OUTLINE

Statistical Data: types, sources and methods of collection

Presentation of data: table, charts and graphs

Error and approximation

Frequency and Cumulative distribution

Measures of location, partition and dispersion

Skewness and kurtosis

Rates, ratio and index numbers

STATISTICS

Statistics can be defined as the scientific procedures involved in collecting, organizing, summarizing, presenting, and analyzing data to draw valid conclusions and make informed decisions. Essentially, statistical methods entail identifying, gathering, and arranging raw facts into meaningful data, followed by summarizing, presenting, and analyzing that data to generate useful and reliable information. Therefore, statistics is the science of collecting, organizing, analyzing, presenting, and interpreting data to support sound decision-making under conditions of uncertainty.

The science of data, known as statistics, can be broadly divided into three interconnected components: descriptive statistics, statistical methods, and statistical inference.

DESCRIPTIVE STATISTICS

Descriptive statistics involves summarizing and providing a clear account of numerical information through reports, charts, and diagrams. Its primary goal is to extract meaningful insights from collected data. The process begins with data gathering either through counting or measurement followed by summarizing key aspects such as measures of central tendency and variability. Appropriate graphs, diagrams, and charts are then used to enhance understanding and support accurate interpretation of the phenomenon under study, while maintaining awareness of the source and milieu of the data.

STATISTICAL METHOD

A statistical method is a tool used to classify data and clarify the relationships between the variables under consideration. This is accomplished through the application of various statistical tools and formulas. These methods range from computing simple data summaries such as the mean, median, and mode to employing advanced modeling techniques used in research, forecasting, and policy formulation.

INFERENTIAL STATISTICS

Inferential statistics involves making deductive statements about a population based on information obtained from a representative sample. It is the process of drawing conclusions or generalizing about a population under specified conditions and assumptions. This branch of statistics encompasses key procedures such as parameter estimation and hypothesis testing, which allow researchers to make informed judgments about population characteristics.

Use of Statistics

- Statistics helps in providing a better understanding and exact description of a phenomenon of nature.
- Statistics helps in the proper and efficient planning of a statistical inquiry in any field of study.
- Statistics helps in collecting appropriate quantitative data.
- Statistics helps in presenting complex data in a suitable tabular, diagrammatic and graphic form for easy and clear comprehension of the data.
- Statistics helps in understanding the nature and pattern of variability of a phenomenon through quantitative observations.
- Statistics helps in drawing valid inferences, along with a measure of their reliability about the population parameters from the sample data.

STATISTICAL DATA

Statistical data are obtained through objective measurement or enumeration of characteristics using precise, unbiased, and reliable instruments. When such data are subjected to statistical analysis, they yield results with a high degree of accuracy and precision.

Data can be described as a collection of unprocessed information obtained through the measurement or counting of a characteristic or phenomenon. These raw facts, when expressed in numerical form, are referred to as **quantitative data**. For example, the ages of students enrolled in STA 111 in a particular session represent quantitative data. When the information is expressed in non-numerical form, it is referred to as **qualitative data**, such as status, sex, religion, or other categorical attributes.

SOURCES OF STATISTICAL DATA

Primary Data: These are data obtained firsthand or collected directly from respondents through methods such as personal interviews, questionnaires, measurements, or observations.

Primary data can be obtained from:

- **Census:** Complete enumeration of all the unit of the population
- **Surveys:** The study of representative part of a population
- **Experimentation:** Observation from experiment carried out in laboratories and research center.

Administrative process e.g. Record of births and deaths.

ADVANTAGES

I. Comprises of actual data needed
II. It is more reliable with clarity
III. Comprises a more detail information

DISADVANTAGES

I. Cost of data collection is high
II. Time consuming
III. There may larger range of non response

Secondary Data: These are data obtained from existing sources such as publications, newspapers, journals, and annual reports. They are typically summarized information originally collected for purposes other than the researcher’s current study. Secondary data can be sourced from:

- Government publications and statistical bulletins
- Newspapers and magazines
- Research journals and academic articles
- Annual reports of organizations and institutions
- Online databases and digital repositories
- Textbooks and reference materials
- International agencies such as the World Bank, IMF, or UN

