Introduction to Statistics

Rasheed A. Adeyemi*,PhD[1]

[1]Department of Statistics
Federal University of Technology, Minna

February 27, 2020

Abstract

These notes were prepared for the module on introduction to Statistics which forms part of the STA117 Undergraduate course at FUT Minna. Real-life data applications and tutorials as well as theoretical practicals are provided.

Recommended textbooks:

- Les Underhill & Dave Bradfield. **INTROSTAT**, Lecture Note on Statistics, University of Cape Town
- Bayo Lawal Applied Statistical Methods in Agriculture, Health and Life Sciences, Publisher: Springer

1 Course Overview

SESSION:	2019/2020
SEMESTER:	First Semester
COURSE	TITLE/CODE: Introduction to Statistics I (STA117)
Names	Offices and Contact of Course Lecturers
Name:	Dr. R.A. Adeyemi
Contact No. :	080xxxxx
Email:	rashid.adeyemi@futminna.edu.ng
Office:	Department of Statistics
	Board Room, Department of Statistics, (Bosso Campus)
	OR Room 329, 3rd Floor, SAAT Complex (GK)
GROUPS	SEET/ SICT and SIPET (1000 Lecture Hall GK Campus)
Name:	Dr. Abubakar Usman
Office:	HOD, Department of Statistics
	Bosso Campus
Contact No.	:080xxxxx
GROUP	SPS and SLS (DH Lecture Hall Bosso Campus)
Name:	Mr. Abdullah. Usman
Contact No. :	080xxxxxxx
Office:	Department of Statistics
	Bosso Campus
GROUP	SPS and SLS (DH Lecture Hall Bosso Campus)
Name:	Dr. A. D. Obaromi
Contact No. :	080xxxxxxx
Office:	Department of Statistics
	Bosso Campus
GROUP	SAAT & SET (LTA SET Faculty)

*Corresponding author: adeyemira@yahoo.ca

Week	Topics
1.	Scope of Statistics, definition, terms, statistical data, data types, scale of measurements, data collection (observations, experiments), survey sampling techniques
2	Presentation of Data : tables, Bar Charts: simple , multiple, component bar charts, Histogram, polygons and graphs
3	Errors and approximation
4	Frequency : relative frequency, cumulative frequency distributions,
5	Measure of location
6	Measure of location continues
7	Measure of dispersion
8	First Assessment
9	Skewness and Kurtosis
10	Rates , Ratios
11	index numbers and weight quantity
12	Rates , Ratio and index numbers
13	Revisions
14	Second Assessment
15	Final Examination

2 COURSEWARE OUTLINE

- **Prerequisite Requirement:** Higher Grade Credit C or Standard Grade A for O'Level Mathematics or NCE Mathematics
- **Objective of the Course:** This course is designed to introduce undergraduate students to a wide range of statistical techniques required for the analysis of quantitative data. The course covers topics Descriptive statistical methods. Measures of central tendency and dispersion. Permutations and Combinations. Basic probability concepts. Discrete random variables and their properties: Bernoulli, Binomial, Poisson, Hypergeometric. Normal distributions. Point and interval estimation. Correlation and simple linear regression. Hypothesis tests for proportions, means and variances.
- **Outcome:** At the end of the course, students are expected to have a better understanding of data classification and appropriate statistical tools for the analysis of quantitative data
- **Lecture Delivery:** Lectures shall mostly be delivered by PowerPoint presentations and class discussions
- **Evaluation :** Two tests (40%); 2-3 hours e-Exam (60%).
- **DP Requirement:** 40% Class mark (CA), 75% attendance at lectures.

3 Definition and Scope

1

Chapter 1 – Terminology

1.1 Definitions

Data/Data set – Set of values collected or obtained when gathering information on some issue of interest.

Examples

1) The monthly sales of a certain vehicle collected over a period.

2) The number of passengers using a certain airline on various routes.

3) Rating (on a scale from 1 to 5) of a new product by customers.

4) The yields of a certain crop obtained after applying different types of fertilizer.

Statistics – Collection of methods for planning experiments, obtaining data, and then organizing, summarizing, presenting, analyzing, interpreting the data and drawing conclusions from it.

Statistics in the above sense refers to the methodology used in drawing meaningful information from a data set. This use of the term should not be confused with **statistics** (referring to a set of **numerical values**) or **statistics** (referring to measures of **description** obtained from a data set).

Descriptive Statistics – Collection, organization, summarization and presentation of data.
To be discussed in chapter 2.

Population – All subjects possessing a common characteristic that is being studied.

Examples

1) The population of people inhabiting a certain country.

2) The collection of all cars of a certain type manufactured during a particular month.

3) All patients in a certain area suffering from AIDS.

4) Exam marks obtained by all students studying a certain statistics course.

Census – A study where every member (element) of the population is included.

2

Examples

Ratio scale – Level of measurement where differences and ratios are meaningful and there is a natural zero. This is the "highest" level of measurement in terms of possible operations that can be performed on the data.

Examples

Variables like height, weight, mark (in test) and speed are ratio variables. These variables have a natural zero and ratios make sense when doing calculations e.g. a weight of 80 kilograms is twice as heavy as one of 40 kilograms.

Summary of 4 measurement scales

Measurement scale	**examples**	**Meaningful calculations**
Nominal	Types of music University faculties Vehicle makes	Put into categories
Ordinal	Motion picture ratings: G- General audiences PG-Parental guidance PG-13 – Parents cautioned R - Restricted NC 17 – No under 17	Put into categories Put into order
Interval	Years: 2009,2010, 2011 Months: 1,2, . . . , 12	Put into categories Put into order Differences between values are meaningfull
Ratio	rainfall humidity income	Put into categories Put into order Differences between values are meaningfull Ratios are meaningfull

Experiment – The process of observing some phenomenon that occurs.

An experiment can be **observational** or **designed**.

1) A designed experiment can be controlled to a certain extent by the experimenter. Consider a study of 4 fuel additives on the reduction in oxides of nitrogen. You may have 4 drivers and 4 cars at your disposal. You are not particularly interested in any effects of particular cars or drivers on the resultant oxide reduction. However, you do not want the results for the fuel additives to be influenced by the driver or car. An appropriate design of the experiment (way of performing the experiment) will allow you to estimate effects of all factors of interest without these outside factors influencing the results.

