800,000
1,600,000
1,400,000
1,200,000
1,000,000
800,000
600,000
400,000
200,000
0
Lusaka
Kitwe
Ndola
Livingstone

THE FREQUENCY HISTOGRAM, RELATIVE FREQUENCY HISTOGRAM AND THE FREQUENCY POLYGON

- **Whereas, pie charts and bar charts refer to qualitative data, these three types of groups are applicable to quantitative data.**

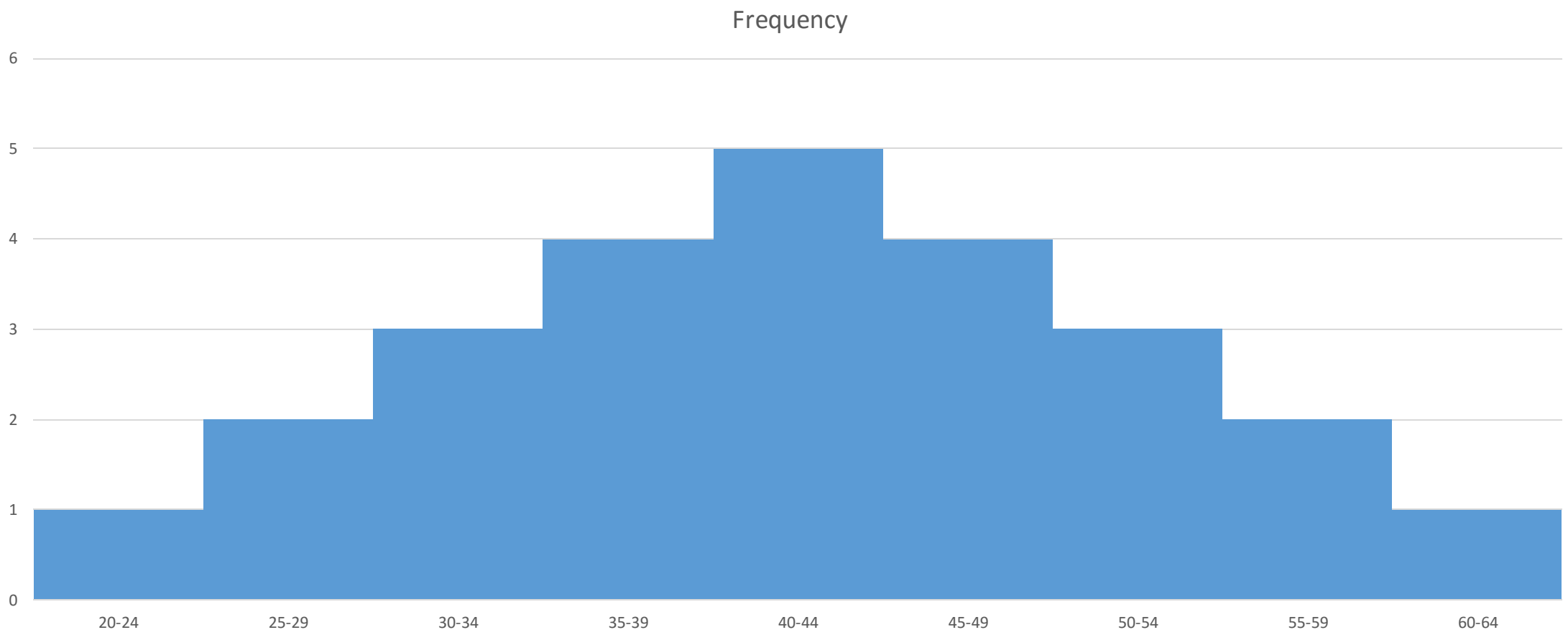
- **In the construction of the frequency histogram, the following guidelines must be followed:-**

- **Label the frequencies along the vertical axis and the class intervals (using the true limits along) the horizontal axis.**

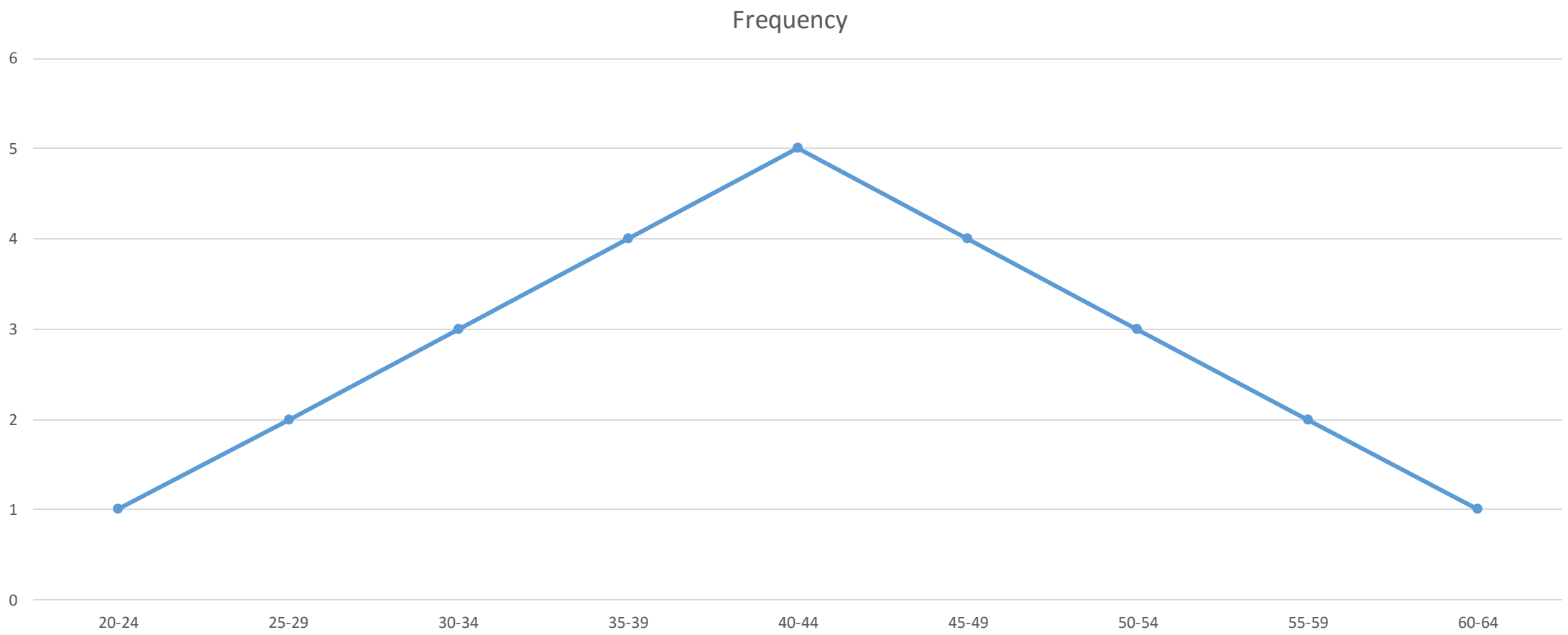
- **Construct a rectangle over each class interval of the interval with a height equal to the frequency (or the number of observations) in the interval.**

- **There should be no spaces between the class intervals (unless you use stated limits).**

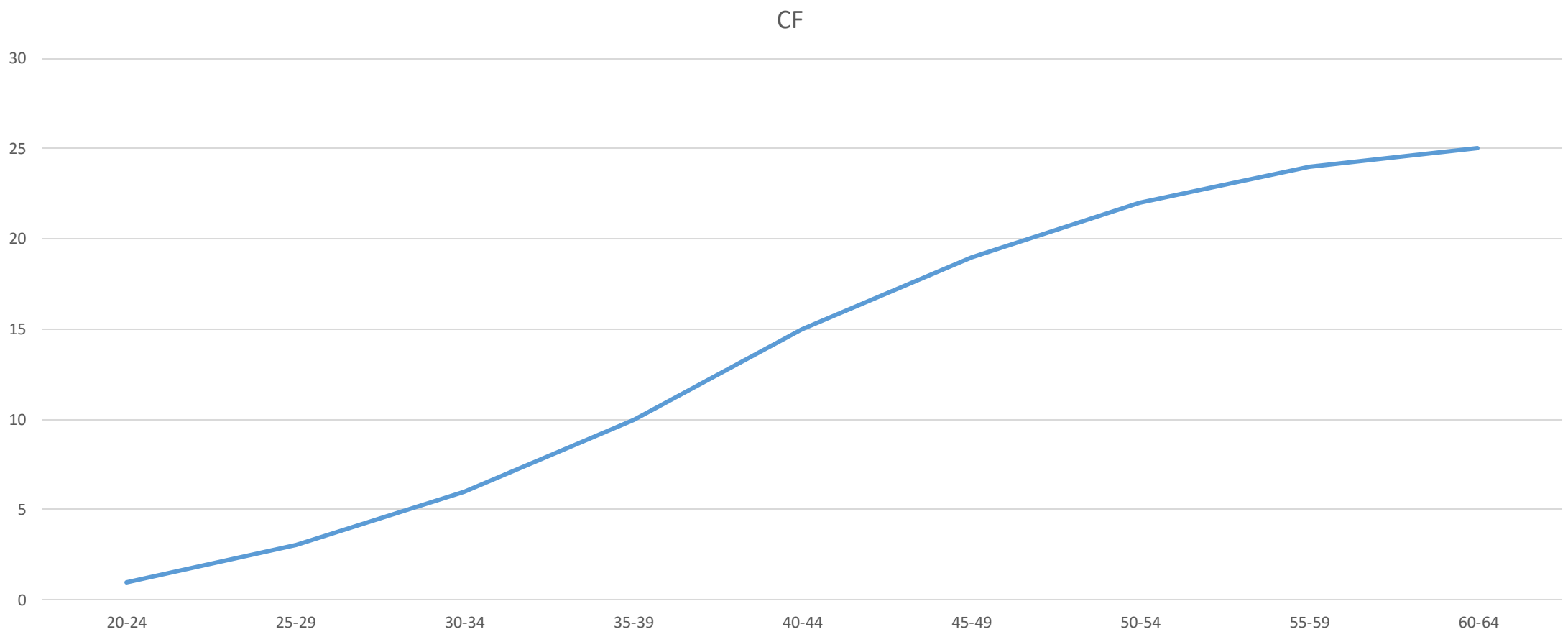
HISTOGRAM


Frequency
6
5
4
3
2
1
0
20-24
25-29
30-34
35-39
40-44
45-49
50-54
55-59
60-64

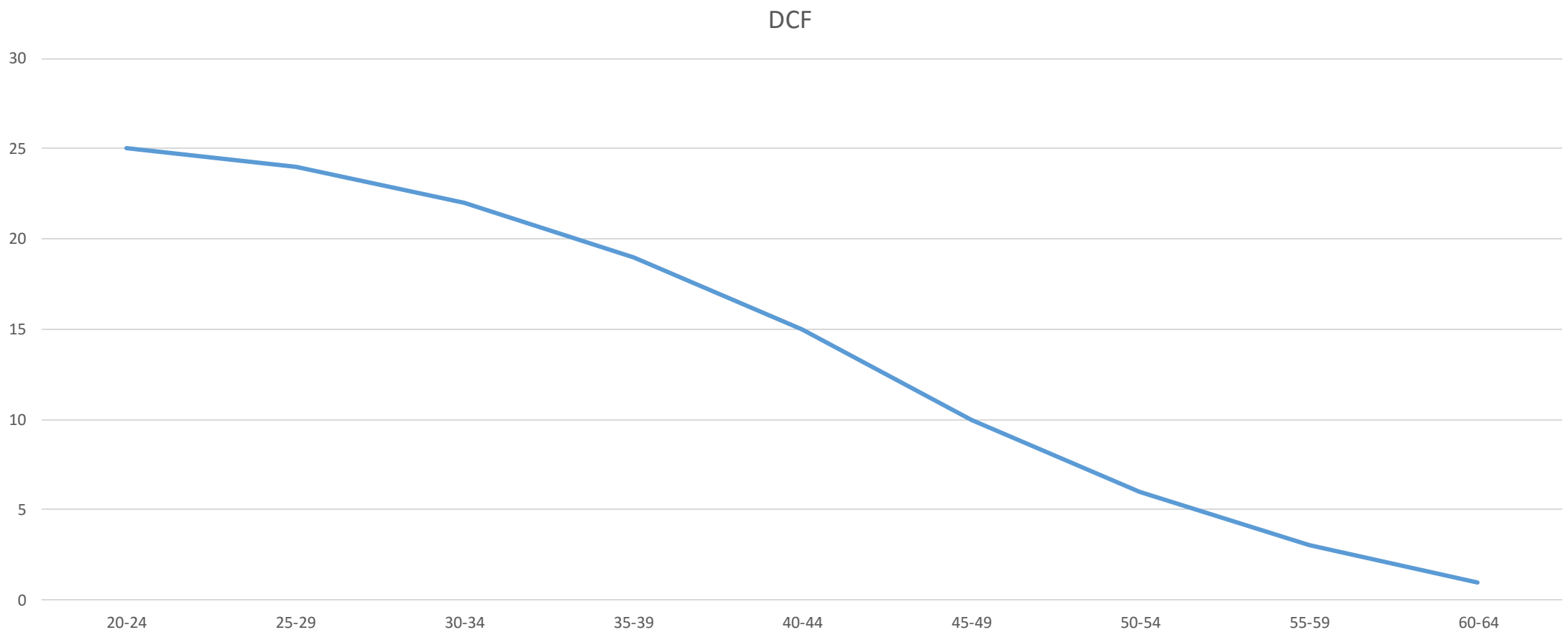
LINE CHART


Frequency
6
5
4
3
2
1
0
20-24
25-29
30-34
35-39
40-44
45-49
50-54
55-59
60-64

CUMULATIVE FREQUENCY



DECUMULATIVE FREQUENCY



NUMERICAL PRESENTATION OF DATA

- **After data has been prepared ad organized, the next step involves doing the calculations .**
- **The calculations include simple counties or frequencies and percentages.**
- **They also include:**
- **Measures of central tendency and dispersion.**

- **These are used to convey a mental image of pictures, objects and other phenomena because graphical methods are inappropriate for statistical inference.**

- **They are also used for expediency for we can use verbal communication to convey the appropriate picture.**

- **Numerical descriptive measures will create a mental picture of a frequency distribution for a set of measurements (we can all picture a man who is 6 feet tall).**

- **Two of the common numerical descriptive measures are those of central tendency and dispersion (or variability).**

- **Numerical descriptive measures for a population are called parameters.**

- **Those for samples are known as statistics.**

- **In problems requiring statistical inference, statistics from samples are used to estimate parameters for the population.**

NUMERICAL PRESENTATION OF DATA

- **Interpretation of the results of the calculation is the next important step in data analysis.**

- **Numbers do not speak for themselves.**

  What do these numbers mean?