ADVANTAGES

I. The outcome is timely
II. The information gathered more quickly
III. It is less expensive to gather.

DISADVANTAGES

I. Most time information are suppressed when working with secondary data
II. The information may not be reliable

METHODS OF COLLECTION OF DATA

There are several methods available for collecting data, and the choice of method depends on the nature of the problem and the type of data required. Some common methods include:

DIRECT OBSERVATION

This method is commonly used in scientific investigations where data are obtained through direct observation, often within controlled experimental settings. It is applied more frequently in the natural sciences, especially during laboratory experiments, than in the social sciences. However, it remains highly useful in studying small communities, institutions, and social groups where behaviours and interactions can be directly observed.

Advantages

- Data collected is devoid of exaggeration.
- It provides reliable result especially when the survey is being supervised by experienced enumerators.
- Supervision by a superior officer is relatively easy during survey.

Disadvantages

- It is laborious.
- The method is time consuming.
- It is not suitable for large groups of respondents.
- It is very expensive

INTERVIEWING

In this method, the person collecting the data, known as the interviewer, asks the respondent (the interviewee) direct questions to obtain the required information. The interviewer meets the interviewee in person and gathers responses verbally. This personal interaction distinguishes the interview method from the questionnaire method, where respondents typically provide written answers without face-to-face contact.

Advantages

- It is best suited to situations where the problems cannot be completely understood by the targeted audience.
- This method is suitable in social anthropological research where the questions cannot be formulated beforehand and one question leads to another.
- It is equally useful in situations where great depth in study is required.
- It provides reliable and genuine response especially when experienced interviewers are involved.

Disadvantages

- The method is time consuming
- It is not suitable for large groups of respondents.
- It is very expensive.
- Errors may be committed while recording responses of respondents during interview.

QUESTIONNAIRE

A questionnaire is a structured set of questions or statements designed to obtain information on one or more variables of interest. Respondents are required to provide answers to these questions or statements, either in written form or through selected response options. Questionnaires may be administered personally by the researcher, distributed to participants, or sent through mail or other delivery methods. Both the interview and questionnaire methods are widely used in the social sciences, where human populations are typically the focus of study.

Advantages

- It is a suitable method where respondents are illiterate.
- The enumerators can see to it that only relevant answers are obtained from the respondents.
- The method ensures great reliability as the accuracy can be checked by supplementary questions.
- The method is very useful in collecting information on exceptional difficult items on a prepared questionnaire.

Disadvantages

- The method can be expensive especially when training of enumerators is involved.
- Errors may be committed during the recording of respondents' responses by the enumerators.

MAILED QUESTIONNAIRE

In this method, a detailed questionnaire is prepared and sent to the targeted respondents for completion and return to the investigator. This approach works well in areas with reliable postal services and typically includes a prepaid postage stamp so respondents can return the questionnaire at no cost. However, with the widespread availability of the internet, traditional postal questionnaires have largely been replaced by online methods. Today, investigators often distribute structured questionnaires electronically, allowing respondents to easily access and submit their responses via the internet.

Advantages

- The costs of running this method are relatively less.
- It allows for consultation on the part of respondents to enunciate valid opinion.
- It is a very useful method of getting information that cannot be sourced conveniently from respondent on face to face with the interviewers.
- It is convenient to use.

Disadvantages

- One of the drawbacks of this method is that it is very difficult to design a questionnaire that can be understood by all and fill appropriately. Often the questions intent is misunderstood leading to inaccurate or even irrelevant responses.
- The method increases the effort a respondent has to put in filling the questionnaire as a result of too many detailed instructions, and this reduces the percentage of response.
- Another fault of this method is that, there is no way of ensuring that questionnaire are filled and returned. This often results in very poor percentage of return of which those responding may not form a fair sample of the population.
- There may not be enough motivation with the targeted respondents and which may cast serious doubts on the validity of responses.
- Respondents may lack the knowledge of the required facts.

TELEPHONE INTERVIEW

This method has become increasingly prominent in data collection worldwide. With the rising number of mobile phone users, enumerators can easily contact identified respondents through phone calls and record their responses. Telephone interviews provide a fast and efficient means of gathering information, especially when face-to-face contact is not feasible.