2) An observational study is not controlled by the experimenter. The characteristic of interest is simply observed and the results recorded. For example

   2.1) Collecting data that compares reckless driving of female and male drivers.
   2.2) Collecting data on smoking and lung cancer.

Parameter – Characteristic or measure of description obtained from a population.

Examples

1) Mean (average) age of all employees working at a certain company.

2) The proportion of registered female voters in a certain country.

Statistic – Characteristic or measure of description obtained from a sample.

Examples

1) The mean (average) monthly salary of 50 selected employees in a certain government department.

2) The proportion of smokers in a sample of 60 university students.

1.2 Sampling methods

When selecting a sample, the main objective is to ensure that it is as representative as possible of the population it is drawn from. When a sample fails to achieve this objective, it is said to be **biased**.

Sampling frame (synonyms: "sample frame", "survey frame") – This is the actual set of units from which a sample is drawn

Example

Consider a survey aimed at establishing the number of potential customers for a new service in a certain city. The research team has drawn 1000 numbers at random from a telephone directory for the city, made 200 calls each day from Monday to Friday from 8am to 5pm and asked some questions.

In this example, the *population* of interest is all the inhabitants in the city. The *sampling frame* includes only those city dwellers that satisfy all the following conditions:

1) They have a telephone.

2) The telephone number is included in the directory.

3) They are likely to be at home from 8am to 5pm from Monday to Friday;

4) They are not people who refuse to answer telephone surveys.

The *sampling frame* in this case definitely differs from the *population*. For example, it under-represents the categories which either have no telephone (e.g. the most poor), have an unlisted number, and who were not at home at the time of calls (e.g. employed people), who don't like to participate in telephone interviews (e.g. more busy and active people). Such differences between the sampling frame and the population of interest is a main cause of bias when drawing conclusions based on the sample.

Probability samples – Samples drawn according to the laws of chance. These include simple random sampling, systematic sampling and stratified random sampling.

Simple random sampling – Sampling in which each sample of a given size that can be drawn will have the same chance of being drawn. Most of the theory in statistical inference is based on random sampling being used.

Examples

1) The 6 winning numbers (drawn from 49 numbers) in a Lotto draw. Each potential sample of 6 winning numbers has the same chance of being drawn.

2) Each name in a telephone directory could be numbered sequentially. If the sample size was to include 2 000 people, then 2 000 numbers could be randomly generated by computer or numbers could be picked out of a hat. These numbers could then be matched to names in the telephone directory, thereby providing a list of 2 000 people.

A random sample can be selected by using a table of random numbers.

Example

Suppose the first 6 random numbers in the table of random numbers are:
10480, 22368, 24130, 42167, 37570, 77921.
Use these numbers to select the 6 wining numbers in a Lotto draw.

The 49 numbers from which the draw is made all involve 2 digits i.e. 01, 02, . . . , 49.
Putting the above numbers from the table of random numbers next to each other in a string of digits gives: 10 48 02 23 68 24 13 04 21 67 37 57 07 79 21 .

The winning numbers can be selected by either taking all pairs of digits between 01 and 49 (discarding any numbers outside this range or repeats) by working from left to right or right to left in the above string.

By working from left to right the winning numbers are: 10, 48, 2, 23, 24 and 13.
By working from right to left the winning numbers are: 21, 7, 37, 21, 4 and 13.

The advantage of simple random sampling is that it is simple and easy to apply when small populations are involved. However, because every person or item in a population has to be listed before the corresponding random numbers can be read, this method is very cumbersome to use for large populations and cannot be used if no list of the population items is available. It can also be very time consuming to try and locate every person included in the sample. There is also a possibility that some of the persons in the sample cannot be contacted at all.

Systematic sampling – Sampling in which data is obtained by selecting every *k*th object, where k is approximately $\frac{N}{n}$.

Examples

1) A manufacturer might decide to select every 20th item on a production line to test for defects and quality. This technique requires the first item to be selected at random as a starting point for testing and, thereafter, every 20th item is chosen.

2) A market researcher might select every 10th person who enters a particular store, after selecting a person at random as a starting point; or interview occupants of every 5th house in a street, after selecting a house at random as a starting point.

3) A systematic sample of 500 students is to be selected from a university with an enrolled population of 10 000. In this case the population size N=10 000 and the sample size n=500. Then every $\frac{10000}{500} = 20^{th}$ student will be included in the sample. The first student in the sample can be randomly selected from an alphabetical list of students and thereafter every 20^{th} student can be selected until 500 names have been obtained.

Stratified random sampling – Sampling in which the population is divided into groups (called strata) according to some characteristic. Each of these strata is then sampled using random sampling.

A general problem with random sampling is that you could, by chance, miss out a particular group in the sample. However, if you subdivide the population into groups, and sample from each group, you can make sure the sample is representative. Some examples of strata commonly used are those according to province, age and gender. Other strata may be according to religion, academic ability or marital status.

Example

In a study investigating the expenditure pattern of consumers, they were divided into low, medium and high income groups.

Income group	**percentage of population**
low	40
medium	45
high	15

A stratified sample of 500 consumers is to be selected for this study.

When sampling is proportional to size (an income group comprises the same percentage of the sample as of the population) the sample sizes for the strata should be calculated as follows.

$$\text{low}: \frac{40*500}{100} = 200, \quad \text{medium}: \frac{45*500}{100} = 225, \quad \text{high}: \frac{15*500}{100} = 75.$$

Convenience Sampling – Sampling in which data that is readily available is used e.g. surveys done on the internet. These include quota sampling.

Quota sampling – Quota sampling is performed in 4 stages.

a) *Stage 1*: Decide which characteristics of the elements/individuals in the population to be sampled are of importance.
b) *Stage 2*: Decide on the categories to be sampled from. These categories are determined by cross-classification according to the characteristics chosen at stage 1.
c) *Stage 3*: Decide on the overall number (quota) and numbers (sub-quotas) to be sampled from each of the categories specified in step 2.
d) *Stage 4*: Collect the information required until all the numbers (quotas) are obtained.

Example

A company is marketing a new product and needs to know how potential customers might react to the product.