- **Interpretation is the process of attaching meaning to the data.**

NUMERICAL PRESENTATION OF DATA

MEASURES OF CENTRAL TENDENCY

These show the tendency of observations (data) to cluster or to converge about the centre.

The common measures of central tendency are:

- **Mean**
- **Median**
- **Mode**

MEASURES OF CENTRAL TENDENCY

- **MEAN**

- **The arithmetic mean of a set of measurements is the sum of observations divided by total number of observations.**

- **Mean = $\frac{\text{Sum of observations}}{\text{Total number of observations.}}$**

MEASURES OF CENTRAL TENDENCY

MEAN

Population mean

The population mean is done by the Greek letter

$$\mu = \frac{\sum xi}{N}$$

MEASURES OF CENTRAL TENDENCY

- **Sample mean**

- **The sample mean is presented by X Bar.**

$$\overline{X} = \frac{\Sigma X}{n}$$

- **Mean = $\frac{1042}{25}$**
- **$= 41.7$ years**

- **For the UNZA data, the mean age is 41.7 years.**

- **Interpretation:**

- **Each UNZA employee is or is expected to be 41.7 years old.**

MEASURES OF CENTRAL TENDENCY

- **MEDIAN**
- **The median of a set of measurements is the middle value when the measurements are arranged in order of magnitude.**
- **EVEN NUMBER OF MEASUREMENTS**
- **In this case the median is the average of the two mid- point values when the measurements are arranged in order of magnitude.**
- **ODD NUMBER OF MEASUREMENTS**
- **In this case the median is simply the mid-point value.**

- **In the UNZA example, there are an odd number of observations.**
- **Therefore 41 years is the age that cuts the distribution in half.**
- **Interpretation:**
- **Half or 50 percent of the UNZA employees are below 41 years.**

MEASURES OF CENTRAL TENDENCY

- **MODE**

- **The mode of a set of measurements is that measurement that occurs most often**

- **This is the measurement with the with the highest frequency.**

- **In the UNZA example, those age 36 years appear three times in the distribution.**

- **Interpretation:**

- **Most or the majority of the UNZA employees are 36 years old.**

THE WEIGHTED MEAN

- **These involve assigning a weight to an observation based on its importance in the distribution.**
- **Instead of each data point contributing equally to the final mean, some data points contribute more "weight" than others.**
- **If all the weights are equal, then the weighted mean equals the arithmetic mean.**
- **The formula for the weighted mean is:**
- $\bar{X} = \frac{\sum W_I X_I}{\sum W_I}$

THE WEIGHTED MEAN

	Alice	Weights	Leonard	
Test	**80**	**0.15**	**72**	
Project	**44**	**0.35**	**44**	
Exam	**72**	**0.50**	**80**	
Final	**65**	**1.00**	**65**	

	Alice	Leonard
Test	12.00	10.80
Project	15.40	15.40
Exam	36.00	40.00
Final	63.40	66.20

MEASURES OF CENTRAL TENDENCY FOR GROUPED DATA

Class Interval	Frequency	x	fx	F
19.5 - 24.5	1	22	22	1
24.5 - 29.5	2	27	54	3
29.5 - 34.5	3	32	96	6
34.5 - 39.5	4	37	148	10
39.5 - 44.5	5	42	210	15
44.5 - 49.5	4	47	188	19
49.5 - 54.5	3	52	156	22
54.5 -59.5	2	57	114	24
59.5 – 64.5	1	62	62	25
Total	25		1,050	

THE MEAN

- **The formula for the mean for grouped data is:**

- $\bar{X} = \frac{\sum f_i x_i}{n}$

- **f_i = frequency or number of observation in class interval, i.**
- **x_i = the mid-point of class interval, i.**

- **= 1050/25**

- **= 42 years**

THE MEDIAN

- **The formula for the median is:-**

- $$\text{Median} = L + \frac{(n/2) - F).i}{f}$$
- **n/2 = The middle item or value.**
- **L = True lower limit of the class interval in which the median item is located.**
- **F = the cumulative frequency of the class interval preceding the one containing the median item.**
- **f = the frequency of the class interval containing the median item.**
- **i= the size of the class interval.**

- $$\text{Median} = 39.5 + \frac{(12.5) - 10).5}{5}$$
- **=42 years**

Class Interval	f	F
19.5 - 24.5	1	1
24.5 - 29.5	2	3
29.5 - 34.5	3	6
34.5 - 39.5	4	10
39.5 - 44.5	5	15
44.5 - 49.5	4	19
49.5 - 54.5	3	22
54.5 -59.5	2	24
59.5 – 64.5	1	25
Total	25	

THE MODE

- **The formula for mode is:**

- **Mode = L + $\frac{(D_1)}{(D_1 + D_2)}$i**

- **L = True lower limit of the class interval where the mode item is = 39.5.**
- **D_1 = f_0-f_1 = 5-4=1**
- **D_2 = f_0 –f_2 = 5-4 =1**
- **i = size of the class interval = 5**

- **Mode 39.5 + $\frac{(1)}{(1 + 1)}$ *5**
- **= 42 years.**

CHOICE OF SELECTION OF AN APPROPRIATE MEASURE OF CENTRAL TENDENCY

- **THE MODE**
- **Measures what is the most common value – the typical value..**
- **It can take on more than one value as in the bimodal or trimodal distribution.**
- **It thus can capture what the mean can not especially in the case of the bimodal or trimodal distribution.**
- **Can be useful in manufacturing.**
- **When a variable is measured on interval or ratio scale, it may be of little value because it cannot be used for further calculations – a situation made worse by the fact that it doesn't take all values into account and can have more than one value.**

- **<u>THE MEAN</u>**

- **It is the most commonly used measure of central tendency because it is easy to use and interpret**

- **It is affected by presence of extreme values in a distribution.**

- **It can take on unrealistic values such as 3.6 children.**

- **It is useful in statistical inference**

- **It takes into account all values in the distribution and can thus be used in further mathematical calculations.**

- **If there are a total of 6,000 students on campus, we can use the sample mean to predict the total amount of Mosi consumed by all the students:-**
-
- **Total sum over due = N X**
-
- **= 6,000 x 8.84**
-
- **= <u>53,040</u>**

- **THE MEDIAN**

- **It is not affected by extreme values in the distribution; it is relatively stable.**

- **It may therefore be more useful in studies of income distribution than the mean.**

- **It does not take unrealistic values like the median number of children being 3 instead of 3.4.**

- **It has the disadvantage of not taking all the values into account; thus it can't be used mathematical calculations.**

COMPARISON OF THE MEAN AND MEDIAN

	89	**89**	**144**
	83	**83**	**83**
	81	**81**	**81**
	77	**77**	**77**
	75	**20**	**75**
MEAN	**81**	**70**	**92**
MEDIAN	**81**	**81**	**81**

MEASURES OF DISPERSION

- **These show the tendency of observations (data) to spread or diverge away from the centre.**
- **They are also known as measures of variability.**
- **These are:**
- **The range**
- **The variance**
- **The standard deviation**

MEASURES OF DISPERSION

RANGE

This is the difference between the largest and smallest observations in a distribution.

The range= 62-22

= 40

Interpretation: The difference between the oldest and youngest student is 40 years.

MEASURES OF DISPERSION

- **THE VARIANCE**

- **This is a better way at measuring dispersion or variability.**

- **The variance of a set of measurements X, X ...Xn with a mean X is the sum of squared deviations from the mean divided by n - 1.**

MEASURES OF DISPERSION

- **POPULATION VARIANCE**

- **The population variance is the average squared deviation from the population mean, as defined by the following formula:**

$$\sigma^2 = \frac{\Sigma (X_i - \mu)^2}{N}$$

MEASURES OF DISPERSION

- **SAMPLE VARIANCE**

- **The sample variance is defined by slightly different formula, and uses a slightly different notation:**

$$s^2 = \Sigma \frac{(x_i - x)^2}{(n - 1)}$$

- **where s^2 is the sample variance, x is the sample mean, x_i is the i^{th} element from the sample, and n is the number of elements in the sample.**

MEASURES OF DISPERSION

x	(x-X)	$(x-X)^2$
22	-19.7	388.09
26	-15.7	246.49
27	-14.7	216.09
31	-10.7	114.49
32	-9.7	94.09
33	-8.7	75.69
35	-6.7	44.89
36	-5.7	32.49
36	-5.7	32.49
36	-5.7	32.49
40	-1.7	2.89
41	-0.7	0.49
41	-0.7	0.49
43	1.3	1.69
44	2.3	5.29
46	4.3	18.49
47	5.3	28.09
48	6.3	39.69
49	7.3	53.29
50	8.3	68.89
51	9.3	86.49
53	11.3	127.69
55	13.3	176.89
58	16.3	265.69
62	20.3	412.09
Total		2,565.45

MEASURES OF DISPERSION

x	f	(x-X)	$(x-X)^2$	$f(x-X)^2$
22	1	-19.7	387.3	387.3
26	1	-15.7	245.9	245.9
27	1	-14.7	215.5	215.5
31	1	-10.7	114.1	114.1
32	1	-9.7	93.7	93.7
33	1	-8.7	75.3	75.3
35	1	-6.7	44.6	44.6
36	3	-5.7	32.3	96.8
40	1	-1.7	2.8	2.8
41	2	-0.7	0.5	0.9
43	1	1.3	1.7	1.7
44	1	2.3	5.4	5.4
46	1	4.3	18.7	18.7
47	1	5.3	28.3	28.3
48	1	6.3	39.9	39.9
49	1	7.3	53.6	53.6
50	1	8.3	69.2	69.2
51	1	9.3	86.9	86.9
53	1	11.3	128.1	128.1
55	1	13.3	177.4	177.4
58	1	16.3	266.3	266.3
62	1	20.3	412.9	412.9
	25			2565.4
Mean	41.7			
Variance	106.9			
Stdev	10.3			

MEASURES OF DISPERSION

- **The variance for the UNZA data is:**

- **= 2,565**
- **24**
- **= 106.89**

MEASURES OF DISPERSION

- **STANDARD DEVIATION**
- **The standard deviation is the square root of the variance.**

- $s = \sqrt{s^2}$

- **This is the average amount of deviation of each observation from the mean.**

- **The larger the standard deviation, the greater is the variation of individual observations from the mean; the smaller it is, the smaller is the variation.**

- **Interpretation: Each of the students is aged 10.3 years below or above the mean of age of 41.7 years.**

MEASURES OF DISPERSION FOR GROUPED DATA

Variance for grouped data

$$s^2 = \Sigma f\frac{(x_i - x)^2}{(n - 1)}$$

MEASURES OF DISPERSION FOR GROUPED DATA

Age group	f	x	(x-X)	$(x-X)^2$	$f(x-X)^2$
20-24	1	22	-20	400	400
25-29	2	27	-15	225	450
30-34	3	32	-10	100	300
35-39	4	37	-5	25	100
40-44	5	42	0	0	0
45-49	4	47	5	25	100
50-54	3	52	10	100	300
55-59	2	57	15	225	450
60-64	1	62	20	400	400
Total	25				2500
MEAN	42.0				
VARIANCE	104.2				
STDEV	10.2				

- **The variance for the UNZA data is:**

- **= 2,500**
- **24**
- **= 104.2**

- **s= √104.2**

- = **10.2 years**

SEMI – INTERQUARTILE RANGE

- **This is a measure of dispersion around the median which is based on the interquartile range.**
-
- **The interquartile range is the distance between 1st and 3rd quartiles in a distribution.**
-
- **The first quartile has ¼ of the frequencies smaller and ¾ larger.**
-
- **The 3rd quartile has ¾ of the frequencies smaller and ¼ larger.**

- **The semi-interquartile range is therefore:**

- **SIR = (Q3-Q1)/2**

- **The formula for the quartile is :-**

- $$\text{First Quartile} = L + \frac{(1/4n - F).i}{f}$$

$$\text{Third Quartile} = L + \frac{(3/4n - F).i}{f}$$

- **1/4n or 3/4n = The quartile value.**
- **L = True lower limit of the class interval in which the quartile value is located.**
- **F = the cumulative frequency of the class interval preceding the one containing the quartile value.**
- **f = the frequency of the class interval containing the quartile value.**
- **i= the size of the class interval.**

SEMI – INTERQUARTILE RANGE

Age group	f	F
20-24	1	1
25-29	2	3
30-34	3	6
35-39	4	10
40-44	5	15
45-49	4	19
50-54	3	22
55-59	2	24
60-64	1	25
Total	25	

MEASURES OF RELATIVE STANDING

- **These show the ranking of observations in a distribution.**
- **These also provide direct assistance in interpreting individual scores.**
- **These measure the relative standing or the rank of a score in a distribution.**

- **PERCENTILES POINTS (OR JUST PERCENTILES) AND PERCENTILE RANKS**

- **Both percentile scores and ranks indicate the position of location of say the value in relation to the other values in a distribution.**