Advantages

- The method is not expensive to manage, since the only cost involved is that of telephone.

- It is practically easy to get a wider coverage using this method, since majority of the populace, both educated and illiterates now uses mobile phone.
- The time taken to obtain responses is usually fast, especially when the targeted respondents are in the right frame of mind.
- It is possible for the investigator to monitor the recruited interviewers for efficiency.
- It is the most convenient method of data collection. The investigator can carry out the interview at his/her comfort zone at any convenient time provided it agrees with that of the respondents.

Disadvantages

- It may be difficult to get elaborate answers to questions asked through telephone.
- Responses received is strongly dependent on the mood of the respondents.
- Occasional delays may occur in getting responses, especially in developing countries where network problem is still an issue of concern.

PRESENTATION OF DATA

When raw data are collected, they are organized by grouping them into classes or categories to determine the number of observations that fall within each group. In many cases, it becomes necessary to present the data using tables, charts, and diagrams to enhance understanding and to clearly illustrate the relationships among the variables under investigation.

FREQUENCY TABLE

A frequency table is a structured arrangement of data into different classes or categories, along with the number of observations (frequencies) in each class.

Procedure for forming frequency distribution

Given a set of observation $x_1, x_2, x_3, ..., x_n$, for a single variable.

- Determine the range (R) = L – S where L = largest observation in the raw data; and S = smallest observation in the raw data.
- Determine the appropriate number of classes or groups (K). The choice of K is arbitrary but as a general rule, it should be a number (integer) between 5 and 20 depending on the size of the data given. There are several suggested guide lines aimed at helping one decided on how many class intervals to employ.

Two of such methods are:

(a) $K = 1 + 3.322(\log_{10} N)$

(b) $K = \sqrt{N}$ where n is the number of observations

- Determine the width (W) of the class interval. It is determined as $W = \dfrac{R}{K}$
- Determine the numbers of observations falling into each class interval i.e. find the class frequencies.

SOME BASIC DEFINITIONS

Variable: A variable is a characteristic or attribute of a population that can assume different values. Variables are generally classified into two main types: continuous variables and discrete variables.

Continuous Variable

A continuous variable is a type of variable that can take any value within a given range. Its values are typically obtained through measurement e.g. height, weight, volume, time, exam score, temperature etc.

Discrete Variable

A discrete variable is a type of variable that changes in distinct steps and typically takes integer values. Its values are usually obtained by counting e.g. number of cars, number of chairs, number of students in a class, number of books.

Class Interval

A class interval is a subdivision of the total range of values that a (continuous) variable can take. It represents a group or category of values within the data e.g. 0-9, 10-19, 20-29, etc. There are three main types of class intervals: Exclusive Class Interval, Inclusive Class Interval and Open-end Class Interval

Exclusive Method

In the exclusive method of classification, class intervals are arranged so that the upper limit of one class is the lower limit of the next class. This method ensures continuity of data without overlap between classes.

Example

Consider the expenditures of some families:

0-1000, 1000-2000, 2000-3000, etc.

In this system, a family with an expenditure of 0 to 999.99 falls into the first class (0-1000), while a family with an expenditure of exactly 1000 belongs to the next class (1000-2000). This approach maintains clear separation between class intervals while covering the entire range of data.

Inclusive Method

In the inclusive method of classification, class intervals are arranged to avoid overlap, and both the lower and upper limits are included in each class. This method is typically used for discrete variables that take only integer values, such as the number of family members, number of workers in a factory, or number of cars.

The inclusive method is not suitable for continuous variables like age, height, or weight, where fractional values may occur; in such cases, the exclusive method is preferred.

Open-End Classes

In open-end class intervals, a class limit is missing either at the lower end of the first class, the upper end of the last class, or both. Open-end classes are often used in practical situations, particularly in economic or medical data, where there are a few extremely high or low values that are far removed from the majority of observations. This approach allows for the inclusion of these outlying values without specifying exact limits.

Class limit: are the boundary values that define a class interval, consisting of the lower and upper class limits. When a class interval does not specify either its lower or upper boundary, it is referred to as an open class interval. Examples include “less than 25” and “25 and above”.

