Dr. Asemota O.J

STA 200A

STATISTICS FOR BIOLOGICAL SCIENCES (1)

Write on both sides of the paper (1)

INTRODUCTION TO BIOSTATISTICS

What is Statistics?

Statistics is all about the conversion of data into useful information. Statistics is therefore a process where we are:

- Collecting data
- Organizing data
- Summarizing data
- Presenting data
- Analyzing data
- Making valid conclusion / inferences.

Hence, Statistics can be defined as the scientific collection, organization, summarization, presentation, analysis of data as well as making valid conclusions or inferences based on the analyses.

To illustrate statistics as a way of thinking, lets consider the conventional criminal court procedures. A crime has been committed and a suspect has been identified. After a police investigation to collect evidence against the suspect, the prosecutor presents summarized evidence to a jury. The jurors interpret the laws regarding convicting beyond a reasonable doubt, and then debate. After the debate, the jurors vote and a verdict is reached: guilty or not guilty. Trial by jury is necessary because the truth is often unknown, at least uncertain. It is uncertain because of variability (every case is different) and also because of possibly incomplete

information. Hence, trial by jury is a way Society deals with uncertainties; its goal is to avoid or minimize mistakes. The trial by jury involves the following steps:

(1) forming an assumption or hypothesis (that every person is innocent until proved guilty)

(2) Gathering data (evidence against the suspect) and

(3) Making decision/conclusion whether the hypothesis should be rejected (guilty) or should not be rejected (not guilty)

Thus, how does biological and medical Sciences deal with uncertainties?

In the context of a trial by jury, let us consider some specific examples:

(1) The Crime is lung cancer and the suspect is Cigarette Smoking

(2) The Crime is leukemia and the suspect is pesticides or

(3) The Crime is breast cancer and the suspect is defective gene.

The process is now called research and the tool to carry out the research is Biostatistics. In a simple way, biostatistics serves as the biomedical version of the trial by jury process.

Dr. Asensty O.J 0.285 6

Candidate's Number..... Page No.: (3)

Do not write on either margin

Question No:.....

Write on both sides of the paper

It helps to deal with uncertainties using incomplete information. Since nature is complex and full of unexplained biological variation, uncertainties are inevitable.

Biological and medical sciences deal with uncertainties through a process called BIOSTATISTICS, consisting of the following steps:

(1) Forming an assumption or hypothesis (from the research question),

(2) Gathering data (from clinical trials, surveys, medical records e.t.c).

(3) Making decision(s)/conclusion(s) (by doing statistical analysis/inference; a guilty verdict is referred to as Statistical significance).

Therefore, Biostatistics can be defined as the branch of Statistics applied to biological or medical Sciences. Biostatistics is also called Biometry The Greek words Bios (life) and metron (measured), hence, biometry simply means measurement of life. It may be stated as the application of statistical methods to the solution of biological problems. Biostatistics covers applications and contributions not only from health, medicine and nutrition but also from fields such as agriculture, genetics, biology, botany, zoology, biochemistry, epidemiology and many others.

USE/ROLE OF STATISTICS IN BIOLOGICAL SCIENCE AND AGRICULTURE

Biostatistics helps in proper interpretation of Scientific data generated in biology, public health and other health sciences. In these sciences, subjects (patients, rats, fishes, cells, e.t.c) exhibit variations in their response to stimuli. These variations may be due to different treatments, it may be due to chance, measurement error, or other characteristics of the individual subjects. Biostatistics is particularly concerned with disentangling these different sources of variation. For example, do results of treating patients with two therapies justify the conclusion that one treatment is better than the other?

In epidemiology, biostatistics is used to determine how diseases develop, progress and spread. For example, biostatisticians use statistics to predict the behaviour of an illness like flu, cholera, Ebola, e.t.c. It is used to predict the mortality rate, the symptoms and even the time of the year people might get it.

Biostatistics helps medical researchers to design studies, decide what data to collect, analyze data from medical experiment, help interprete the results of the analysis. Infact, all medical researchers uses biostatistics from beginning to end.

Thus, biostatistics is integral to the advancement of knowledge in biology, health policy, clinical

medicine, public health, genomics, agriculture and other disciplines.

Statistics plays an important role in agriculture whether it is the question of compiling agricultural statistics, conduct of agricultural experiments along with the techniques of drawing valid inferences about treatments/~~variation~~ varieties/animals etc or in animal/crop improvement program. For example, the agricultural experimenter may be interested in establishing relations to explain the behaviour of the variable under study as a function of certain input factors. Even though the functional form is unknown, the method of least squares in statistics can be used to estimate the functional relationship.

Agriculture statistics ascertain the crop production, crop yield and qualities of the crop production. It also furnishes information about the different methods which can be adopted for improving crop output. Agriculture statistics is also useful in providing facilities for livestocks, crops and provision of pesticides and fertilizers.

Candidate's Number.....

Question No:.....

Write on both sides of the paper

BIOLOGICAL VARIATIONS AND UNCERTAINTIES

One of the outstanding features of human beings and all living things, is that no two individuals are exactly alike. Even identical twins become distinguishable due to environmental influences. Variations are present not only between individuals but also within individuals from time-to-time.

Biological variations are those differences between cells, individual organisms, or group of organisms of any species.

FACTORS CONTRIBUTING TO VARIATIONS AND UNCERTAINTIES

The following are some of the factors contributing to variations and uncertainties

(1) Biological ~~Variability~~ Factors

Age, Sex, height and weight are some of those biological factors which are considered natural in health set-up. It is a known fact that health parameters of children are entirely different from those of adults. Anatomical, physiological, biochemical and almost all kinds of measurements differ from age-to-age. Biological variability is seen not only between subjects but within subjects also. Sometimes a person responds exceedingly well at one time but fails at other time.

(2) Observer ~~Variability~~ Factor

Observers tend to differ in their assessment of th

Same subjects. X-rays are particularly notorious in this respect. Variation in blood pressure readings due to differences in hearing acuity or in interpretation of Korotkov sounds. For example, a blood pressure level of 160/95 mmHg in a person of age 50 years may be considered sufficient to warrant intervention by one physician but not so by another physician in the same subject. Such variations on the part of the observer, researcher or investigator are facts of life and cannot be wished away. They are inherent in human conscience and efforts are made from time-to-time to reconcile and come to a consensus.

(3) Instrument ~~Variability~~ Factor

Two instruments, when used on the same person, are very likely to give different readings. Laboratories tend to differ in their chemicals, reagents, techniques, processing time, etc which ultimately lead to different results. Another everyday example is that of differences in weight on beam balance and on Spring balance.

(4) Environmental ~~Variability~~ Factor

While our anatomy and physiology can be traced mostly to hereditary factors, the pathology is mostly caused by environmental factors. Smoking is seen as an important ingredient for several types of carcinomas and heart conditions. Tensions and Stresses in the home or work environment substantially affects one's ability to cope with, say infections. Love, affection and prayers sometimes do miracle in the

/16

recovery of a patient.

(5) Incomplete information

Health is a sensitive issue yet the health managers seldom get complete information on a community or a subject.

Even while interviewing a perfectly healthy person, it is doubtful that he/she would reveal the complete and truthful picture, either the person would forget part of the information or would intentionally suppress it.

CONTROLLING VARIATION

(1) Randomisation.

The allocation of subjects to various treatments should be random. This increases the likelihood of chance sources of variation being equally distributed all over, thus, cancelling out in the final analysis

(2) The second strategy is to choose a design such that the effect of all known sources of variation, like age, sex, body mass index, education, e.t.c can be directly estimated. This means that no important source of variation should be allowed to confound with the others.

(3) Variation can also be eliminated by studying an appropriate control. For clinical trials and animal experiments, the designs like randomized block, and ~~Complete random~~ Cross-over are used with such additional features as double blind.

FREQUENCY DISTRIBUTIONS

A frequency distribution is a table showing the number of observations or frequencies at different values or ranges of values of the variable.
It is often used to organize or arrange a body of data. A frequency distribution break up the data into groups or classes and shows the number of observations in each class.

In constructing a frequency table for grouped data, we first determine a set of class intervals that cover the range of the data (i.e include all the observed values). The class intervals are usually arranged from lowest numbers at the top of the table to highest numbers at the bottom of the table and are defined so as not to overlap. We then tally the number of observations that fall in each interval and present that number as a ~~frequency~~ frequency, called a class frequency. Some frequency tables include a column that represents the frequency as a percentage of the total number of observations; this column is called the relative frequency percentage. The completed frequency table provides a Frequency distribution.

GUIDELINES FOR CREATING FREQUENCY DISTRIBUTION FOR GROUPED DATA

(1) Determine the range of values - the difference between the highest and lowest values.
(2) Decide how many intervals to use (usually between 5 and 15 intervals are acceptable, unless the data is very large).
(3) Decide on the width of the interval. To determine the

width of the interval, divide the range by the number of class intervals selected. Round this result as necessary. i.e $w = R/K$

where R is the range, and K is the number of intervals. In addition, a width should be chosen so that it is convenient to use or easy to recognize, such as a multiple of 5 (or 1, for example, if the data set has a narrow range).

(4) Be sure that the class categories do not overlap!

(5) Most of the time, use equally spaced intervals, which are simpler than unequally spaced intervals and avoid interpretation problems. Sometimes wider intervals are needed where the data are sparse.

EXAMPLES:

(1) The following are weights in pounds of 57 Children at a day-care center in Gwagwalada.

68	63	42	27	30	36	28	82	79	27
22	23	24	25	44	65	43	25	74	51
36	42	28	81	28	25	45	12	57	51
12	32	49	38	42	27	31	50	38	21
16	24	69	47	23	22	43	27	49	28
23	19	46	30	43	49	12			

Construct a frequency distribution for the data.

Solution

(i) Range = $79 - 12 = 67$

(ii) If five intervals are used, we would have

$w = 67/5 = 13.4$

Sturges $K = 1 + 3.322 \log_{10} N$

and if 15 intervals are used, we would have

$W = 67/15 = 4.5$

Between these two values, 4.5 and 13.4, there are two convenient or conventional numbers: 5 and 10. Since the sample size is not large, a width of 10 should be an apparent choice because it results in fewer intervals (the usual concept of "large" is "100 or more").

(11) Since the smallest number is 12, we may begin our first interval at 10. Using the discussions above, we have the following seven intervals:

10–19, 20–29, 30–39, 40–49, 50–59, 60–69, 70–79

A Frequency Table of Weights

Weight (lb)	Tally	Frequency	(%) Relative Frequency
10–19	𝍸	5	8.8
20–29	𝍸 𝍸 ||||	19	33.3
30–39	𝍸 𝍸	10	17.5
40–49	𝍸 𝍸 |||	13	22.8
50–59	||||	4	7.0
60–69	||||	4	7.0
70–79	||	2	3.5
TOTAL		57	100.0

(2) A study was conducted to investigate the possible effects of exercise on the menstrual cycle. From the data collected from that study, we obtained the menarchal age (in years) of 56 female swimmers who began their swimming training after they had reached menarche; these

Served as controls to compare with those who began the training prior to menarche. The data is given below

14.0	16.1	13.4	14.6	13.7	13.2	13.7	14.3
12.9	14.1	15.1	14.8	12.8	14.2	14.1	13.6
14.2	15.8	12.7	15.6	14.1	13.0	12.9	15.1
15.0	13.6	14.2	13.8	12.7	15.3	14.1	13.5
15.3	12.6	13.8	14.4	12.9	14.6	15.0	13.8
13.0	14.1	13.8	14.2	13.6	14.1	14.5	13.1
12.8	14.3	14.2	13.5	14.1	13.6	12.4	15.1

Construct a frequency distribution table for the data

Solution

(i) Range = 16.1 − 12.4 = 3.7

(ii) If 5 intervals are used,

$w = 3.7/5 = 0.74$

and if we considered 15 intervals,

$w = 3.7/15 = 0.25$

Between 0.25 and 0.74, there are two convenient numbers, 0.25 and 0.5. However, 0.25 would create many intervals (15) for such a small data set. Hence, 0.5 seems to be a convenient number to use as the width.

(iii) Since the smallest number is 12.4, we may begin our interval at 12.0,

A Frequency Distribution for Menarchal Age (in Years).

Age (Years)	Tally	Frequency	Relative frequency (%)
12.0 – 12.4	I	1	1.8
12.5 – 12.9	||| III	8	14.3
13.0 – 13.4	|||	5	8.9
13.5 – 13.9	||| ||| II	12	21.4
14.0 – 14.4	||| ||| ||| I	16	28.6
14.5 – 14.9	IIII	4	7.1
15.0 – 15.4	||| II	7	12.5
15.5 – 15.9	II	2	3.6
16.0 – 16.4	I	1	1.8
Total		56	100.0

NOTE

A relative frequency distribution is obtained by dividing the number of observations in each class by the total number of observations in the data as a whole. The sum of the relative frequencies (%) equals 100.

CUMULATIVE FREQUENCY DISTRIBUTION

A cumulative frequency distribution shows, for each class, the total number of observations in all classes up to and including that class. A cumulative frequency table provides another way to display a frequency distribution. In a cumulative frequency table we list the class intervals and the cumulative relative frequency

in addition to the relative frequency.

The CUMULATIVE RELATIVE FREQUENCY OR CUMULATIVE PERCENTAGE gives the percentage of cases less than or equal to the upper boundary of a particular class interval. The cumulative relative frequency can be obtained by summing the relative frequencies in a particular row and in all preceding class intervals.

Example

Construct a Cumulative Relative frequency distribution for the data on weight of 57 children.

Solution

A CUMULATIVE RELATIVE FREQUENCY DISTRIBUTION

Weight Interval (lb)	Tally	Frequency	Cumulative Frequency	Relative (%) frequency	Cumulative (%) Relative frequency
10 – 19	𝍸	5	5	8.8	8.8
20 – 29	𝍸 𝍸 𝍸 ||||	19	24	33.3	42.1
30 – 39	𝍸 𝍸	10	34	17.5	59.6
40 – 49	𝍸 𝍸 |||	13	47	22.8	82.4
50 – 59	||||	4	51	7.0	89.4
60 – 69	||||	4	55	7.0	96.4
70 – 79	||	2	57	3.5	99.9 ≈ 100
Total				100	

From the cumulative relative frequency table above, we can say that 59.6% of the children in the data set have a weight of 39.5 lb or less. We can also say

(1) What is the percentage of children having weight less than 70 lb
(2) What is the percentage of children having weight less than 70 lb but at least 20 lb

= 96.6 − 8.8 = 87.8

that 96.4% of the children weigh 69.5 lb or less, and so on.

GRAPHICAL METHOD

A second way to display data is through the use of graphs. Graphs give the reader an overview of the essential features of the data.
Graphs are designed to provide visually an intuitive understandary of the data. Effective graphs are simple and clean. Hence, it is important that the graph be self-explanatory (i.e., have a descriptive title, properly labeled axes, and an indication of the units of measurement).
Some examples of graphical methods for displaying data includes: histograms, frequency polygon, Cumulative frequency polygons, bar-charts, pie-charts, etc.