Stage 1: It is decided that age (the 3 groups under 20, 20-40, over 40) and gender (male, female) are the characteristics that will determine the sample.

Stage 2: The 6 categories to be sampled from are (male under 20), (male 20-40), (male over 40), (female under 20), (female 20-40) and (female over 40).

Stage 3: The numbers (sub-quotas) to be sampled are (male under 20) - 40, (male 20-40) - 60, (male over 40) - 25, (female under 20) - 35, (female 20-40) - 65 and (female over 40) -30. The total quota is the total of all the sub-quotas i.e. 255.

Stage 4: Visit a place where individuals to be interviewed are readily available e.g. a large shopping center and interview people until all the quotas are filled.

Quota sampling is a cheap and convenient way of obtaining a sample in a short space of time. However, this method of sampling is not based on the laws of chance and cannot guarantee a sample that is representative of the population from which it is drawn.

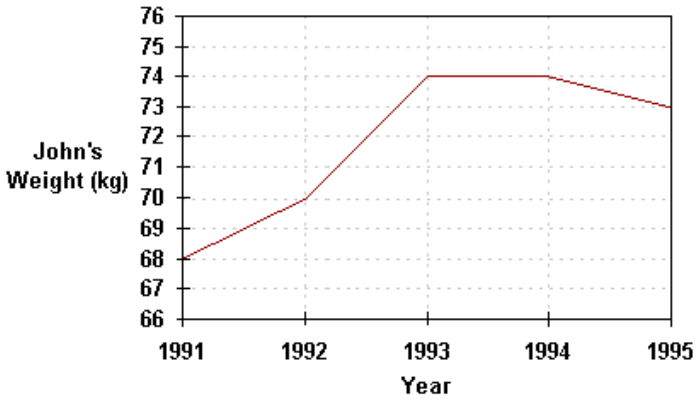
When obtaining a quota sample, interviewers often choose who they like (within criteria specifications) and may therefore select those who are easiest to interview. Therefore sampling bias can result. It is also impossible to estimate the accuracy of quota sampling (because sampling is not random).

Chapter 2 – Descriptive Statistics (Exploratory Data Analysis)

All the data sets used in this chapter will be regarded as samples drawn from some population. One of the main purposes of studying a sample is to get information about the population. The main focus here is on summarizing and describing some features of the data.

2.1 Graphs and diagrams

Line graph – A line graph is a graph used to present some characteristic recorded over time.
Example



The graph above shows how a person's weight varied from the beginning of 1991 to the beginning of 1995.

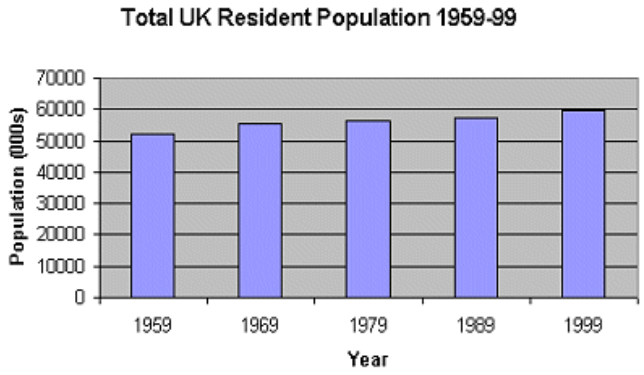
Bar charts

A bar chart or bar graph is a chart consisting of rectangular bars with heights proportional to the values that they represent. Bar charts are used for comparing two or more values that are taken over time or under different conditions.

Simple Bar Chart

In a simple bar chart the figures used to make comparisons are represented by bars. These are either drawn vertically or horizontally. Only totals are represented. The height or length

of the bar is drawn in proportion to the size of the figure being presented. An example is shown below.

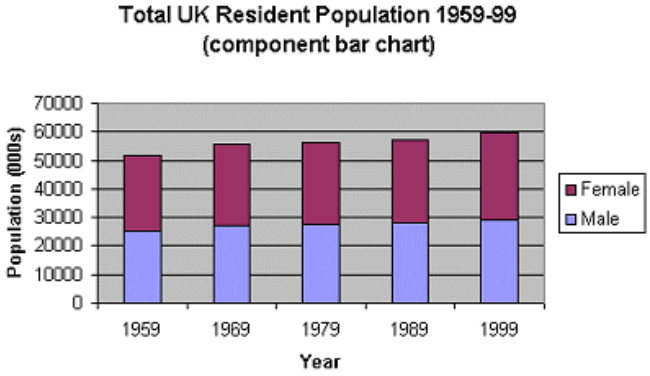


Component Bar Chart

When you want to draw a bar chart to illustrate your data, it is often the case that the totals of the figures can be broken down into parts or components.

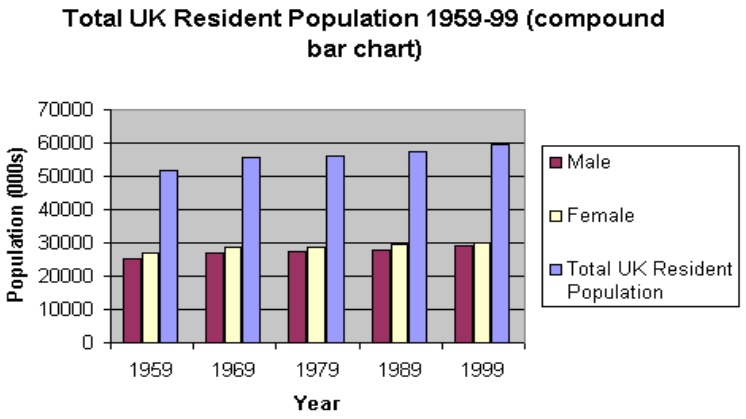
Year	Total	Male	Female
1959	51 956 000	25 043 000	26 913 000
1969	55 461 000	26 908 000	28 553 000
1979	56 240 000	27 373 000	28 867 000
1989	57 365 000	27 988 000	29 377 000
1999	59 501 000	29 299 000	30 202 000

You start by drawing a simple bar chart with the total figures as shown above. The columns or bars (depending on whether you draw the chart vertically or horizontally) are then divided into the component parts.