- **If we are comparing SS240 test scores or incomes of individuals, measures of relative standing can show, for example, the percentage of students who scored above a certain mark or conversely the score below which a certain percentage of the students fall.**

MEASURES OF RELATIVE STANDING

- **PERCENTILE SCORES (POINTS)**

- **Percentile score, or simply the percentile, is the score below which a percentage of the scores in a distribution fall and above which the reminder falls.**

- **By definition, the percentile of a set of measurements arranged in order of magnitude is that value that has p% of the values below it and (100-p) above it.**
-
- **For example, if 10% are below score A, then (100-10% or 90%) are above it.**

- **EQUIVALENCE**
-
- **1st quartile – 25th percentile**
- **2nd quartile – median – 50th percentile**
- **3rd quartile – 75th percentile**
-

MEASURES OF RELATIVE STANDING

- **TO FIND THE SCORE, GIVEN A PERCENTILE**

- **If you want to know the score below which a percentage falls, use this approach.**

- **For example, you may want to know the income below which 70 percent of the of the population fall or the score below which 75 percent of students fall in an examination.**

MEASURES OF RELATIVE STANDING

- **The formula for the percentile score is:-**

- $$PS = L + \frac{(P(n) - F)}{f}$$

- **P(n) = the percentage item or value corresponding to the required percentile.**
- **P= is the proportion corresponding to the required value,**
- **n=the sample Size.**
- **If a 25^{th} percentile is required in a sample at 100, then P(n) =0 .25 x 100**
- **L = True lower limit of the class interval in which the percentile is located.**
- **F = the cumulative frequency of the class interval preceding the one containing the containing the percentile.**
- **f= the frequency of the class interval containing the percentile item.**
- **i = the size of the class interval.**

EXAMPLE OF SS 242 EXAM

Grade	Class interval	Number of students	Percent
D	26 - 35	36	3.7
D+	35 - 39	23	2.3
C	40 - 45	65	6.6
C+	46 - 55	213	21.7
B	56 - 65	359	36.6
B+	66 - 75	233	23.8
A	76 - 85	51	5.2
Total		980	100.0

MEASURES OF RELATIVE STANDING

- **EXAMPLE 1:**
- **THE PERCENTILE SCORE**
- **Given the following distribution of SS241 scores:-**
- **TASK:- To find the score below which 50% of the students in SS241 fell.**

MEASURES OF RELATIVE STANDING

Grade	Class interval	Number of students	CF	Percent
D	26 - 35	36	36	3.7
D+	35 - 39	23	59	6.0
C	40 - 45	65	124	12.7
C+	46 - 55	213	337	34.4
B	56 - 65	359	696	71.0
B+	66 - 75	233	929	94.8
A	76 - 85	51	980	100.0

MEASURES OF RELATIVE STANDING

- **MEDIAN**

- **Find the student corresponding to the 50^{th} percentile**

- **P(n) = 0.50 x 980 = 490.**

MEASURES OF RELATIVE STANDING

MEDIAN		**59.8**
N/2	**490**	
L	**55.5**	
F	**337**	
f	**359**	
i	**10**	

MEASURES OF RELATIVE STANDING

- **TASK:- To find the score below which 60% of the students in SS241 fell.**

- **Find the student corresponding to the 60^{th} percentile:**

- **P(n) = 0.60 x 980 = 588**

MEASURES OF RELATIVE STANDING

SCORE	**62.5**
.6N	**588**
L	**55.5**
F	**337**
f	**359**
i	**10**

MEASURES OF RELATIVE STANDING

- **PS = 55.5 + $\frac{(588 - 337)}{359}$ 10**
- **= 55.5 + 7.0**
- **= 62.5**

MEASURES OF RELATIVE STANDING

- **THE PERCENTILE RANK**
- **TO FIND THE %, GIVEN THE SCORE**
- **This is used when you have a score and you want to know the percentage below or above that score.**
- **It is a reverse way of looking at or measuring the relative standing of scores.**
- **In definitional terms, the percentile rank is simply the percentage in a**
- **distribution falling below a certain score or value**

MEASURES OF RELATIVE STANDING

- **The formula for the percentile ranks is:-**
-
- $$PR = \frac{f(x-L) + F_i}{Ni}$$
-
- **x = the score below which a percentage (%) falls.**
-
- **f = the frequency in the interval where, X is the score is located.**
-

MEASURES OF RELATIVE STANDING

- **L = the true lower limit of the interval where x is located.**
- **F = the cumulative frequency of the class interval preceding the one containing the percentile.**
- **N= the total sum of frequencies**
- **i= the size of the class interval**

MEASURES OF RELATIVE STANDING

- **EXAMPLE 2**
- **THE PERCENTILE RANK**
- **TASK: To find the percentage of students who scored below Mr. Zimba's score of 62.5**

MEASURES OF RELATIVE STANDING

Grade	Class interval	Number of students	DCF	Percent
D	26 - 35	36	980	100.0
D+	35 - 39	23	944	96.3
C	40 - 45	65	921	94.0
C+	46 - 55	213	856	87.3
B	56 - 65	359	643	65.6
B+	66 - 75	233	284	29.0
A	76 - 85	51	51	5.2

MEASURES OF RELATIVE STANDING

Percentile rank	
x	62.5
L	55.5
F	337
f	359
i	10
N	980

MEASURES OF RELATIVE STANDING

- **PR= $\frac{359(62.5- 55.5) + 337*10}{980 * 10}$**

- **= $\frac{5880}{9800}$**

- **= $\underline{0.60 \text{ or } 60\%}$**

INTRODUCTION TO PROBABILITY

- **INTRODUCTION TO PROBABILITY**
- **The concept of probability is basic to social, economic and political reasoning because complete certainty of mathematical or logical argument is not possible when we come to the real world.**
- **Thus as researchers we often talk of the chances or likelihood occurring.**
- **When we do so we are actually using probability.**

INTRODUCTION TO PROBABILITY

- **EXAMPLES:-**

- **Likelihood of:**

- **Loadshedding**
- **Water shortages**
- **Rainfall**
- **Defaulting on loans**

INTRODUCTION TO PROBABILITY

- **BASIC CONCEPTS AND VOCABULARY IN PROBABILITY**
- **Probability refers to the numerical evaluation of likelihood of occurrence of an event.**
- **EXPERIMENT**
- **A situation with a defined set as outcomes – flipping a coin, boy, girl.**

- **EXAMPLE**
- **Flipping a coin produces two outcomes – head or tail.**

INTRODUCTION TO PROBABILITY

- **EVENTS OR SAMPLE POINTS**
- **One of the possible outcomes of an experiment are called sample points or events.**
- **EXAMPLE**
- **In flipping a coin, Heads is an event of sample point.**

INTRODUCTION TO PROBABILITY

- **SAMPLE SPACE**
- **A list of all possible outcomes as an experiment is sample space**
- **HEADS, TAILS constitutes sample space.**

- **MUTUALLY EXCLUSIVE EVENTS**

- **In this, the occurrence of one event precludes the occurrence of any other.**

- **EXAMPLE**

- **In a flip of a coin, heads and tails are mutually exclusive events.**
- **Death and life are mutually exclusive.**

- **SIMPLE EVENT**
- **An event that contains a single outcome“**
- **If heads results in a coin toss that is a simple event.**

- **COMPOUND EVENT**

- **An event that consists of two, or more, simple events; for example: A or B; A and B and C.**

- **This refers to any event that can be decomposed into two or more events. A car accident – survival, death, injury**

- **The events are not necessarily mutually exclusive. You can survive an accident but still be injured.**

- **INDEPENDENT EVENTS**

- **The occurrence of an event does not in any way affect the likelihood of the occurrence of another event – the two events are independent.**

- **EXAMPLE**

- **Weight and intelligence**

DEPENDENT EVENTS

The occurrence of an event influences the likelihood of the occurrence of another event – the two events are dependent.
Reading > exam success

KINDS (TYPES) OF PROBABILITY

- **Probability has in many applications in many areas.**

- **Businessmen will always weigh chances of success just as opinion pollers think in terms of probabilities.**

- **There are three major types of probability**

- **A priori**
- **Posterior or experimental**
- **Subjective**

A PRIORI PROBABILITY

- **A priori means “before the event.”**
- **This assumes all possible events, E_1, are known and have equal likelihood of occurrence.**
- **Given that all possible outcomes of a given event are equally likely, the probability of any specified outcome is equal to the ratio of the number of ways that outcome could be achieved to the total number of ways that all possible outcomes can be achieved.**

STEPS IN COMPUTING PROBABILITY

- **This means therefore that to compute a probability, you follow these steps:**
- **Step 1: Determine the number of possible ways that the outcome you are interested in can occur. (Interested in number of heads in one coin toss or flip)**
- **Step 2: Determine the number of all possible outcomes (The number of all possible outcomes in In a coin toss or flip is heads, tails)**
- **Step 3: Divide the first number by the second.**
- **The answer you get gives the probability that the event in question will occur.**

STEPS IN COMPUTING PROBABILITY

- **EXAMPLE OF COIN FLIP/TOSS**
-
- **P (HEADS) = # OF HEADS CAN RESULT (H)**
- **# OF POSSIBLE OUTCOMES FROM ONE TOSS (H,T)**

- **= (H) = 1**
- **(H,T) 2**
-
- **= 0.50**

A PRIORI PROBABILITY

- **It follows in other words :**

- **If an event, E, can happen A different ways and cannot happen in A' different ways, then in general the probability of an event, E, will occur is:**

- **$P(E) = \frac{A}{A + A'}$**

A PRIORI PROBABILITY

- **The probability that event E will not occur is therefore**

- **(P(E') = A'**
- **A +A'**

- **Since the sum of ways something can occur + the sum of ways something cannot occur is equal to the total number of vents, then**

- **$A + A^1 = n$**

A PRIORI PROBABILITY

- **Therefore $P(E) = \frac{A}{n}$**

- **And $P(E') = \frac{A'}{n}$**

- **Therefore, $P(E) + P(E') = 1$**

AXIOMS OF PROBABILITY

1. The probability of an event occurring is always between 0 and 1.

- **$0 \leq P(E) \leq 1$**

- **Impossibility = 0**
- **Absolute certainty = 1**

AXIOMS OF PROBABILITY

2. Sum of probabilities of a mutually exclusive events is equal to 1.

- **P(E1) + P(E2) + ….P(En) = 1**

3. The probability that event E will not occur is 1 minus the probability that it will occur.

- **P(E’) = 1 – P(E)**

POSTERIOR OR EXPERIMENTAL PROBABILITY

- **This is also known as EXPERIMENTAL (RELATIVE) PROBABILITY**
- **This is based on actual observations and data obtained through sample surveys, records, experiments, and other sources of data.**
- **This is based on a limited number of observations, based on say, probability sampling on the basis of which relative frequencies can be computed.**
- **Posterior probabilities are also known as long – run probabilities.**

EXPERIMENTAL PROBABILITY

Grade	Class interval	Number of students	Probability
D	26 - 35	36	0.04
D+	35 - 39	23	0.02
C	40 - 45	65	0.07
C+	46 - 55	213	0.22
B	56 - 65	359	0.37
B+	66 - 75	233	0.24
A	76 - 85	51	0.05
Total		980	1.00

EXPERIMENTAL PROBABILITY

- **The probability of failing SS 241 is only 0.03 or less than 1 chance out of 10.**

- **The probability of getting a B is 0.37 or approximately 4 out 10 chances.**

- **There is 0.05 probability or 1 chance out of 20 of getting an A.**

SUBJECTIVE PROBABILITY

- **This is not based on actual observation**
- **Rather it is based on personal conviction that an event will occur.**
- **It is based on a person's mind and not with or physical event.**
- **Intuition is therefore key in this type of probability.**
- **EDUCATED GUESS**
- **Competitors price reduction – 10**
- **Increase in cost of raw materials – 90**
-
- **Incorporate this in statistical decision making later**.

INTRODUCTION TO THE MATHEMATICS OF PROBABILITY

- **The discussion relies on the application of posterior probabilities.**
- **GENERAL RULES OF PROBABILITY**
- **Statisticians refer to the probability of one event or another as the union of two probabilities.**
- **This can be expressed symbolically as:**
- **P(A U B)**
- **This can also be expressed as P of A union B.**

INTRODUCTION TO THE MATHEMATICS OF PROBABILITY

- **Statisticians refer to the probability of two events occurring simultaneously as a joint probability.**
- **This expressed symbolically as:**
- **$P(A \cap B)$.**
- **This can also be expressed as P of A intersection B.**

INTRODUCTION TO THE MATHEMATICS OF PROBABILITY

- **A condition probability is probability of one event given that another has occurred.**
- **This expressed symbolically as:**
- **P(A/B)**
- **This can also be expressed as the probability of A given B or P of A given B.**

INTRODUCTION TO THE MATHEMATICS OF PROBABILITY

- **A condition probability is probability of one event given that another has occurred.**
-
- **This expressed symbolically as:**
-
- **P(A/B)**
-
- **This can also be expressed as the probability of A given B or P of A given B.**

INTRODUCTION TO THE MATHEMATICS OF PROBABILITY

- **There are primarily two rules in the mathematics of probability:**

- **The additional rule**
- **The multiplication rule**

INTRODUCTION TO THE MATHEMATICS OF PROBABILITY

- **THE GENERAL RULE OF ADDITION FOR ANY TWO EVENTS**

- **This applies whenever we want to know the probability that either of two events occurred.**

- **The general rule for any two events is expressed symbolically as:**

- **$P(A \cup B) = P(A) + P(B) - P(A \cap B)$**

- **In other words, the probability of either of two events [$P(A \cup B)$] is equal to the probability that one event will occur [P(A)] plus the probability that the other event [P(B)] will occur minus the probability that both events will occur simultaneously [$P(A \cap B)$].**

INTRODUCTION TO THE MATHEMATICS OF PROBABILITY

- **THE ADDITION RULE FOR MUTUALLY EXCLUSIVE EVENTS**
- **If the events are mutually exclusive then rule is:**
- **$P(A \cup B) = P(A) + P(B)$**