1. Histogram

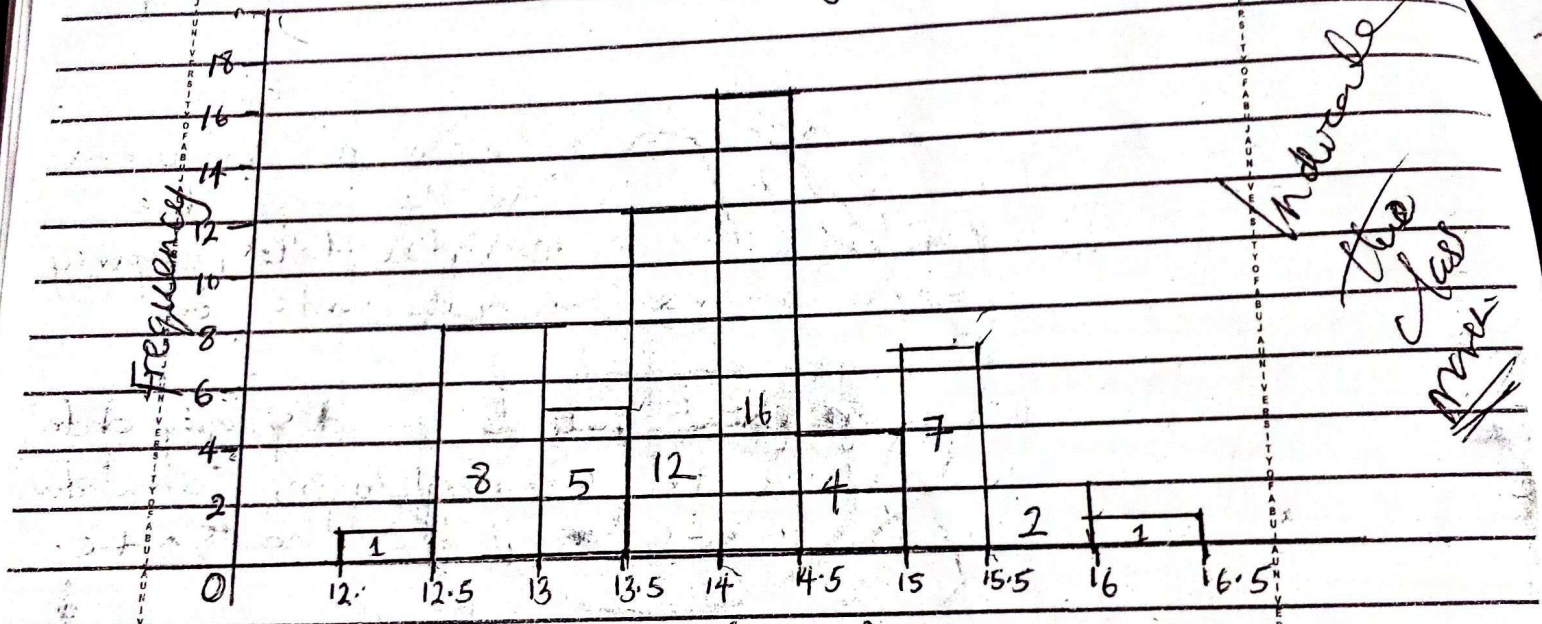
A histogram is a bar graph of a frequency distribution, where classes are listed on the horizontal (x) axis and frequencies along the vertical (y) axis. The appropriate labelling is important. A histogram presents us with a graphic picture of the distribution of measurements. This picture consists of rectangular bars joining each other, one for each class interval. However, if disjoint class intervals are used (such as the example on the weight of 57 children), the horizontal axis is marked with

true boundaries.

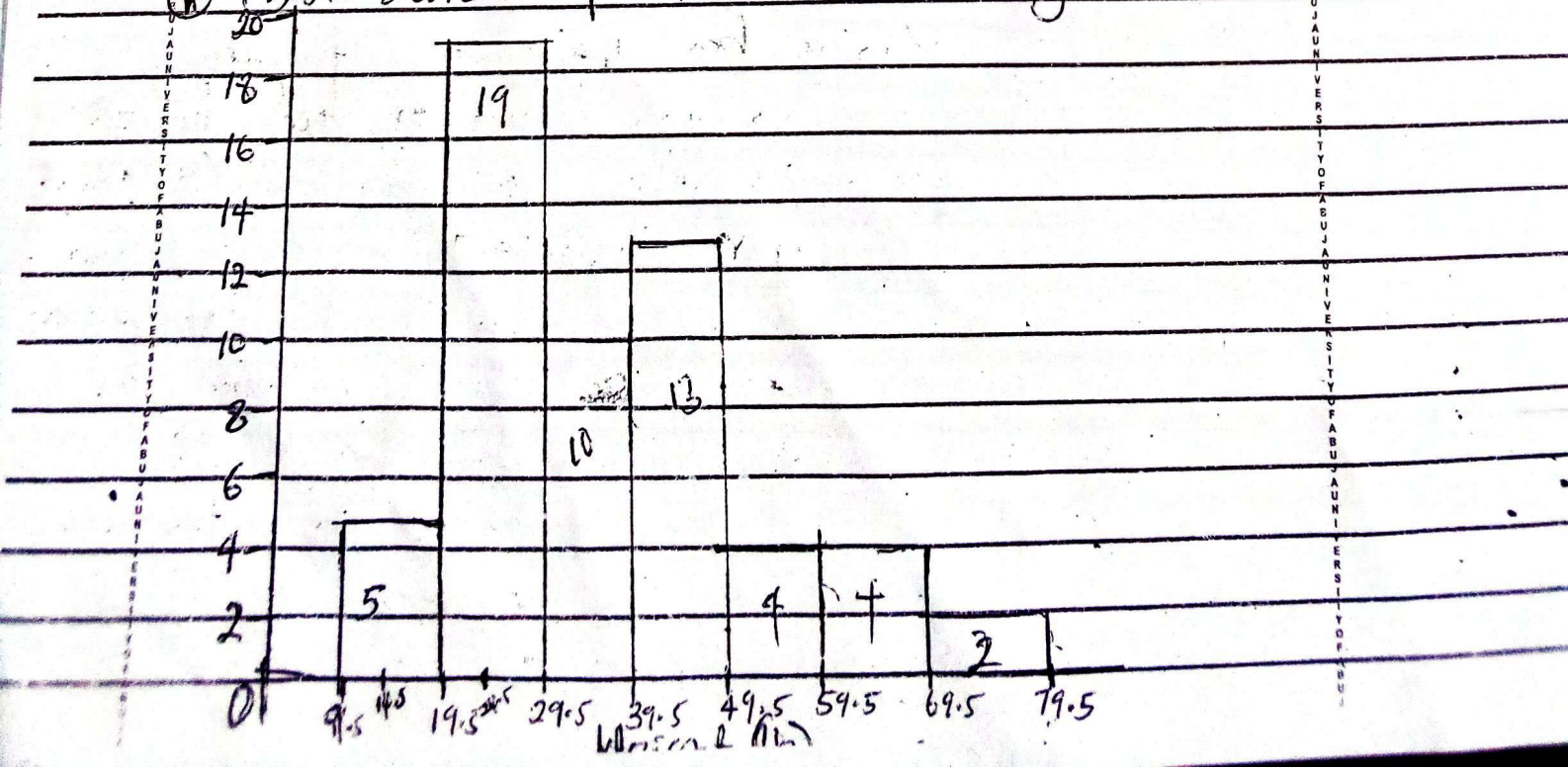
A true boundary is the average of the upper limit of one interval and the lower limit of the next-higher interval.

Example 1

Histogram

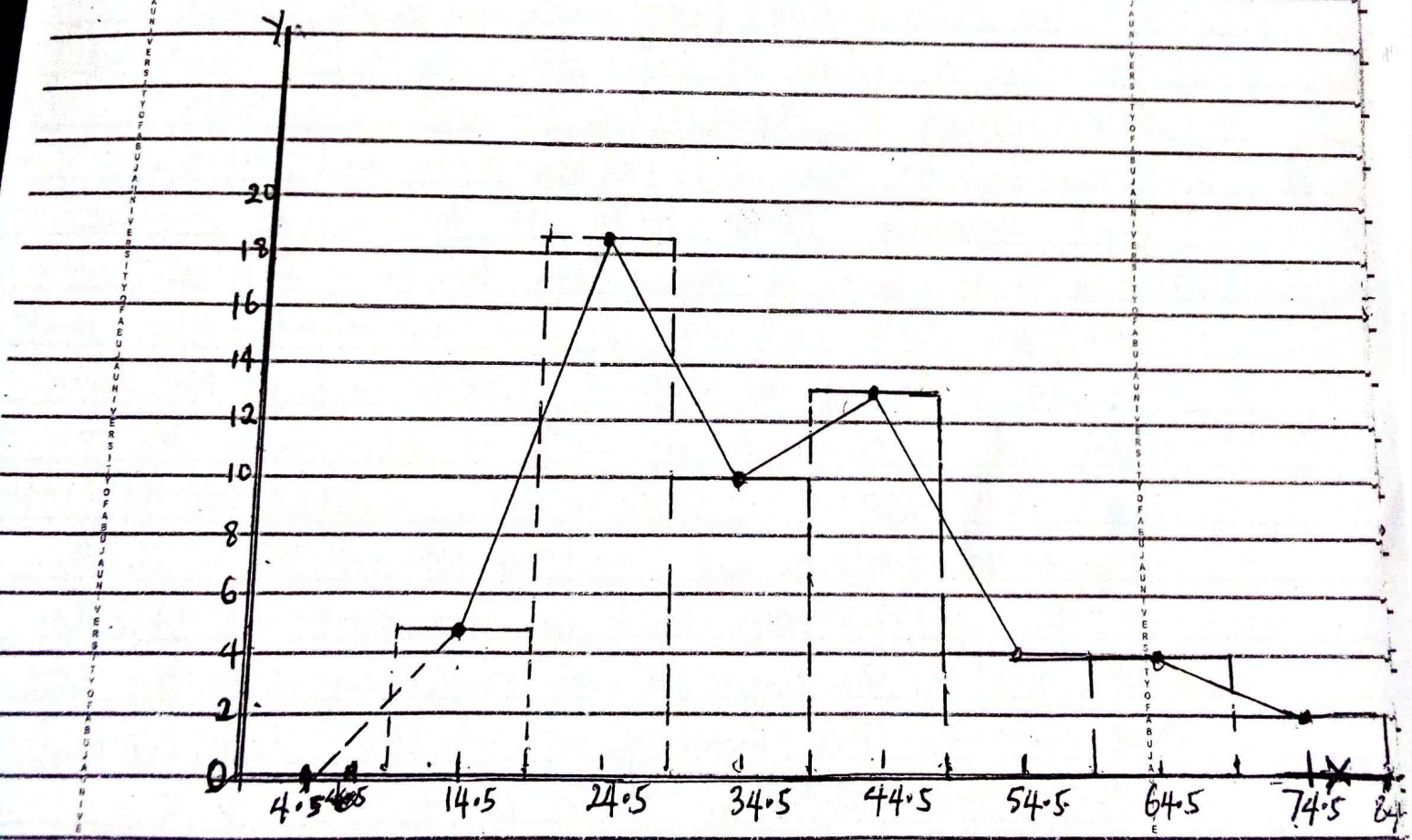


(4) Distribution of Menarchal Age.



② Frequency Polygon

A frequency polygon is a line graph of the class frequency plotted against the midpoint of the class interval (classmark). It can be obtained by connecting the midpoints of the tops of the rectangles in the histogram. When these midpoints are connected by straight lines, it results to a polygonal shape, hence the name ~~frequency~~ frequency polygon.



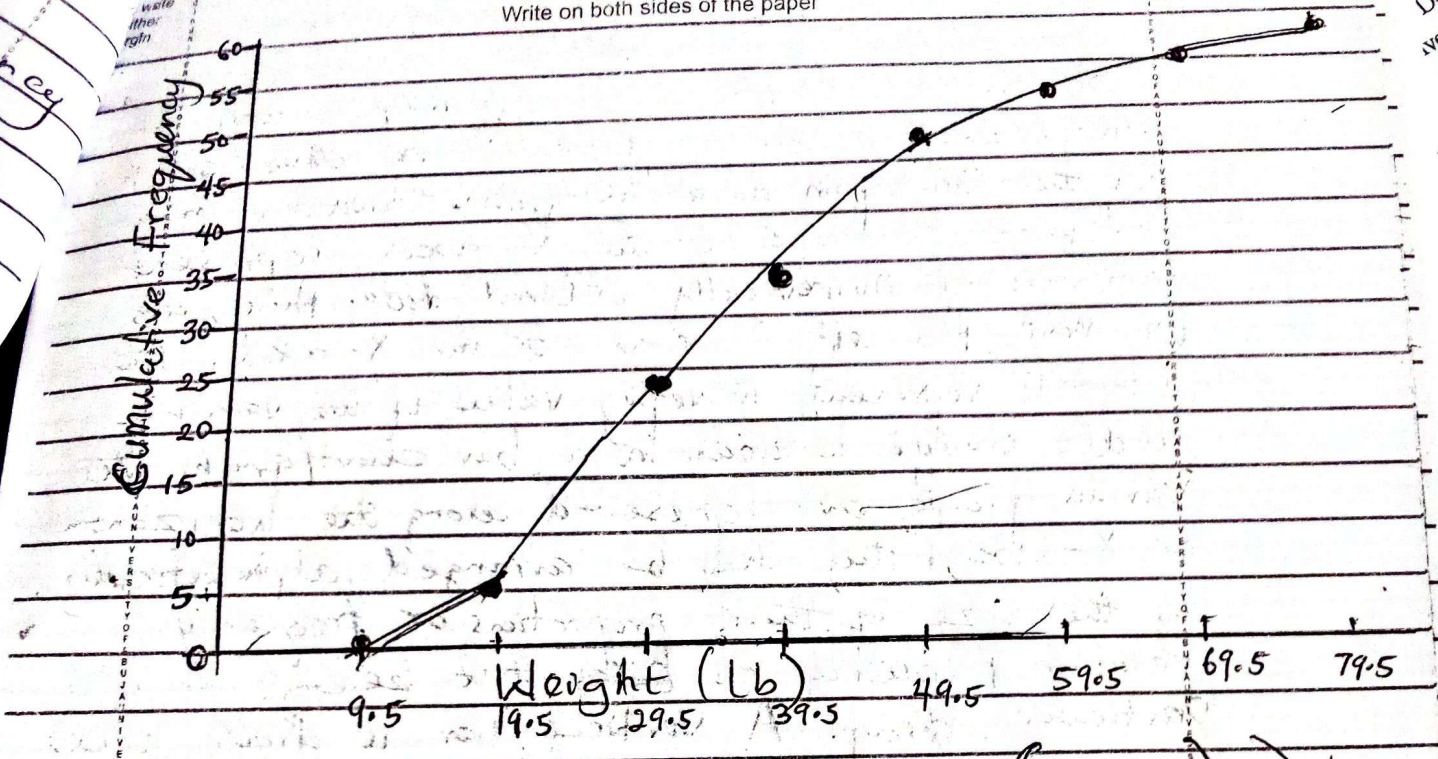
Frequency polygon for Height (lb) Data

Note: It is customary to ~~add the extension~~ extend the frequency polygon to the next-lower and -higher classmarks, which have a corresponding class frequency of zero.