Class boundaries: are the dividing points that separate one class interval from the next. They mark the exact limits between adjacent classes. For example, the class boundaries for the interval 10–19 are 9.5 and 19.5.

Cumulative frequency: is obtained by adding the frequency of a given class to the total of all frequencies in the preceding classes.

Example 1

Thirty students were given ten multiple choice questions to answer and the number of questions they failed were recorded below:

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

Solution

x	Tally	Frequency(f)
0	III	3
1	II	2
2	III	3
3	III	3
4	IIII	4
5	~~IIII~~	5
6	IIII	4
7	II	2
8	III	3
9	I	1
Total		30

The scores of the test are called the variables (x) and are gotten by arranging the scores in either ascending or descending order. The number of time each variable occurred is called the frequency.

Example 2

The following are the marks of 50 students in STA 111:

48 70 60 47 51 55 59 63 68 63 47 53 72 53 67 62 64 70 57 56 48 51 58 63 65 62 49 64 53 59 63 50 61 67 72 56 64 66 49 52 62 71 58 53 63 69 59 64 73 56.

(a) Construct a frequency table for the above data.
(b) Answer the following questions using the table obtained:
(i) How many students scored between 51 and 62?
(ii) How many students scored above 50?
(iii) What is the probability that a student selected at random from the class will score less than 63?

Solution

(a) Range (R) (Highest-Lowest) = $73-47=26$

No of classes (K) = $\sqrt{n}=\sqrt{50}=7.07\approx 7$

Class size = $W=\frac{R}{K}=\frac{26}{7}=3.7\approx 4$

Frequency Table

Mark	Tally	Frequency
47-50	~~IIII~~II	7
51-54	~~IIII~~II	7
55-58	~~IIII~~II	7
59-62	~~IIII~~III	8
63-66	~~IIII~~ ~~III~~I	11
67-70	~~III~~ I	6
71-74	IIII	4
		$\sum f=50$

(b) (i) 22 (ii) 43 (iii) $\frac{29}{50}=0.58$

Example 2

The following is the record of marks obtained by 50 students in statistics test:

38	29	14	38	29	35	48	33	28	32	26	37
46	19	45	47	11	26	38	42	18	10	19	32
44	26	40	16	33	46	31	47	23	47	17	33
27	51	43	24	31	37	41	50	46	27	32	20
37	35										

Construct a group frequency table using Sturge’s formula for the above data.

Solution

Range (R) (Highest-Lowest) = $53-10=43$
No of classes (K) = $1+3.322\log_{10} N$

$$K = 1 + 3.322\log_{10} 50$$
$$= 1 + 3.322 \times 1.69897 = 1 + 5.64397$$
$$= 6.643 \approx 7$$

Class size $= W = \frac{R}{K} = \frac{43}{7} = 6.14 \approx 6$

Class Interval	**Tally**	**Frequency**
10-15	III	3
16-21	~~IIII~~ I	6
22-27	~~IIII~~ II	7
28-33	~~IIII~~ ~~IIII~~ I	11
34-39	~~IIII~~ III	8
40-45	~~IIII~~ I	6
46-51	~~IIII~~ IIII	9
Total		**50**

Example 3
The following data represent the ages (in years) of people living in a housing estate in Abeokuta.

18 31 30 6 16 17 18 43 2 8 32 33 9 18 33 19 21 13 13 14 14 6 52 45 61 23 26 15 14 15 14 27 36 19 37 11 12 11 20 12 39 20 40 69 63 29 64 27 15 28.

Present the above data in a frequency table showing the following columns; class interval, class boundary, class mark (mid-point), tally, frequency and cumulative frequency in that order.

Solution
Range (R) (Highest-Lowest) = $69 - 2 = 67$
No of classes (K) = $\sqrt{n} = \sqrt{50} = 7.07 \approx 7$
Class size $= W = \frac{R}{K} = \frac{67}{7} = 9.5 \approx 10$

Class interval	Class boundary	Class Mark	Tally	Frequency	Cumulative Frequency
2-11	1.5-11.5	6.5	~~IIII~~ II	7	7
12-21	11.5-21.5	16.5	~~IIII~~ ~~IIII~~ ~~IIII~~ ~~IIII~~ I	21	28
22-31	21.5-31.5	26.5	~~IIII~~ III	8	36
32-41	31.5-41.5	36.5	~~IIII~~ II	7	43
42-51	41.5-51.5	46.5	II	2	45
52-61	51.5-61.5	56.5	II	2	47
62-71	61.5-71.5	66.5	III	3	50

Exercise
Below are the data of weights of 40students women randomly selected in Ogun state. Prepare a table showing the following columns; class interval, frequency, class boundary, class mark, and cumulative frequency.

