SMA_160 INTRODUCTION TO PROBABILITY AND STATISTICS

INTRODUCTION TO STATISTICS

Variable

Characteristic or attribute that can assume different values

Random Variable

A variable whose values are determined by chance.

Population

All subjects possessing a common characteristic that is being studied.

Census

The collection of data from *every* element in a population.

Sample

A subgroup or subset of the population.

Parameter

Characteristic or measure obtained from a population.

statistic (not to be confused with Statistics)

Characteristic or measure obtained from a sample.

Descriptive statistics refers to the quantitative or qualitative description of a sample measurement and characteristics.

Inferential statistics

Generalizing from samples to populations using probabilities. Inferential statistics refers to the application of the sample statistics to the Population parameters from which the samples were drawn subject to the stated significant levels.

Discrete Variables (Data)

Variables (data) which assume a finite or countable number of possible values. Usually obtained by counting.

Continuous Variables (Data)

Variables (data) which assume an infinite number of possible values. Usually obtained by measurement.

Observational Study

A study in which the subjects are observed and studied, but no attempt is made to manipulate or modify the subjects.

Experiment

A study in which a treatment is applied, and then its effects on the subjects are studied.

Sampling Error

The difference between the sample result and the true population result that occurs because of chance variation.

Non-sampling Error

An error that occurs because sample data is incorrectly collected, recorded, or analyzed.

Scales of Measurement

In statistics, there are four data measurement scales: nominal, ordinal, interval and ratio data types.

- Nominal- under nominal scale the items are differentiated by a simple naming system. Nominal items are usually *categorical*. Therefore nominal scales are just used for labeling variables, without any quantitative value. Questions like what is your gender? Where do you live are nominal data. Nominal can be categorized as nominal with order or without order eg cold, warm, hot and male or female respectively
- Ordinal- under ordinal scale the items are set into some kind of *order* by their position on the scale. Ordinal items are usually categorical.eg teams can be ranked as first, third and fifth etc regardless of the score between each consecutive position.
- Interval-Interval data (sometimes called *integer*). Just like the ratio scale it is measured along a scale in which each position is equidistant from one another.eg Altitudes (the height above sea level), **Celsius temperature** in which the difference between any two consecutive values is the same.
- Ratio- under the ratio scale, items are measured along a regular scale in which each position point is equidistant from one another, therefore numbers can be compared as multiples of one another and have an absolute zero (reference point) eg weight and height (Both have absolute zero such that no numbers exist below the zero point)

Frequency Distribution

Statistical data obtained by means of census, sample surveys or experiments is raw, unorganized and usually contains some errors. Before these are analyzed and used as a basis for inferences about the phenomenon under investigation or as a basis for decision making, they must be cleaned, summarized and the pertinent information extracted. One way of presenting data for analysis is construction of Frequency distribution tables.

Frequency table or a frequency distribution is constructed by dividing the overall range of values into a number of classes and then counting the number of observations that fall into each of these classes or intervals. Insofar as possible, equal class intervals are preferred. But the first and last classes can be open-ended to cater for extreme values.

Exercise 1

A random sample of 100 KPLC employees was selected and their annual OT for one year was recorded as follows:

Table: 2. KPLC employees

55	82	83	109	78	87	95	94	85	67
80	109	83	89	91	104	90	103	67	52
107	78	86	29	72	66	92	99	60	75
88	112	97	88	49	62	70	66	88	62
72	85	81	78	77	41	105	92	94	74
78	75	87	83	71	99	56	69	78	60
119	39	104	86	67	79	98	102	82	91
46	120	73	125	132	86	48	55	112	28
42	24	130	100	46	57	31	129	137	59
102	51	135	53	105	110	107	46	108	117

By using the class interval 20-39, 40-59 and so forth construct the frequency distribution, cumulative frequency distribution, relative frequency distribution and relative cumulative frequency distribution in one table.

Definition of terms

Class boundary is the precise point that separates one class from another, rather than being a value indicated in one of the classes. A class boundary is typically located midway between the upper limit of a class and the lower limit of the next higher class adjoining it. Therefore the class boundary separating the class 60-79 and the class 80-99 is halfway between 79 and 80, that is, at the point 79.5. This is the upper class boundary and lower class boundary for 60-70 and 80-99 classes respectively.

Class interval: is the width of a class. The class interval of a class is computed by subtracting the class boundaries.

Class midpoint or class mark: is the point dividing the class into equal halves on the basis of class interval. This point can be obtained by adding the lower and upper limits (boundaries) of a class and dividing by 2.

Relative frequency of a class: it is the ratio of the frequency of any class to sum of the frequencies.

Cumulative frequency distribution: shows the number of items of a series that are less than (or more than) certain specified values.

Measure of Central Tendency

A value that would describe the 'centre' of a distribution would be visually located near the spot where most of the data seem to be concentrated. Consequently, values that fulfil this role are called measures of central tendency.
The most common measures of the central tendency of a data set are arithmetic mean or simply as mean, median and mode.

Calculating mean, median and mode for individual (Ungrouped) data

Calculating mean, median and mode for grouped data

The following table shows the daily wages of a random sample of construction workers. Calculate its mean, median and mode.

Table 4: Workers daily wage

Daily Wages	Number of Workers
200 - 399	5
400 - 599	15
600 - 799	25
800 - 999	30
1000 - 1199	18
1200 - 1399	7
Total	100

Solution

Daily Wages	Number of Workers f_i	Class Mark x_i	$f_i x_i$	Cum. frequency F
200 - 399	5	299.5	1,497.5	5
400 - 599	15	499.5	7,492.5	20
600 - 799	25	699.5	17,489.5	45
800 - 999	30	899.5	26,985.5	75
1000 - 1199	18	1,099.5	19,791.0	93
1200 - 1399	7	1,299.5	9,096.5	100
Total	100		82,350.0	

$$\bar{x} = \frac{\sum_{i=1}^{6} f_i x_i}{\sum_{i=1}^{6} f_i} = \frac{82,350.0}{100} = 823.5$$

$$M_d = L + \left(\frac{\frac{1}{2}\sum f - F_a}{f_w} \right) c_i$$ Where: L is the lower real limit of the middle class

f_w is the frequency of the middle class
F_a is the cumulative frequency above the middle class
c_i is the class interval of the middle class

$$= 799.5 + \frac{0.5(100) - 45}{30}(200) = 832.8$$

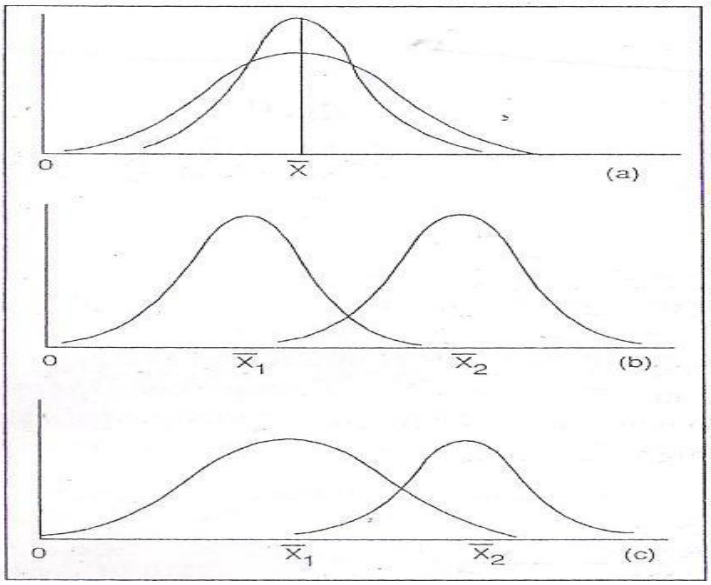
Mode $= L + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2} \right) c_i$

Where: L is the lower real limit of the modal class
f_1 is the frequency of the middle class
f_0 is the frequency of the class preceding modal class
f_2 is the frequency of the class succeeding the modal class and
c_i is the class interval of the modal class

$$\text{Mode} = 799.5 + \frac{30 - 25}{2(30) - 25 - 18}(200) = 858.3$$

Measure of data variation (Dispersion)

The figure below represents frequency distribution with some of the characteristics we need to understand. The two curves in (a) represent two distributions with the same mean $\bar{X}$, but with different varations. The two curves in (b) represent two distributions with the same variations but with unequal means, $\bar{X}_1$ and $\bar{X}_2$, finally, (c) represents two distributions with unequal means and unequal variations.