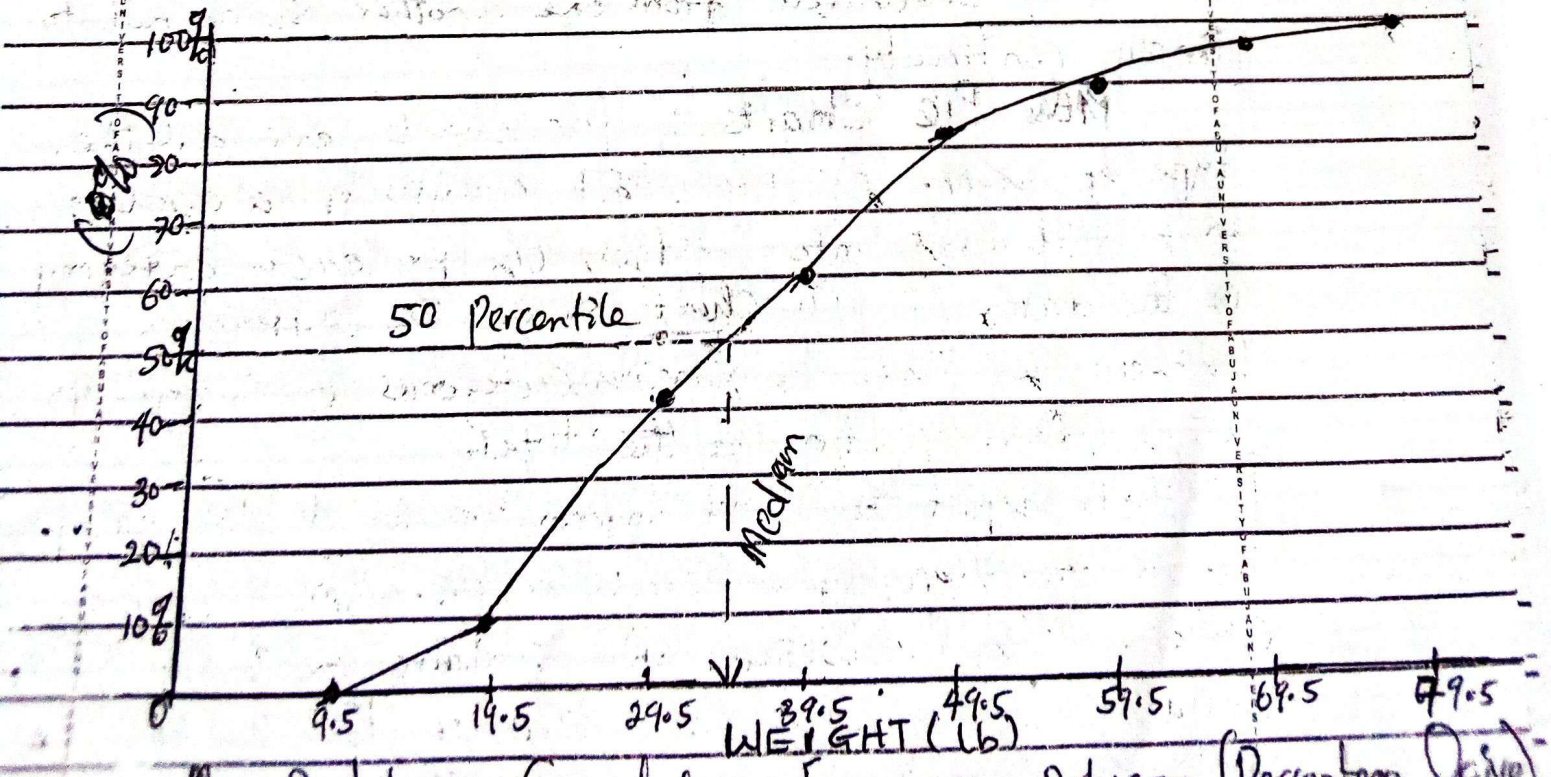
③ Cumulative Frequency Polygon or Ogive

A graph showing the cumulative frequency plotted against the upper class boundary is called a cumulative frequency polygon or ogive. To construct such a graph, we place a point with the horizontal axis marked at the upper class boundary and a vertical axis marked at the corresponding cumulative frequency. Relative percentage.

The cumulative frequency graph or Ogive provides a class of important statistics known as PERCENTILES OR PERCENTILE SCORES. The 90th percentile, for example, is the numerical value that exceeds 90% of the values in the data set and is exceeded by only 10% of them. Also, the 80th percentile is that numerical value that exceeds 80% of the values contained in the data set and is exceeded by 20% of them, and so on. The 50th percentile is commonly called the MEDIAN. To get the median, we start at the 50% point on the vertical axis and go horizontally until meeting the cumulative frequency graph; the projection of this intersection on the horizontal axis is the median. Other percentiles are obtained similarly.



A Cumulative Frequency polygon (Ogive) for Weight



A Relative-Cumulative Frequency polygon (Percentage Ogive)

wiley Pg(25) → 75
Intro (75) → 82

Page No. 20

Candidate's Number

Question No.

Do not write on either margin

Write on both sides of the paper

④ Bar graphs and Pie Charts

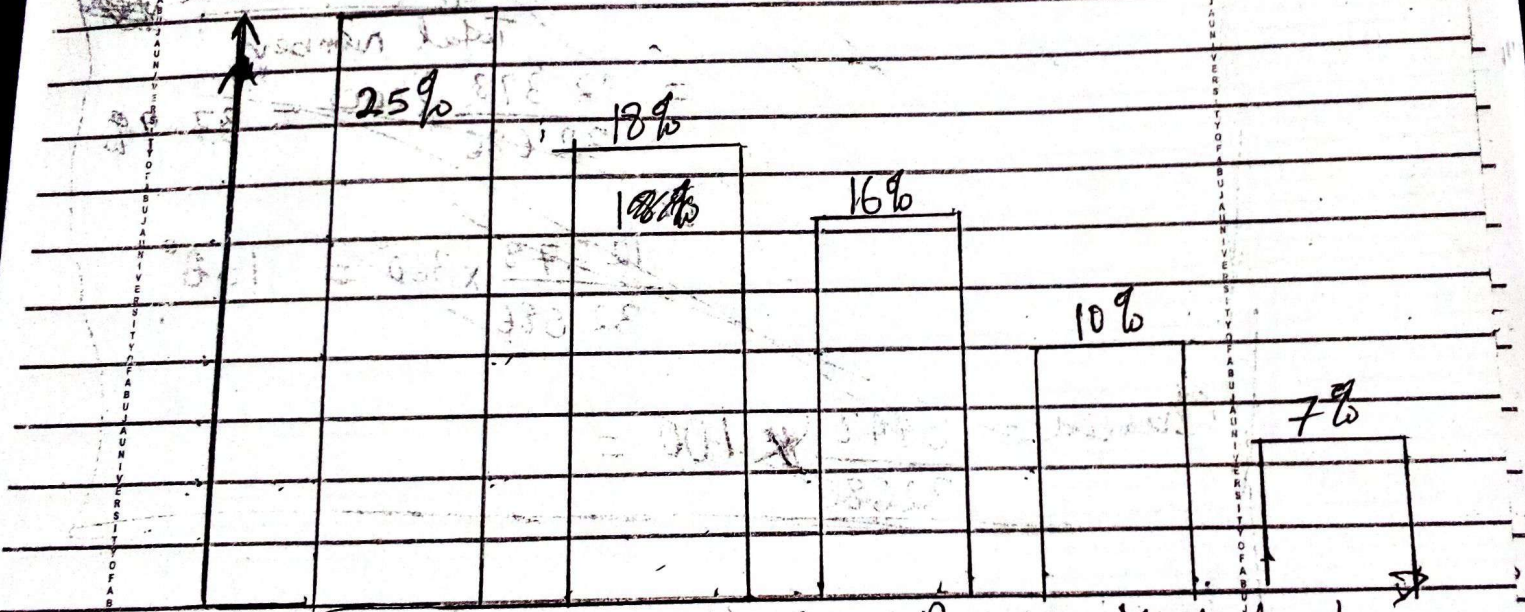
Bar graphs and pie charts are useful tool for summarizing categorical data. A bar graph has the same form as a histogram. However, in a histogram the values on the X-axis represent intervals of numerically ordered data. Hence, as we move from left to right on the X-axis, the intervals represent increasing values of the variable under study. Whereas, in a bar chart/graph, the various groups are represented along the horizontal or X-axis; they may be arranged alphabetically, by the size of their proportions or frequency relative frequency of cases that belongs to that particular group. A vertical bar is drawn above each group such that the height of the bar is the proportion associated with that group. The bars should be seperated from one another so as not to imply continuity.

Pie Charts: Pie charts are another type of graph. A pie chart consists of a circle; that is divided into wedges, one for each category of the data. A pie chart shows the differences between the sizes of various categories or subgroups as a decomposition of the total.

Pie charts depict the same information as do bar graphs, but in the shape of a circle or pie. Since a circle contains 360°, a wedge that contains 50% of the cases would have an angular measurement of 180°. The proportions of various categories in a piechart should add up to 100%

Examples

① A study was conducted by the University of Abuja Teaching Hospital on students without a recent medical check up. According to the report, 25% of Ijaws said that they had not seen a doctor or a dentist in the last two years, followed by 18% of Igalas, 16% of Idomas, 10% of Gwaris and 7% of Nupes. Use a bar chart to present the data.



A Bar Chart: Students Without a Recent Medical Checkup.

②

Causes of Death	Number of Deaths
Heart disease	12,378
Cancer	6,448
Cerebrovascular disease	3,958
Accidents	1,814
Others	8,088
Total	32,686

The table above shows the number of deaths

due to a variety of causes among U.S.A residents for the year 2015.
Use a pie chart to represent the data.

Solution

First, we calculate the proportion of deaths due to each category.

Death due to Heart disease = No of heart disease / total number

$= \frac{12\,378}{32\,686} \times 100 = 37.9\%$

or

$\frac{12\,378}{32\,686} \times 360^\circ = 136^\circ$

Cancer $= \frac{6448}{32686} * 100 =$

Heart disease = Number of heart disease / Total number of death.

$= \frac{12,378}{32\,686} = 0.379$

In Percentage $= 0.379 \times 100 = 37.9\%$

In degrees $= 0.379 \times 360^\circ = 136.4^\circ$

Cancer $= \frac{6448}{32686} = 0.197$

GLOSSARY

(1) Class Intervals

A symbol defining a class, such as 10–19, is called a class interval. The terms class and class interval are often used interchangeably, although the class interval is actually a symbol for the class.

(2) Open class interval

A class interval that has either no upper class limit or no lower class limit indicated is called an Open class interval. For example, referring to age groups of individuals, the class interval, "20 years and over" is an open class interval.

(3) Class Limit.

The end numbers, 10 and 19, are called class limits; the smaller number (10) is the lower class limit, and the larger number (19) is the upper class limit.

(4) Class Boundaries

Class boundaries refer to the true class limits. They are obtained by adding the upper limit of one class interval to the lower limit of the next-higher class interval and dividing by 2.

(5) Class Size or Width.

The size or width, of a class is the

In percent = $0.197 \times 100 = 19.7\%$

In degree = $0.197 \times 360^\circ = 70.9^\circ$

Cerebrovascular disease = $\frac{3958}{32686} = 0.121$

In % = $0.121 \times 100 = 12.1\%$

In degree = $0.121 \times 360^\circ = 43.6^\circ$

Accidents = $\frac{1814}{32686} = 0.055$

In % = $0.055 \times 100 = 5.5\%$

In degree = $0.055 \times 360^\circ = 19.8^\circ$

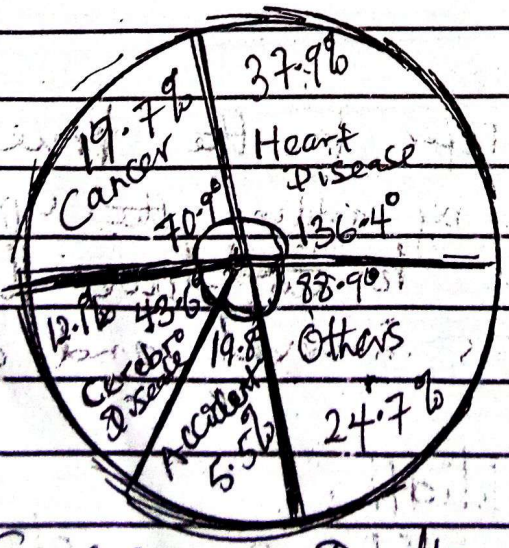
Others = $\frac{8088}{32686} \rightarrow 0.247$

In Percentage = $0.247 \times 100 = 24.7\%$

In degrees = $0.247 \times 360^\circ = 88.9^\circ$

Total Percentage = $37.9 + 19.7 + 12.1 + 5.5 + 24.7 \simeq 100\%$

Total degrees = $136.4 + 70.9 + 43.6 + 19.8 + 88.9 \simeq 360^\circ$



Pie Chart: Causes of Death in U.S.A for 2015.

SOURCES OF DATA

There are two common means of gathering data for research purpose. These are primary and Secondary ~~data~~ sources.

1. Primary data: These are data collected by the investigator or researcher himself/herself for a specific purpose. They They are first hand data. The method of primary data Collection includes; questionnaire, interview, experiments, observation, survey e.t.c. primary data are highly reliable and misunderstanding is usually avoided in such data.

2. Secondary data: These are data abstracted from a source. The investigators are usually not responsible for the original collection of the data. Sources of Secondary data include; published statistics (e.g UNO, WHO, National Bureau of Statistics, CBN, Ministry of Health e.t.c), abstraction from records.

Secondary data are faster and less expensive but they are highly prone to error. Hence, they are not as reliable as primary data.

difference between the lower and upper class boundaries. It is also referred to as the class length.

(6) Class Mark

The class mark is the midpoint of the class interval and is obtained by adding the lower and upper class limits and dividing by 2. Hence, the class mark of the interval 10-19 is (10+19)/2 = 14.5. Then classmark is also called class midpoint.

MEASURES OF CENTRAL TENDENCY

Measures of central tendency are numbers that tells us where the majority of values in the distribution are located. They represent the center of the probability distribution from which the data were drawn or sampled. These measures are also called measures of location or averages. The most common measures of central tendency or averages are; arithmetic mean, median, mode, geometric mean, and the harmonic mean.

1 The Arithmetic Mean

The arithmetic mean is the sum of the individual values in a data divided by the number of values in the data set. It is denoted by $\bar{X}$ and is defined as

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

If the numbers $X_1, X_2, \ldots X_n$ occur $f_1, f_2, \ldots f_k$ times (frequencies), or for a large data set (e.g. more than 30 observations when performing calculations by hand), summing the individual observations become tedious, so we use grouped data. The mean using grouped data is

$$\bar{X} = \frac{\sum_{i=1}^{n} f_i X_i}{\sum_{i=1}^{n} f_i} = \frac{\sum fX}{\sum f} = \frac{\sum fX}{\sum N}$$

Page No. 28

Where X_i is the class mark of the ith interval and f_i is the frequency of observations in the ith interval, $N = \sum f$ is the total frequency (i.e the total number of cases).