Multiple (compound) Bar Chart

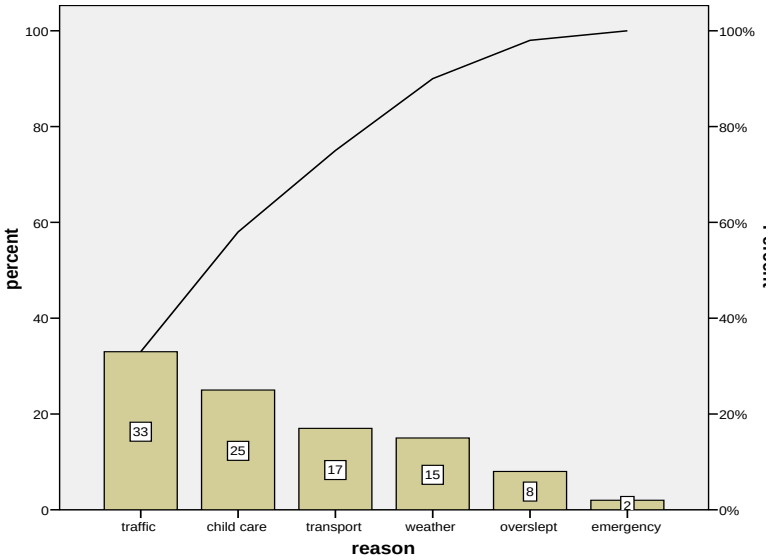
You may find that your data allows you to make comparisons of the component figures themselves. If so, you will want to create a multiple (compound) bar chart. This type of chart enables you to trace the trends of each individual component, as well as making comparisons between the components.



Pareto chart

A Pareto chart is a special type of bar chart where *the values being plotted are arranged in descending order.* The graph is accompanied by a line graph which shows the cumulative totals of each category, left to right.

The graph below is a Pareto chart that shows the percentage of late arrivals at a place of work organized according to cause of late arrival (from the most common to the least common cause). The line shows the accumulated percentages.



Dot Plot

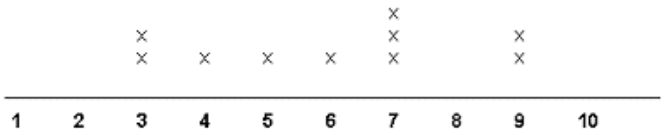
This is diagram where a line is drawn according to a scale that is appropriate for the data set and the values (in the data set) plotted at their positions on the scale. If the same value occurs more than once, the multiple values are plotted on top of each other at the same point on the scale. For small data sets (few values) this plot can provide useful information regarding data patterns.

Example

Imagine that a medium-sized retailer, thinking of expanding into a new region identifies a business that it considers as being ready for takeover. It finds the following annual profit figures (in tens of thousands of pounds) for the target retailer's last ten years trading:

9 9 7 7 7 6 5 4 3 3

To draw a dot plot we can begin by drawing a horizontal line across the page to represent the range of values of all the numbers; then we can mark an 'x' above the appropriate value along the line as follows:

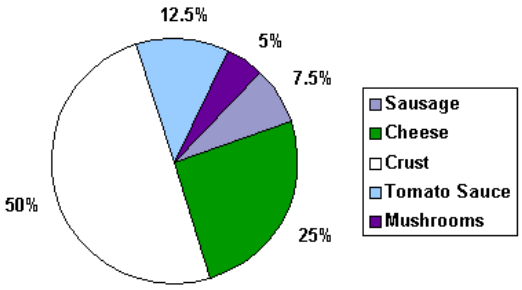


Pie Chart

A Pie chart is a diagram that shows the subdivision of some entity/total into subgroups. The diagram is in the form of a circle which is divided into slices with each slice having an area according to the proportion that it makes up of the total.

Example

The pie chart below shows the ingredients used to make a sausage and mushroom pizza.



The degrees needed for each slice is found by calculating the appropriate percentage of 360 e.g. for sausage the degrees are 0.125×360 = 45 and for cheese 0.25×360 =90 etc.
The complete calculations are shown in the table below.

Ingredient	Percentage	Degrees
Sausage	7.5	0.075 × 360 = 27
Cheese	25	0.250 × 360 = 90
Crust	50	0.50 × 360 = 180
Tomato sauce	12.5	0.125 × 360 = 45
Mushrooms	5	0.050 × 360 = 18

Stem-and-leaf plot

A stem-and-leaf plot is a device used for summarizing quantitative data in a table/graphical format to assist in visualizing the shape of a data set.

Examples

1) To construct a stem-and-leaf plot, the values must first be sorted in ascending order. Here is the sorted set of data values that will used in the example:

44 46 47 49 63 64 66 68 68 72 72 75 76 81 84 88 106

Next, it must be determined what the stems will represent and what the leaves will represent. Typically, the **leaf** contains the last digit of the number and the **stem** contains all of the other digits. In the case of very large or very small numbers, the data values may be rounded to a particular place value (such as the hundredths place) that will be used for the leaves. The remaining digits to the left of the rounded place value are used as the stems.

In this example, the leaf represents the “ones” place and the stem the rest of the number (“tens” place or higher).

The stem-and-leaf plot is drawn with two columns separated by a vertical line. The stems are listed to the left of the vertical line. It is important that each stem is listed only once and that no numbers are skipped, even if it means that some stems have no leaves. The leaves are listed in increasing order in a row to the right of each stem.

```
4 | 4 6 7 9
5 |
6 | 3 4 6 8 8
7 | 2 2 5 6
8 | 1 4 8
9 |
10 | 6
```

key: 5|4=54
leaf unit: 1.0
stem unit: 10.0

Conclusion: The 12 of the 17 values are greater or equal to 63 and less or equal to 88.

2) **Two data sets** can be compared by drawing a **back-to-back stem-and-leaf plot**.

As an example, suppose the fat contents (in grams) for eating English breakfasts and cold meat sandwiches are to be compared. The fat contents are shown below.

Sandwiches: 6, 7, 12, 13, 17, 18, 20, 21, 21, 24, 26, 28, 30, 34

Breakfasts: 12, 14, 15, 16, 18, 23, 25, 25, 36, 36, 38, 41, 44, 45

A back-to-back stem-and-leaf plot is shown below.