The measures of central tendency are, therefore, insufficient. They must be supported and supplemented with other measures. A measure of variation is designed to state the extent to which the individual measures differ on an average from the mean. Hence for an adequate summary and characteristics description of a set of data we need to determine the data variation.

The most common measures of variability or dispersion are the ***range, mean deviation, interquartile range, deciles, percentiles, variance and standard deviation.***

Example 1

Consider the following measurements, in grams, for two samples of strawberry jam bottled by companies A and B:

Table 5: Strawberry

Sample for Company A	31	32	32	33	32
Sample for Company B	28	29	32	35	36

Both samples have the same mean, 32 grams. It is obvious that company A, in comparison with company B, bottles strawberry jam with a more consistent content. We say that the variability of the observations is smaller for company A. Therefore in buying strawberry jam we would feel more confident that the bottle we select will be closer to the advertised average content if we buy from company A.

The ***range*** of a set of numbers is the difference between the largest (L) and the smallest (S) number in the set. Therefore we have range = L-S and the Co-efficient of range $= \dfrac{L-S}{L+S}$

Absolute Mean deviation is the average of the absolute deviation of the numerical data

Variance is the average of the squared deviations from the arithmetic mean.

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

Standard deviation of a population is the positive square root of the variance
Using the values in table 4 determine the variance and standard deviation

Solution

Daily Wages	Number of Workers f_i	Class Mark x_i	$f_i(x_i - \bar{x})^2$
200 - 399	5	299.5	1, 372,880
400 - 599	15	499.5	1,574,640
600 - 799	25	699.5	384,400
800 - 999	30	899.5	173,280
1000 - 1199	18	1,099.5	1,371,168
1200 - 1399	7	1,299.5	1,586,032
Total	100		6,462,400

$$\text{Variance}\,(s^2) = \frac{6462400}{99} = 65,276.77$$

$$\text{Standard deviation} = \sqrt{65276.77} = 255.49$$

$$s = \sqrt{\frac{\sum x^2 f}{\sum f} - \left(\frac{\sum xf}{\sum f}\right)^2}$$

Quartiles- Quartile divides the data set into 4 equal parts
Lower quartile (Q_1) and upper quartile (Q_3) are computed as;

$$Q_1 = L + \left(\frac{\frac{1}{4}\sum f - F_a}{f_w}\right)c_i \text{ and } Q_3 = L + \left(\frac{\frac{3}{4}\sum f - F_a}{f_w}\right)c_i \text{ respectively.}$$

Where: L is the lower real limit of the class containing lower/upper quartile score
f_w is the frequency of the lower/upper quartile class
F_a is the cumulative frequency above the lower/upper quartile class
c_i is the class interval of the lower/upper quartile class

Deciles- Divides the data set into 10 equal parts

Percentile- divides the data set into 100 equal parts

The median formula is adjusted to determine deciles and percentiles.

Moments, Skewness and Kurtosis

a) Moments

In statistics moments refer to a quantitative measure of the shape of a set of points representing mass. Represented by the Greek letter μ (mu) moments give a summary description of a distribution characteristics. The rth moment (raw moment) is denoted by $\mu_r^{'} = E(x^r)$, such that if we have a set of discrete data, $S = \{5,7,9\}$ then the first raw moment is $S_1^{'} = \mu_1^{'} = \frac{(5^1 + 7^1 + 9^1)}{3} = 7$, the second raw moment is $S_2^{'} = \mu_2^{'} = \frac{(5^2 + 7^2 + 9^2)}{3} = 51.67$ and then the rth moment is; $S_r^{'} = \mu_r^{'} = \frac{(5^r + 7^r + 9^r)}{3}$.

Determine $\mu_3^{'}$ and $\mu_4^{'}$

Moments about the mean (central moments) are obtained as;

$\mu_1 = \frac{\sum f(x-\bar{x})}{\sum f}$ (1st moment about the mean), $\mu_2 = \frac{\sum f(x-\bar{x})^2}{\sum f}$ (2nd moment …)

and

$\mu_r = \frac{\sum f(x-\bar{x})^r}{\sum f}$ The rth moment of a variable *x* about the mean $(\bar{x})$, such that using the above set of discrete data where $S = \{5,7,9\}$ the first central moment about the mean is

$\mu_1 = \frac{\{1(5-7)^1 + 1(7-7)^1 + 1(9-7)^1\}}{3} = 0$. The second moment about the mean will be

$\mu_2 = \frac{\{1(5-7)^2 + 1(7-7)^2 + 1(9-7)^2\}}{3} = 2.67$ which is equal to the variance of the data.

The first moment about the mean tells us about the sample mean, second about the variance, third about the skewness i.e if $(\mu_3 \neq 0)$ then the data is skewed and the fourth moment about the mean tell us about the kurtosis.

NB: The following relationship holds true

a) The moments about the mean (central moments) and the raw moments;

$\mu_1 = 0$,

$\mu_2 = \mu_2^{'} - (\mu_1^{'})^2$,

$\mu_3 = \mu_3^{'} - 3\mu_1^{'}\mu_2^{'} + 2(\mu_1^{'})^3$ and $\mu_4 = \mu_4^{'} - 4\mu_1^{'}\mu_3^{'} + 6\mu_2^{'}(\mu_1^{'})^2 - 3(\mu_1^{'})^4$

Using $S = \{5, 7, 9\}$ determine μ_3 and μ_4.

b) The betas and the central moments;

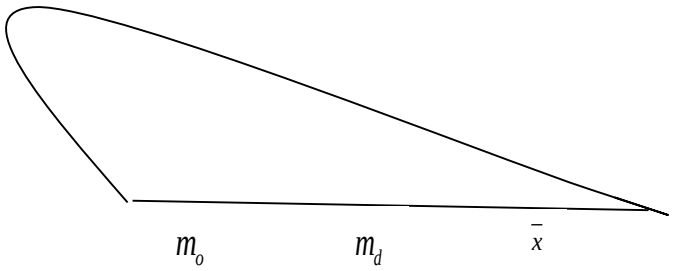
$\beta_1 = \frac{(\mu_3)^2}{(\mu_2)^3}$ and $\beta_2 = \frac{\mu_4}{(\mu_2)^2}$ The first beta (β_1) is used to measure the data skewness while

The second beta (β_2) measures the kurtosis of the plotted data curve as discussed below.