Examples

1. Consider the data 8, 3, 5, 6, 7.
Compute the arithmetic mean.

$$\bar{x} = \frac{\sum x}{n} = \frac{8+3+6+6+7}{5} = \frac{29}{5} = 5.8$$

(2) The data below represents the plasma glucose values (mg/dl) for a sample of 100 adults.

74, 75, 77, 78, 78, 78, 80, 81, 81, 81, 82, 82, 8
83, 83, 83, 84, 85, 86, 86, 86, 86, 87, 87, 87, 88,
88, 88, 88, 88, 88, 89, 89, 89, 89, 89, 89, 89,
90, 90, 90, 90, 90, 90, 90, 90, 90, 90, 91, 91, 92,
92, 92, 92, 93, 93, 94, 94, 94, 94, 95, 95, 95, 9
95, 96, 96, 97, 97, 97, 98, 99, 99, 99, 99, 100, 101
104, 105, 106, 108, 108, 113, 113, 115, 118, 120, 121, 122, 12
124, 128, 132, 134, 140, 151, 153 156, 164

Using a class interval of 70-79, 80-89, ... 160-16
Construct a frequency distribution table and Compute the mean.

Class Interval	Class Mark (x)	Tally	f	fx	cf
70 – 79	74.5	𝍸 I	6	447	6
80 – 89	84.5	𝍸 𝍸 𝍸 𝍸 𝍸	33	2788.5	39
90 – 99	94.5	𝍸 𝍸 𝍸 𝍸	37	3496.5	76
100 – 109	104.5	𝍸 II	7	731.5	83
110 – 119	114.5	IIII	4	458	87
120 – 129	124.5	𝍸 I	6	747	93
130 – 139	134.5	II	2	269	95
140 – 149	144.5	I	1	144.5	96
150 – 159	154.5	III	3	463.5	99
160 – 169	164.5	I	1	164.5	100
Σ			100	9710	

$$\bar{X} = \frac{\Sigma fX}{\Sigma f} = \frac{9710}{100} = 97.1$$

(2) MEDIAN

The median is a measure of central tendency which appears in the middle of an ordered sequence of values or observation. The median refers to the 50% point in a frequency distribution of a population. That is, half of the observations in a set of data are lower than the median value and half are greater than it. In other words, the median is the central point of the distribution. To compute the median, we must first arrange the data in ascending or descending order. If we have odd number of observations, the median is represented by the $(n+1)^{th}/2$ ordered observations.

On the other hand, if the number of observations in the data is even, the median is represented by the mean or average of the two middle values in the ordered data.

For a grouped data, the Median is given by;

$$\text{Median} = L_1 + \left(\frac{\frac{N}{2} - cf'}{fm}\right)c$$

Where

L_1 = Lower class boundary of the median class

N = total frequency

cf' = Cumulative frequency preceding the median class

fm = frequency of the median class

c = the class width or size of the median class interval

Examples

1. The data set, 6, 3, 4, 8, 8 has a median of 6. i-e 3, 4, (6) 8, 8

2. The set 5, 5, 9, 11, 15, 18 has a median $\frac{9+11}{2} = 10$.

(3) For the data on plasma glucose values, calculate the median.

Solution:

$N/2 = 100/2 = 50$,

The 50th position belongs to 90–99, hence the median class is 90–99.

$L_1 = 89.5$, $cf' = 39$, $fm = 37$, $c = 10$

(81)

$L_1 + \left(\frac{N}{2} - (f)\right)C$ / f_m

$\therefore$ Median $= 89.5 + \left(\frac{\frac{100}{2} - 39}{37}\right) \times 10$

$= 89.5 + \left(\frac{11}{37}\right) \times 10$

$= 89.5 + 2.97$

Median $= 92.47$

$L_1 + \left(\frac{N}{2} - (f)\right)C$ / f_m

(c) MODE

The mode of a set of numbers is that value which occurs with the greatest frequency; that is, it is the most common value. The mode may not exist, and even if it does exist, it may not be unique. A distribution having only one mode is called UNIMODAL DISTRIBUTION.

When a distribution is portrayed graphically, the mode is the peak in the graph. A distribution with two modes is called a Bimodal Distribution; while a distribution with three or more modes (peaks) is called a Multimodal Distribution. Such multimodal distributions are of interest to epidemiologists because they may indicate different causal mechanisms for biological phenomena, for example, in the age of onset of diseases such as tuberculosis meningococcal diseases, Ebola diseases etc.

For a grouped data, the modal class is the class that contains the highest frequency of cases. It can be computed using the formula;

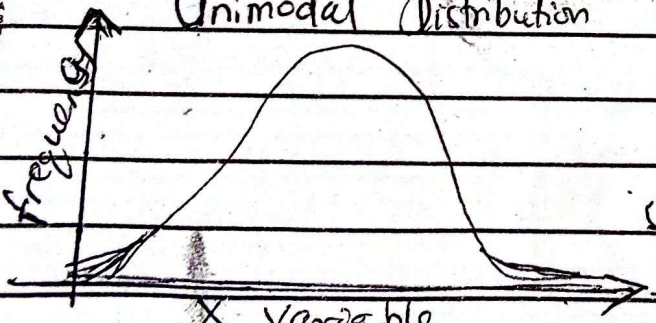
$$\text{MODE} = L_1 + \left(\frac{D_1}{D_1 + D_2}\right) C$$

L_1 = lower class boundary of the modal class,
D_1 = frequency of the modal class minus the frequency of the previous class (next-lower class)
D_2 = frequency of the modal class minus the frequency of the following class (next-higher class)
C = size or width of the modal class interval

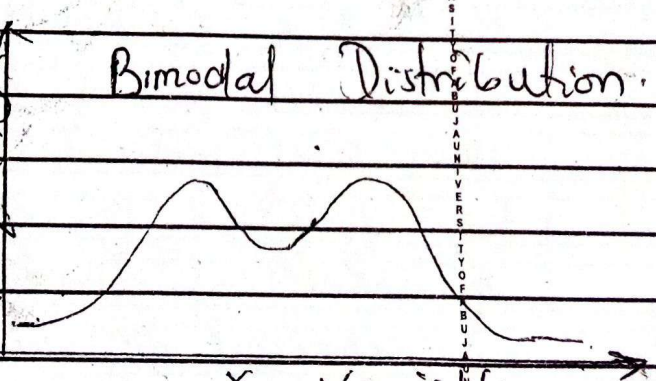
Example.
Using the previous data,
$L_1 = 89.5$,
$D_1 = 37 - 33 = 4$, $D_2 = 37 - 7 = 30$, $C = 10$

$$\text{Mode} = 89.5 + \left(\frac{4}{4+30}\right)10$$
$$\text{Mode} = 89.5 + (4/34) \times 10$$
$$= 89.5 + 1.18$$
$$\text{Mode} = 90.68$$

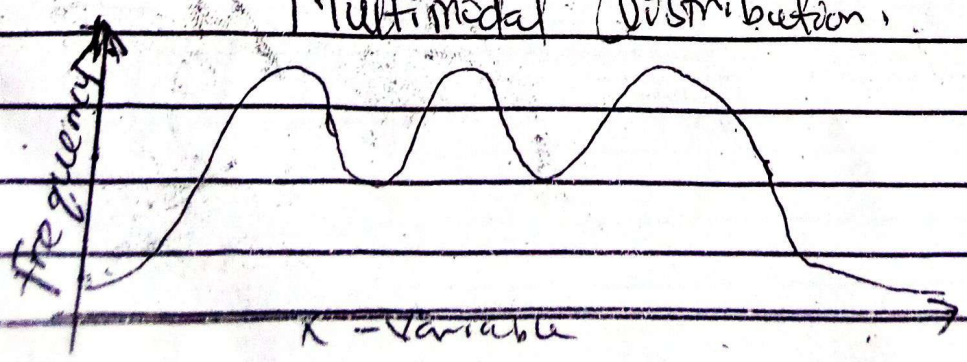
Unimodal Distribution



Bimodal Distribution.



Multimodal Distribution.



D THE GEOMETRIC MEAN (GM)

The geometric mean (GM) of a set of n positive numbers $x_1, x_2, \ldots, x_n$ is obtained by multiplying the set of values and then finding their nth root. All the values must be non-zero and greater than 1. The formula for the geometric mean is given by;

$$GM = \sqrt[n]{x_1 x_2 x_3 \ldots x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i} \quad \ldots (1)$$

Applying log transformation to (1),

$$\log GM = \frac{\sum_{i=1}^{n} \log x_i}{n} \quad \ldots (2)$$

However, for grouped data, the geometric mean is;

$$\log GM = \frac{\sum_{i=1}^{n} f_i \log x_i}{\sum_{i=1}^{n} f_i} = \frac{\sum f \log x}{N} \quad \ldots (3)$$

where $x_1, x_2, \ldots x_n$ represent the classmarks and $f_1, f_2, \ldots f_n$ are the corresponding frequencies.

A geometric mean is preferred to an arithmetic mean when there are extreme observations (very large numbers in the data set). These higher values would tend to inflate or distort an arithmetic mean.

For example, considering the data set, {8, 5, 4, 12, 15, 7, 28} The arithmetic mean ($\bar{x}$) = 11.3 due to the extreme observation 28 which is an inflated value not very close to any of the values in the data. However, the geometric mean (GM) is 9.3 which is closer to the data set.

a value less affected by the large measurement. Hence, in practice, when greatly differing values occur in a data set, as in some biomedical applications, the geometric mean becomes appropriate. To illustrate, a common use for the geometric mean is in determination whether the fecal coliform levels exceed a safe standard. The fecal coliform bacteria are used as an indicator of water pollution and unsafe swimming conditions at beaches. For instance, the standard may be set at a 30-day geometric mean of 100 fecal coliform units per 100 ml of water. When the water actually is tested, most of the individual tests may fall below 100 units. However, on a few days some of the values could be as high as 5000 units. Consequently, the arithmetic mean would be distorted by these extreme values. By using the geometric mean, one obtains an average that is closer to the average of the lower values. In addition, when data do not conform to a normal distribution or for very positively skewed distribution, the geometric mean is especially useful. A log transformation of the data will produce a symmetric distribution that is normally distributed.

(2)

EXAMPLE

For the data, 3, 5, 6, 6, 7, 10, and 12. Calculate the geometric mean.

$$GM = \sqrt[7]{3 \cdot 5 \cdot 6 \cdot 6 \cdot 7 \cdot 10 \cdot 12} = \sqrt[7]{453,600} = 6.43$$

OR

$$\log GM = \frac{1}{7}(\log 3 + \log 5 + \log 6 + \log 6 + \log 7 + \log 10 + \log 12)$$

$$\log GM = 0.8081 \; ; \quad GM = \text{antilog}(0.8081) = 6.43$$

$\frac{1}{H} = \frac{1}{n}\Sigma\frac{f}{x}$

$n = H\Sigma\frac{f}{x}$

$H = \frac{n}{\Sigma\frac{f}{x}}$

35

E HARMONIC MEAN (H)

The harmonic mean H of a set of n numbers $X_1, X_2, \ldots, X_n$ is the reciprocal of the arithmetic mean of the reciprocals of the numbers. It is given by;

$$H = \frac{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{X_i}}{\frac{1}{N}} = \frac{N}{\sum_{i=1}^{n}\frac{1}{X}}$$

For a grouped data

$$H = \frac{N}{\sum_{i=1}^{n}\frac{f}{X}}$$

EXAMPLE

For the data set, 3, 5, 6, 6, 7, 10, and 12. Find the harmonic mean.

$$H = \frac{n}{\sum_{i=1}^{n}\frac{1}{X_i}} = \frac{7}{\left(\frac{1}{3}+\frac{1}{5}+\frac{1}{6}+\frac{1}{6}+\frac{1}{7}+\frac{1}{10}+\frac{1}{12}\right)}$$

$$H = \frac{7}{(0.3333+0.2000+0.1667+0.1667+0.1429+0.1000+0.0833)}$$

$$H = \frac{7}{1.1929}$$

$$H = 5.87$$

The harmonic mean is not used commonly. But it is used when computing the mean dealing with rates of types x per d (eg kilometre per hour, gas per litre).

mean X median ① What is the Avg selling of houses in Abuja?

② What is the Average mean Nigerian worker?

Examples

Candidate's Number........ Page No........

Question No:........

Write on both sides of the paper

Do not write on either margin

Which Measures Should You Use?

Each measures of central tendency has strengths and weakness. The mode is difficult to use when a distribution has more than one mode, especially when these modes have the same frequencies.

The median is useful in describing a distribution that has extreme values at either end; for example, distribution of income and selling prices of houses. Since a few extreme values at the upper end will inflate the mean, the median will give a better picture of central tendency.

Finally, the mean is more useful for statistical inference than the mode or median. For example, the mean is useful in calculating the variance.

* The choice of a particular measure of central tendency depends on the shape of the population distribution. For normally distributed data, the arithmetic mean is the most appropriate measure of central tendency. However, if a log transformation creates normally distributed data, then the geometric mean is appropriate to the raw data.

Eg

₦21000, ₦10000, ₦2000, ₦2000, ₦2000, ₦2000, ₦1000, ₦1000, ₦1000, ₦1000, ₦1000

Rearranging

₦1000, ₦1000, ₦1000, ₦1000, ₦1000, ₦2000, ₦2000, ₦2000, ₦2000, ₦10,000, ₦21,000.

spread of the data
If data are concentrated near the mean, S.D will be small. We will use data values for our mean

Candidate's Number..... Page No. 37

Question No:.....

Write on both sides of the paper

Do not write on either margin

Do not write on either margin

MEASURES OF DISPERSION

Dispersion is the degree to which numerical data tend to spread about an average value. Various measures of dispersion (or variation) have been developed. The most common measures include the range, mean absolute deviation, Standard deviation, percentiles and Semi-interquartile range.

1 RANGE

The range is defined as the difference between the largest and smallest value in a distribution. When we have small number of values, one can easily identify the largest and smallest values. However, for a large data set, a simple way to identify these values is to sort the data set in ascending order. If X_n and X_1 denote the largest and smallest values respectively, then the range (R) is;

$$R = X_n - X_1$$

Example

For the data set, 200, 195, 225, 90, 140,

the range is; 90, 140, 195, 200, 225 (Sorted)

$$R = 225 - 90 = 135$$

2 MEAN ABSOLUTE DEVIATION

This measure involves first computing the mean of a set of data and then calculating the deviation of each observation from the mean. We then take the sum of the absolute values of each deviation and calculate their mean.

The formula for the mean absolute deviation of a set of n numbers $x_1, x_2, x_3, \ldots x_n$ is given by

$$\text{Mean Absolute Deviation (M.A.D)} = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

Where n is the number of observations in the data set, $\bar{x}$ is the mean of the numbers and $|x_i - \bar{x}|$ is the absolute value of the deviation of x_i from $\bar{x}$.