INTRODUCTION TO THE MATHEMATICS OF PROBABILITY

		E4 Female	E_5 Male	Total
E_1	Female condom	450	500	950
E_2	Male condom	300	800	1,100
E_3	Abstinence	100	350	450
	TOTAL	850	1,650	2,500

INTRODUCTION TO THE MATHEMATICS OF PROBABILITY

- **ADDITION RULE FOR ANY TWO EVENTS WHICH ARE NOT NECESSARILY MUTUALLY EXCLUSIVE**

EXAMPLE

- **The probability of preferring the female condom or being female:**

- $P(E_1 \cup E_4) = P(E_1) + P(E_4) - P(E_1 \cap E_4)$

- **Joint of occurrence must be subtracted to avoid double calculating $P(E_1)$ OR $P(E_4)$**

$P(E_1 \cup E_4) = \frac{950}{2500} + \frac{850}{2500} - \frac{450}{2500}$

- $= 0.38 + 0.34 - 0.18$

$= \underline{0.54}$

INTRODUCTION TO THE MATHEMATICS OF PROBABILITY

- **ADDITION RULE FOR MUTUALLY EXCLUSIVE EVENTS**
-
- **The probability of being male or female**
-
- **P(E4 or E5) = 0.34 + 0.66**
- **= <u>1.0</u>**
-
- **The probability of being male is =**

-
- $P(E_5) = 1 - P(E_4) = 1 - 0.34 = \underline{0.66}$
-
- Probability being female is
-
- $P(E_4) = 1 - P(E_5) = 1 - 0.88 = \underline{0.34}$
-

INTRODUCTION TO THE MATHEMATICS OF PROBABILITY

- **THE GENERAL RULE OF MULTIPLICATION FOR ANY TWO EVENTS**
-
- The general of multiplication is applied when you want to know the joint probability of two events (i.e. the probability that both events will occur).
-
- The general rule for any two events that are not necessarily independent is expressed symbolically as:
-
- **$P(A \cap B) = P(A).\ P(B/A)$**
-
- This is the same thing as:
-
- **$P(A \cap B) = P(A).\ P(A/B)$**

INTRODUCTION TO THE MATHEMATICS OF PROBABILITY

- **MULTIPLICATION FOR ANY TWO EVENTS**

EXAMPLE

- **The joint probability of being female and preferring abstinence.**

- $\mathbf{P(E_4 \cap E_3) = P(E_4).\ P(E_3/E_4)}$

- = 850 * 100
- 2500 850

- = 100
- 2500

- = 0.04

INTRODUCTION TO THE MATHEMATICS OF PROBABILITY

		E4 Female	E_5 Male	Total
E_1	Female condom	450	500	950
E_2	Male condom	300	800	1,100
E_3	Abstinence	100	350	450
	TOTAL	850	1,650	2,500

INTRODUCTION TO THE MATHEMATICS OF PROBABILITY

- **THE MULTIPLICATION RULE FOR INDEPENDENT EVENTS**

- **In the case of two independent events, a special rule of multiplication applies.**

- **Two events are independent if the probability of one event is not affected by whether or not the other event occurs.**

- **$P(A \cap B) = P(A)\ P(B)$**
- **Therefore, $P(A).\ P(B/A) = P(A)\ P(B)$**
- **The probability of P(B) remains the same regardless of whether P(A) has occurred.**

INTRODUCTION TO THE MATHEMATICS OF PROBABILITY

		E4 Female	E_5 Male	Total
E_1	Female condom	570	380	950
E_2	Male condom	660	440	1,100
E_3	Abstinence	270	180	450
	TOTAL	1,500	1,000	2,500

INTRODUCTION TO THE MATHEMATICS OF PROBABILITY

- **THE MULTIPLICATION RULE FOR INDEPENDENT EVENTS**

EXAMPLE

- **Probability of being male and preferring abstinence.**

- **$P(E_5 \cap E_{3)} = P(E_5)\ P(E_3)$**
- **= 1000/2500 * 450/2500**
- **= 0.40 * 0.18**
- **= <u>0.07</u>**

- = $P(E_5 \cap E_{3)} = P(E_5)\ P(E_3) = P(E_3)P(E_3/E_5)$
- = 1000/2500 * 180/1000
- = 0.40 * 0.18
- = <u>0.07</u>

- The probability of $P(E_5)$ remains the same regardless of whether $P(E_3)$ has occurred.

- **CONDITIONAL PROBABILITY**

- $P(A/B) = \frac{P(A).\ P(B/A)}{P(B)}$

INTRODUCTION TO THE MATHEMATICS OF PROBABILITY

- **CONDITIONAL PROBABILITY**

- **The probability of preferring the male condom given that one is female:**

- **$P(E1/E4) = \frac{P(E1) * P(E4/E1)}{P(E4)}$**

- **$= \frac{(950/2500) * (450/950)}{(850//2500)}$**

- **$= 450/850$**

- **$= 0.53$**

TYPES OF DISTRIBUTIONS

- **OBSERVED OR EMPIRICAL DISTRIBUTION**

- **Based on actual observations**

- **Example**
- **Frequency distribution**

- **PROBABILITY DISTRIBUTION**

- **Theoretical distribution of all possible events and the probabilities of occurrence of each event (classical probability).**

- **EXPECTED DISTRIBUTION**

- **Product of the probabilities of the occurrence of each event and its total number of events.**

TYPES OF DISTRIBUTIONS

EVENT	PROBABILITY DISTRIBUTION	EXPECTED DISTRIBUTION	EMPIRICAL OR OBSERVED DISTRIBUTION
HEADS	0.5	10	12
TAILS	0.5	10	8
TOTAL	**1.0**	**20**	**20**

COMMON PROBABILITY DISTRIBUTION

- **BINOMIAL**
- **POISSON**
- **NORMAL**

STATISTICAL INFERENCE

- **STATISTICAL INFERENCE is a process of drawing conclusions about a whole population on the basis a sample.**
- **It means inferring a population value on the basis of a sample value.**
- **A sample value is known as a statistic while a population value is known as a parameter.**
- **Statistical inference is there the process of inferring a parameter based on our knowledge of the statistic.**

STATISTICAL INFERENCE

- **This process of making inferences carries with it some risks.**
- **For one thing, we may not have much or no information about the values in the population.**
- **Thus there is always a risk or chance or probability that we may make a wrong inference.**
- **To estimate the risk of a wrong inference, we require an understanding and the help of a theoretical probability distribution known as the normal curve on the normal distribution.**

STATISTICAL INFERENCE

- **The normal curve is a theoretical representation of the manner in which most traits or variables which occur at random distribute themselves in nature.**

- **In virtually all populations, and for any variable or trait, there is a tendency, if we take its measurements that such measurements will tend to bunch up and create a hump at the center where most observations will concentrate or converge.**

- **The remainder of the measurements will taper off at the extremes of the distribution.**

STATISTICAL INFERENCE

- **This will be case for almost all the variables in the population such as age, examination scores, weight, height, etc.**

- **A large number of empirical distributions in the form of frequency distribution, frequency histogram of the variables always display this in the population.**

- **This means therefore that if we take a sample that resembles the population, the variables in that sample will tend to have similar distribution as that in the population.**

STATISTICAL INFERENCE

- **In other words, if a variable in the population is normally distributed, we assume that even the variable in a sample derived from that population will be normally distributed.**

- **This, however, is only possible if the sample resembles or is adequately representative of the population.**

- **This means using probability sampling in selecting the elements from the sample to the population.**

- **If we do this, then we can be justified with a certain degree of probability to say something about the population values on the basis of our sample values.**

STATISTICAL INFERENCE

- **But since the sample values cannot be exactly the same as population values, there is a risk of making an erroneous inference.**

- **To make it possible to measure this risk of a wrong inference requires knowledge of the normal distribution and its characteristics.**

- **Statistical inference is not limited to making inference from statistics to parameter, it also involves testing hypotheses.**

STATISTICAL INFERENCE

- **CHARACTERISTICS OF THE NORMAL CURVE**
- **IT IS UNIMODAL**
- **it has an identical mean, median and mode in the middle where most observations of the random variable are clustered.**
- **IT IS SYMMETRICAL.**

- **both sides of the centre make up 50% on both sides of the baseline giving a total area under the curve which is equal to unity.**

- **IT IS CONTINUOUS**

- **It is continuous. I.e. the variable can take all numerical values like weights, distances, heights etc discrete variables.**

STATISTICAL INFERENCE

- **IT HAS A STANDARD DEVIATION WHICH EARMARKS THE DISTANCE ON THE BASELINE FROM THE CENTER POINT, THE MEAN, μ.**

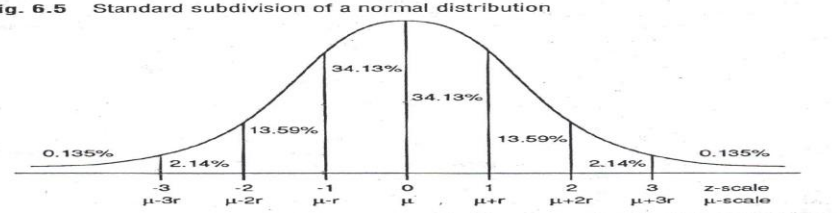
- **This is done in such a way that the area between the curve and the baseline is expressed in proportions or percentages partitioned so that:**

- **About 68% of the area lies with one standard deviation of the mean (i.e. μ±σ)**

- **About 95% of the area lies with two standard deviation of the mean (i.e. $\mu \pm 2\sigma$)**

- **About 99.7% of the area lie within three standard deviation the mean (i.e. $\mu \pm 3\sigma$)**

Fig. 6.5 Standard subdivision of a normal distribution



STATISTICAL INFERENCE

- **THE STANDARD NORMAL DISTRIBUTION**
- **For any variable in a sample, it is possible check if it is normally distributed and to even construct one or many ordinary normal curves.**
- **One way of checking for normality is to do the following:**
- **Check if the observations are symmetrical (i.e. 50% below and above the mean).**
- **Construct a histogram or polygon and see if observations are clustered around the center.**
- **Check if the mean, median, and mode are the same or nearly the same.**

STATISTICAL INFERENCE

- **One way problem with ordinary normal curves is that you many ordinary normal curves even for one variable, each with mean and standard deviation. curves for every combination of σ and μ.**

- **This poses challenges in comparing different normal curves especially when variable is measured in different units.**

STATISTICAL INFERENCE

- **A standard normal distribution closely resembles the ordinary normal curve so that even the relationships between the standard deviations and the areas will hold up reasonably well.**

- **The major difference between an ordinary normal curve and a standard normal distribution is that for the ordinary normal curve there are many different normal curves for every combination of σ and μ.**

- **The standard normal curve always has has a mean, μ, of 0 and a standard deviation, σ, of 1.**

- **Standardization is necessary to facilitate comparisons of variables measured in different units on the same footing.**
-
- **Standardization is achieved through the use of STANDARD SCORES also known as STANDARD NORMAL DEVIATES or Z – SCORES.**

STATISTICAL INFERENCE

- **This expresses values or scores in terms o STANDARD DEVIATION UNITS, using the formula:**

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

- **In the absence of information on population values, this is the formula:**

- **$z = \frac{x - X}{s}$**

- **Conversion of an ordinary normal curve into these standard units is the basis for the tables of the normal curve.**

THE STANDARD NORMAL DISTRIBUTION

x	(x-X)	$(x-X)^2$	z (Population)	z (Sample)
20	-30	900	-1.22	-1.16
40	-10	100	-0.41	-0.39
50	0	0	0.00	0.00
50	0	0	0.00	0.00
90	40	1,600	1.63	1.55
70	20	400	0.82	0.77
30	-20	400	-0.82	-0.77
60	10	100	0.41	0.39
10	-40	1,600	-1.63	-1.55
80	30	900	1.22	1.16
		6,000		
Mean	50		0.00	0.00
Standard deviation	24.49 (Population)		1.00	1.00
	25.82 (Sample)			

- **HOWE TO USE THE TABLES OF THE NORMAL CURVE**

- **The z – scores on the extreme left column while the proportions or areas under the curve are to the right.**
-
- **A value of z –score of 1 is equal to an area of 0.3414 while a value of a z-score of 3 corresponds to an area under the curve of 0.4987.**

- **Standard scores to the left of the centre or the mean are smaller; standard scores to the right are larger.**

- **For any z-score, score to the left is represent smaller values while those to the right represent larger values.**

THE STANDARD NORMAL DISTRIBUTION

- **RULES TO FOLLOW WHEN WORKING WITH Z – SCORES**

- **If the z – scores are located on opposite sides of the centre, add the corresponding proportions.**

- **If the z – scores are located on the same side of the centre, subtract the corresponding proportions.**

THE STANDARD NORMAL DISTRIBUTION

				Sample
	x	x-X	$(x-X)^2$	z
JB	20	-30	900	-1.16
PP	40	-10	100	-0.39
JC	50	0	0	0
MP	50	0	0	0
PB	90	40	1600	1.55
JC	70	20	400	0.77
PM	30	-20	400	-0.77
ML	60	10	100	0.39
AK	10	-40	1600	-1.55
PC	80	30	900	1.16
			6000	
	Mean	50		0.00
	Standard deviation	24.49		1.00

	x	(x-X)	$(x-X)^2$	z (Population)
JB	20	-30	900	-1.22
PP	40	-10	100	-0.41
JC	50	0	0	0
MP	50	0	0	0
PB	90	40	1600	1.63
JC	70	20	400	0.82
PM	30	-20	400	-0.82
ML	60	10	100	0.41
AK	10	-40	1600	-1.63
PC	80	30	900	1.22
			6000	
	Mean	50		0
	Standard deviation	24.49		1

THE STANDARD NORMAL DISTRIBUTION

OPPOSITE SIDES				
PP+PB	Z	Proportion		
PP	-0.41	0.1591		
PB	1.63	0.4485		
		0.6076		