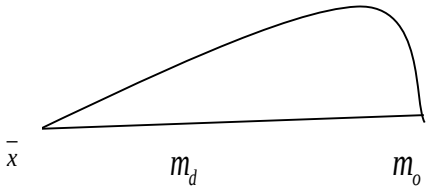
b) Skewness

Asymmetrical data is said to be skewed distribution. The distribution is either skewed to the right or left otherwise it is symmetrical distribution (Normally Distributed)

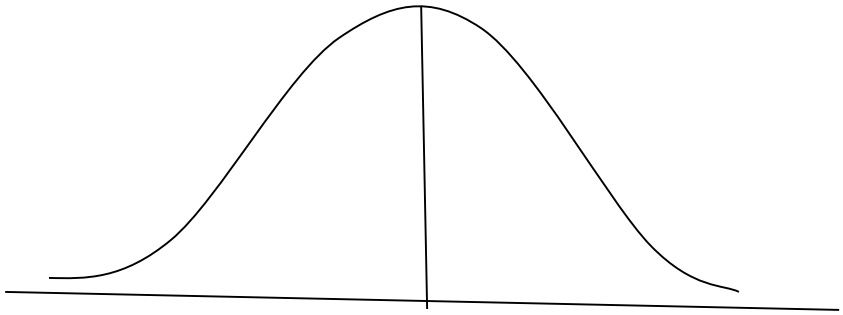
Positively skewed distribution is tailed to the right such that $m_o < m_d < \bar{x}$



Negatively skewed distribution is tailed to the left and $\bar{x} < m_d < m_o$



Normally distributed data has the three measures of central equal such that $\bar{x} = m_o = m_d$



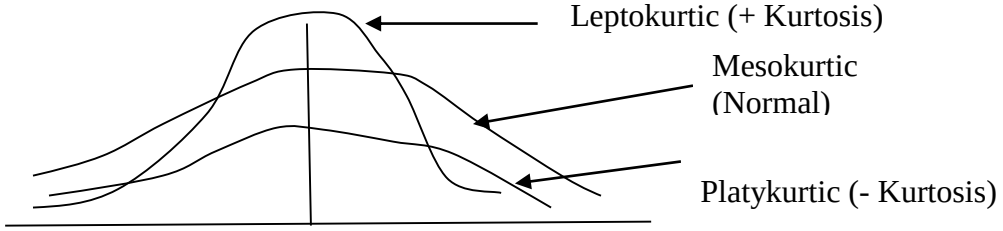
Measures of Skewness

i. Karl pearson's coefficient of skewness (S_{kp})

$S_{kp} = \frac{Mean - Mode}{\sigma}$ or $S_{kp} = \frac{3(Mean - Median)}{\sigma}$ the second one is recommended since

c) Kurtosis

This is a non-dimensional measure of the relative peakness or flatness of a data distribution i.e relative to a normal distribution.



According to Kar Pearson β_2 is used to determine the degree of peakness of a curve relative to the normal curve. Where $\beta_2 = \frac{\mu_4}{(\mu_2)^2}$, such that when $\beta_2 = 3$ the curve is mesokurtic (normal), when $\beta_2 < 3$ the curve is platykurtic and when the $\beta_2 > 3$ the curve is leptokurtic.

STATISTICAL DATA REPRESENATIONS

Data can be represented in form of; bar charts, pie charts, boxplots (box and whiskers plots), histograms, stem and leaf, scatter diagram among others.

Examples

Stem and leaf

The following record represents the long jump results (in meters) of inter-house competitions in a certain school within Machakos County:

2.3, 2.5, 2.5, 2.7, 2.8 3.2, 3.6, 3.6, 4.5, 5.0

And here is the stem-and-leaf plot:

Stem	Leaf
2	3 5 5 7 8
3	2 6 6
4	5
5	0

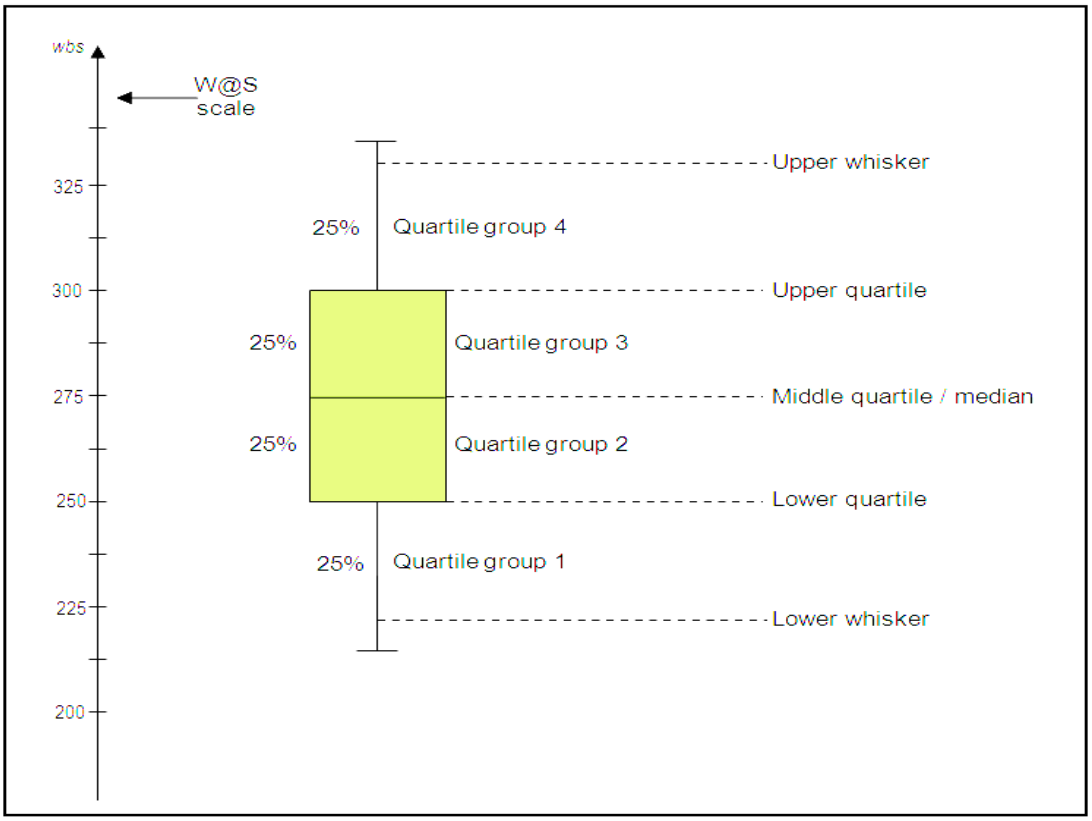
Stem "2" Leaf "3" means **2.3**

Note:

- Say what the stem and leaf mean (Stem "2" Leaf "3" means **2.3**)
- In this case each leaf is a decimal
- It is OK to repeat a leaf value
- 5.0 has a leaf of "0"

Box-and-Whisker Plots:

Under this exploration technique statistics assumes that the data points are clustered around some central value, the "box". To create a box-and-whisker plot, the data is numerically ordered. The box divides the entire data set into quarters, called "quartiles".



Understanding and interpreting box plots

Box plots enable us to study the distributional characteristics of a group of scores as well as the level of the scores.

The median (middle quartile) marks the mid-point of the data and is shown by the line that divides the box into two parts.

Upper quartile-Seventy-five percent of the scores fall below the upper quartile.

Lower quartile-Twenty-five percent of scores fall below the lower quartile

Inter-quartile range-The middle "box" represents the middle 50% of scores for the group. The range of scores from lower to upper quartile is referred to as the inter-quartile range.

Whiskers-The upper and lower whiskers represent scores outside the middle 50%. Whiskers often (but not always) stretch over a wider range of scores than the middle quartile groups.