Breakfasts		Sandwiches
	0	6 7
2 4 5 6 8	1	2 3 7 8
3 5 5	2	0 1 1 4 6 8
6 6 8	3	0 4
1 4 5	4	

key: 2|4=24 for sandwiches and 2|4=42 for breakfasts
leaf unit: 1.0
stem unit: 10.0

Conclusion: The fat content in English breakfasts appears to be higher than that in sandwiches.

2.2 Sigma and subscript notation

The symbol **sigma** $\sum$ (Capital S in Greek alphabet) is used to denote "the sum of" values.

Suppose the symbol x is used to denote some variable of interest in a study. In order to distinguish between values of this variable, **subscripts** are used.

x_1 – first value in the data set which has a subscript 1.
x_2 – second value in the data set which has a subscript 2.
.
.
x_n – nth value in the data set which has a subscript n.

The sum of these values is written in shorthand notation as

$$x_1 + x_2 + \ldots + x_n = \sum_{i=1}^{n} x_i .$$

If it is understood that the range of subscript indices over which the summation is taken involves all the x values, the summation can be written as just

$$x_1 + x_2 + \ldots + x_n = \sum x .$$

Example 1: Suppose $x_1 = 70$, $x_2 = 74$, $x_3 = 66$, $x_4 = 68$, $x_5 = 71$. Then

$$\sum_{i=1}^{5} x_i = x_1+x_2+ \ldots + x_5 = 70+74+66+68+71 = 349.$$

The sum of the squares of a set of values are written as

$$\sum_{i=1}^{n} x_i^2 = x_1^2 + x_2^2 + \cdots + x_n^2 \text{ or } \sum x^2 \text{ for short.}$$

Example 2: For the data set in example 1,

$$\sum_{i=1}^{5} x_i^2 = 70^2 + 74^2 + 66^2 + 68^2 + 71^2 = 24397.$$

Note that $\sum_{i=1}^{n} x_i^2 \neq (\sum_{i=1}^{n} x_i)^2$

e.g. for the abovementioned data $\sum_{i=1}^{5} x_i^2 = 24397 \neq 349^2 = 121801$.

The summation notation can also be used to write the sum of products of corresponding values for 2 different sets of values.

$$\sum_{i=1}^{n} x_i y_i = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n$$

Example: Consider the following values.

i	1	2	3	4	5	6
x_i	11	13	7	12	10	8
y_i	8	5	7	6	9	11

For this data

$$\sum_{i=1}^{6} x_i y_i = (11\times8) + (13\times5) + (7\times7) + (12\times6) + (10\times9) + (8\times11)$$
$$= 88 + 65 + 49 + 72 + 90 + 88$$
$$= 452.$$

Note that $\sum_{i=1}^{n} x_i y_i \neq (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)$

e.g. for the abovementioned data $\sum_{i=1}^{6} x_i = 61$ and

$\sum_{i=1}^{6} y_i = 46$ $(\sum_{i=1}^{6} x_i)(\sum_{i=1}^{6} y_i) = 2806 \neq \sum_{i=1}^{6} x_i y_i$.

The summation notation is used extensively in specifying calculations in statistical formulae.

2.3 Frequency distributions and related graphs

Frequency distribution

A frequency distribution is a table in which data are grouped into classes and the number of values (frequencies) which fall in each class recorded.
The main purpose of constructing a frequency distribution is to get insight into the distribution pattern of the frequencies over the classes. Hence, the name frequency distribution is used to refer to this pattern.

Example 1
In a survey of 40 families in a village, the number of children per family was recorded and the following data obtained.

1	0	3	2	1	5	6	2
2	1	0	3	4	2	1	6
3	2	1	5	3	3	2	4
2	2	3	0	2	1	4	5
3	3	4	4	1	2	4	5

number of children	Tally	frequency (f)
0	|||	3
1	||||| ||	7
2	||||| |||||	10
3	||||| |||	8
4	||||| |	6
5	||||	4
6	||	2
Total		40

Note: The sum of the frequencies = sample size i.e. $\sum f = n.$

Example 2

Consider the following data of low temperatures (in degrees Fahrenheit to the nearest degree) for 50 days. The highest temperature is 64 and the lowest temperature is 39.

Data Set - Low Temperatures for 50 Days				
57	39	52	52	43
50	53	42	58	55
58	50	53	50	49
45	49	51	44	54
49	57	55	64	45
50	45	51	54	58
53	49	52	51	41
52	40	44	49	45
43	47	47	43	51
55	55	46	54	41

Constructing a frequency distribution

The classes into which the above values can be sorted can be found by following the steps shown below.

1. Find the maximum (=64) and minimum (=39) values and calculate the

   range = maximum – minimum = 64 – 39 = 25.

2. Decide on the number of classes. Use Sturges' rule which states that

   No. of classes = k
   = the rounded up value of (1 + 1.44 ln *n*)
   = 1 + 1.44 × ln(50)
   = 6.63

   i.e. k = 7.

3. Calculate the class width such that no. of classes × class width > range

   i.e. 7× class width > 25.

   This suggests a class width of 4.

4. Find the lower value that defines the first class. This is usually a value just below the minimum value in the data set. Since the minimum value for this data set is 39, the lowest class can have a minimum value one below this i.e. 38.

5. Find the lower values that define each of the classes that follow by successively adding the class width to the lower value of class.

   lower value of the second class = 38 + 4 = 42.

   lower value of the third class = 42 + 4 = 46 etc.

   The frequency distribution below shows the data values sorted into the classes

   38 – 41, 42 – 45, 46 – 49, 50 – 53, 54 – 57, 58 – 61, 62 – 65

   The table below shows the classes and their frequencies for the temperatures data set.

class limits	*f*
38 – 41	4
42 – 45	10
46 – 49	8
50 – 53	15
54 – 57	9
58 – 61	3
62 – 65	1
Total	50

The values in the above example that define the classes of the frequency distribution are called class limits. The classes of the type 38 – 41, 42 – 45,… in which both the upper and lower limits are included are called " inclusive classes" . For example, the class 38 – 41 includes all the values from 38 to 41.

In spite of great importance of classification in statistical analysis, no hard and fast rules can be laid down for it.