THE STANDARD NORMAL DISTRIBUTION

SAME SIDE				
JB+PP				
JB	**-1.22**	**0.3888**		
PP	**-0.41**	**0.1591**		
		0.2297		

THE STANDARD NORMAL DISTRIBUTION

- **PERCENTILES AND Z – SCORES**

- **PP has 34% below him. Therefore is on the 34th percentile.**
- **PB is on the 94th percentile**
- **JB is on the 11th percentile**
- **JC is on the 79th percentile**

THE STANDARD NORMAL DISTRIBUTION

- **SOLVING FOR A SCORE GIVEN A Z-SCORE**

- **What does a z-score of 1.34 correspond to?**

- **1.34 = x -50/24.49**
- **X = 83**

STATISTICAL INFERENCE

- **THE USE OF THE NORMAL CURVE IN STATISTICAL INFERENCE**

- **One of the other important uses of the standard normal curve is in the estimation of population parameters on the basis of sample statistics and hypotheses testing.**

- **For example, by using the standard normal deviate/or standard score, we can compute the distance between the parameter (e.g. the population mean) and the statistic (e.g. the sample mean) to determine how far away the statistic is from the parameter.**

- **To do, this however, we first have to establish that the distribution of the values of a parameter resembles a normal distribution.**

- **This requires some knowledge of the sampling distribution, the LAW OF LARGE NUMBERS and CENTRAL LIMIT THEOREM**

STATISTICAL INFERENCE

- **THE LAW OF LARGE NUMBERS**

- **A basic finding about parameters is that if one draws a sufficiently large sample size from a population, the distribution of the statistic will become similar to the distribution of the parameter.**

- **In other words, a population parameter such as the average of UNZA students can be estimated from the sample statistics as the sample size increases.**

STATISTICAL INFERENCE

- **CENTRAL LIMIT THEOREM**

- **Another important finding about parameters is that if one draws samples of the same size, repeatedly from a population and computes the sample statistics, say the MEAN, for each of the samples, the SAMPLING DISTRIBUTION of the means will be approximately normal and the mean of this distribution will be equal to the population mean.**

- **This is borne out by example on the handout.**

- **In statistical inference, these two laws are important.**

- **As a rule of the thumb, the same principles of the sampling distribution hold true for samples of large size regardless of how the population is distributed.**

- **Sufficiently large samples of 100 or more are large enough to yield normally distributed sampling distribution means.**

STATISTICAL INFERENCE

- **THE IMPORTANCE OF THE SAMPLING DISTRIBUTION**

- **The sampling distribution of means is an important distribution to know because it is the one that is made use of in significance tests and hypothesis testing all of which are very relevant in statistical inference.**

- **The sampling distribution of means is conceptually similar to the ordinary normal distribution in several respects.**

- **The only distinguishing characteristic is that with an ordinary normal distribution we normally talk of a number of observations, Xi values, for example grades in DEM 2414.**

STATISTICAL INFERENCE

- **In the case of the sampling distribution instead of talking of Xi values as our points of observations, we talk in terms of X is sample means, as our points of observation.**
-
- **Since these observations, Xi samples are normally distributed, all the characteristics of the normal curve apply to the distribution of sample means.**
-
- **This means that if we draw any sample and compute its mean, there is 68% probability that this mean will be within $\mu \pm \sigma$ of the true population mean, the population; there is a 95% probability that it will be $\mu \pm 2\sigma$ of the true population mean; and 99% probability within $\mu \pm 3\sigma$ of the true population mean.**
-

- **THE STANDARD ERROR OF THE MEAN**

- **The standard deviation of the sampling distribution of the means is called the standard error of the mean.**
-
- **This is expressed in a formula:-**
-
- $SE = \frac{\sigma_x}{\sqrt{n}}$
-
- **Where σ_x = standard deviation of the population**
-
- **n =sample size**
-

- **Since the standard deviation of the population means is often unknown, sample statistics are used instead:**
-
- $SE = \frac{s_x}{\sqrt{n}}$
-

- **Knowledge of the standard error of the sample mean can be used estimated how accurately the sample mean estimates the population mean.**
-
- **The smaller the standard error, the more accurately the sample mean estimates the population mean and vice versa.**

- **ESTIMATION OF PARAMETERS**
-
- **Rarely do we have all the information we need on the population.**
-
- **In addition, it is tedious and time consuming to draw several samples from population sample means.**
-
- **Instead we only have to draw one sample from the population because of various constraints.**
-
- **With these limitations in mind, there are two types of parameter estimates can be made on the basis of information from one sample namely:**
-
- **Point estimates**
- **Interval estimates**
-

INFERENCES ABOUT ONE VARIABLE

These include the point estimate and interval estimate.

THE POINT ESTIMATE

- **This is the simplest example of statistical inference.**
- **As long as the sample is representative, you simply use the mean of the sample to represent the population mean.**

- **POINT ESTIMATES**

- **In this case, the sample statistical is simply taken to estimate the population parameter.**

- **For example, the average age of students from a sample is used to represent the average age of all UNZA students.**

- **Thus the point estimate of a population mean is:**

| $\mu = \bar{x}$ | or | $\bar{x} = \mu$ |

- **The error of estimation is given by:**

$| \bar{x} - \mu |$ or $| \mu - \bar{x} |$

- **In other words, the error of estimation is the absolute difference between what we think the parameter value is and what is actually is.**
-
- **EXAMPLE**
-
- **Given a random sample of n 100 UNZA students, with a mean X = 24.7 years spent on campus and a standard deviation of s=.1.02 years, our point estimate of u is 24.7 years.**

- **WHAT DETERMINES THE GOODNESS OF AN ESTIMATE?**
- **UNBIASEDNESS**
- **An estimate is unbiased if the mean of the sampling distribution equals the population parameter.**
- **$| \bar{x} = \mu |$**
- **Given a sufficiently large sample size the mean of the sampling distribution will equal the population parameter and will hence be unbiased.**

CONSISTENCY

- **An estimate is consistent when it gets closer to the parameter as the sample size becomes larger and larger.**

EFFICIENCY

- **An estimate is efficient when the standard error of the statistic for the sampling distribution is small.**

- **The smaller the standard error, the more efficient the estimate.**

- **CONFIDENCE INTERVAL ESTIMATES**

- **Since the point estimate can sometimes be far off the mark, it is therefore better to know the interval estimate.**

- **This provides a range of values likely to include an unknown population value.**

- **According to the central limit theorem and the sampling distribution, and properties of the normal curve, the interval, $\mu \pm 2\sigma$ includes 95% of the sample mean, X's in repeated sampling.**

- **Conversely, in the absence of information on the population mean, the interval, $X \pm 1.96$ will contain the population mean with 0.95 probability.**

- **The formula for the confidence interval estimate is:**
-
- **Confidence interval = sample statistic ± margin of error**
-
- **Margin of error = Critical value * Standard error**
- **Critical value = z from the standard normal distribution.**
- **Standard error = s √ n**
-
- **s = standard deviation from the sample**
- **n = sample size**

- **The sample statistic is often the mean based on sample data for a particular variable like age.**

THE CONFIDENCE INTERVAL ESTIMATE

- **THE MARGIN OF ERROR**

- **This is a measure of the difference between the estimate from the sample and the population value. The formula for the margin of error is:**

- **Margin of error = Critical value * Standard error**

- **A small margin of error means the sample value is estimating the population value more accurately than a large margin error.**

- **Ideally the margin of error should be ±5%.**

- **THE CONFIDENCE COEFFICIENT**

- **In interval estimate, we also want to know the confidence or certainty we can have that the interval estimate contains the population value.**

- **The computation of the confidence interval estimate therefore requires information on the confidence coefficient.**

- **The commonly used confidence coefficient used is 95% and the critical value associated to this is 1.96. therefore we now have:**

- **The confidence coefficient gives the probability than an interval encompasses the parameter to be estimated.**

- **For any critical value, a confidence coefficient can be found.**

- **For example, for a z – score of 1.96, the confidence coefficient is 0.95.**

- **Using the formula for the standard error:**
-
- **SE = $\frac{s_x}{\sqrt{n}}$**
- = $\frac{10.2}{\sqrt{(25)}}$
- = $\frac{\mathbf{10.2}}{5}$
- =**2.04**

THE CONFIDENCE INTERVAL ESTIMATE

- **Given a standard error of 2.04, the margin error is:**

- **Margin of error = 1.96 * 2.04**
  - **= 3.9984**

- **Therefore, the confidence interval estimate:**
-
- **= 42 ± 3.9984**
- **= 38.0016 - 45.9984**

- **Interpretation:**

- **We can be 95 percent confident or certain that the mean age for all UNZA employees is between 38 and 46 years.**

CONFIDENCE INTERVAL FOR A PROPORTION

- **In a random sample of 100 students, 68% have not been to a VCT centre.**

- **Establish the confidence interval estimate of UNZA students who do not go for VCT.**

CONFIDENCE INTERVAL FOR A PROPORTION

- **Sample statistic = 68%**
- **Standard deviation for a proportion= √ p(1-p)**
- **= √ .68(1-.68)**
- **= √ .68(.32)**
- **= √ 0.2176**
- **= 0.47**

CONFIDENCE INTERVAL FOR A PROPORTION

- **COMPUTATION PROCEDURE FOR THE CONFIDENCE INTERVAL ESTIMATE**
-
-
- **Sample statistic = 68%**
- **Standard deviation for a proportion= √ p(1-p)**
- **= √ .68(1-.68)**
- **= √ .68(.32)**
- **= √ 0.2176**
- **= 0.47**

CONFIDENCE INTERVAL FOR A PROPORTION

- **The standard error of a proportion $= \sqrt{\frac{p(1-p)}{n}}$**
-
-
- **$= \sqrt{\frac{0.68(.32)}{100}}$**
-
-
- **$= 0.047$**

CONFIDENCE INTERVAL FOR A PROPORTION

- **Critical value for 95% confidence = 1.96**
- **Sample size = 100**
- **Margin of error = 1.96 * .047**
- **= 1.96 * 0.047**
- **= ± 0.092**
- **= 0.68 ± 0.092**
- **= 0.588 - 0.772**
-
- **= 59% and 77%**
-
- **Interpretation:**

- **We can be 95 percent confident or certain that the percentage of students not going for VCT is between 59 and 77 percent.**

HYPOTHESIS TESTING

- **Hypothesis testing is another way of making inferences.**
- **The test involves comparison of the data mean from the sample (the observed value) with the mean of the sampling distribution (the expected value).**
- **It rests on the a priori knowledge about the variable of interest.**
- **It involves asking the question: is the population mean equal to a specified value of μ_o.**

HYPOTHESIS TESTING

- **When the sample mean deviates substantially from the mean of the sampling distribution (or population), one rejects the hypothesis that states that the sample mean and the population mean are equal.**

- **One then leans towards accepting the hypothesis that states that the sample mean and population mean are different.**

HYPOTHESIS TESTING

- **In general, the test of a null hypothesis is a question of probability.**
- **More specifically, it is a question of conditional probability.**
- **The question is: What is the probability of my results given that the null hypothesis is true?**
- **In order to answer this question, one should follow some basic steps:**

HYPOTHESIS TESTING

- **SPECIFY THE INDEPENDENT AND DEPENDENT VARIABLES ALONG WITH THEIR RESPECTIVE LEVELS OF MEASUREMENT.**

- **Indicate the characteristics of the independent and dependent variables.**

HYPOTHESIS TESTING

- **STATE THE NULL AND ALTERNATIVE HYPOTHESES AND INDICATE WHETHER THE TEST IS ONE-TAILED OR TWO-TAILED.**
- **ALTERNATIVE OR RESEARCH HYPOTHESIS**
- **This is a hypothesis deduced from theory or other considerations.**
- **The alternative (or experimental) hypothesis states the result that you would expect if your independent variable had an effect.**

.
.
.

HYPOTHESIS TESTING

- **A one-tailed or directional alternative hypothesis one that states the direction of the relationship.**

- **For example, you could state direction by claiming that one group should have higher scores than another.**

HYPOTHESIS TESTING

- **NULL OR STATISTICAL HYPOTHESIS**

- **The null hypothesis is the outcome that you would expect simply by chance alone.**
- **This is the hypothesis to be directly tested and contradicts the theoretical or research hypothesis.**

- **Rejection of the null hypothesis increases the probability that the theoretical hypothesis could be correct.**

- **Or that the outcome of our research is probably not due to chance.**

HYPOTHESIS TESTING

- **NULL OR STATISTICAL HYPOTHESIS**

- **Acceptance of the null hypothesis means that the theoretical or research hypothesis could very well be mistaken.**

- **Or that the outcome of our research is probably due to chance.**

HYPOTHESIS TESTING

- **ASSUMPTIONS IN HYPOTHESIS TESTING**
- **These assumptions refer to the conditions that must be fulfilled for us to adopt a particular statistical test for testing our hypotheses.**
- **These assumptions concern the following issues:**
- **Sample size: If the sample is sufficiently large n> 30.**
- **Normality: The parameter will be assured to be normally distributed if n> 30.**
- **Random sampling: Normality is anly likely to be achieved if sample is representative of the population.**

HYPOTHESIS TESTING

- **Sample design: Whether probability or nonprobability sampling has been used.**
-
- **Scale of measurement: Is this nominal, ordinal, interval or ratio**
-
- **These assumptions determine the appropriateness of the statistical test – parametric or non-parametric – to use.**

HYPOTHESIS TESTING

- **OBTAINING THE SAMPLING DISTRIBUTION**

- **If normality of the distribution of the parameter is assumed, the standard normal distribution and tables are used in the selection of critical values.**

- **In large sample tests to estimate μ_o, the parameter, we use the z-distribution.**

- **In small sample tests, we use the t-distribution.**

HYPOTHESIS TESTING

- **CHOOSING THE SIGNIFICANCE LEVEL AND CRITICAL REGIONS**
- **In choosing the significance level and critical regions, two types of errors can be committed in relation to the choice between our theoretical and our null hypothesis.**
- **i) TYPE I ERROR**
- **This refers to the error of rejecting a null hypothesis when it is in fact true.**
- **The probability of committing a Type I error is known as the level of significance.**
- **The level of significance is designated as ALPHA (α)**

HYPOTHESIS TESTING

- **TYPE II ERROR**

- **This arises from the mistake of accepting a null hypothesis when it is false and the research hypothesis true.**

- **This is denoted by the symbol BETA (*β*).**

HYPOTHESIS TESTING

- **THE POWER OF A STATISTICAL TEST**

- **The probability of *not* committing a Type II is called the power of a hypothesis test.**

  Power = 1- BETA (*β*).