Revision Exercise

a) Differentiate between descriptive statistics and inferential statistics
b) Highlight four levels of variable measurement scales in statistics
c) The table below shows the frequency distribution of sales made by 100 shops

Sales Ksh '000'	*Number of Shops*
100-119	2
120-139	*"a"*
140-159	20
160-179	19
180-199	*"b"*
200-219	21
220-239	1

Given that the mean is Ksh 177,100, determine

i. The values of "*a"* and "*b"*
ii. The Median
iii. The Standard deviation
iv. Karl Pearson's coefficient of skewness ($_{Skp}$)

d) The number of days the college nurse was called for emergencies per month for the last 10 months were; 2,3,4,0,5,6,7,4,3,2. Determine the

i. Mean
ii. Mean Absolute Deviation.
iii. Variance
iv. Standard deviation

e) Differentiate the following terms as they apply in scientific research
i. Sample and a population
ii. Skewness and Kurtosis of a data distribution
iii. Sample statistic and Population parameter
iv. Sampling error and Non-Sampling error

f) The table below shows the wages of 80 employees of XYZ Company

Wages Ksh '000'	Number of Employees
10-15	5
15-20	*x*
20-25	17
25-30	20
30-35	*y*

35-40	16
40-45	4

Given that the median wage is Ksh 27,000, determine

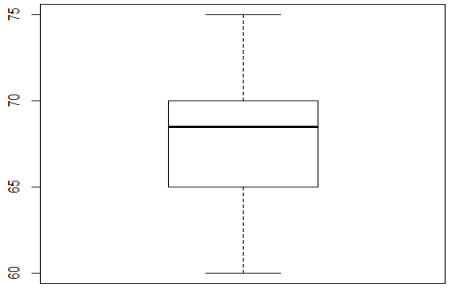
i. The values of *x* and *y*
ii. The mean
iii. The inter-quartile wage
iv. Karl Pearson's coefficient of Skewness ($_{Skp}$)

g) Highlight five properties of a good measure of data variation

h) Define the term *variable* as used in statistics, giving two examples.

i) Explain in words each of the following terms as used in Statistics:
(i) mean;
(ii) median.

j) Estimate the sample median and quartiles using the box plot given below



k) The data given below represents the age in years of employees of an organisation.

28, 30, 33, 37, 37, 38, 42, 43, 43, 44, 45, 48, 48, 51, 55

Use the data to construct a box and whisker plot.

RELATIONSHIPS

A distribution in which there is only one variable is referred to as univariate distribution, eg the age of the students of a class. A distribution involving two discrete variables is called a bivariate frequency distribution.

Correlation & Regression Analysis

The measure of correlation called the coefficient of correlation denoted by the symbol (r) summarizes in one figure the direction and degree of correlation. Thus, correlation can be defined as the covariate analysis of two or more variables.

Correlation analysis involves;

1) Determination of any relationship existence; $-1 \le r \le 1$whereby +1 $\rightarrow$perfect positive correlation, -1$\rightarrow$perfect negative correlation and 0 $\rightarrow$ may mean there is no linear correlation or no correlation at all.
2) Testing its significance to ensure the correlation is not by a mere chance due to pure random sampling or investigator's bias in selecting the sample; and finally

Correlation can be classified in several different ways: Linear or non-linear, Simple, or multiple.

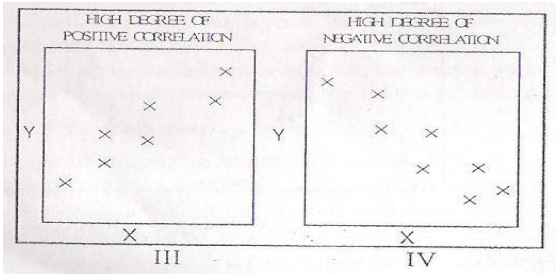
Methods of calculating correlation

The following are the important methods of ascertaining whether two variables are correlated or not:

I. Scatter Diagram Method;
II. Karl Pearson's Coefficient of Correlation;

I. Scatter Diagram Method

This is a dot chart also referred to as called dotogram, for each pair of *X* and *Y* values we put dots and thus obtain as many points as the number of observations. By looking to the scatter of the various points, one can form an idea as to whether the variables are related.



2. Karl Pearson's coefficient of correlation

The product- moment coefficient correlation popularly known as Pearsonian coefficient of correlation, is mostly widely used in practice. It gives both the degree and direction of the relationship between two variables. If the two variables understudy are X and Y, then

$$r = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2}\sqrt{\sum(Y-\bar{Y})^2}} \qquad \ldots(i)$$

$$The\ above\ formula\ can\ be\ written\ as:$$

$$r^* = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}} \qquad \ldots(ii)$$

$$where\ x = (X-\bar{X}) and\ y = (Y-\bar{Y})$$

This formula is to be used only where the deviations are taken from actual means and not from assumed means.

The coefficient of correlation can also be calculated from the original set of observations (i.e., without taking deviations from mean) by applying the following formula

$$r^{**} = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{\sum X^2 - \frac{(\sum X)^2}{N}}\sqrt{\sum Y^2 - \frac{(\sum Y)^2}{N}}} = \frac{N\sum XY - \sum X \sum Y}{\sqrt{N\sum X^2 - (\sum X)^2}\sqrt{N\sum Y^2 - (\sum Y)^2}} \quad \ldots \qquad (iii)$$

$\sum(x_i - \bar{x})(y_i - \bar{y})/n$ = covariance of the two variables x and y

It measures their joint variation. When x and y are not related its value is close to zero.

The position $(\bar{x}, \bar{y})$ is known as the centroid of all the points.

Illustration 2

Find correlation coefficient between the sales and expenses from the data given below:

Firm:	A	B	C	D	E	F	G	H	I	J
Sales (Ksh, 000):	50	50	55	60	65	65	65	60	60	50
Expenses (Ksh. 000):	11	13	14	16	16	15	15	14	13	13

Exercise

i. The following data relate to the age of 10 employees from company ABC Ltd and the number of days which they reported sick in a month:

Age:	20	30	32	35	40	46	52	55	58	62
Sick days:	11	12	10	13	14	16	15	17	18	19

By letting the age and sick days be presented by variable X and Y respectively, calculate Karl Pearson's coefficient of correlation and interpret its value.

ii. Find the coefficient of correlation by Karl Pearson's method between X and Y and interpret its value.

X	57	42	40	33	42	45	42	44	40	56	44	43
Y	10	60	30	41	29	27	27	19	18	19	31	29

Assumptions of the Pearsonians Coefficient

The Karl Pearson's coefficient of correlation is based on the following assumptions:

1. There is linear relationship between the variables, i.e., when the two variables are plotted on a scatter diagram, a straight line will be formed by the points so plotted.
2. The two variables understudy are affected by a large number of independent causes so as to for a normal distribution. Variables like height, weight, price, demand and supply are affected by such forces that a normal distribution is formed.
3. There is a cause and effect relationship between the forces affecting the distribution of the item in the two series. If such a relationship is not formed between the variables, i.e., if the variables are independent there cannot be any correlation. For example, there is no relationship between income and height because the forces that affected these variables are common.

Properties of the Coefficient Of Correlation

The following are the important properties of the coefficient of correlations, *r:*

1. The coefficient of correlation lies between-1 and +1,($-1 \leq r \leq +1$)
2. The coefficient of correlation is independent of change of origin and scale

3. The coefficient of correlation is the geometric mean of two regression coefficients. Symbolically : $r = \sqrt{b_{xy} \times b_{yx}}$
4. If X and Y are independent variables then coefficient of correlation is zero. However the converse is not true. The closeness of the relationship is not proportional to r. if the value of *r* is 0.8, it does not indicate a relationship twice as close as that of 0.4. it is in fact very much closer.
5. The probable error of the coefficient of correlation is obtained as follows:

$P.E.r^* = 0.6745\frac{1-r^2}{\sqrt{N}}$, assuming a coefficient of correlation 0.80 computed from a sample of 16 pairs of items, we have

P.E.r $=0.6745\frac{1-(0.8)^2}{\sqrt{16}} = 0.06$

1. If the value of *r* is less than the probable error, there is no evidence of correlation, i.e., the value of r is not at all significant.
2. If the value of *r* is more than six times the probable error, the existence of existence of correlation practically certain, i.e., the value of *r* is significant.
3. By ± the probable errors from the coefficient of correlation get respectively the upper and lower limits within which coefficient of correlation in the population can be expected to lie. Symbolically, $\rho = r \pm P.E.r$

*******The measure of probable error can be properly used only when the data approximates to a normal frequency curve (bell-shaped curve), the statistical measure for which the P.E is computed must have been calculated from sample and finally tthe sample must have been selected in an unbiased manner and the individual items must be independent. However, these conditions are generally not satisfied and as such the reliability of the correlation efficient is determined largely on the basis of exterior tests of reasonableness which are often of statistical character.