The following points must be kept in mind for classification:

1) The classes should be clearly defined and should not lead to any ambiguity.
2) Each of the given values in the data set should be included in one of the classes.
3) The classes should be of equal width, otherwise the different class frequencies will not be comparable. If the class widths are unequal, then comparable figures can

be obtained by dividing the value of the frequencies by the corresponding widths of the class intervals. The ratios thus obtained are called ' frequency density'.

4) The number of classes should not be too large nor too small.

Continuous Frequency Distribution

If we deal with a continuous variable, it is not possible to arrange the data in the class intervals of above type. Let us consider the distribution of age in years. If class intervals are 15 – 19, 20 – 24 then persons with ages between 19 and 20 years are not taken into consideration. In such a case we form the class intervals as 0 – 5, 5 – 10, 10 – 15, 15 – 20,…… Here all the persons with any fraction of age are included in one group or the other. In the above classes, the upper limits of each class are excluded from the respective classes and are included in the immediate next class and are known as 'exclusive classes'. The upper and lower class limits of the new exclusive type classes are known as class boundaries.

If *d* is the gap between the upper limit of any class and the lower limit of the succeeding class, the class boundaries for any class are then given by :

Upper class boundary = upper class limit + (d/2)
Lower class boundary = Lower class limit – (d/2)

Example 2 continued (temperature data)
The frequency distribution below includes the class boundaries.

class limits	class boundaries	*f*
38 – 41	37.5 – 41.5	4
42 – 45	41.5 – 45.5	10
46 – 49	45.5 – 49.5	8
50 – 53	49.5 – 53.5	15
54 – 57	53.5 – 57.5	9
58 – 61	57.5 – 61.5	3
62 – 65	61.5 – 65.5	1
	Total	50

Example 3
The monthly expenditures (thousands of rands) of 60 households are shown on the next page. The values of this data set were accurately recorded (not rounded).

7.21741	7.8989	6.85461	10.31167	8.48253	5.17069
5.09063	8.16412	5.67094	7.7394	7.87423	5.41634
9.37265	10.14436	7.15675	10.31107	8.86571	10.1734
5.99276	6.5738	7.06965	8.82439	7.47467	9.50018
4.90014	5.50273	8.12516	5.51933	7.43641	10.95599
5.87188	9.36936	9.83773	10.18893	5.12028	9.60018
8.56534	9.27719	8.37107	7.03318	10.78344	9.08941
6.85749	7.7887	9.68159	6.75009	8.0521	8.19638
10.17312	7.51527	11.31383	8.5765	7.48021	8.39881
7.37565	7.28159	8.81773	5.53182	5.98515	7.71778

The frequency distribution shown below is a summary of this data set.

classes	f
4.5 – 5.5	5
5.5 – 6.5	7
6.5 – 7.5	13
7.5 – 8.5	13
8.5 – 9.5	9
9.5 – 10.5	10
10.5 – 11.5	3
Total	60

For this distribution lower (upper) class limit = lower (upper) class boundary for each of the classes.
A value that falls on the boundary of 2 classes is allocated to the higher of the two classes e.g. 5.50000 is allocated to the class 5.5 – 6.5 (not 4.5 to 5.5).

Class midpoints

The midpoint of class (x_{mid}) can be calculated from

$$x_{mid} = \frac{Lower\ class\ limit\ (boundary) + Upper\ class\ limit\ (boundary)}{2}$$

Examples

1) For the frequency distribution in example 2 (temperature data), the class midpoints are given on the following page.

class limits	class boundaries	f	midpoints
38 – 41	37.5 – 41.5	4	39.5
42 – 45	41.5 – 45.5	10	43.5
46 – 49	45.5 – 49.5	8	47.5
50 – 53	49.5 – 53.5	15	51.5
54 – 57	53.5 – 57.5	9	55.5
58 – 61	57.5 – 61.5	3	59.5
62 – 65	61.5 – 65.5	1	63.5

2) For the frequency distribution in example 3 (expenditure data), the class midpoints are given below.

classes	midpoints
4.5 – 5.5	5
5.5 – 6.5	6
6.5 – 7.5	7
7.5 – 8.5	8
8.5 – 9.5	9
9.5 – 10.5	10
10.5 – 11.5	11

Cumulative frequencies

The "less than" cumulative frequency of a class is the number of values in the sample that are less than or equal to the upper class boundary of the class.

Examples

1) For the frequency distribution in example 2 (temperature data) the cumulative frequencies are calculated as shown below.

class boundaries	f	cumulative frequency	calculations
37.5 – 41.5	4	4	4
41.5 – 45.5	10	14	4+10
45.5 – 49.5	8	22	4+10+8
49.5 – 53.5	15	37	4+10+8+15
53.5 – 57.5	9	46	4+10+8+15+9
57.5 – 61.5	3	49	4+10+8+15+9+3
61.5 – 65.5	1	50	4+10+8+15+9+3+1

2) For the frequency distribution in example 3 (expenditure data) the cumulative frequencies are calculated as shown below.

classes	f	cumulative frequencies	calculations
4.5 – 5.5	5	5	5
5.5 – 6.5	7	12	5+7
6.5 – 7.5	13	25	5+7+13
7.5 – 8.5	13	38	5+7+13+13
8.5 – 9.5	9	47	5+7+13+13+9
9.5 – 10.5	10	57	5+7+13+13+9+10
10.5 – 11.5	3	60	5+7+13+13+9+10+3
Total	60		

Relative and percentage frequencies

- Relative frequency = frequency/sample size i.e. Rf = $\frac{f}{n}$.
- The percentage frequency of a class is calculated from relative frequency × 100.

Examples

1) The relative and percentage frequencies for the frequency distribution in example 2 (temperature data) are shown below.

class boundaries	f	relative frequency	percentage frequency
37.5 – 41.5	4	0.08	8
41.5 – 45.5	10	0.2	20
45.5 – 49.5	8	0.16	16
49.5 – 53.5	15	0.3	30
53.5 – 57.5	9	0.18	18
57.5 – 61.5	3	0.06	6
61.5 – 65.5	1	0.02	2

2) The relative and percentage frequencies for the frequency distribution in example 3 (expenditure data) is shown on the following page.