If the $x_1, x_2, \ldots x_n$ occur with frequencies $f_1, f_2, \ldots, f_n$, respectively, the mean absolute deviation is given by;

$$M.A.D = \frac{\sum_{i=1}^{n} f_i |x_i - \bar{x}|}{n} = \frac{\sum f |x - \bar{x}|}{\sum f}$$

Example

Find the mean deviation of the set; 2, 3, 6, 8, 11

x_i	$x - \bar{x}$	$\lvert x_i - \bar{x} \rvert$
2	-4	4
3	-3	3
6	0	0
8	2	2
11	5	5
$\sum x = 30$		$\sum \lvert x_i - \bar{x} \rvert = 14$

$n = 5$

$\bar{x} = 6$

$$\therefore M.A.D = \frac{14}{5} = 2.8$$

(2)

For the group data on weight of 100 Students

Weight (kg)	Class Mark (x)	$\lvert X-\bar{x}\rvert = \lvert X-67.45$	f	$f\lvert x-\bar{x}\rvert$
60 – 62	61	6.45	5	32.25
63 – 65	64	3.45	18	62.10
66 – 68	67	0.45	42	18.90
69 – 71	70	2.55	27	68.85
72 – 74	73	5.55	8	44.40
$\sum$			N=100	226.50

$$\therefore M.A.D = \frac{\sum f\lvert X-\bar{x}\rvert}{N} = \frac{226.50}{100} = 2.26 \text{ kg.}$$

(3) VARIANCE AND STANDARD DEVIATION

The variance and Standard deviation are useful for comparing data sets that are measured in the same units. A data set that has a "large" variance in comparison to one that has a "small" variance is more variable than the latter one. Instead of using the absolute value of the deviations about the mean, both the variance and standard deviation use squared deviations about the mean, defined for the ith observation as $(x_i - \bar{x})^2$.

The Standard deviation is the squared root of the variance. When it is necessary to distinguish the variance and standard deviation of a population from the variance and Standard

deviation of a Sample drawn from the population, we often use the Symbols σ^2 and σ (Sigma) for the population Variance and Standard deviation respectively, and S^2 and S for the Sample Variance and Standard deviation respectively.

For a set of n data sets $x_1, x_2, \ldots x_n$ (Ungrouped data), the Variance and Standard deviation are given by;

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

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

where n is the Sample size and $\bar{x}$ is the Sample mean.

For the grouped data, the formulas are given by

Variance: $S^2 = \frac{\sum_{i=1}^{n} f(x_i - \bar{x})^2}{n} = \frac{\sum_{i=1}^{n} f(x - \bar{x})^2}{n}$

Standard deviation: $S = \sqrt{\frac{\sum_{i=1}^{n} f(x_i - \bar{x})^2}{n}} = \sqrt{\frac{f(x - \bar{x})^2}{n}}$

Where $n = \Sigma f$

A shortcut formula that is computationally faster than the difference Score formula is given by;

EXAMPLES

(1) The blood cholesterol level of 10 randomly sampled individuals are 276, 304, 316, 188, 214, 252, 333, 271, 245 and 198. Calculate the variance and standard deviation.

	X	$X-\bar{X}$	$(X-\bar{X})^2$	X^2
1	276	16.3	265.69	76176
2	304	44.3	1962.49	92416
3	316	56.3	3169.69	99856
4	188	-71.7	5140.89	35344
5	214	-45.7	2088.49	45796
6	252	-7.7	59.29	63504
7	333	73.3	5372.89	110889
8	271	11.3	127.69	73441
9	245	-14.7	216.09	60025
10	198	-61.7	3806.89	39204
Σ	2597		22210.10	696,651

$$\bar{X} = \frac{\Sigma X}{n} = \frac{2597}{10} = 259.7$$

$$\text{Variance} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n} = \frac{22210.10}{10} = 2,221.01$$

$$S.D = \sqrt{\frac{\Sigma(X_i-\bar{X})^2}{n}} = \sqrt{2221.01} = 47.13$$

For Ungrouped data,

$$S^2 = \frac{\sum X^2}{n} - \bar{X}^2 = \frac{\sum X^2 - n\bar{X}^2}{n}$$

$$S = \sqrt{\frac{\sum X^2}{n} - \bar{X}^2} = \sqrt{\frac{\sum X^2 - n\bar{X}^2}{n}}$$

For grouped data,

$$S^2 = \frac{\sum fX^2}{n} - \bar{X}^2 = \frac{\sum fX^2 - n\bar{X}^2}{n}$$

$$S = \sqrt{\frac{\sum fX^2}{n} - \bar{X}^2} = \sqrt{\frac{\sum fX^2 - n\bar{X}^2}{n}}$$

where $n = \sum f$ (for the grouped data)

Note — Standard deviation & Variance

Sometimes the standard deviation of a sample is data are defined with $(n-1)$ replacing n in the denominator of the formula because the resulting values represent better estimates of the standard deviation and Variance of a population from which the sample is drawn. However, for large n ($n > 30$), there is practically no difference between the two definitions.

Candidate's Number.....

Question No: Q3 Page No.:.....

Write on both sides of the paper

Do not write on either margin

(2) For the group data on weight of 100 UNIABUJA students, Compute the variance and standard deviation.

Weight (kg)	Class Mark (X)	X^2	$X-\bar{X}$	$(X-\bar{X})^2$	f	$f(X-\bar{X})^2$	fX^2
60-62	61	3721	-6.45	41.6025	5	208.0125	18 605
63-65	64	4096	-3.45	11.9025	18	214.2450	73 728
66-68	67	4489	-0.45	0.2025	42	8.5050	188 538
69-71	70	4900	2.55	6.5025	27	175.5675	132 300
72-74	73	5329	5.55	30.8025	8	246.4200	42 632
$\sum$					100	852.7500	455,803

$$S^2 = \frac{\sum f(X-\bar{X})^2}{n} = \frac{852.7500}{100} = 8.5275$$

Standard deviation: $S = \sqrt{S^2} = \sqrt{8.5275} = 2.92\ kg$

OR

$$S^2 = \frac{\sum fX^2 - n\bar{X}^2}{n} = \frac{455,803 - (100)(67.45)^2}{100}$$

$$S^2 = \frac{455,803 - 454,950.25}{100} = \frac{852.75}{100} = 8.5275$$

$$S = \sqrt{S^2} = \sqrt{8.5275} = 2.92\ kg.$$

Alternatively, Since $n = 10$ is less than 30, we can use,

$$\text{Variance} = \frac{\sum(x_i - \bar{x})^2}{n-1} = \frac{22210.10}{10-1} = \frac{22210.0}{9}$$

$$\text{Variance} = 2467.79$$

$$S.D = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}} = \sqrt{2467.79} = 49.68$$

Using the Shortcut formula:

$$\text{Variance: } S^2 = \frac{\sum x^2 - n\bar{x}^2}{n} \text{ or } \frac{\sum x^2 - n\bar{x}^2}{n-1}$$

$\sum x^2 = 696651$, $n = 10$, $n-1 = 9$, $\bar{x} = 259.7$

$$\text{Variance: } S^2 = \frac{696651 - (10)(259.7)^2}{9}$$

$$S^2 = \frac{696651 - 674440.9}{9} = \frac{22210.1}{9}$$

$$S^2 = 2467.79$$

$$\text{Standard deviation} = \sqrt{S^2} = \sqrt{2467.79} = 49.68$$

Candidate's Number.....

Question No:..... Page No.:.....

Write on both sides of the paper

Do not write on either margin

(4) **COEFFICIENT OF VARIATION (CV)**

The Coefficient of Variation is defined as the ratio of the Standard deviation to the absolute value of the mean. The Coefficient of Variation is well defined for any variable that has a non zero mean. The purpose of the coefficient of Variation is to make comparisons between different distributions. For instance, if we want to see whether the distribution of the length of the tails of mice is similar to the distribution of the length of elephants' tail, we cannot meaningfully compare their actual Standard deviations. Since the Standard deviation of elephants' tails would be larger than that of the mice, simply because of the much larger measurement scale being used. However, these very differently sized animals might very well have Similar Coefficients of Variation with respect to their tail lengths.

Given a data set with Sample mean $\bar{X}$ and Standard deviation S, the Coefficient of Variation (CV) is given by

$$CV(\%) = \left(\frac{S}{\bar{X}}\right) \times 100$$

For the previous example, the CV is,

$$CV(\%) = \left(\frac{2.92}{67.45}\right) \times 100 = 4.33\%$$

Do not write on either margin

2021

Abuja.

Q_1, Q_2, Q_3 (4 equal p-ts)
$D_1 \ldots D_9$ (10 deciles)
$P_1 \ldots P_{99}$ (percentile.

Candidate's Number.....

Page No.: .....

Question No:.....

Write on both sides of the paper

Do not write on either margin

(2) A biologist has two types of rats, A and B. The rats have mean lifetimes of $\bar{X}_A = 1495$ days, and $\bar{X}_B = 1875$ days, and standard deviation of $S_A = 280$ days and $S_B = 310$ days. Which rat has the greater variability in their lifetimes?

$$CV_A(\%) = \left(\frac{S_A}{\bar{X}_A}\right) \times 100 = \frac{280}{1495} = 18.7\%$$

$$CV_B(\%) = \left(\frac{S_B}{\bar{X}_B}\right) \times 100 = \frac{310}{1875} = 16.5\%$$

Thus, rat A has the greater variability in lifetimes.

NOTE:

The standard deviation is useful in telling or reporting the research findings better.

Two basic rules are:

(1) 67% of the group is within 1 SD of the mean.

(2) 95% of the group is within 2 S.D of the mean.

For example, two towns (say, in Abuja and Lagos) could each have average household incomes of ₦100,000 a year. But if Lagos has S.D of ₦5000 and Abuja has S.D of ₦30,000, you have an additional important information about those two towns that can enrich your research.

(1) In Lagos 95% (nearly all households) earn between ₦90,000 and ₦110,000, ($\bar{X} \pm 2$ S.D), which signifies an incredibly homogeneous group of families

SHAPE OF FREQUENCY DISTRIBUTION

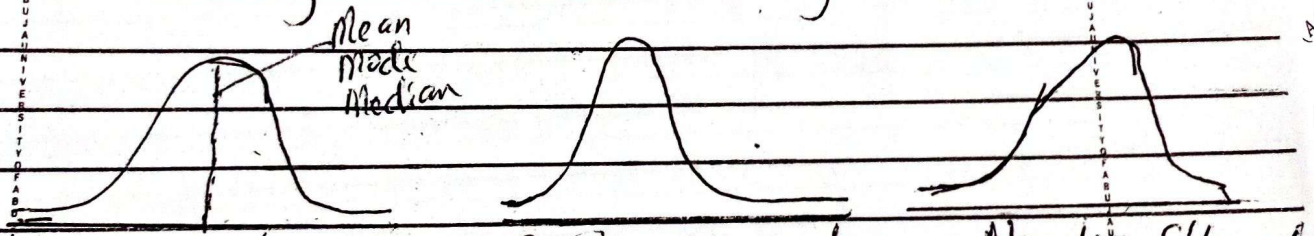
Skewness and Kurtosis are used to describe the Shape of a distribution.

(1) SKEWNESS: Skewness is a statistical measure for lack of Symmetry (asymmetry) of a distribution.

A distribution is positively skewed if the right tail is longer (positive skewness).

A distribution is negatively skewed if the left tail is longer (Negative Skewness).

A distribution with zero skewness is said to be Symmetrical or Normally distributed.



Symmetrical | Positive Skewed | Negative Skewed

Skewness can be measured by the Pearson Coefficient of Skewness given by;

$$SK = \frac{3(\bar{x} - \text{Median})}{s}$$

where $\bar{x}$ = mean, s = standard deviation.

Skewness can also be measured using the third Moment, and is given by;

$$SK = \frac{\sum f(x - \bar{x})^3}{s^3}$$

in the town.

(2) By ~~Past~~ contrast, in Abuja; 95% of the households earn ~~₦~~ between ₦40,000 and ₦160,000 a year. Hence, ~~expect to find~~ Substantial ~~wealthy~~ wealthy neighborhoods, and a Substantial low-income neighborhoods in ~~Abuja~~ Abuja — but not in Lagos.

The C.V of the two towns can also reveal the variability.

Adamu, O.J

Candidate's Number .....

Question No: Probability Page No: 19364

Write on both sides of the paper (1)

Introduction

In Statistics, sample data are used to learn or make inferences about the population. Since a sample is only a part of a population, any inferences we make from the sample involve uncertainty. If the uncertainty is too great, decisions suggested by sample data might be too risky to justify.

Our entire world is filled with uncertainty. We make decisions affected by uncertainty in our every day activities. For example, if the weather forecast says that "there is a 70% chance of rain", should we take an umbrella with us?

Often the words, chance and likelihood occur when we discuss uncertainty. Each is essentially a synonym for the word probability.

Probability is very important in inferential statistics and is generally regarded as the heart of inferential statistics.

SOME DEFINITIONS

Before defining what probability is, some preliminary definitions are required.

(i) Random Experiment – This is an experiment whose outcome is not certain, that is, the outcome is not perfectly predictable. Examples are; the tossing of a coin, the number of no-shows for a scheduled flight, the throwing of a dice, e.t.c.

(ii) OUTCOME – An Outcome is what we observe when we perform an experiment. For example, Suppose

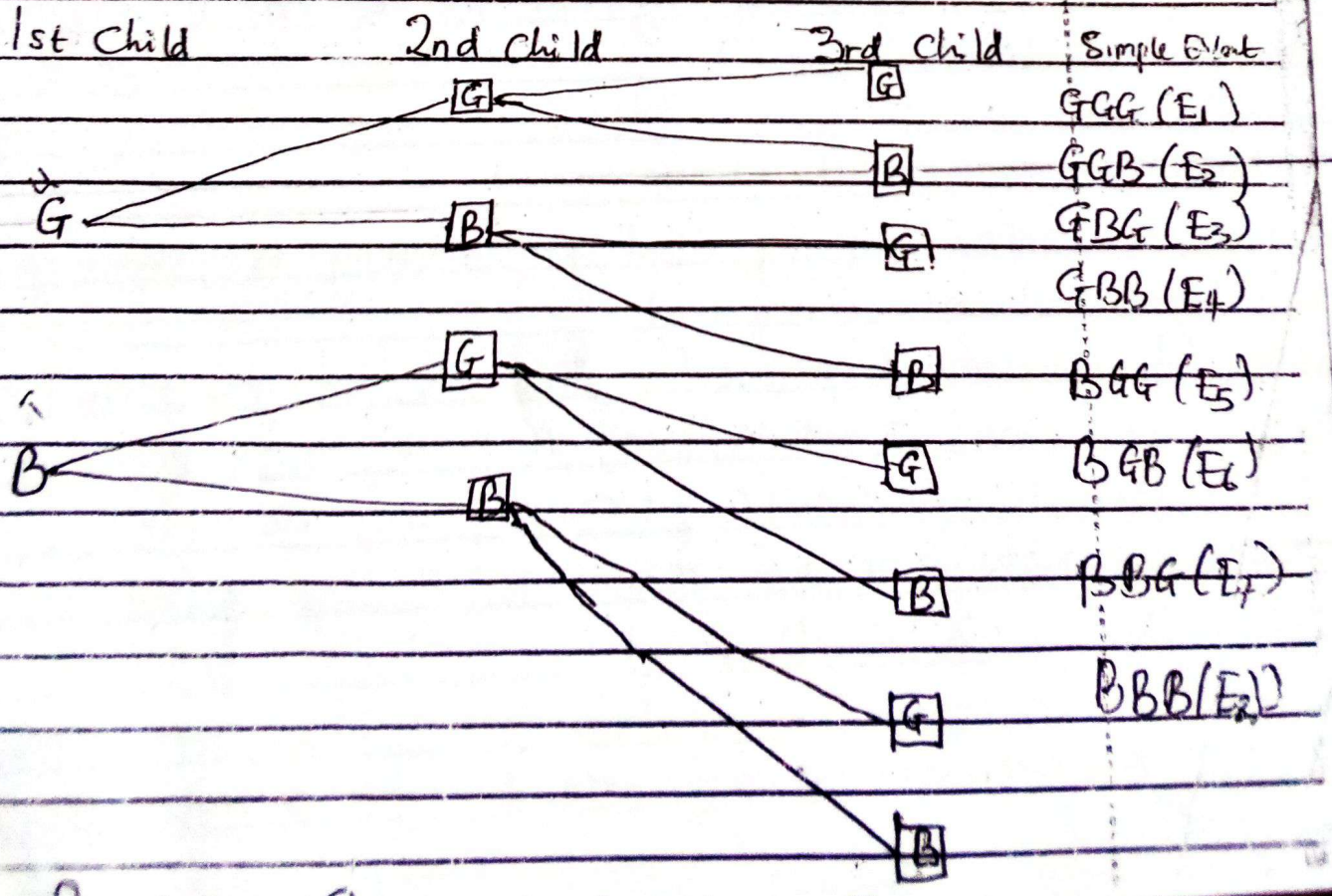
experiment is the toss of a coin. There are two possible outcomes; either we observe head or we observe tail.