Illustration 7

If *r*=0.6 and *N*=64, find out the probable error of the coefficient of correlation and determine the limits for *r*.

Solution P.E.*r*$=0.6745\frac{1-r^2}{\sqrt{N}}$;*r*=0.6 and *N =64*

. P.E.r=0.6745$\frac{1-(0.6)^2}{\sqrt{64}} = \frac{0.6745\times0.64}{8} = 0.054$

Limits of $r = 0.6 \pm 0.054$ *or =0.546 to 0.654*

Merits (r)

The correlation coefficient summarizes in one figure the degree of correlation and direction.

Limitations (r)

1. The correlation coefficient always assumes linear relationship of the variables.
2. The value of the coefficient is unduly affected by the extreme values.
3. As compared to other methods of finding correlation, this method is more time-consuming.
4. Subject to misinterpretation

Coefficient of Determination*

The coefficient of determination is equals to r^2. It expresses the proportion of the variance in Y due to X, that is, the ratio of the explained variance to the total variance. eg if r=0.9, r^2 will be 0.81 and this would mean that 0.81 per cent of the variation in the dependent variable has been explained by the independent variable. The maximum value of r^2 is a unit because it is possible to explain all of the variation in Y, but it is not possible to explain more than all of it.

Regression Analysis

Introduction

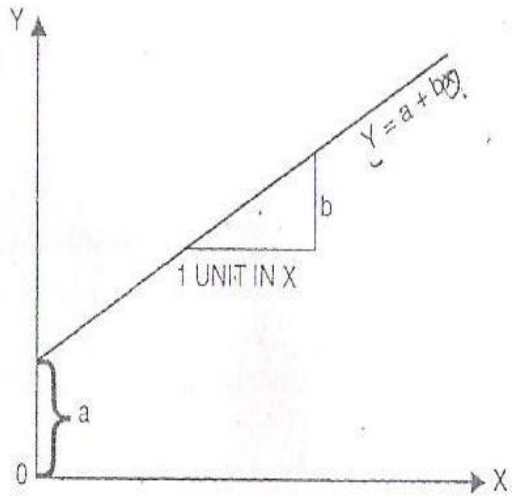
Regression was first used by Francis Galton (1877) in his fathers' vs sons' heights relationship study. He described the relationship by using a 'regression Line'. The term is still used to describe that a line drawn from a group of points to represent the trend, although most of the modern writers use the term *estimating line* or *predicting line* instead of *regression line*.

Regression analysis establishes the relationship between the dependent variable and the regressors by obtaining the rate of change of the response variable due to a unit change of the independent variable(s). It enables the analyst to estimate (or predict) the unknown values of one variable from known values of another variable. It can also be used to obtain

a measure of the error (standard error) involved in using the regression line as a basis for estimations. We can use regression analysis to estimate correlation between two variables.

The Linear Bivariate Regression Model (Simple Regression)

The average relationship between X and Y can be adequately described by a linear equation $Y = a + bX$ whose geometrical presentation is a straight line as in the diagram below:



In this equation *a* and *b* are the population regression coefficients.

Regression Equations

Regression equations are algebraic expressions of the regression lines. The regression equation of Y on X is expressed as $Y_e = a + bX$ or $y = \beta_0 + \beta_1 x_i + \varepsilon_i$

i. $E(\varepsilon_i) = 0 \ \ \forall i = 1, 2 \cdots n \ \Leftrightarrow E(y_i) = \beta_0 + \beta_1 x_i$

ii. $Var(\varepsilon_i) = \sigma^2 \ \ \forall i = 1, 2 \cdots n$

$\Leftrightarrow Var(y_i) = E[y_i - E(y_i)]^2 = (y_i - \beta_0 - \beta_1 x_i)^2 = E(\varepsilon_i^2) = \sigma^2$ referred to as the assumption of homoscedasticity or homogeneous variance or constant variance and

iii. $Cov(\varepsilon_i \varepsilon_j) = 0, \forall i = i \neq j \ , \Leftrightarrow \text{cov}(y_i y_j) = 0$ i.e variables are uncorrelated

Regression Coefficients

The parameter ' β_0 ' determines the *level* of the fitted line (i.e., the distance of the line directly above or below the origin). The parameter ' β_1 ' determines the *slope* of the line,

i.e., the change in *Y* for unit change *X*. The linear regression with response Y and several predictors X will have a model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$ which will dealt with later.

Estimation of β_0, β_1 and σ^2

There several methods of estimating the parameters including the method of least squares, maximum likelihood method among others.

Method of least squares

The line of best fit is drawn in such a manner that the sum of the vertical deviations of the actual *Y* values from the estimated *Y* values is least. i.e $(\hat{Y} - Y)$ is minimum.

With a little algebra and differential calculus, it can be shown that the following two equations, if solved simultaneously, will yield values of the parameters *a* and *b* such that the least squares requirement is fulfilled.

$$\sum Y = Na + b\sum X \qquad \ldots(i)$$

$$\sum XY = a\sum X + b\sum X^2 \qquad \ldots\ldots(\text{ii})$$

These equations are usually called the normal equations. In the equations $\sum X$, $\sum Y$, $\sum XY$, $\sum X^2$ indicate totals which are computed from the observed pairs of values of two variables *X* and *Y* which the least squares estimating line is to be fitted and *N* is the total number of observed pairs of values.

Regression Equation Calculation

Calculate the regression equations of Y on X and X on Y on X from the following data:

X:	1	2	3	4	5
Y:	2	5	3	8	7

Solution

X	Y	X^2	Y^2	XY	$\hat{y}$	
1	2	1	4	2		
2	5	4	25	10		
3	3	9	9	9		
4	8	16	64	32		
5	7	25	49	35		
$\sum X = 15$	$\sum Y = 25$	$\sum X^2 = 55$	$\sum Y^2 = 151$	$\sum XY = 88$		

a) Regression equation of X on Y is given by $X = a + bY$

The equations are: $\sum X = Na + b\sum Y$ --------(i)

$\sum XY = a\sum Y + a\sum Y^2$----------(ii)

Substituting the values, we get

$$15 = 5a + 25b$$

$$88 = 25a + 15b$$

Solving (i) and (ii) we get $a = 0.5 \ and \ b = 0.5$

Hence the required regression equations of X and Y is given by $X = 0.5 + 0.5Y$

b) Regression equations of Y on X is given by : $Y = a + bX$

The normal equations are: $\sum Y = Na + b\sum X$

$$\sum XY = a\sum Y + a\sum X^2$$

Substituting the values, we get $25 = 5a + 15b$

$$88 = 15a + 55b$$

Solving (iii) and (iv), we get $a = 1.10 \ and \ b = 1.3$

Hence, the required regression equation of Y on X is given by $Y = 1.10 + 1.30X$

c) Using a calculator resolve (a) and (b) above

Exercise

Calculation of regression equations of Y on X given;

X	9	7	5	10	4	5	3	2
Y	45	42	41	60	30	34	25	20

Predict Y when X is 8 and 20

d) Equivalently by carrying out a partial derivative we can also determine the betas, given that

$\hat{\varepsilon}'\hat{\varepsilon} = \sum_{i=1}^{n}\hat{\varepsilon}^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum(\hat{y}_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$ then

$$\frac{d(\hat{\varepsilon}'\hat{\varepsilon})}{d\beta_0} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \qquad \text{.....(i)}$$

$$\frac{d(\hat{\varepsilon}'\hat{\varepsilon})}{d\beta_1} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i = 0 \qquad \text{.....(ii)}$$

Solving (i) and (ii) above we have

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ **……. (iii)**

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum xy}{\sum x^2}$$ **……..……(iv)**

Hence the estimated regression equation will be $\hat{y} = \hat{b}_0 + \hat{b}_1 x$

Determine the regression equation of Y on X given that

$\sum xy = 130$, $\sum x^2 = 2400$, $\bar{x} = 60$ and $\bar{y} = 4$.

e) Regression equation of Y on X: $Y - \bar{Y} = b_{yx}(X - \bar{X})$ where $b_{yx} = r\frac{\sigma_y}{\sigma_x}$

PROBABILITY

Introduction and concepts

Probability is the likelihood quantitative measure of an event occurring.
Probability is the basis upon which the discipline of statistics has been developed and applied in many fields associated with chance occurrences such as politics, business, weather forecasting, and scientific research.