96 84 75 80 64 105 87 62 105 101 108 106 110 64 105 117 103 76 93 75 110 88 97 69 94 117 99 114 88 60 98 77 96 96 91 73 82 81 91 84

Use your table to answer the following question

i. How many women weight between 71 and 90?

ii. How many women weight more than 80?

iii. What is the probability that a woman selected at random from Ogun state would weight more than 90?

GRAPHICAL PRESENTATION OF DATA

Graphs provide a visual representation of the relationship between variables. In statistics, various types of graphs and charts are used depending on the nature of the data and the purpose of the analysis. The importance of diagrams and graphs in data presentation cannot be overstated. They make information easier to understand and interpret, helping readers quickly grasp the significance of the data.

In the social and management sciences such as business, economics, and accounting modern methods of presenting financial information, including profit-and-loss accounts and balance sheets, now often use graphical displays. These visual formats allow shareholders and stakeholders, whether or not they have accounting expertise, to easily understand financial performance.

Likewise, disciplines like political science, sociology, geography, and education rely on diagrams and charts such as bar charts and pictographs to present information on issues like voter statistics, crime rates, monthly rainfall, or school enrollment.

Overall, the purpose of these diagrams and graphs is to present data in a clear, attractive, and engaging manner, giving readers a broad and accessible understanding of the information displayed.

BAR CHART

A bar chart is a statistical graph that displays data using rectangular bars whose lengths or heights are proportional to the values they represent. Each bar is separated from the next by equal spaces.

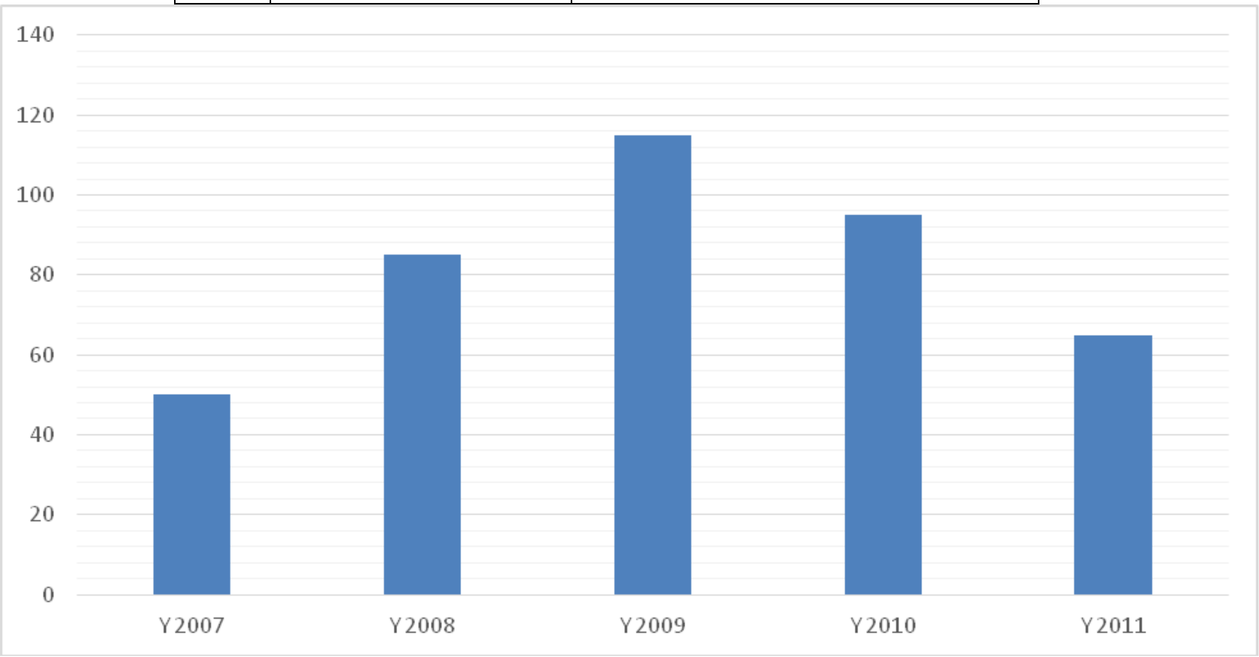
Example 1

The following data is the record of NCE mathematics students for 5 years. Draw a bar-chart to represent the information.