- **Power may be defined as the probability of correctly rejecting the null hypothesis.**

HYPOTHESIS TESTING

- **COMPUTING THE TEST STATISTIC**

- **From the sample data, we calculate test statistic, the observed value, the mean and standard deviation for computing our test statistic either z- or t – score.**

HYPOTHESIS TESTING

- ***DECISION MAKING***

- ***Compare the Calculated Value to the Critical Value.***

- **For this step, indicate whether the calculated value is greater or less than the critical.**

HYPOTHESIS TESTING

- **If the calculated value is larger, then it has a probability lower than alpha.**
- **In this case, you reject the null hypothesis and doubt whether it is really true.**
- **If the calculated value is smaller than the critical value, then the probability of getting the calculated value when the null hypothesis is true is high.**
- **In this case, you fail to reject the null hypothesis (note that this is different from accepting the null hypothesis!).**

HYPOTHESIS TESTING

- **CONCLUSION**

- ***State Your Conclusion in Words.***

- **Lastly, you should go beyond the mathematics to answer the question that was posed.**

- **For example, if purpose of the study was to test to see if two groups scored differently, then step seven should state whether or not the two groups differed significantly on the dependent variable.**

- **TESTING THE HYPOTHESES ABOUT A MEAN FOR LARGE SAMPLE: SINGLE SAMPLE SITUATION**

- **AGE AT MARRIAGE**

- **A demographer questions a CSO report that that puts the mean age of marriage among UNZA female students at 22 years. To prove his point, he collects data based on a random sample of 100 UNZA female students and finds a mean age of 23 years with a standard deviation of 4.4 years.**
- **Is the demographer on the right track?**

- **STATEMENT OF HYPOTHESES**

- **Ho: $\mu = 22$**

- **Hi: $\mu > 22$**

- **ASSUMPTIONS**

- **The subjects are randomly and independently selected.**
- **The population distribution is normal in form**
- **The scale (or level) of measurement is interval**

- **DECISION RULES**

- **Given 0.05 level of significance for a large sample, use the z – distribution and a directional test, the critical value is 1.65. Therefore,**

- **If Zobs < 1.65, accept Ho.**

- **If Zobs ≥ 1.65, reject Ho.**

- **COMPUTATION**
-
- **$z = \frac{\bar{x} - \mu}{SE}$**
- **$SE = \frac{s_x}{\sqrt{n}}$**

- **$Z = \frac{23 - 22}{\frac{4.4}{\sqrt{100}}}$**
-
- **$= \underline{2.27}$**

- **<u>DECISION</u>**
- **Since Zobs of 2.27 is larger than 1.65, reject Ho.**
- **CONCLUSION**
- **It is highly likely that the demographer’s claims are true.**

- **TESTING A HYPOTHESES ABOUT A PROPORTION: LARGE SAMPLE SITUATION**

- **HIV PREVALENCE**

- **A researcher working for an NGO informs the Ministry of Health that the HIV prevalence among UNZA students is 20 percent. A demographer from the Department of Population Studies disputes this and argues that with the increasing promiscuity among students the situation is actually worse. To prove his point, he collects data based on a random sample of 400 students and establishes an HIV prevalence rate of 24 percent.**

- **Who is telling the truth?**

- **STATEMENT OF HYPOTHESES**

- **Ho: P μ = 0.20, the claim is false**
-
- **Hi: P μ > 0.20, the claim is true**

- **ASSUMPTIONS**

- **The subjects are randomly and independently selected.**
- **The population distribution is normal in form**
- **The scale (or level) of measurement is interval**
-

- **DECISION RULES**

- **Given 0.05 level of significance for a large sample, use the z or normal) distribution, with a critical value of +1.645. Therefore,**

- **If Zobs < +1.645, accept Ho.**

- **If Zobs ≥ +1.645, reject Ho.**

- **COMPUTATION**
-
- **Zobs = $\frac{ps - P\ \mu}{\sqrt{\frac{(p)(q)}{n}}}$**
-
-
- **Zobs = $\frac{0.24 - 0.20}{\sqrt{\frac{(0.24)(0.76)}{400}}}$**
-
- **= $\underline{1.87}$**

- **DECISION**
- **Since Zobs of 1.87 is greater than 1.645, reject Ho.**
- **CONCLUSION**
- **It is highly likely that the Demographer's claim is true.**

- **TESTING A HYPOTHESIS CONCERNING DIFFERENCE BETWEEN MEANS FOR LARGE SAMPLES**

-
- **MEAN AGE BY INSTITUTION**
-
- **The mean age of random sample of 30 mature age students is CBU is found to be 28 years with a standard deviation of 14 years. At UNZA, the mean age of a random sample of 40 mature age students is 27 years with a standard deviation of 10 years.**

- **Do you agree with a claim by students' union member's claim that there is a significant difference in the mean age of recruitment of mature age students at the two universities?**

- **STATEMENT OF HYPOTHESES**

- **Ho: $\mu_1 = \mu_2$**

- **Hi: $\mu_1 \neq \mu_2$**

- **ASSUMPTIONS**

- **The subjects are randomly and independently selected.**
- **The groups are independent from one another**
- **The population variances are equal or homogeneous**
- **The population distribution is normal in form**
- **The scale (or level) of measurement is interval**

- **DECISION RULES**

- **This is a non-directional test, with a critical value of 1.96. Therefore,**

- **If -1.96 <z < -+1.96, accept Ho**
- **If z $\geq$ +1.96 or z $\leq$ -1.96, reject Ho**

- **COMPUTATION**

- **$Zobs = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma^2}{n1} + \frac{\sigma^2}{n2}}}$**

- **$\frac{28 - 27}{\sqrt{\frac{14^2}{30} + \frac{10^2}{40}}}$**

- **$Zobs = \frac{28 - 27}{\sqrt{\frac{14^2}{30} + \frac{10^2}{40}}}$**

- **$= \underline{0.33}$**

- **DECISION**
- **Since the observed zobs falls within the area of acceptance, accept Ho.**
- **CONCLUSION**
- **The student union member’s claim is false .**

- **TESTING A HYPOTHESIS CONCERNING DIFFERENCE BETWEEN PROPORTIONS FOR LARGE SAMPLES**

-
- **CONTRACEPTIVE KNOWLEDGE BY REGION**
- **A research agency takes a sample of 1,000 people on the Copperbelt and finds that 200 know about modern contraceptives. In Lusaka, research also reveals that based on a random sample of 1,091 people, 240 respondents know about modern contraceptives.**

- **Would you agree with the conclusion that Copperbelt residents are more knowledgeable about modern contraceptives than Lusaka residents?**

- **STATEMENT OF HYPOTHESES**

- **Ho: $p\mu_1 = p\mu_2$**

- **Hi: $p\mu_1 > p\mu_2$**

- **ASSUMPTIONS**

- **The subjects are randomly and independently selected.**
- **The groups are independent from one another**
- **The population variances are equal or homogeneous**
- **The population distribution is normal in form**
- **The scale (or level) of measurement is interval**

- **DECISION RULES**
-
- **This is a directional test, with a critical value of 1.65. Therefore,**

- **If zobs < +1.65, accept Ho**
- **If zobs $\geq$ +1.65, reject Ho**

- COMPUTATION

- **Zobs = $\frac{P_1 - P_2}{\sqrt{\frac{pq}{n1} + \frac{pq}{n2}}}$**
- **$\frac{0.20 - 0.22}{\sqrt{\frac{0.20(0.80)}{1000} + \frac{0.22(0.78)}{1091}}}$**
- **= -1.12**

- **DECISION**
- **Since the observed zobs falls within the area of acceptance, accept Ho.**
- **CONCLUSION**
- **The research agency’s claim is not likely to be correct.**

HYPOTHESIS TESTING FOR SMALL SAMPLES

- **THE T- DISTRIBUTION**
- **This is used when there are small samples because in some situations it is not possible to obtain a large.**
- **For example, a study of the handicapped on campus might not yield a sample larger than 30.**
- **Louis Gosset found that he was falsely rejecting the null hypothesis at a much higher rate:**
- **He therefore came up with what is now known as the student's t distribution.**

- **PROPERTIES OF THE STUDENT OR T - DISTRIBUTION**
- **The t distribution like that of z is symmetrical about the mean O.**
- **The t distribution is more variable .**
- **There are, however, many different t distribution.**

- **Each distribution if specified by a parameter called degrees of freedom (df)**

- **The degree of freedom is:-**

- **df = n – 1 for a single sample or group or df=n1+n2-2 for two sample or two group situation.**

- **As n (or equivalently) increases, the distribution of t approaches the distribution of z.**

- **USING TABLES FOR T**
- **Depending on the test if H_1 , assess whether it is directional or not.**
- **Decide on the level of significance.**
- **Then determine the degrees of freedom and choose the critical value.**
- **The critical value is fond at the intersection of the degrees of freedom and the significance level.**
- **For example, for a sample size of 10, the degrees of freedom are 10-1 = 9 at 5% level of significance, the critical value is :**

- **HUMAN RESOURCE MANAGEMENT EXAMPLE**

- **At a management meeting of ZTK Investment Holdings, the Managing Director (MD) argues that the employees in the company are aging because he believes that the mean age is about 44 years.**
- **On the basis of this, he tells the Human Resource Manager (HRM) that the older employees should be retrenched and new younger workers be recruited.**

- **The HRM, however, argues that there is no need to recruit new younger staff because the workers are actually younger than the Managing Director is suggesting.**

- **To prove his point the HRM engages a consultant who finds on the basis of a random sample of 10 that the mean age is 42 years with a standard deviation of 6 years,**

- **Using a 5 percent level of significance, who between the MD and the HRM is right?**

EXAMPLES

- **STATEMENT OF HYPOTHESES**

- **Ho: U=44 years**
-
- **Hi: U<44 years**

- **<u>ASSUMPTIONS</u>**

- **The subjects are randomly and independently selected.**
- **The population distribution is normal in form**
- **The scale (or level) of measurement is interval**

- **DECISION RULES**

- **Given 0.05 level of significance for a small sample, use the t – distribution, with 9 df and a directional test, the critical value is 1.833. Therefore,**
-
- **If tobs > - 1.833, accept Ho.**
-
- **If tobs ≤ -1.833, reject Ho.**

- **COMPUTATION**
-
- $$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$
-
- $$t = \frac{42 - 44}{\frac{6}{\sqrt{10}}}$$
-
- $$= -1.49$$
-

- **Decision**
- **Since tobs of –1.49 is larger than -1.729, accept Ho.**
- **Conclusion**
- **It is highly likely that the MD is right.**

- **TESTING A HYPOTHESES FOR A DIFFERENCE BETWEEN MEANS: SMALL SAMPLE SITUATION**

- **The monthly incomes of two groups of salesmen are being investigated to see if there is a difference in the average income received.**

- **Random samples of 12 and 9 are taken from the two groups.**

- **Test the hypothesis that salesmen from Zambia Breweries are better paid than those from Copperbelt Bottling Company.**

- **Use 5% level of significance. The data are presented below:**

	Zambia Breweries	Copperbelt Bottling Company
n	12	9
x	1,060	970
s	63	76

- **STATEMENT OF HYPOTHESES**

- **Ho: $U_1 = U_2$**

- **Hi: $U_1 > U_2$**

- **ASSUMPTIONS**

- **The subjects are randomly and independently selected.**
- **The groups are independent from one another**
- **The population variances are equal or homogeneous**
- **The population distribution is normal in form**
- **The scale (or level) of measurement is interval**

- **DECISION RULES**

- **This is a directional test, df=n1+n2-2= 19 df, with a critical value of 1.729. Therefore,**
-
- **If $t < +1.729$, accept Ho**
- **If $t \geq +1.729$, reject Ho**

$$t = \frac{X_1 - X_2}{\sqrt{\left[\frac{(n-1)_1(s_1)^2 + (n-1)_2(s_2)^2}{n_1 + n_2 - 2}\right]\left[\frac{1+1}{n_1 \; n_2}\right]}}$$

$$t = \frac{1060 - 970}{\sqrt{\left[11\frac{(63)^2 + 8(76)^2}{12 + 9 - 2}\right]\left[\frac{1+1}{12\ 9}\right]}}$$

$$= \underline{2.97}$$

- **DECISION**
- **Reject Ho.**
- **CONCLUSION**
- **There is sufficient evidence to conclude that salesmen at Zambia Breweries are better paid than those from Copperbelt Bottling Company.**

ANALYSIS OF VARIANCE (ANOVA)

- **In the case of the t-test and the z-test for the difference between means, we are concerned with difference between two means.**

- **With analysis of variance (ANOVA), we have to deal with three or more means.**

- **For example, we may want to compare crime rates by three residential types, namely, low, medium, and high.**

ANALYSIS OF VARIANCE (ANOVA)

- **But instead of dealing directly with means, ANOVA also refereed to as the F-test involves working directly with variances.**
-
- **For this reason, two independent estimates of a common variance are required.**
-
- **One of these, based upon variability between groups is called between group variance.**
-
- **The other, based upon variability within groups is called within groups variance.**