Some Basic Concepts and definitions

- **Sample space:** is a set of all possible distinct outcomes of an experiment. Eg the discrete sample space for rolling a six sided die; $S = \{1, 2, 3, 4, 5, 6\}$

- **Event:** is a sample point or subset of a sample space;
  The probability of an event A denoted P(A) is $0 \le P(A_i) \le 1$ and $\sum_i P(A_i) = 1$

- **Fundamental Principle of Counting;**

  a) Additional rule of counting for events that are not mutually exclusive we have
  $$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
  Where $A \cup B$ is the union of two sets A and B, i.e the set of elements that belong to A or to B or to both. While $A \cap B$ is the intersection of two sets A and B, i.e the set of elements that are common to A and B. For mutually exclusive events, we have $P(A \cup B) = P(A) + P(B)$ If A and A' are complementary events then $P(A) = 1 - P(A')$

Some Basic Concepts and definitions

- **Sample space:** is a set of all possible distinct outcomes of an experiment. Eg the discrete sample space for rolling a six sided die; $S = \{1, 2, 3, 4, 5, 6\}$

- **Event:** is a sample point or subset of a sample space;
  The probability of an event A denoted P(A) is $0 \le P(A_i) \le 1$ and $\sum_i P(A_i) = 1$

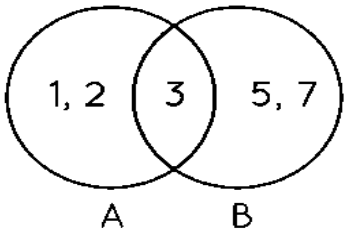
- **Fundamental Principle of Counting;**

  b) Additional rule of counting for events that are not mutually exclusive we have
  $$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
  Where $A \cup B$ is the union of two sets A and B, i.e the set of elements that belong to A or to B or to both. While $A \cap B$ is the intersection of two sets A and B, i.e the set of elements that are common to A and B. For mutually exclusive events, we have $P(A \cup B) = P(A) + P(B)$ If A and A' are complementary events then $P(A) = 1 - P(A')$

SETS

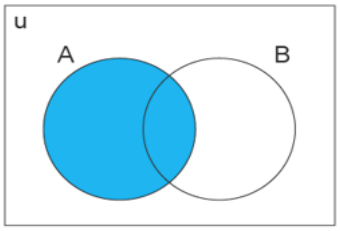
Sets are a collection of distinct elements, which are enclosed in curly brackets, separated by commas. The list of items in a set is called the elements of a set

Symbols	**Meaning**	**Example**
{ }	Symbol of set	
U	Universal set	
n(X)	Cardinal number of set X	
b ∈ A	'b' is an element of set A	
a ∉ B	'a' is not an element of set B	
∅	Null or empty set	
A U B	Set A union set B	
A ∩ B	Set A intersection set B	
A ⊆ B	Set A is a subset of set B	
B ⊇ A	Set B is the superset of set A	

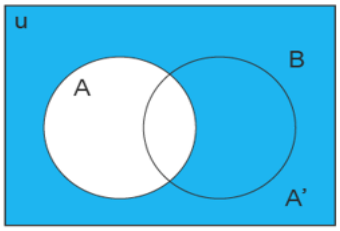


Set A = {1, 2, 3}
Set B = {3, 5, 7}
Elements of set A are 1, 2, 3
Element of set B are 3, 5, 7
Common element of set A and B is 3.

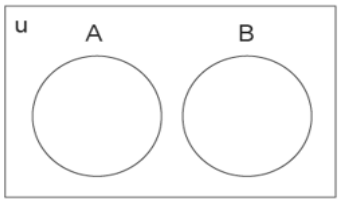
Operations of Sets and Venn Diagrams



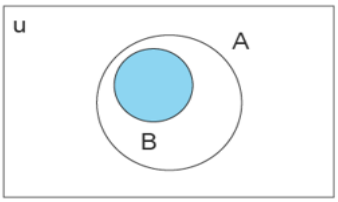
Set A



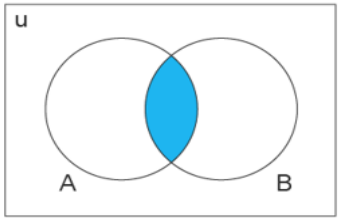
A' is the complement of A



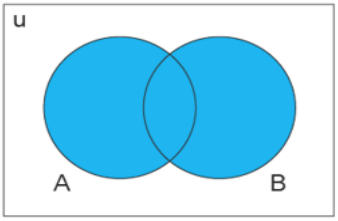
A and B are disjoint sets



$B \subset A$
B is proper subset of A



$A \cap B$
A and B are overlapping sets



$A \cup B$
A and B are overlapping sets

c) Multiplication rule of counting is such that if a task can be performed in n_1 ways and for each of these a second task in n_2 ways and for each of the k^{th} in n_k ways, then the entire sequence of k tasks can be performed in $n_1.n_2 \cdots n_k$ ways

- **Permutation and Combinations**

a) A permutation of n distinct objects is the choice of r objects from a set of n objects without replacement regarding the order.

${}_nP_0 = 1,\quad {}_nP_1 = n,\quad {}_nP_n = (n)(n-1)(n-2).....(2)(1) = n!$ and the number of permutations of n distinct objects taking r at a time is

$${}_nP_r = (n)(n-1)...(n-r+2)(n-r+1)$$

$$= \frac{\big((n)(n-1)...(n-r+2)(n-r+1)\big)\big((n-r)(n-r-1)...(2)(1)\big)}{(n-r)(n-r-1)...(2)(1)} = \frac{n!}{(n-r)!}$$

e.g. The number of 3-letter words formed from 5 letters is given ${}_nP_r = \frac{n!}{(n-r)!}$ then

${}_5P_3 = \frac{5!}{(5-3)!} = 60$

The number of distinct permutations of n objects of which n_1 are alike of the first kind, n_2 are alike of the second kind,....., n_k are alike of the kth kind and $n_1 + n_2 + ... + n_k = n$ is

$$\frac{n!}{(n_1!)(n_2!)...(n_k!)}$$

Conditional Permutation

i. How many 4 digit number (without repetition) that can be made from the digits 1-7 if 4 must be part of the number.
$r \times {}^{n-1}P_{r-1} = 4 * {}^6P_3 = 480$

ii. How many different 3 letter word can be made out of the 5 vowels if letter A will never be included
${}^{n-1}P_r = {}^4P_3 = 24$

iii. How many ways can the five vowels be arranges if;
- Two vowels, e and i are always together; m!(n-m+1)!=2!*4!=48
- Two vowels, e and i are never together; n!-[m!(n-m+1)!]= 5!-48=72

iv. How many different words can you form from the word MISSISSIPPI
$\frac{n!}{p!q!r!} = \frac{11!}{4!2!4!} = 34650$

b) Combination is the process of choosing r objects at a time from n distinct objects regardless of the order such that ${}_nC_r = \frac{n!}{r!(n-r)!}$

Therefore with regard to combination ABC=BAC
Note some special cases ${}_nC_0 = {}_nC_n = 1$, ${}_nC_1 = {}_nC_{n-1} = n$, ${}_nC_r = {}_nC_{n-r}$ and

${}_nC_r = \binom{n}{r} = \frac{{}_nP_r}{r!} = \frac{n!}{(n-r)!r!}$

Illustrations:

i. In how many ways can the three letters A, B, C be arranged? Since the three are unique we have ABC, ACB, BAC, BCA, CAB and CBA hence six ways.