(iii) SAMPLE SPACE:- The set of all possible outcomes of an experiment is called Sample space and it is normally denoted by S. The Sample space for a random experiment consists of the entire set of possible SIMPLE EVENTS. For example,

(a) When a coin is tossed, the Sample space consists of two simple events; $S = \{H, T\}$.

(b) In the case of rolling a dice, the Sample space is given by; $S = \{1, 2, 3, 4, 5, 6\}$

(c) Imagine an experiment involving couples with three children where we are interested in finding out the possible gender outcomes of the children. This is illustrated below.



POSSIBLE EVENTS

The Sample Space of this experiment consists of the eight simple events, i.e; S = {GGG, GGB, GBG, GBB, BGG, BGB, BBG, BBB}.

The following are some examples of events of this Sample Space and their simple events that share the common characteristics.

A: at least two girls (E_1, E_2, E_3, and E_5)
B: exactly two girls (E_2, E_3 and E_5)
C: All the same gender (E_1 and E_8)
D: One or fewer girls (E_4, E_6, E_7 and E_8)

(IV) SIMPLE EVENT – Any member of the Sample Space is called a simple event. For example, when a dice is rolled a simple event can be either a "1" or a "2", or a "3", e.t.c.

(V) COMPLEMENTARY EVENTS – Two events are said to be complementary if they contain no common simple events, and together they make up the entire Sample Space. With reference to the above example on gender combinations, events A and D are known as complementary events. Hence, A is the complement event to D, and D is the complement event to A. Also, when a dice is rolled, the simple events of even numbers (2, 4, 6) and odd numbers (1, 3, 5) are complementary events. Suppose A is any event of a Sample space S. Then the complement of event A, denoted by $\bar{A}$ or A', is the event containing all simple events in Sample Space that are not in A.

(vi) Mutually Exclusive Events.

Two events are mutually exclusive if the occurrence of one rules out the occurrence of the other - that is, they cannot both happen simultaneously. For example, the simple experiment of tossing a single coin results in two mutually exclusive outcomes - either a head or a tail can occur, but both cannot occur simultaneously. Also, the Nutritional status groups classified as Normal or Malnutrition is mutually exclusive. In addition, Sexes are mutually exclusive events. If a person is classified as female she cannot be a male, and vice versa.

(vii) INDEPENDENT EVENTS

Two events A and B are statistically independent if the outcome of event A does not influence the outcome of event B. An example is a repeated tosses of a coin. If a coin is tossed and it lands as head, this outcome does not influence the outcome of the next toss which can either be heads or tails with the same probability.

Alternatively, if the probability of event B is affected by the knowledge regarding the occurrence of A, then the events A and B are statistically dependent

Card drawing from a pack of cards without replacement demonstrates this condition.

Definition

Given the above preliminary definitions, let us now try to define what the word probability means. There are various definitions of probability. We shall consider three different definitions or interpretation

(1) CLASSICAL DEFINITION: This approach to probability applies when possible simple events are mutually exclusive and equally likely.

It states that, if the simple events of an experiment are equally likely, the probability of any event equals the proportion of simple events that satisfy that event out of the set of all possible simple events

For example, Consider the roll of a single die. If we assume that the die is perfectly balanced, then the six possible face values (1, 2, 3, 4, 5, 6) are equally likely to lie face up. Consequently, the probability of observing an even number when the die is rolled is 3/6.

Hence, If the sample space consists of N(S) equally likely events, then the probability of occurrence of the event A is given by:

$$P(A) = \frac{N(A)}{N(S)}$$

where N(A) is the number of times in which the event A occurs.

(2) RELATIVE FREQUENCY DEFINITION

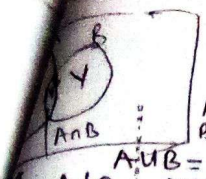
According to the frequentist view, the probability of an event can be defined as the proportion of the time (i.e relative frequency) with which the event occurs over an infinite number of repetitions of the experiment under identical conditions.

For example, if a fair die is rolled 60 times and the event "1" occurs 9 times, then its relative frequency is $9/60 = 0.15$. If we increased the number of times the die is rolled to a larger number, e.g. to 600, and we observe that the event "1" occurs 95 times, then its relative frequency is $95/600 = 0.1583$.

(3) EMPIRICAL PROBABILITY

The empirical probability is an approximate probability of a given event obtained by observing the relative frequency with which the event occurred over a finite number of repetitions of the experiment under nearly identical conditions. It is based on the empirical observation of experimental outcomes.

For example, consider the problem of determining the proportion of defective laptops produced by a Computer Company.



from Venn D.
$A = X + A \cap B$
$B = Y + A \cap B$
$A + B = X + Y + 2(A \cap B)$
By defn
$A \cup B = X + Y + (A \cap B)$
$\therefore A + B = A \cup B + (A \cap B)$
$(A \cup B) = A + B - A \cap B$
$n(A \cup B) = n(A) + n(B) - n(A \cap B)$
Divide be $n(S)$
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Properties

① $0 \leq P(A) \leq 1$
② $P(S) = 1$
③ $P(A') = 1 - P(A)$
$P(\phi) = 0$

Dr. ASanda

Candidate's Number
Question No:
Page No 19364 (9)
Write on both sides of the paper

PROBABILITY LAWS

No matter how one approaches probability (the classical, relative frequency or empirical interpretation), a set of fundamental rules must be satisfied.

(1) $0 \leq P(A) \leq 1$, where $P(A)$ is read "probability of event A". This rule states that all probabilities are numbers between 0 and 1, inclusively.

(2) $P(S) = 1$, where $P(S)$ is read "probability of the certain event". This rule states that, one of the simple events of the sample space is certain to occur when the experiment is performed.

(3) LAW OF ADDITIVITY: If two events, say A and B are mutually exclusive, then

$P(A \text{ or } B) = P(A) + P(B)$, & $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

This rule states that the probability of occurrence of any one of two or more mutually exclusive events is the sum of their individual probabilities.

Example 1

What is the probability that a card drawn at random from a well-shuffled standard pack will be either a Spade or a Club?

spades, Hearts, Diam. & Club

Solution

$N(Sp) = 13$, $N(C) = 13$, $N(S) = 52$

$P(Sp) = N(Sp)/N(S) = 13/52 = 1/4$

$P(C) = N(C)/N(S) = 13/52 = 1/4$

The two outcomes are mutually exclusive

$\therefore P(Sp \text{ or } C) = P(Sp) + P(C) = 1/4 + 1/4 = 1/2$

A(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K

Example 2

In a certain college with 800 students, there were 320 normals, 200 with anaemia, 160 with B-complex deficiency and 120 with Vitamin A deficiency. The data is presented in the table below;

S/NO	Clinical Status	No Examined
1	Normals	320
2	Vitamin A Def	120
3	B-Complex Def	160
4	Anaemia	200
	Total	800

(i) What is the probability of getting children either with Vitamin A deficiency or Vitamin B-Complex deficiency?

(ii) What is the probability that the children are likely to be normal?

Solution

(i) The events are mutually exclusive,

∴ The probability of Children with Vit A or Vit B Def is given by;

$$P(A \text{ or } B) = P(\text{Vit } A) + P(\text{Vit } B)$$

$$P(\text{Vit } A) = 120/800 = 0.15$$

$$P(\text{Vit } B) = 160/800 = 0.20$$

$$\therefore P(\text{Vit } A \text{ or Vit } B) = 0.15 + 0.20 = 0.35$$

(ii) $P(\text{Normal children}) = 320/800 = 0.40$

(c) Dr Asemota, O.J

(4) MULTIPLICATION LAW OF PROBABILITY

The probability that an independent event will occur jointly is the product of the probabilities of each event.

If A, B, and C are independent events, then the probability that A, B and C will occur is

$P(ABC) = P(A \cap B \cap C) = P(A) \times P(B) \times P(C)$

So also $P(AB) = P(A \cap B) = P(A) \times P(B)$

$P(Ac) = P(A \cap C) = P(A) \times P(C)$

$P(BC) = P(B \cap C) = P(B) \times P(C)$

More generally, $P(ABC \ldots N) = P(A) \times P(B) \times P(C) \ldots P(N)$

NOTE!

The law is applied to two or more events occuring together but they must NOT be associate. They must be independent of each other.

EXAMPLES

(1) If a die is thrown twice in succession, what is the probability of getting 5 and 2?

SOLUTION

$N(S) = 6, \quad N(5) = 1, \quad N(2) = 1$

$P(5) = P_1 = N(5)/N(S) = 1/6$

$P(2) = P_2 = N(2)/N(S) = 1/6$

$\therefore P(5,2) = P(5 \cap 2) = P(5) \times P(2) = 1/6 \times 1/6 = 1/36$

(2) The table below indicates the number of New borns by blood "Rh" factor and Sex.

Rh Factor	Male	Female	Total
+	45	45	90
−	5	5	10
Total	50	50	100

(i) What is the probability of newborn being female with Rh^+?

(ii) What is the probability of new born male being Rh^-?

SOLUTION

(i) $N(F) = 50$, $N(S) = 100$, $N(Rh^+) = 90$

$$P(F) = \frac{N(F)}{N(S)} = \frac{50}{100} = \frac{1}{2} = 0.5$$

$$P(Rh^+) = \frac{N(Rh^+)}{N(S)} = \frac{90}{100} = 9/10$$

Since Sex and Rh are independent of each other, the probability of new born being female with Rh^+ is,

$$P(F \text{ and } Rh^+) = P(F \cap Rh^+) = P(F) \times P(Rh^+) = \frac{1}{2} \times 9/10$$
$$= 9/20 = 0.45$$

(ii) $N(M) = 50$, $N(Rh^-) = 10$

$$P(M) = 50/100 = 1/2$$
$$P(Rh^-) = 10/100 = 1/10$$
$$P(M \text{ and } Rh^-) = P(M \cap Rh^-) = P(M) \times P(Rh^-)$$
$$= 1/2 \times 1/10 = 1/20 = 0.05$$

(3) There is a list of hospital admitted patients by age for one day. There were 1000 patients attended. OF all these patients, children below 5 years were 600; 200 were between the ages of 5 and 45 years and 200 were with age of 45 years and above. What is the probability that the population attending the hospital to be of

(i) Preschool Age

(ii) Preschool age and adults of 45 years and above

(iii) less than 45 years of age.

SOLUTION

(i) $N(PSA) = 600, \quad N(S) = 1000$

$$P(PSA) = \frac{N(PSA)}{N(S)} = \frac{600}{1000} = \frac{6}{10} = 0.6$$

(ii) $N(\text{Adults} \geq 45) = 200$

$$P(Ad \geq 45) = \frac{N(Ad \geq 45)}{N(S)} = \frac{200}{1000} = \frac{1}{5}$$

These two events are mutually exclusive; since you cannot fall within the two age brackets at the same time.

$$P(PSA \cup A \geq 45) = P(PSA) + P(A \geq 45) = \frac{6}{10} + \frac{2}{10} = \frac{8}{10}$$
$$= 0.8$$

(iii) less than 45 years of Age.

$$N(\text{Patients} < 45) = 600 + 200 = 800$$

$$\therefore P(\text{patients} < 45) = 800/1000 = 0.8$$

(5) The probability that an event A does not occur is often denoted by $P(\bar{A})$ or as $P(A')$ and is given by;

$$P(A') = 1 - P(A)$$

The law states that, the probability of the complement of an event equals 1 minus the probability of the event. This relationship is known as the probability rule for complementary events.

CONDITIONAL PROBABILITY

Conditional probability refers to the probability that an event B occurs given that an event A has occurred. This is denoted as, $P(B/A)$, i.e, the probability of B given A. This law is very useful in medicine and is popularly known as Baye's rule.

The Bayes' rule states that,

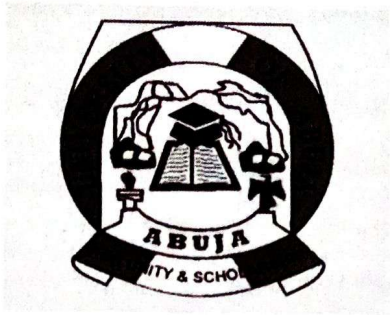
$$P(A/B) = \frac{P(B/A) \times P(A)}{P(B)}$$

Examples

(1) A review of records of a hosip hospital reve the following

Total number of patients ... 18,176

Number of patients reporting Complants of pains in abdomen, Vomiting and Constipation of long duration ... 530



STA 200A

Statistics for Biological Sciences

Dr. ASEMOTA, O. J
Department of Statistics
University of Abuja (UofA), Abuja.

CONDITIONAL PROBABILITY

Conditional probability refers to the probability that an event B occur given that an event A has occurred. This is denoted by P(B/A) i.e. the probability of B given that A. this law is very useful in medicine and is popularly known as Bayes' rule state that .

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

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

EXAMPLE

1. A review of records of a hospital reveal the following:

Total number of patients ---18,176;

Number of patients reporting complaint of pains in abdomen, vomiting and constipation of long duration ---- 530;

Number of patients diagnosed as abdominal TB cases ---- 41;

Number of abdominal TB case who had those 3 complaint ---- 32.

Using Bayes' rule, calculate P(Disease/Compliant) and Interpret the result.

Solution

Using Bayes' rule

$$P(Disease/Compliant) = \frac{P(Complaints/Disease) \times P(Disease)}{P(Complaints)}$$

$P(Disease/Compliant) = {}^{32}/_{41} = 0.7805$

$P(Disease) = {}^{41}/_{18176} = 0.002256$

$P(Compliant) = {}^{530}/_{18176} = 0.02916$

$$P(Disease/Compliant) = \frac{0.7805 \times 0.002256}{0.02916} = 0.06$$

By the presence of complaints in 32 out of the 41 cases of TB, a lay person can erroneously conclude that those complaints are good predictor of the disease.