Solution

S/N	YEARS	NUMBER OF STUDENTS
1	2007	50
2	2008	85

3	2009	115
4	2010	95
5	2011	65
	TOTAL	**400**



Bar chart showing the number of enrolment in the NCE mathematics department.

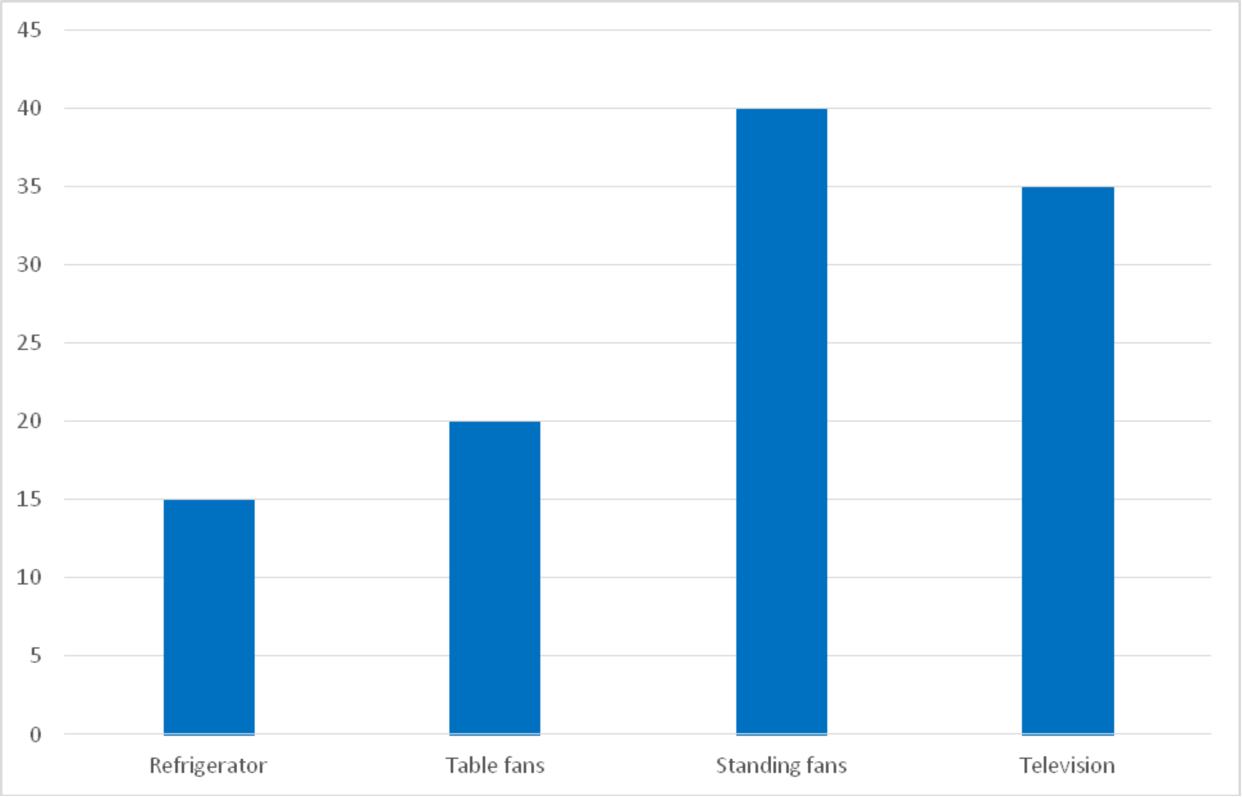
Example 2

A company produces the following items on daily basis

S/N	ITEMS	QUANTITY
1	Refrigerator	15
2	Table fans	20
3	Standing fans	40
4	Television	35
	TOTAL	**110**

Draw a bar-chart to represent the information.