THE UNDERLYING LOGIC OF VARIANCE

- **The underlying logic in ANOVA is to determine if the differences among the group means is significant by comparing them to the variation within groups.**

- **This takes the form of a comparison of the variance between groups with the variance within groups I a ratio called the F- ratio.**

- **F – ratio = Variance between groups /Variance within groups**

- **DETERMINING THE SIGNIFICANCE OF THE F – RATIO**

- **Usually the larger the F-ratio, the greater is the significance of the difference among sample means.**

- **A lower F-ratio means little significance among the sample means.**

- **EXAMPLE**

- **The objective is to establish if there are statistically significant differences in the performance of students from three tutors in DEM 2414 at 5% level of significance.**
- **To do this, the course coordinatorselects 4 students randomly from each tutorial group.**

	Tutor 1	Tutor 2	Tutor 3
	80	70	63
	92	81	76
	87	78	70
	83	74	58
MEAN	85.5	75.75	66.75

- **STATEMENT OF HYPOTHESES**

- **$H_{0:}\ \mu_1 = \mu_2 = \mu_3$**
- **$H_1: \mu_1 \neq \mu_2 \neq \mu_3$**

- **No directional hypotheses are given in ANOVA.**

- **ASSUMPTIONS**

- **Subjects are independently and randomly selected**
- **Groups are independent from one another**
- **Population variances are homogeneous**
- **The population distribution is normal**
- **The scale of measurement is interval**

- **DECISION RULES**
- **USING THE TABLES OF THE F-DISTRIBUTION**
- **ANOVA unlike the other tests uses an F-distribution.**
- **In testing a hypotheses concerning three means, the computed F – ratio has to be compared with the F – critical values given in the table.**
-
- **To get the F – critical, we must first have the degrees of freedom for the two variance estimates and a specified level of significance.**

- **For the variance between, the degrees of freedom are (NUMERATOR):**
-
- **Dfb = j-1**
- **= 3 -1 = 2**
- **For the variance within, the degrees of freedom are (DENOMINATOR):**
-
- **Dfw = n – j**
- **= 12 -3 = 9**
-
- **j = number of groups, samples or categories**
- **n = sample size.**
-

- **To get the F – critical, we look at the point of intersection between the numerator and the denominator.**

- **For example, the degrees of freedom for dfb=3-1=2df**

- **Therefore F-critical is 4.26**

- **If F – ratio < F – critical, accept Ho**
- **If F – ratio ≥ F – critical, reject Ho**

- **COMPUTATION**

- **COMPUTATION OF THE GRAND MEAN**

- **Sum all observations in all columns and rows and divide by the total number of observations thus:**

Tutor 1	Tutor 2	Tutor 3
80	70	63
92	81	76
87	78	70
83	74	58

- $\bar{X} = \sum\sum X/n$
-
- $= 912/12$
-
- $=\underline{76}$

- **COMPUTATION OF THE TOTAL SUM OF SQUARES**

- **Find the sum of squared deviations of the observations from the GRAND MEAN**

Tutor 1	$(x- X)^2$	Tutor 2	$(x- X)^2$	Tutor 3	$(x- X)^2$
80	16	70	36	63	169
92	256	81	25	76	0
87	121	78	4	70	36
83	49	74	4	58	324
85.5	**442**	**75.75**	**69**	**66.75**	**529**

- **COMPUTATION OF THE SUM OF SQUARES BETWEEN**
- **This first involves the computation of the CATEGORY MEANS for each CATEGORY or COLUMN.**
- **Then the SQUARED DEVIATIONS of the CATEGORY MEANS from the GRAND MEAN is calculated.**
- **Then each SQUARED DEVIATION is multiplied by the CATEGORY SIZE as a WEIGHTING FACTOR.**
- **The TOTAL SUM OF SQUARES BETWEEN is found by summing up all these products.**

	n	$(X- X)^2$	$n(X- X)^2$
Tutor 1	4	90.25	361
Tutor 2	4	0.0625	0.3
Tutor 3	4	85.563	342
Total			704

- **COMPUTATION OF THE VARIANCE BETWEEN GROUPS**

- **This is found by dividing the TOTAL SUM OF SQUARES BETWEEN by the DEGREES OF FREEDOM BETWEEN.**

- **VB = SSB/DFB = 703.5/2= 351.75**

- **COMPUTATION OF THE TOTAL SUM OF SQUARES WITHIN GROUPS**

- **This is done by subtracting the TOTAL SUM OF SQUARES BETWEEN from the TOTAL SUM OF SQUARES.**

- **SSW = TSS – SSB**
- **SSW = 1040 – 703.5 = 336.5**

- **COMPUTATION OF THE VARIANCE WITHIN GROUPS**

- **This is found by dividing the TOTAL SUM OF SQUARES WITHIN GROUPS by the DEGREES OF FREEDOM WITHIN GROUPS.**

- **VW = SSW/DFW = 336.5/9 = 37.39**

- **COMPUTATION OF THE F – RATIO**

- **This is done by dividing the VARIANCE BETWEEN and the VARIANCE WITHIN,**
- **F – ratio = VB/VW = 351.75/37.39 = 9.41**

- **DECISION**

- **Reject Ho.**

- **CONCLUSION**

- **It is likely that there are differences in the way tutors are handling the course.**

INFERENTIAL STATISTICS

THE P – VALUE APPROACH TO HYPOTHESIS TESTING

- **In situations where software is used in testing hypothesis, it is often not necessary to follow the procedures in the preceding slides.**

- **The software will generate its own statistics and it is up to the researcher to do the interpretation.**

- **This requires the understanding of the so –called p – values and how to interpret them.**

- **The P – values are all about probability and are just another way of talking about levels of significance.**

- **For example, the probability that the observed differences or relationship could have occurred by chance is called the significance level.**

- **The significance level is normally written as p< followed by the probability that the result could have occurred by chance.**

- **The most commonly used significance level is p<.05**

INFERENTIAL STATISTICS

- **Thus p<.05 means that the probability due to chance is less than 1 in 20.**

- **If p<.05, it means you will be wrong 5% of the time when coming to your conclusion.**

- **Therefore, in trying to establish if the relationship or difference is statistically significant use these simple rules:**
    - **If the observed or computed p value is greater than .05 (p>.05), then the observed difference or relationship is more likely due to chance.**
    - **Decision: Accept H0**
- **If the observed or computed p value is less than or equal to 0.05 ($p \leq .05$), then the observed difference or relationship is not likely to be due to chance.**
    - **Decision: Reject H0**

CORRELATION ANALYSIS

- **Correlation refers to the existence of a relationship between two variables such that a change in one of the variables is accompanied by a change in the other variable**
- **If the changes in the variables are moving in the same direction, we have a positive correlation.**
- **If the changes are moving in opposite directions, we have a negative correlation.**
- **The Pearson product moment correlation coefficient, r, is used to measure the strength of the relationship between two variables.**
- **It also measures how well the data fit a straight line.**

- **PROPERTIES OF THE CORRELATION COEFFICIENT**

- **A correlation coefficient lies between -1 and +1.**
- **A value of +1 indicates a perfect positive correlation.**
- **A value of -1 indicates a perfect negative correlation**
- **A value of zero indicates the non-existence of a relationship.**

- **CORRELATION AND CAUSALITY**
- **Strong relationships should not be mistaken for causality.**
- **Sometimes you can have spurious relationships.**
- **For example, a strong relationship may exist between size of feet and performance in DEM 2414.**
- **There is also a possibility of spurious non-correlation when expected strong correlation turns out to be low or negative as in the case of income and the number of hours worked.**

INTERPRETATION OF THE SIZE OF r

- TABLE

CORRELATION	NEGATIVE		POSITIVE	
NONE	-0.09	0.00	0.00	0.09
SMALL	-0.30	-0.10	0.10	0.30
MEDIUM	-0.50	-0.30	0.30	0.50
HIGH	-1.00	-0.50	0.50	1.00

- =
- **COMPUTATION PROCEDURES**
-
-
- $$r = \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[n\Sigma X^2 - (\Sigma X)^2][n\Sigma Y^2 - (\Sigma Y)^2]}}$$
-

	Y	X	XY	X^2	Y^2
1	549	5.50	3,019.5	30.25	301,401
2	605	9.00	5,445.0	81.00	366,025
3	589	4.00	2,356.0	16.00	346,921
4	590	8.00	4,720.0	64.00	348,100
5	575	9.50	5,462.5	90.25	330,625
6	555	3.00	1,665.0	9.00	308,025
7	560	7.00	3,920.0	49.00	313,600
8	527	1.50	790.5	2.25	277,729
9	650	8.50	5,525.0	72.25	422,500
10	600	7.50	4,500.0	56.25	360,000
11	560	9.50	5,320.0	90.25	313,600
12	536	6.00	3,216.0	36.00	287,296
13	550	2.50	1,375.0	6.25	302,500
14	525	1.50	787.5	2.25	275,625
Total	7,971	83	48,102	605	4,553,947

$$r = \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[n\Sigma X^2 - (\Sigma X)^2][n\Sigma Y^2 - (\Sigma Y)^2]}}$$

REGRESSION ANALYSIS

- **This is a descriptive tool by which the linear dependence of one variable on another is determined.**
- **Example – predicting income on the basis of educational attainment.**
- **With multiple regression, the linear dependence of one variable on several variables can be determined.**
- **The most important aspects of the regression technique include:**
  - **Finding the best linear prediction equation**
  - **Evaluating its prediction accuracy or**
  - **In multiple regression, controlling for confounding factors in order to evaluate the specific contribution of a variable or set of variables.**

REGRESSION ANALYSIS

- **In simple regression analysis, values of a dependent variable are predicted from the equation:**

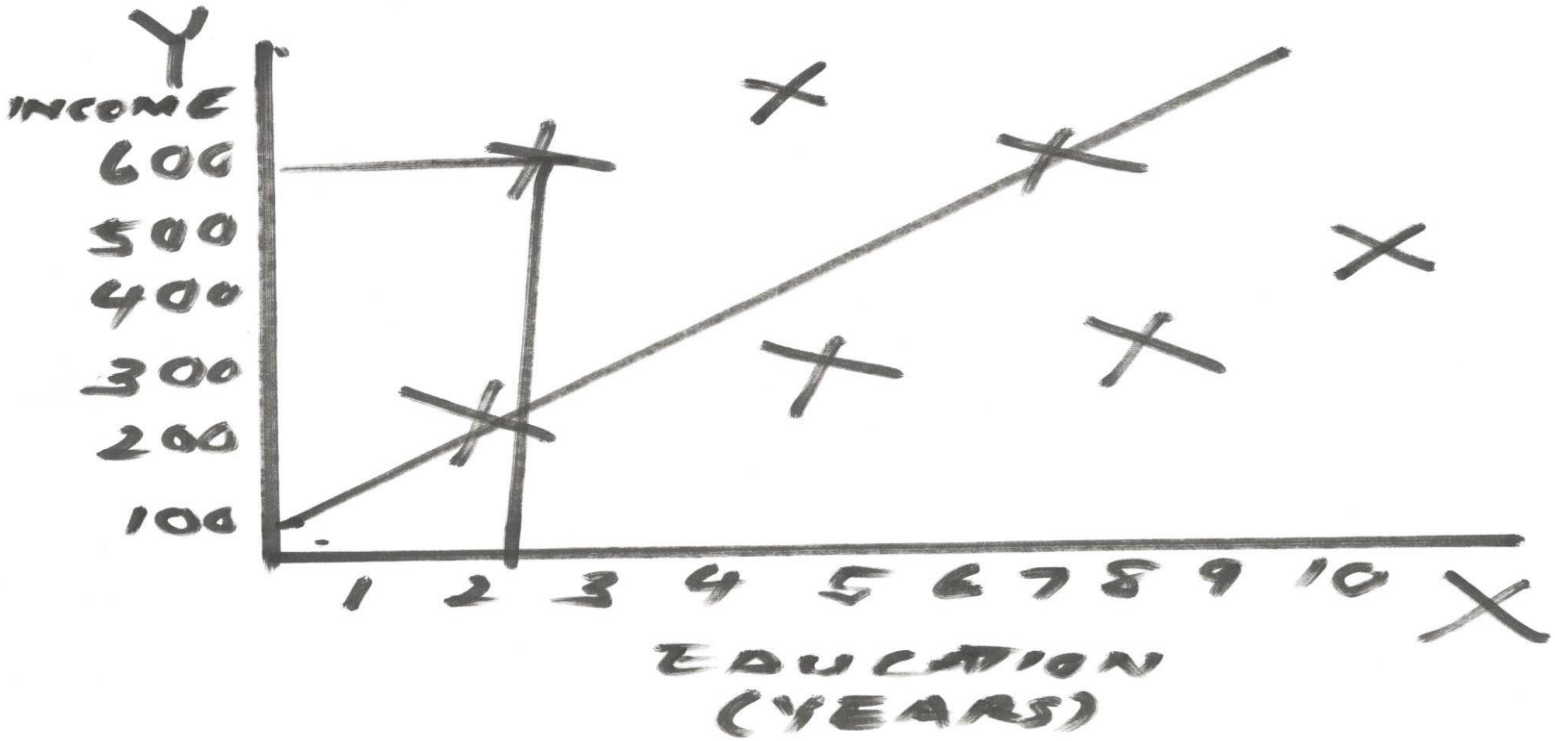
- **Y = A +BX**
- **Y' = Predicted value**
- **B = Constant by which X is multiplied**

- **A = A constant added to each case**

- **In multiple regression analysis the prediction equation is:**
- **Y = A + B1X1 + B2X2 ……BnXn**

THE GRAPHICAL APPROACH TO REGRESSION ANALYSIS

- **In simple regression analysis, the values of the dependent variable can be predicted by using the eyeball fitting technique.**
- **This is done by plotting data for X and Y on a scatter gram.**
- **The dependent variable is located on the vertical axis whilst the independent variable is on the horizontal axis.**