However our result indicates that 6% P(0.06) of the cases with the complaint have that diseases. Thus the probability, P(Complaint/Disease) is entirely different from P(Disease /Complaints).

2. Consider a study to assess opinions about abortion (pro or against) and religion belief (religion versus non-religion) in young women. A random sample of 1000 women (aged 18 to 30) are selected and each woman is asked if she is in favour of or against abortion and if she is religious or not. The summary of the investigation is provided in the table below.

s/n	Against abortion	Pro abortion	Total
Religious	280	120	400
Non-religion 190	190	410	600
Total	470	530	1000

i. Using Bayes', rule calculate the conditional probability of being against abortion given that a woman is religious

ii. Is religion belief and opinion about abortion independent ?

Hints: if two events are independent $P(B/A) = P(A/B)$ = P(A) and $P(A \cap B) = P(A) \times P(B)$

Solution

i. Using Bayes' rules

$$P(Against\ Abortion/Religious) = \frac{P(Religious/Against\ Abortion) \times P(Against\ Abortion)}{P(Religous)}$$

$$P(Against\ Abortion/Religious) = \frac{280}{470}$$

$$P(Against\ Abortion) = \frac{470}{1000}$$

Therefore $P(Against\ Abortion/Religious) = \frac{\frac{280}{470} \times \frac{470}{1000}}{\frac{400}{1000}}$

$$= \frac{280}{400} = 0.7$$

ii. $P(Against\ Abortion) = \frac{470}{1000}$

$$P(Against\ Abortion/Religious) = \frac{280}{400}$$

Since $P(Against\ Abortion) \neq P(Against\ Abortion/Religious)$, it follows that opinnions about abortion and religious belief are NOT independent in this study.

FACTORIALS

Factorial is a special multiplication operator. The factorial sign "!" indicates a special repeated multiplication which is used frequently in statistical application.

Some examples:

$3! = 3.2.1 = 3 \times 2 \times 1 = 6$

$4! = 4.3.2.1 = 4 \times 3 \times 2 \times 1 = 24$

In general,

$n! = n \cdot (n-1) \cdot (n-2) \ldots 2 \cdot 1 = n \times (n-1) \times (n-2) \times \ldots \times 2 \times 1$

Factorial can be applied to know the number of ways n-objects/events can be ordered.

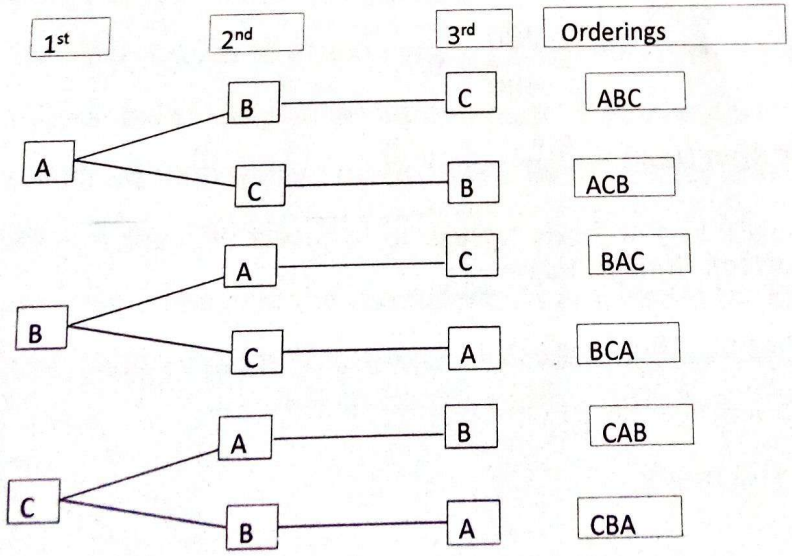
Note

$0! = 1$

$1! = 1$

Example

1. If we have 3 objects or events A, B, and C. how many different ways there are for ordering them?



Using Factorial,

$$n! = 3! = 3 \times 2 \times 1 = 6$$

The number of ways of arranging n-different objects is $= \frac{n!}{P! \cdot q! \cdot r!}$

For example:

How many different ways can letters of the word RUMOUR be arranged?

$$n = 6, \qquad n(R) = 2, \qquad n(U) = 2$$

$$= \frac{6!}{2! \cdot 2!} = 180$$

PERMUTATION

The word "permutation" simply means "ORDER OF ARRANGEMENT".

The number of permutations of r-object from a set of n-different objects is denoted by "n_{P_r}" and given by:

$$n_{P_r} = \frac{n!}{(n-r)!}$$

Example

Consider that there are 4 different objects, A,B,C and D and only 2 are selected at a time (i.e. AB, BA, AC, CA, AD, etc.). how many permutations are there for 2 objects chosen from a set of 4 different objects?

$$n_{P_r} = \frac{n!}{(n-r)!}$$

$$4_{P_2} = \frac{4!}{(4-2)!} = \frac{4!}{2!} = 12$$

COMBINATIONS

The concept "combination" does not deal with order and the concept "permutation" deals with order.

Consider the situation where there are 4 objects (A,B,C and D). we should select two of these at a time.

Unlike in permutations, we want to identify the number of all different pairs of objects irrespective of their ordering. The unordered arrangement of r-objects selected from a set of n-objects are called combinations.

The number of combinations of r objects selected from a set of n-objects is given by:

$$n_{C_r} = \frac{n!}{r!\,(n-r)!}$$

Example

i. The number of combinations of 2 objects taken from a set of 4 objects is:

$$4_{C_2} = \frac{4!}{2!\,(4-2)!} = \frac{4!}{2!\cdot 2!} = \frac{4\times 3\times 2\times 1}{(2\times 1)\times(2\times 1)} = 6$$

ii. There are 25 rats in a "rat colony", how many possible random samples of 20 rats can be taken from the colony?

Solution

Possible samples of 20 out of 25 is ;

$$25_{C_{20}} = \frac{25!}{5!\,(25-5)!} = \frac{25!}{5!\cdot 20!} = 53130$$

iii. In a class of fifteen short students obesity was detected. A medical examiner wanted to assess the incidence in all possible samples of three. How many samples of size three are there in these 15 students?

Solution

$$15_{C_3} = \frac{15!}{3!\,(15-3)!} = 455$$

Exercise

There are 20 albino rats of similar body weight which were given to a scientist. How many possible samples of 5 from 20 he could have?

DISTRIBUTIONS

RANDOM VARIABLE: A variable is considered to be random if each of its values is the result of a random observation.

A random variable is any numerical quantity whose values are determined by chance.

Random variables are often denoted by capital letters (usually X, Y, Z) while the observed values of a random variable X are denoted as x (lower case).

Suppose we are interested in the numbers of girls in a family of two children the simple events representing all possible gender combination for two children are BB, BG, GB and GG. In term of the variables numbers of girls the four sample event yield;

Simple Event	No of Girls
BB	0
BG	1
GB	1
GG	2

Thus, a **RANDOM VARIABLE** is a transformation of the simple events of a chance experiment into numerical quantities that represent all possible result for a phenomenon of interest

X: is the number of girls in the family of two children

X is a variable because it assumes numerical quantities as values. And the values X are subject to uncertainty; that is why X is referred to as a random variable.

DISCRETE RANDOM VARIABLE: A random variable is said to be discrete if the possible values of the random variable are countable. The following are examples of discrete random variables:

i. X_1: the number of patient visiting a hospital in a given day. The possible values of X_1 are 0, 1, 2, 3… up to some conceivably very large integer number .

ii. X_2: the number of telephone calls received by a business office in a given hour.

CONTINUSES RANDOM VARIABLE: If the possible values are uncountable, the random variable is said to be continuous. The following are examples of continuous random variables:

iii. X_3: The percentage increase (or decrease) in profit of a company when compared to last year. The possible value of X_3 can be negative (a decrease) or positive value (an increase) and are theoretically without $(-\infty < X_3 < \infty)$.

iv. X_4: The fill amount in a container. If we assume that the maximum amount possible for the container is 20kg, then the possible values of X_4 lie in interval 0 to 20 i.e.$(0 \leq X_4 \leq 20)$.

PROBABILITY DISTRIBUTION: The probability distribution of a random variable X is a representation of the probabilities for all values of that *X* can take on. It can be in form of a table, graph or mathematical expression it is denoted by *P(X)* or *F(X)* for discrete and continuous random variables respectively.

For example, let the random variable *X* represent the number of heads in two tosses of coin. We can represent the probability distribution in a tabular form as ;

$$X: Number\ of\ heads$$
$$S = \{HH, HT, TH, TT\}$$

X	0	1	2
P(X)	¼	2/4	¼

Probability distribution for the number of heads in two tosses of a coin

EXPECTED VALUE OF A RANDOM VARIABLE: The expected value of a random variable is it average value from a very large number of repeated experiment. It is denoted by E(X), it is the mean of the experiment. In other words the expected values is the long run average values of the random variables the concept of expected values is very useful as an aid to decision making.

Example

Suppose you are given a coin which you have up to 3 chances to tosses head. The game ends as soon as you toss heads or after the three attempt whichever comes first. If the first, second or third toss is head you receive $2, $4 or $8 respectively. However if you fail to toss heads in 3 attempt, you lose $20. Would you like to play this game? How might you objectively determine whether you should play the game?

Solution

One approach is to see how you would fare in the long run if you play the game many times that is, find the expected value, the long run average or loss,

Let *X* represents the amount we win or lose any times we play the game. The possible values of *X* are \$2, \$4, \$8 and -\$20. The probability of the first value is the same as the probability of tossing heads, which is ½. The probability of winning \$4 is the same as the probability of tossing tails first and then tossing heads. These two events are statistically independent so the probability of "tails and heads" is $(½)(½) = \frac{1}{4}$. By containing this approach we determine the remaining two probability to be 1/8 and 1/8 respectively. Thus, the possible value of the random variable *X* and their probabilities are as follows;

Values	Sequences	Probabilities
2	H	P(X=2)=1/2
4	T, H	P(X=4)=1/4
8	T, T, H	P(X=8)=1/8
-20	T, T, T	P(X=-20)=1/8

The expected value of *X* is given by

$$E(X) = \sum_X P(X)$$

$$E(X) = \left(\$2 \times \frac{1}{2}\right) + \left(\$4 \times \frac{1}{4}\right) + \left(\$8 \times \frac{1}{8}\right) + \left(-\$20 \times \frac{1}{8}\right) = \$0.50$$

This means that, if we play this game many times we expect to win \$0.50 per games on the average.

Note

In game of chance in which the random variable represents the amount won or lost the game is said to be **FAIR** if the expected value of the random variable is 0. If the expected value is positive, you are assured of to win in the long run. I f it is negative, you will lose in the long run

BINOMIAL DISTRIBUTION

The binomial distribution is one of the most useful discrete probability distributions. It describes the distribution of discrete events or qualitative data when an event can have only one or two possible outcome. For example male or female, yes or no, twins or not twins, positive or negative, survival or death. If the probably of occurrence remain constant, then the probability distribution of the possible out with repeated instance is observed by the binomial distribution.

This was developed by Professor James Bernoulli. In general let's agree to label an event's occurrence a "success" and its nonoccurrence a "failure". Thus "failure" is the complement of "success". Repetitions of a random experiment under identical condition is called *trials*. Suppose we perform experiment "n" times independently or under identical condition (that is, n independent trials). Since the sample space is precisely the same each time we perform the experiment, the probability success is the same for every trial.

If P is the probability of occurrence in one trial q is the probability of non-occurrence in the same trials ($q = 1 - p$) n is the number of trials and the random variable X is the number of success out of n independent trials. Then X has a binomial distribution with the probability function given by

$$P(X = x) = P(x) = n_{C_x} P^x q^{n-x}$$

$$= \frac{n!}{(n-x)!\,x!} P^x q^{n-x}; \qquad x = 0,1,2,\ldots,n; \quad 0 \leq P \leq 1$$

Mean

The mean of the binomial distribution of event is $E(x) = np$

Variance

$$Var(x) = npq = (np(1-p); \qquad since\ q = 1 - p$$

It follows, therefore that the standard deviation of x is $SD(x) = \sqrt{np(1-p)}$

CONDITIONS UNDER WHICH A RANDOM VARIABLE HAS A BINOMIAL DISTRIBUTION

i. The experiment consist of n identical trials
ii. The outcome of each trials is either success or failure
iii. The n trials are statistically independent
iv. The probability of success p is constant from trial to trial
v. The random variable X represent the number of success

EXAMPLES OF BINOMIAL DISTRIBUTION

1. It is known that 30 percent of a specific population are immune to polio virus. If a specific sample of size 10 is selected from this population what is the probability that it will contain exactly four immune persons?

Solution

$$\text{Probability of immune person to be } p = 30\% = 0.3$$

$$q = \text{probably of person to be not immune } = 1 - p = 0.7$$

$$n = 10$$

$$x = 4$$

the probability of obtaining $x = 4$ immune persons from $n = 10$ is

$$P(x) = n_{C_x} P^x q^{n-x}$$

$$= 10_{C_4}(0.3)^4(0.7)^{10-4} = 10_{C_4}(0.3)^4(0.7)^6$$

$$= 0.20$$

2. It is observed that 24 percent of a specific population have blood group B for a sample of size 20 drawn from this population find the probability that;

i. Exactly 3 persons with blood group B will be found

ii. Exactly 5 persons with blood group B will be found

iii. fewer than 3 persons with blood group B will be found.

Solution

a) $n = 20;\ p = 24\% = 0.24;\ q = 0.76;\ x = 3$

the probability of obtaining $x = 3$ *with blood group B from* $n = 20$ *is*

$$P(x) = n_{C_x} P^x q^{n-x}$$

$$P(3) = 20_{C_3}(0.24)^3(0.76)^{20-3}$$

$$= 20_{C_3}(0.24)^3(0.76)^{17}$$

$$= 0.1484$$

b) $n = 20;\ p = 24\% = 0.24;\ q = 0.76;\ x = 5$

The probability of obtaining $x = 5$ with blood group B from n=20 is

$$P(5) = 20_{C_5}(0.24)^5(0.76)^{20-5}$$

$$= 20_{C_5}(0.24)^5(0.76)^{15} = 0.2012$$

c) Few than 3 with blood group "B" include either 1 or 2 or both.

The probability of obtaining $x = 1$ from 20 is; $P(1) = 20_{C_1}(0.24)^1(0.76)^{19}$

The probability of obtaining $x = 2$ from 20 is; $P(2) = 20_{C_2}(0.24)^2(0.76)^{18}$

Therefore the probability of obtaining fewer than 3 with blood group B is;

$$20_{C_1}(0.24)^1(0.76)^{19} + 20_{C_2}(0.24)^2(0.76)^{18}$$

3. In a rural project 30% of children under 6 years of age were found with severe forms of under nutrition. If only 3 children were selected at random from the rural area what is the probability of 3, 2, 1 and not being severely undernourished?