Similarly ${}_3P_3 = \frac{3!}{(3-3)!} = \frac{3!}{0!} = 6$

ii. Find the possible permutations of the following 5 letters: A, A, A, B, C

There are five objects of which three are alike.

$\therefore$ The answer $= \frac{{}_5P_5}{3!} = \frac{5!}{3!} = 60$

iii. How many 7-letter words can be formed using the letters of the word 'BENZENE'?

(Since there are 1 B, 3 E, 2 N and 1 Z) The number of 7-letter words that can be formed is $\frac{7!}{(1!)(3!)(2!)(1!)} = 420$

iv. Find the possible combinations of 5 distinct objects taken 3 at a time. The answer is

$$= \frac{5!}{3!(5-3)!} = 10$$

v. The number of 3-person committees that can be formed from a group of 4 persons is ${}_4C_3 = \frac{4!}{3!(4-3)!} = 4$

vi. A committee of 3 persons is to be constituted from a group of 2 men and 3 women. In how many ways can this be done? How many of these committees would consist of 1 man and 2 women? [10], [6]

vii. A group consists of 4 girls and 7 boys. In how many ways can a team of 5 members be selected if the team has
(i) no girls [21]

(ii) at least one boy and one girl [441]

(iii) at least three girls [91]

viii. A box contains 8 eggs, 3 of which are rotten. Three eggs are picked at random. Find the probabilities of the following events.

(a) Exactly two eggs are rotten.
(b) All eggs are rotten.
(c) No egg is rotten.

Solution:

(a) The 8 eggs can be divided into 2 groups, namely, 3 rotten eggs as the first group and 5 good eggs as the second group.

Getting 2 rotten eggs in 3 randomly selected eggs can occurred if we select randomly 2 eggs from the first group and 1 egg from the second group.

The number of this outcome is $\left({}_3C_2\right)\left({}_5C_1\right)=15$

Total number of possible outcomes of selecting 3 eggs randomly from the total 8 eggs is ${}_8C_3=56$.

Thus the probability of having exactly two rotten among the 3 randomly selected eggs is $\dfrac{\left({}_3C_2\right)\left({}_5C_1\right)}{{}_8C_3}=\dfrac{15}{56}$

(b) Similarly, the probability of having all 3 rotten eggs is

$$\frac{\left({}_3C_3\right)\left({}_5C_0\right)}{{}_8C_3}=\frac{1}{56}$$

(c) The probability of having no rotten egg is

$$\frac{\left({}_3C_0\right)\left({}_5C_3\right)}{{}_8C_3}=\frac{10}{56}=\frac{5}{28}$$

ix. Three items are selected at random from a manufacturing process. Each item selected is inspected and classified as either defective (D) or non-defective (N). Determine the probability that out of four bunches selected there is at least one defective iten.

Its sample space is = $\begin{Bmatrix} DDD & DDN & DND & NDD \\ DNN & NDN & NND & NNN \end{Bmatrix}$

The event that the number of defectives in above example is greater than one is such that

$P(at\ \ least\ \ 1D) = 1 - P(No\ \ defective) = 1 - \frac{1}{8} = \frac{7}{8}$

x. How many different license plate containing two letters following by three digits with the first digit not zero can be printed? (repetition allowed)

	1st Letter	2nd Letter	1st Digit	2nd Digit	3rd Digit
Number of Choices	A - Z (26)	A - Z (26)	1 - 9 (9)	0 - 9 (10)	0 - 9 (10)

Number of different licence plates that can be printed is

(26)(26)(9)(10)(10) = 608,400

xi. 180 students took examinations in Communication skills and Basic Mathematics. Their results were as follows:

Number of students passing Communication skills = 80

Number of students passing Basic Mathematics = 120

Number of students passing at least one subject = 144

Find the probability that a randomly selected student passed both subject.

Solution

Then we can rewrite the above results as:

Probability that a randomly selected student passed Communication skills = $\frac{80}{180} = \frac{4}{9}$

Probability that a randomly selected student passed Basic Mathematics $= \frac{120}{180} = \frac{2}{3}$

Probability that a randomly selected student passed at least one subject $= \frac{144}{180} = \frac{4}{5}$

Let E be the event of passing English, and M be the event of passing Mathematics, such that $P(E) = \frac{4}{9}$; $P(M) = \frac{2}{3}$; $P(M \cup E) = \frac{4}{5}$

As $P(M \cup E) = P(E) + P(M) - P(M \cap E)$

$\therefore \quad P(M \cap E) = P(E) + P(M) - P(M \cup E) = \frac{4}{9} + \frac{2}{3} - \frac{4}{5} = \frac{14}{45} = 0.31$

xii. What is the probability of getting a total of '7' or '11' when a pair of dice is tossed?

Solution

Total number of possible outcomes = (6) (6) = 36

Possible outcomes of getting a total of '7' :{ 1, 6; 2, 5; 3, 4; 4, 3; 5, 2; 6, 1}

Possible outcomes of getting a total of '11': {5,6; 6,5}

Let A be the event of getting a total of '7', and B be the event of getting a total of '11'.

The probability of getting a total of '7' or '11' is $P(A \cup B)$.

$P(A \cup B) = P(A) + P(B) - P(A \cap B) = P(A) + P(B)$ ...A and B are mutually exclusive

$$= \frac{6}{36} + \frac{2}{36} = \frac{2}{9}$$

Conditional Combination

i. In how many different ways can 5 boys and 5 girls form a circle such that the boys and girls are alternating? After fixing one boy there will 4 ways of placing boys;4! There will 5 places to be filled by girls in 5 ways; 5! Hence by the principle of multiplication there 4!*5!=2880

ii. How can the above be seated if no two boys are allowed to seat together? Leaving one seat vacant between two boys there are 4! Ways and the remaining 5 seated can be filled by girls in 5! Ways hence there 4!*5!=2880

iii. How many ways can 8 persons be arranged in a circle? (8-1)!=7!= 5040

Conditional Probability

Let A and B be two events. The conditional probability of event A given that event B has occurred, denoted by $P(A/B)$ is defined as

$$P(A/B) = \frac{P(A \cap B)}{P(B)} \qquad \text{Provided that P(B) > 0.}$$

Similarly, the conditional probability of B given that event A has occurred is defined as

$P(B/A) = \frac{P(A \cap B)}{P(A)}$, provided P(A) > 0.

Some properties of conditional probability are:

- Let A and B be the events of a sample space S of an experiment.
  Then P(S|B) = P(S|A) = 1
- Let A and B be the events of a sample space S of an experiment and let E be an event such that P(E) ≠ 0. Then, P[(A U B)|E] = P(A|E) + P(B|E) – P[(A ∩ B)|E]
- P[(not A)|B] = 1 – P(A|B)

If A and B are disjoint events then P(A ∩ B) = 0

Illustration 1

The university has observed that 75% of all government sponsored students use part of their HELB money to pay fees, 80% get more from their guardians, and 65% use both the HELB money and guardians. What are the probabilities that:

(a) A student clears his/her fees using his/her HELB money after guardians'?
(b) A student clears his/her fees using his/her guardian money after HELB?

Solution

Let A be the event that a student clears fees using HELB, and B the event that he/she uses guardians'.