Y
INCOME
600
500
400
300
200
100
1 2 3 4 5 6 7 8 9 10
X
EDUCATION
(YEARS)

- **A ruler can then be used to draw a straight line that most accurately displays the linear trend of the data.**

- **The values of Y can then be predicted on the basis of X values on the scatter gram.**

- **The possibility of making incorrect readings is quite high as this depends largely on one’s visual acuity.**

THE METHOD OF LEAST SQUARES

- **A better and preferred method of making predictions is the method of least squares.**

- **This involves minimization or reduction of the sum of squared residuals along the regression line.**

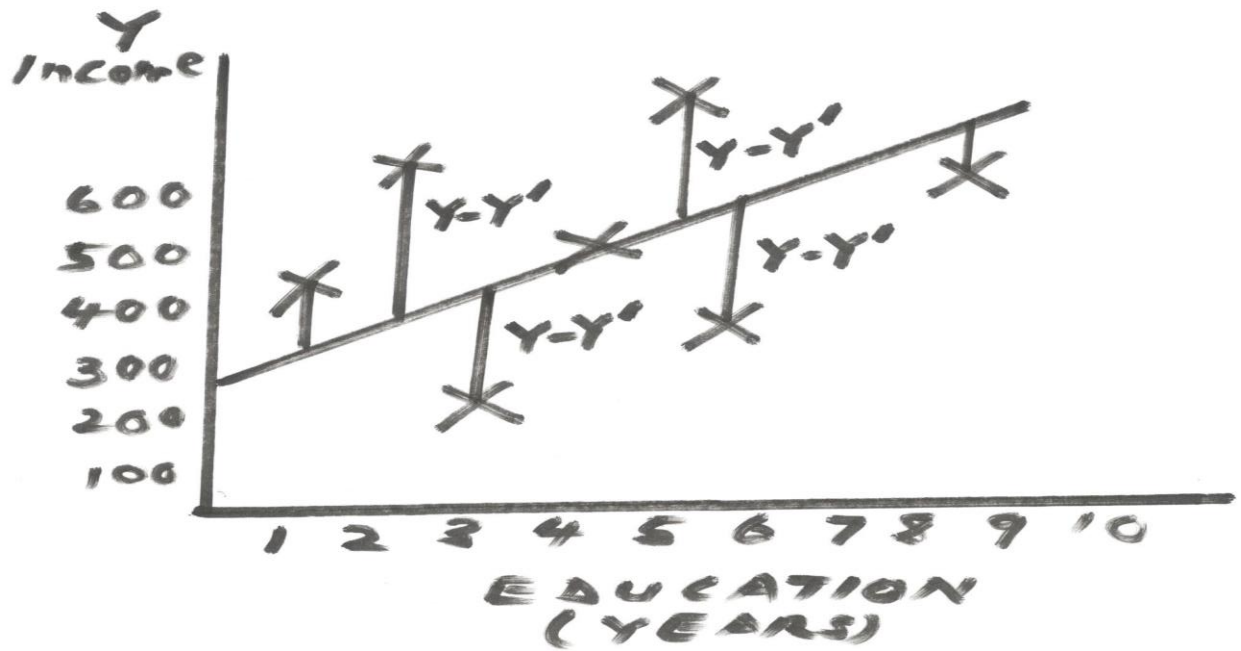
- **If Y’ is the predicted value for a given value of X, then the error of prediction is represented by the residual:**

- **Residual = Y – Y’**

- **In other words, the residual is simply the difference between the actual value, Y, and what we predict it to be, Y’.**

THE METHOD OF LEAST SQUARES

- **The method of least squares attempts to minimize the sum of squared residuals $(Y – Y')^2$ for all sample points.**
- **To do this, it uses the best prediction equation:**
- **$Y' = A + BX$**
- **The predicted values, Y', fall along the regression line.**
- **The vertical distances, Y – Y', from the regression line represent the residuals (errors of prediction).**


Y
Income
600
500
400
300
200
100
Y-Y'
Y-Y'
Y-Y'
Y-Y'
Y-Y'
1 2 3 4 5 6 7 8 9 10
EDUCATION
(YEARS)

THE METHOD OF LEAST SQUARES

- **On the basis of the prediction equation, the following are the interpretations of the constants:**
- **The constant A (the Y intercept) is the point at which the regression line crosses the Y – axis and represents the value of Y when X is X=0.**
- **Example –If examining the relationship between income and education, this would be the income level when education is zero.**
- **The constant B (the regression coefficient) is the slope of the regression line and indicates the expected change in Y with a change of one unit in X.**
- **Example – If examining the relationship between income and education, this would be expected increase in income for each additional year spent in school.**

- **EXAMPLE**
-
- Years spent in tertiary education (X) and earnings (Y)
-
- **COMPUTATION PROCEDURES**
-
- The optimum values for A and B are obtained thus:
-
- $$B = \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{n\Sigma X^2 - (\Sigma X)^2}$$
-
- $$A = Y - BX$$

- **EXAMPLE**
-
- **To examine the relationship between salary and years of working experience.**
-
-

	Y	X	XY	X^2
1	549	5.5	3019.5	30.25
2	605	9	5445	81
3	589	4	2356	16
4	590	8	4720	64
5	575	9.5	5462.5	90.25
6	555	3	1665	9
7	560	7	3920	49
8	527	1.5	790.5	2.25
9	650	8.5	5525	72.25
10	600	7.5	4500	56.25
11	560	9.5	5320	90.25
12	536	6	3216	36
13	550	2.5	1375	6.25
14	525	1.5	787.5	2.25
Total	7971	83	48102	605

673428		661593	11835
8470		6889	1581
			7.485769

				Y’
0			0.00	524.98
1			7.49	532.46
2			14.97	539.95
3			22.46	547.43
4			29.94	554.92
5			37.43	562.41
6			44.91	569.89
7			52.40	577.38
8			59.89	584.86
9			67.37	592.35
10			74.86	599.83
11			82.34	607.32
12			89.83	614.81
13			97.31	622.29
14			104.80	629.78
15			112.29	637.26
16			119.77	644.75
17			127.26	652.24
18			134.74	659.72
19			142.23	667.21
20			149.72	674.69

LINEARITY

The mean values for X and Y all lie on a straight line which is the regression line.

The relationship between X and Y must be linear so that the independent variable, X, changes the dependent variable Y tends to change systematically in a straight like manner.

If there is no linearity as in he case of curvilinear relationship modifications to the equation may be necessary

NORMALITY

This means that for any fixed value of the independent variable, X, the distribution of the dependent variable, Y, is normal.

This means that not all individuals with the same level of education have the same income.Instead there is a normal distribution for each level of education.

EQUALITY OF VARIANCE (HOMOSCEDASTICITY)

This means that the average size of the residuals along the regression line is roughly constant all along the regression line

INDEPENDENCE

This means that the observations of Ys are statistically independent of each other.

That is the observations are not in any way influenced by other observations.

If observations are re drawn from each of the four families, then the twelve observations are not independent.

INTERVAL SCALE OF MEASUREMENT

Either interval or ratio scales are accepted although nominal scale variables can be transformed into dummy variables.

Logistic regression can use nominal scale as a dependent variable.

RANDOM SAMPLING

The subjects have to randomly selected

NON PARAMETRIC TESTS

- **NON – PARAMETRIC TESTS**
- **These are tests that apply to data when assumptions of normality do not apply.**
- **They are also used in situations where data is measured on either ordinal or nominal scale.**
- **They are called nonparametric because they do not involve inferences about the mean.**
- **One simply hypothesizes that populations are identical.**
-

- **ADVANTAGES**
- **Assumptions of normality can be relaxed.**
- **They are easy to apply.**

- **COMMON NON-PARAMETRIC TESTS**

- **Rank – order correlation**
- **Signed – rank test**
- **Mann – Whitney “U”test**
- **Kruskal – Wallis “H”test**

- **SPEARMAN'S RANK ORDER CORRELATION**

- **This is used when two pairs of scores are ranked or measured on the ordinal scale.**

- **EXAMPLE**

- **Suppose we want to establish if there is a relationship between performance in MAT 1110 and DEM 2414 we have the following scores on the next slide.**

- **The formula for this correlation cofficeient is:**

- $r = 1 - \frac{6\sum D^2}{N(N^2-1)}$
- :

MAT 1110	DEM 2414
64	64
68	76
60	56
76	80
20	28
24	44
32	52
40	32
44	40
52	48

- **PROCEDURE**

- **Rank the marks in either an ascending or descending order.**
- **If the marks are tied give them the average of the ranks they would have had if they were not tied.**
- **Find the difference in ranks (D).**
- **Square the differences in ranks (D2)**
- **Sum the squared differences.**
- **Substitute them in the formula and solve for r using the formula below**

MAT 1110	R	DEM 2414	R	D	D^2
64	3	64	3	0	0
68	2	76	2	0	0
60	4	56	4	0	0
76	1	80	1	0	0
20	10	28	10	0	0
24	9	44	7	2	4
32	8	52	5	3	9
40	7	32	9	-2	4
44	6	40	8	-2	4
52	5	48	6	-1	1
∑					22

- **r =1 – $\frac{6*22}{10(10^2-1)}$**
- **$= 1 - \frac{132}{10*99}$**
- **$= 1 - \frac{132}{990}$**
- **= 1 -** 0.13
- **= <u>0.87</u>**

- **CONTINGENCY TABLES**
- **CHI – SQUARE TEST OF INDEPENDENCE**
- **Research problems in the social sciences frequently involve more than one variable.**
- **If measurements (or observations) are taken on two or more variables, we say that we have bivariate (or multivariate) data.**
- **Bivariate data is often arranged in a two way table – with one variable along the columns and another variable down the rows.**

- **The objective of this arrangement is to determine whether the two variables are related (or dependent on one another) or to predict one variable on the basis of knowledge of the other variable.**

- **The two way tables are sometimes called contingency tables because the alternative (or research) hypothesis that the two variables are dependent; that is there is contingency between the two variables.**

	Male	Female	Total
MMD			
UPND			
UNIP			
FDD			
PF			
Total			

- **EXAMPLE**

- **To what extent does religious affiliation influence attitudes towards abortion?**

Attitude	Protestant	Catholic	Total
For	126	99	225
Against	71	162	233
Total	197	261	458

- **HYPOTHESES**

- **H_o: There is no relationship between religious affiliation and attitudes towards abortion**

- **H_1: There is a relationship between religious affiliation and attitudes towards abortion**

- **ASSUMPTIONS OF CHI-SQUARE**
- **The subjects for each group are randomly and independently selected**
- **The groups are independent**
- **Each observation qualifies for one and only one category**
- **The sample size must be fairly large such that no expected frequency is less than 5 for r and c greater than 2 or less than 10 if r=c=2**
- **The scale of measurement must be nominal or ordinal**

- **DECISION RULES**

- **Given 5% level of significance with df=(r-1)(c-1) = (2-1(2-1)= 1 df.**

- **If X≤ 3.84, accept Ho**
- **If X ≥ 3.84, reject Hi**

- **COMPUTATION**

- **A test of independence of two variables arranged in the two way table makes use of the statistic:**

- $X^2 = \sum \sum (O_{ij} - E_{ij})^2$
- E_{ij}

- **In this equation, E_{ij}, is the expected number of measurements falling into the ij cell (the cell of the ith row and the jth column).**

- **The formula for E_{ij} is:**

- $E_{ij} = \frac{\text{(row total*column total)}}{N}$

Cell	O	E	O-E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
11	126	96.78	29.22	853.84	8.82
12	99	128.22	-29.22	853.84	6.66
21	71	100.22	-29.22	853.84	8.52
22	162	132.78	29.22	853.84	6.43
Σ					30.43

- **DECISION**
- **Reject Ho**

- **CONCLUSION**

- **There is a relationship between religious affiliation and attitudes towards abortion**

- **PRINCIPLES OF TABLE READING**

- **Decide on the independent variable and dependent variable.**
- **Percentage in terms of the independent variable (down the columns)**
- **Compare the percentages in terms of the categories of the depenent variable (across the rows)**

	Protestant	Catholic	Total
For	126 (63.96%)	99 (37.93%)	225
Against	71 (36.04%)	162 (62.07%)	233
Total	197	261	458

Attitude	Protestant	Catholic	Total
For	63.96%)	37.93%	225
Against	36.04%)	62.07%	233
Total	100%	100%	458

- **CHI – SQUARE GOODNESS OF FIT TEST**

- **This is used to determine the extent to which our expectations match with reality. Just how well do observed data fit or agree with our expectations**

- **For example, on the expectation that the number of customers is equal on all days at Shoprite, management assigns the same number of 15 till operators on all days from Monday to Sunday.**

- **Complaints are however heard from many customers that service is very slow.**

- **To deals with this problem management hires a consultant to investigate this.**

- **The consultant the consultant collects data over a 24-week period observing the number of customers on each day. The total number of customers observed over the period is 9,792. Thus we would expect 1,632 on each day.**

- Is management justified in assigning only 15 till operators in the shop?
-
- **HYPOTHESES**
-
- Ho: The number of customers is evenly spread over six working days
- Hi: The number of workers is not evenly spread
-

- **ASSUMPTIONS**
-
- Nominal scale
- No random sampling
- Sample is large
-
- **DECISION RULES**
-
- For k-1 degrees of freedom at 5% level of significance,
-
- If X less than 11.07, accept Ho
- If X greater than 11.07, reject Ho

Day	O	E	O-E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
Monday	1525	1632	-107.00	11449.00	7.02
Tuesday	1711	1632	79.00	6241.00	3.82
Wednesday	1655	1632	23.00	529.00	0.32
Thursday	1497	1632	-135.00	18225.00	11.17
Friday	1603	1632	-29.00	841.00	0.52
Saturday	1801	1632	169.00	28561.00	17.50
Σ					40.35

- **DECISION**
- **Reject Ho**
- **CONCLUSION**
- **The customers are not evenly spread over six working days.**