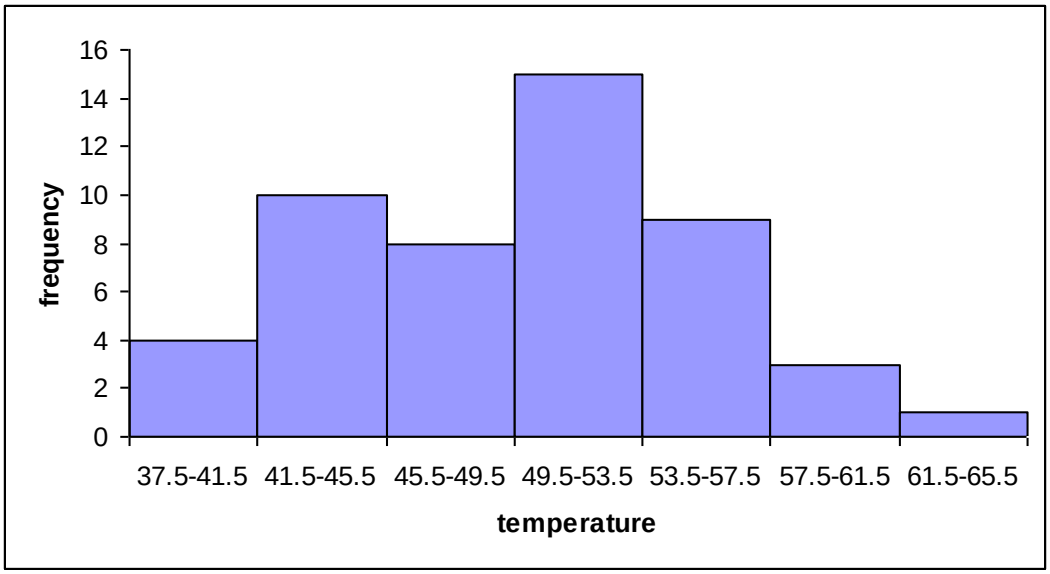
classes	f	relative frequency	percentage frequency
4.5 – 5.5	5	0.083	8.3
5.5 – 6.5	7	0.117	11.7
6.5 – 7.5	13	0.217	21.7
7.5 – 8.5	13	0.217	21.7
8.5 – 9.5	9	0.15	15
9.5 – 10.5	10	0.167	16.7
10.5 – 11.5	3	0.05	5
Total	60	1	100

Histogram

A histogram is the graphical representation of a frequency distribution. The frequency for each class is represented by a rectangular bar with the class boundaries as base and the frequency as height.

Example

A histogram of the frequency distribution in example 2 (temperature data) is shown below.



Frequency polygon

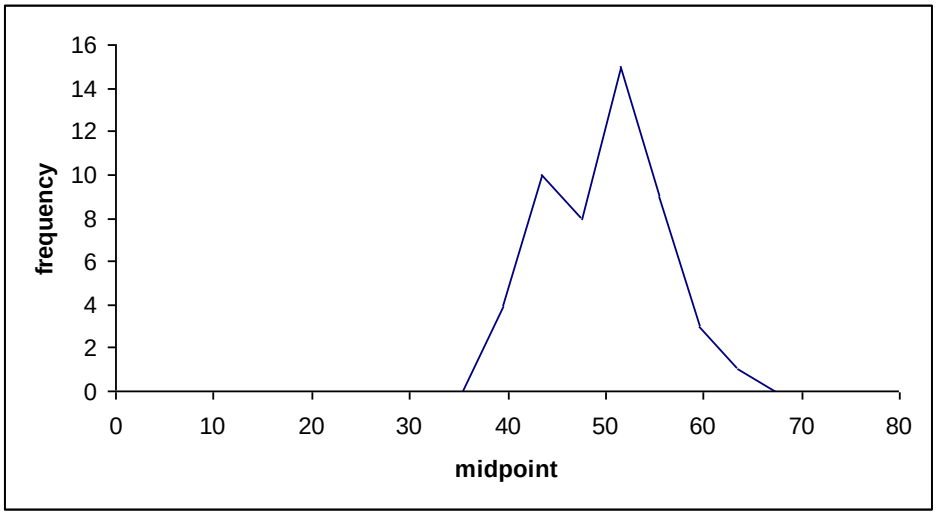
This is also a graphical representation of a frequency distribution. For each class the class midpoint is plotted against the frequency and the plotted points joined by means of straight lines.

Example

For the temperature data the following values are plotted.

midpoint	35.5	39.5	43.5	47.5	51.5	55.5	59.5	63.5	67.5
f	0	4	10	8	15	9	3	1	0

The plot is shown below.



Note:
The two plotted values at the lower and upper ends were added to anchor the graph to the horizontal axis. The lower end value is a plot of 0 versus the midpoint of the class below the first (lowest) class (35.5). This midpoint is obtained by subtracting the class width (4) from the midpoint of the lowest class (39.5). The upper end value is a plot of 0 versus the midpoint of the class above the last class (67.5). This midpoint is obtained by adding the class width (4) to the midpoint of the last (highest) class (63.5).

The histogram and frequency polygon are equivalent graphical representations of the pattern of the frequencies shown in the frequency distribution.

The the histogram can provide an estimate of the probability (chance) that a value drawn at random from the data set will lie between two values.

Examples

1) For the frequency distribution in example 2 (temperature data), the estimated chance that a randomly drawn value will be at least 45.5 but less than 57.5 is $\frac{8+15+9}{50} = 0.64$.

2) For the frequency distribution in example 3 (monthly expenditure), the estimated chance that a randomly drawn value will be at least 7.5 is $\frac{13+9+10+3}{60} = 0.583$.

"Less than" ogive

This is the graph of the cumulative frequencies versus the upper class boundaries.

Example

For the "less than" ogive of the frequency distribution in example 2 (temperature data)

the following values are plotted.

class boundary	37.5	41.5	45.5	49.5	53.5	57.5	61.5	65.5
cumulative frequency	0	4	14	22	37	46	49	50

The material developed here may be used as a whole unit, or parts of it may be extracted and taught in various courses. The elementary concepts may be incorporated into general mathematics classes in grades seven to twelve, and the more difficult parts may be used in advanced algebra classes in the high school. Depending upon the amount of material used, several days or several weeks may be allotted to teach the unit.