It is given that: $P(A) = 0.75$; $P(B) = 0.80$; $P(A \cap B) = 0.65$

(a) P(clears by HELB having used guardians') $= P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{0.65}{0.80} = 0.8125$

(b) P(Clears by guardians' having used HELB) $= P(B/A) = \frac{P(A \cap B)}{P(A)} = \frac{0.65}{0.75} = 0.8667$

Illustration 2

In a survey among few Machakos town residents indicated that, 60% read Daily Nation newspaper, 40% read Standard newspaper and 20% read both. If a person is chosen at random and found to have read Standard newspaper determine the probability that he also reads Daily Nation newspaper.

Solution:

P(read Daily Nation newspaper) =P(A) = 0.60

P(read Standard newspaper)= P(B)=0.40

P(read both) = (A ∩ B) = 0.20

Probability of a person reading Daily Nation newspaper having already read Standard newspaper is given by –

P(A|B) = P(A ∩ B)/P(B) = 20/40 = 0.5

Exercise

In a class, 40% of the students like Basic Mathematics, 25% of students like Statistics and 15% like both units. If a student is chosen at random, determine the probability that his interest in Statistics was influenced by interest in basic Mathematics [⅜.]

Multiplicative Rule

$P(A \cap B) = P(A)P(B/A)$

or $= P(B)P(A/B)$

Statistically Independent events: the occurrence or non-occurrence of one event has no effect on the probability of occurrence of the other event.

Two events A and B are independent iff $P(A \cap B) = P(A)P(B)$

Illustration

A pair of fair dice is thrown twice. What is the probability of getting totals of 7 and 11?

Solution

Let A_i be the event of getting '7' in the ith throw and B_j be the event of getting '11' in the jth throw.

P(Getting totals of 7 and 11) $= P(A \cap B) = P(A_1 \cap B_2) + P(B_1 \cap A_2)$

$$= P(A_1)P(B_2/A_1) + P(B_1)P(A_2/B_1)$$

$$= P(A_1)P(B_2) + P(B_1)P(A_2) \text{ .... } A_i, B_j \text{ are independent}$$

$$= \left(\frac{6}{36}\right)\left(\frac{2}{36}\right) + \left(\frac{2}{36}\right)\left(\frac{6}{36}\right) = \frac{1}{54}$$

Theorem of Total Probability

If the events $B_1, B_2, ..., B_k$ constitute a partition of the sample space S such that $P(B_i \neq 0)$ for i = 1, 2, ... , k, then for any event A of S

$$P(A) = P(A \cap B_1) + P(A \cap B_2) + ... + P(A \cap B_k)$$

$$= P(B_1)P(A/B_1) + P(B_2)P(A/B_2) + ... + P(B_k)P(A/B_k)$$

Illustration

Suppose in a certain local university 50% of the students are admitted in engineering courses and 15% of these are females; 30% of the students are admitted in business courses and 40% of these are females; and finally, 20% are admitted in tourism and hospitality and 60% of these are female students. If a student is picked at random from the university, find the probability that it is a female student.

Let A be the event that the student is a female,

B_1 be the event that the student is from Engineering department,

B_2 be the event that the student is in business department, and

B_3 be the event that the student is from tours and hospitality. Then

$$P(A) = P(A \cap B_1) + P(A \cap B_2) + P(A \cap B_3)$$

$$= P(B_1)P(A/B_1) + P(B_2)P(A/B_2) + P(B_3)P(A/B_3)$$

$$= (0.50)(0.15) + (0.30)(0.40) + (0.20)(0.60) = 0.315$$

Baye's Theorem

It is a probability of purporting that one proposition is true given that another proposition is true. It separates the tests from the events since a times tests are flawed (false ±) If $E_{1,}E_2,...,E_k$ are mutually exclusive events such that $E_1 \cup E_2 \cup ... \cup E_k$ contains all sample points of S, then for any event D of S with $P(D) \neq 0$,

$$P(E_i / D) = \frac{P(E_i \cap D)}{P(D)} = \frac{P(E_i \cap D)}{\sum_{j=1}^{k} P(E_j \cap D)}$$

$$= \frac{P(E_i)P(D/E_i)}{P(E_1)P(D/E_1) + P(E_2)P(D/E_2) + ...P(E_k)P(D/E_k)}$$

The probabilities $P(E_i)$ are referred to as marginal or prior probabilities while the conditional probabilities $P(E_i \mid D)$ are referred to as the posterior probabilities.

Illustrations;

xiii. Using the previous example the probability that the selected student was a female but from the business department is such that

$$P(B_i \mid A) = \frac{P(B_2)P(A/B_2)}{P(B_1)P(A/B_1)+P(B_2)P(A/B_2)+P(B_3)P(A/B_3)}$$

$$= \frac{0.12}{0.315} = 0.38$$

xiv. Suppose a box contains 2 red balls and 1 white ball and a second box contains 2 red ball and 2 white balls. One of the boxes is selected by chance and a ball is drawn from it. If the drawn ball is red, what is the probability that it came from the 1st box?

Solution

Let A be the event of drawing a red ball and B be the event of choosing the 1st box.

Given: $P(B) = P(B') = \frac{1}{2}$; $P(A/B) = \frac{2}{3}$; $P(A/B') = \frac{2}{4}$

P(Coming from the 1st box/the drawn ball is red) $= P(B/A)$

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

$$= \frac{P(B)P(A/B)}{P(B)P(A/B)+P(B')P(A/B')} = \frac{(1/2)(2/3)}{(1/2)(2/3)+(1/2)(2/4)} = \frac{4}{7}$$

xv. The following are factories A, B and C outputs and their defective rates

xvi.

Factory	Daily production (output) %	Probabilities of defective bunch
A	0.35= P(A)	0.015=P(D|A)
B	0.35=P(B)	0.010=P(D|B)
C	0.30=P(C)	0.020=P(D|C)

If an item was randomly selected from the daily output and found to be defective, determine the probability that it was produced by factory C

xvii. Based on the recent observations 5% of the MUC female students have low hemoglobin. The college clinic laboratory can detect 80% of the low HgB when it is there. It also gives 12% false positive results. A female student in MUC volunteers for HgB test in the college clinic laboratory, determine the following probabilities;

(a) That the test result will be positive (Low HgB)

(b) That, given a positive result, she has a low HgB;

(c) That, given a negative result, she has a normal HgB;

(d) That she was misclassified

solution

	Low HgB (5%)	Normal HgB (95%)
Tests positive	80%	12%
Tests negative	20%	88%

True positive=5%*80%, True negative 95%*88%, False negative=5%*20% and False positive=95%*12%

Let T=test positive, H=Hemoglobin low and M=misclassified

∴

(a) P(T)=P(T|H)P(H)+P(T|H')P(H')=5%*80%+95%*12%=0.154

(b) P(H|T)÷P(T)=0.04÷0.154=0.260

(c) P(T')=P(T'|H)P(H)+P(T'|H')P(H')=5%*20%+95%*88%=0.846
P(H'|T')=P(T'|H')P(H')÷P(T')=0.836÷0.846=0.988

(d) P(M)= P(T|H')+P(T'|H)=5%*20%+95%*12%=0.124

Sensitivity or specificity? A matter of choice

If the criteria for a positive test result are stringent then there will be few false positives but the test will be insensitive. Conversely, if criteria are relaxed then there will be fewer false negatives but the test will be less specific

Analyzing repeatability

The repeatability of measurements of continuous numerical variables such as blood pressure can be summarized by the standard deviation of replicate measurements or by their coefficient of variation (standard deviation mean). When pairs of measurements have been made, either by the same observer on two different occasions or by two different observers, a scatter plot will conveniently show the extent and pattern of observer variation. For qualitative attributes, such as clinical symptoms and signs, the results are first set out as a contingency table.