Solution



PIE CHART

A pie chart is a circular diagram divided into sections, called sectors, where each sector's central angle is proportional to the frequency of the item it represents. It visually displays the entire dataset as a single circle, with each portion showing its share of the whole.

shows the totality of the data being represented using a simple circle.

Example

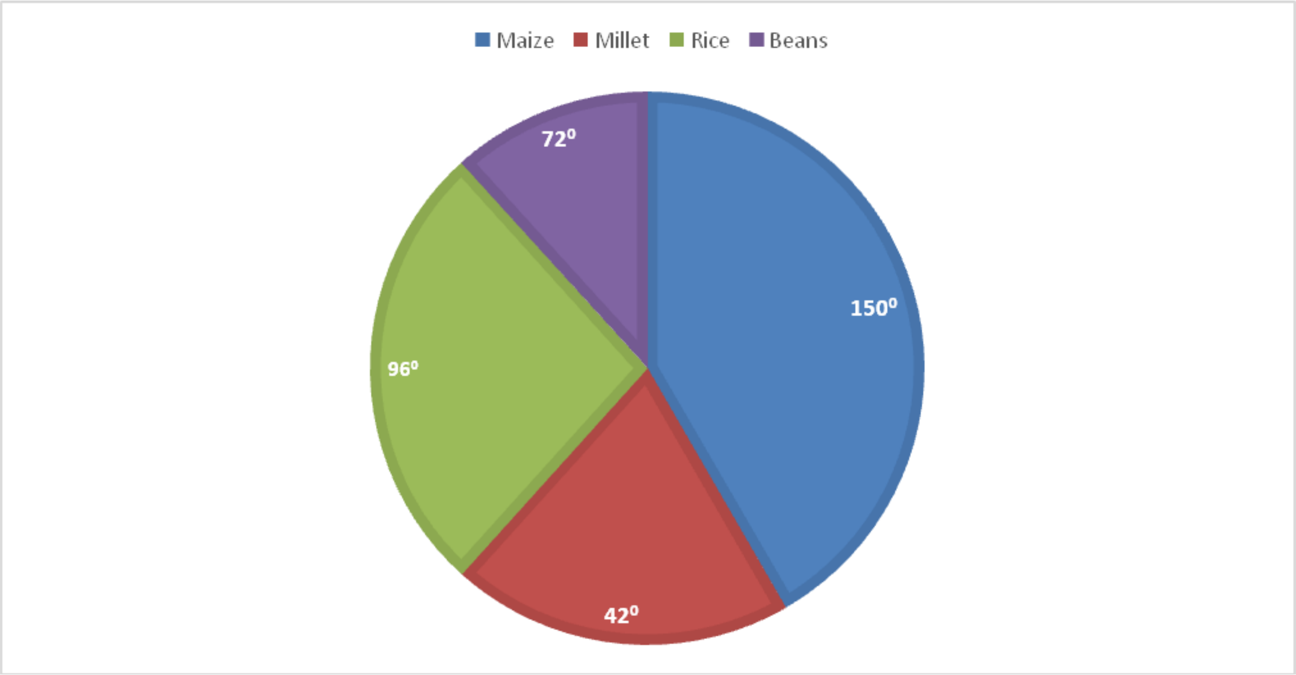
The record of a farm product (in bags) of a peasant farmer is given in the table below:

Item	Maize	Millet	Rice	Beans
Quantity	25 bags	12 bags	16 bags	7 Bags

Draw a pie-chart to represent the above information

Solution:

Items	Quantity	Sectorial angles
Maize	25	$\frac{25}{60} \times 360 = 150^0$
Millet	12	$\frac{12}{60} \times 360 = 72^0$
Rice	16	$\frac{16}{60} \times 360 = 96^0$
Beans	7	$\frac{7}{60} \times 360 = 42^0$
TOTAL	**60**	$\mathbf{360^0}$


Maize
Millet
Rice
Beans
72⁰
150⁰
96⁰
42⁰