Solution

P = probability of children to be undernourished 30% = 0.3
Q = the probability of children not to be undernourished = 1 – p = 1 = 0.3 = 0.7
n = 3;

i. The probability of all 3 children to be undernourished is x = 3;

$P(x) = n_{C_x} P^x q^{n-x}$

$= 3_{C_3}(0.3)^3(0.7)^{3-3} = 3_{C_3}(0.3)^3(0.7)^0 = 0.027$

$P(x = 3) = 0.027$ *or* 2.7%

ii. x=2

$P(x = 2) = 3_{C_2}(0.3)^2(0.7)^{3-2} = 3_{C_2}(0.3)^2(0.7)^1 = 0.189$

iii. x=1

$P(x = 1) = 3_{C_1}(0.3)^1(0.7)^{3-1} = 3_{C_2}(0.3)^1(0.7)^2 = 0.441$

Therefore $P(x = 1) = 0.441$ *or* 44.1%

The probability of all children to be normal is $x = 0$

$P(x = 0) = 3_{C_0}(0.3)^0(0.7)^{3-0} = 3_{C_0}(1)(0.7)^3 = 0.343$

Therefore $P(x = 0) = 0.343$ *or* 34.3%

4. If the mortality rate for a certain disease is 0.10 and 10 people are with that disease what is the probability that none will survive?

Solution

$n = 10, p = 0.10, q = 1 - p = 1 - 0.10 = 0.9$

$x = 0$ *(none will survive)* $= P(x = 0)$

$P(x = 0) = 10_{C_0}(0.1)^0(0.9)^{10} = (0.9)^{10} = 0.349$

Exercise

If the probability of recovery from a severe infection is 0.4 and 5 children are with the severe infection, find the probability that

(a) Five will recover
(b) Four will recover
(c) 3 or more will recover

Poisson Distribution

This distribution was developed by a French mathematician, Simeon Denis Poisson in 1837. This has been used extensively in the field of biology and medicine.

If x is the number of occurrences of some random event in an interval of time or space the probability that x will occur is given by

$$f(x) = \frac{e^{-\lambda}\lambda^{x}}{x!},\ x = 0,1,2,3,\ldots.$$

Where λ (lamba) is called the parameter of the distribution and is the mean number of occurrences of the random event in the interval and e in the constant equivalent to 2.7183.

Some Example of the Poisson Random Variable

i. The number of radioactive decays over a period of time
ii. The number of accident per day on Abuja to Kaduna road
iii. The number of still births per week in university of Abuja teaching hospital
iv. The number of telephone calls per hour
v. The number of cars arriving at the university of Abuja permanent site per hour
vi. The number of typographical errors per page in a book
vii. The number of particle emitted by a radioactive sources in a unit of tome
viii. The number of customer arriving at a service providing center per unit of time

The occurrences of the event are independent of the occurrence of the occurrence in an interval of time or space has no effect on the probability of a second occurrence of the event in the same per any other interval.

An interesting features of the Poisson distribution is the fact that the mean and variance are equal. $Mean = \lambda; Variance = \lambda$

Example on Poisson Distribution

i. Hospital records showed that on an average there are three emergency admission per day. Find the probability that exactly two emergency admission will occurs in a given day.

Solution

If x is a random variable denoting the number of day emergency admission, then x is a Poisson random variable with probability distribution;

$$f(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}, x = 0,1,2,3,\ldots;\ \lambda = 3, \qquad x = 2$$

$$= \frac{e^{-3}3^{2}}{2!} = 0.225$$

Therefore the probability that 2 emergency admissions will occur is equal to 0.225.

ii. What is the probability that no emergency will occur admission will occurrence a particular day?

$$\lambda = 3; \quad x = 0 \ (no\ emergency\ admission)$$
$$= \frac{e^{-3}3^0}{0!} = 0.05$$

iii. What is the probability of having 3 or 4 emergency cases on a particular day?

$$f(3or4) = f(3) + f(4)$$
$$f(3) = \frac{e^{-3}3^3}{3!} = 0.225$$
$$f(4) = \frac{e^{-3}3^4}{4!} = 0.1688$$
$$f(3or4) = 0.225 + 0.1688 = 0.3938$$

2. In a study of certain organism mean number of organism per sample was found to be 2. Assuming the number of organism to be of Poisson distribution. Find the probability that

i. The next sample taken will contain one or more organism
ii. The next sample taken will contain exactly 3 organism
iii. The next ample taken will contain fewer than 5 organism

Solution

i. One or more organism implies that $x = 1, 2, 3, 4 \ldots$

$$f(x \geq 1) = f(1) + f(2) + f(3) + f(4) + \ldots$$

This will be cumbersome to compute. However, since the set of one or more organ is the complement of the set of no organisms.

$$f(x \geq 1) = 1 - f(x = 0) = 1 - f(0); \ \lambda = 2, x = 0$$
$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!}$$
$$f(0) = \frac{e^{-2}2^0}{0!} = 0.135$$

Therefore $f(x \geq 1) = 1 - f(0) = 1 - 0.135 = 0.865$

ii. Sample taken will contain exactly 3 organism

$$f(x=3)=\frac{e^{-\lambda}\lambda^{x}}{x!}=\frac{e^{-2}2^{3}}{3!}=0.18$$

iii. Sample taken will contain fewer than 5 organism

$$f(x=0,1,2,3,4)$$
$$=f(x=0)+f(x=1)+f(x=2)+f(x=3)+f(x=4)$$
$$f(x=0)=0.135$$
$$f(x=1)=\frac{e^{-\lambda}\lambda^{x}}{x!}=\frac{e^{-2}2^{1}}{1!}=2\times e^{-2}$$
$$f(x=2)=\frac{e^{-\lambda}\lambda^{x}}{x!}=\frac{e^{-2}2^{2}}{2!}=2\times e^{-2}$$
$$f(x=3)=0.18$$

$$f(x=4)=\frac{e^{-\lambda}\lambda^{x}}{x!}=\frac{e^{-2}2^{4}}{4!}$$

Exercise: Complete the computations

3. Emergency hospital admission in USA with heart diseases are found to follow Poisson distribution an investigation show that there are four admission per day (mean). Find the probability that exactly 3 emergency cases will occur in a given day?

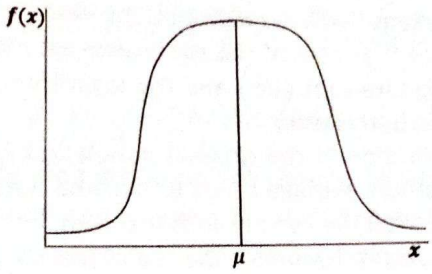
Solution

$$\lambda=4;\quad x=3$$
$$f(x)=\frac{e^{-\lambda}\lambda^{x}}{x!}$$
$$f(3)=\frac{e^{-4}2^{3}}{3!}=\frac{e^{-4}\times 32}{3}=0.1954$$

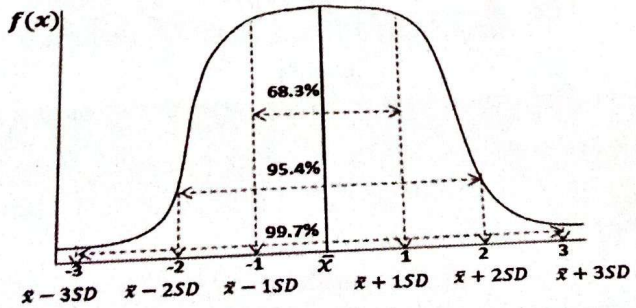
The Normal Distribution

The normal distribution is also known as the Gaussian mathematical Karl Friedrich Gauss the normal distribution is without doubt the most important and most widely used continues probability distribution.

The graphical appearance of the normal probability function is bell shaped curve. The curve signifies that the frequency is highest for middle value of the variable and the decline on both sides is similar (symmetrical).



The mean, median and mode value are same. The curve describe that ideal distribution of continuous values. In the ideal normal distribution, the value lying between the point '1' SD (standard deviation) below and '1' SD above the mean value (i.e. $mean \pm SD$) will include 68.27% of all values. The range $mean \pm 2\,SD$ (from $mean - 2SD$ to $mean + 2SD$) includes approximately 95% of values and the rang $mean \pm 3SD$ include about 99.7% of such values. This is depicted in the figure below.



Preference for the Use of Normal Distribution

The normal distribution has dominated statistical practice as well as the theory because of the following reasons:

i. The distribution of many variable are approximately normal such as height, age, weight, hemoglobin, PCV, body mass index, etc.;
ii. The normal distribution has been extensively and accurately tabulated. Hence the table is readily available;
iii. With measurement whose distributions are not normal a simple transformation of the scale of the measurement may induce approximate normality. The square-root ($\sqrt{x}$) and the logarithm ($\log x$) are often used as transformation in this way;
iv. Even if the distribution in the original population is far from normal, the distribution of sample average tends to become normal as the sample size increase. This is called the central limit theorem.

This is the most important reason for the use of normal distribution.

The Normal Probability Density Function

The probability density function of a continuous random variable does not provide the probability that the random variable takes on a particular value, that probability is in fact 0. Instead, we find that probability that a random variable lies in a given interval by finding the area under the density function and within the boundaries of that interval.

If a continuous random variable X is normally distributed then its distribution or probability density function is give by

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{1}{-2\sigma^2}(x-\mu)^2}; \quad -\infty \leq x \leq \infty$$

Or

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\left\{\frac{(x-\mu)^2}{2\sigma^2}\right\}}; \quad -\infty \leq x \leq \infty$$

Where π and e are constants (i.e. $\pi = 3.1416, e = 2.7183$); μ and σ^2 are the parameters of the distribution (i.e. the values which define the bell shape of the normal distribution.

$$Mean = \mu$$

Variance: The variance of a normally distributed random variable x is given by;

$$Var(x) = \sigma^2$$

THE STANDARD NORMAL DISTRIBUTION

When the mean is subtracted from each observation and this difference is divided by the standard deviation of the observation we obtain a new standardized variable called STANDARD NORMAL VARIATE and is denoted by the symbol Z. It measures how such an observation is bigger or smaller than mean in units of standard deviation. The standard normal variate Z is formally define as;

$$Z = \frac{Individual\ Observation - Mean}{Standard\ deviation} = \frac{x - \mu}{\sigma}\ or$$

$$Z = \frac{x - \bar{x}}{S}(for\ sample\ data)$$

Where μ and $\bar{x}$ denote the population and sample mean respectively; σ and S represent the population and sample standard deviation respectively. When this standardization is applied to the values of a normally distributed continuous random variable X, we obtain a new continuous random variable Z which is also normally distributed with $\mu = 0$ and $\sigma^2 = 1$. this distribution is called the standard normal distribution.

Hence

$$Mean\ of\ Z\ =\ 0$$
$$Var\ (z)\ =\ 1$$

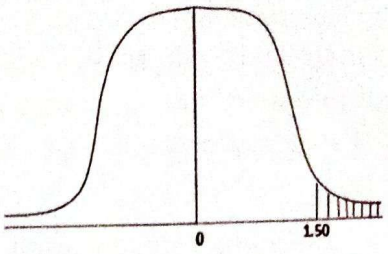
To find the proportion of the individuals who will exceed any particular observation (x) in a standard population, we have to refer to the Z value in the table of the standard normal distribution.

Examples

1. Calculate the percentage of individuals exceeding the observation $(X = calories)$ for the given values of Z in the table below.

s/n	Values of Z	Proportion of indicual exceeding X	Percentage of individual exceeding X	Percentage of indivual not exceeding X
1	1.50	0.0668	6.68	93.32
2	2.00	0.0228	2.28	97.72
3	2.575	0.0050	0.50	99.50
4	3.00	0.0013	0.13	99.87

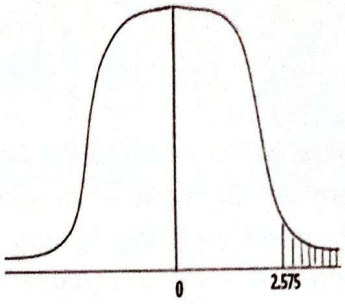
1. $0.500 - 0.4332 = 0.0668$



2. Since 2.575 is b/w 2.57 and 2.58

$$Z(2.57) = 0.4949$$
$$Z(2.58) = 0.4951$$
$$Z(2.575) = \frac{0.4949 + 0.4951}{2}$$
$$= 0.4950$$

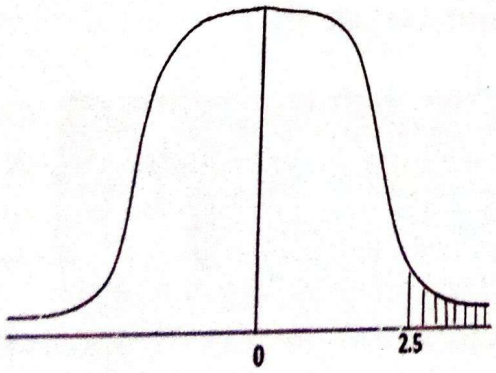
The shaded area: $0.5000 - 0.4950 = 0.0050$



2. Pulse rate of healthy male adults follow normal distribution with a mean 75/ minutes and a SD of 4/minute. Find out the percentage of individuals having pulse rate of 85/minutes?

Solution

$$x = 85, \bar{x} = 75, SD = 4$$
$$Z = \frac{x - \bar{x}}{SD} = \frac{85 - 75}{4} = 2.5$$



$$Z(2.5) = 0.4938$$

Shaded Area

$$= 0.5000 - 0.5000 - 0.4938$$
$$= 0.0062$$

The percentage of individual having pulse rate beyond 85 minutes is 0.62% while the percentage falling below 85minutes is $1-0.0062 = 0.9938 =$ 99.38%

3. The weight of 500 students follows a normal distribution. If mean weight is 60kg and the SD is 5kg

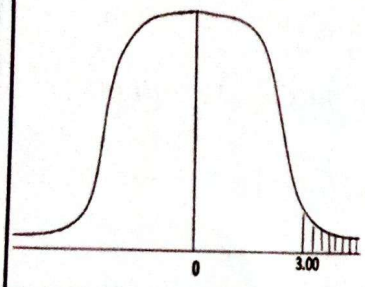
i. What are the chance of weight above 75kg

ii. What parentage of students will have weight above 68kg?

iii. How many of the students will have weight between 68 and 75kg

Solution

i. $Z = \frac{x-\bar{x}}{SD} = \frac{15-60}{5} = 3.00$



$Z(3.00) = 0.4987$

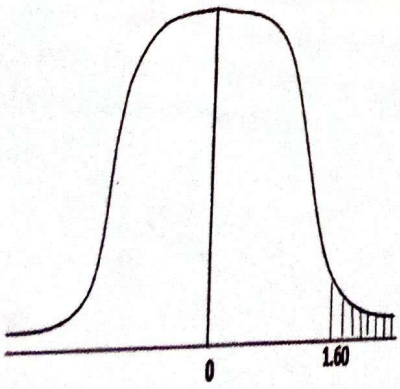
Therefore the $shaded\ area = 0.5000 - 0.4987$

$= 0.0013$

Hence, 0.13% of the students will have weight above 75kg

ii. $X = 68, \bar{x} = 60, SD = 5$

$$Z = \frac{x-\bar{x}}{SD} = \frac{68-60}{5} = \frac{8}{5} = 1.60$$



$Z(1.60) = 0.4452$

The $shaded\ area = 0.5000 - 0.4452$

$= 0.0548$

Therefore 5.48% of the students will be above 68kg