Definitions of Statistical Terms

Statistics is a branch of mathematics in which groups of measurements or observations are studied. The subject is divided into two general categories— *descriptive statistics* and *inferential statistics*. In descriptive statistics one deals with methods used to collect, organize and analyze numerical facts. Its primary concern is to describe information gathered through observation in an understandable and usable manner. Similarities and patterns among people, things and events in the world around us are emphasized. Inferential statistics takes data collected from relatively small groups of a population and uses inductive reasoning to make generalizations, inferences and predictions about a wider population.

Throughout the study of statistics certain basic terms occur frequently. Some of the more commonly used terms are defined below:

A *population* is a complete set of items that is being studied. It includes all members of the set. The set may refer to people, objects or measurements that have a common characteristic. Examples of a population are all high school students, all cats, all scholastic aptitude test scores.

A relatively small group of items selected from a population is a *sample*. If every member of the population has an equal chance of being selected for the sample, it is called a *random sample*. Examples of a sample are all algebra students at Central High School, or all Siamese cats.

Data are numbers or measurements that are collected. Data may include numbers of individuals that make up the census of a city, ages of pupils in a certain class, temperatures in a town during a given period of time, sales made by a company, or test scores made by ninth graders on a standardized test.

Variables are characteristics or attributes that enable us to distinguish one individual from another. They take on different values when different individuals are observed. Some variables are height, weight, age and price. Variables are the opposite of *constants* whose values never change.

Exercises:

1.) Tell whether each of the following is a variable or a constant:
a.) Scores obtained on a final examination by members of a statistics class.
b.) The cost of clothing purchased each year by secretaries.
c.) The number of days in the month of June.
d.) The time it takes to do grocery shopping.
e.) The age at which one may become a voter in the United States of America.
2.) Fill in the missing word to make a true statement.
a.) ____ are measurements obtained by observation.
b.) A ____ is a complete set of items.
c.) _____takes data collected from a small group and makes predictions about a wider sample.
d.) When every member of a set has an equal chance of being selected as part of a sample, the sample is called a ____ ____.

e.) Characteristics that vary from one individual to another are ____.
f.) The study that deals with methods of collecting, organizing and analyzing data is ____ ____.

Frequency Distributions

Groups of data have little value until they have been placed in some kind of order. Usually measurements are arranged in ascending or descending order. Such a group is an *array* or *distribution*. A *frequency distribution* is a table in which measurements are tallied and the *frequency* or total number of times that each item occurs is recorded.

***Example 1*:**

The frequency distribution below shows data obtained in a survey asking a group of people to name their favorite among several kinds of cars. Use the table to answer the following questions:

a.) How many people are included in the sample?
b.) What percent of the people surveyed preferred Chevrolets?
c.) What is the ratio of people who prefer Oldsmobiles to those who prefer Buicks?
d.) If the number of Subarus were increased by three, what would the percent of increase be?

(figure available in print form)

Solution:

a.) 25 + 18 + 15 + 12 + 10 = 80 80 people are included in the survey.
b.) 25 Ö 80 = .3125 = 31.25% 31.25% of the people surveyed preferred Chevrolets.
c.) 15:10 = 3:2 The ratio of people who prefer Oldsmobiles to those who prefer Buicks is 3:2.
d.) 3:18=x:100 1:6 =x:100 6x=100 x =16 2/3 The percent of increase is 16 2/3%.

When the number of measurements in a survey is large, or when the range, that is, the difference between the highest and lowest measurements in the survey, is great, it is usually more efficient to arrange the data in intervals and show the number of items within each group. The number of intervals used in a frequency distribution may vary. However, it has been found that ten to twenty intervals are most practical.

The following steps may be used to set up a frequency distribution:

1.) Select an appropriate number of intervals for the given data.
2.) Find the difference between the highest and lowest measurements in the data. Add one to the result end divide the sum by the number of intervals. If the quotient is not an integer, round it to the nearest odd integer. This will be the size or width of each interval and will be designated by the symbol w.
3.) The lowest number in the bottom interval will be the lowest measurement in the given data. Add (w-1) to this measurement to obtain the highest number in the bottom interval. The next interval begins at the integer following the highest number in the bottom group. Continue in this manner for each successive higher interval until every measurement has been placed in its proper group.
4.) After the intervals have been established, a tally mark is placed by the interval for each measurement in the group. The frequency, or number of measurements in each interval, is indicated with a numeral.

***Example 2*:**

Make a frequency distribution of the following scores obtained by 40 students on a mathematics test.

86 82 56 73 87 89 72 86 88 76
72 69 84 85 62 97 70 78 84 93
70 60 91 76 83 94 65 72 92 81
98 78 88 76 96 89 90 83 74 80

***Solution*:**

Use ten intervals.

Highest Score—Lowest Score = 98Ð56 = 42 (42 + 1) = 10 = 43 dv 10 = 4.3 Round to 5. The size of each interval is 5.

Scores	Tally	Frequency
96Ð100	111	3
91Ð95	1111	4
86Ð90	111	8
81Ð85	11	7
76Ð80	1	6
71Ð75	1111	5
66Ð70	111	3
61Ð65	11	2
56 Ð60	11	2

Although it is not necessary, it is often helpful for use in further analysis to have additional information in a frequency distribution. This additional information may include the midpoint of each interval, the percentage of the numbers in the frequency column relative to the total frequencies, the cumulative frequency of successive summation of entries in the frequency column, and the percentage of the cumulative frequency.

***Example 3*:**

In the frequency distribution for example 2 find (a) the midpoint of each interval; (b) the percentage of each frequency relative to the total frequencies; (c) the cumulative frequency; and (d) the percentage of cumulative frequency relative to the total frequencies.

***Solution*:**

(a) Since the width of each interval is 5, the third score is the midpoint of the interval. For example, the lowest interval contains the scores 56, 57, 58, 59, 60. 58 is the midpoint of this interval.
(b) To find the percentage of each frequency divide the frequency by the total number of measurements and change the resulting decimal to a percent. The frequency of the lowest interval is 2. The total number of measurements is 40. 2 dv 40 = .05 = 5%
(c) The cumulative frequency at any interval may be obtained by successively adding the frequencies of all the groups from the lowest interval up to and including the given interval. The cumulative frequency of the interval 76-80 is 2+ 2+ 3+ 5+ 6 =18.
(d) To obtain the percentage of cumulative frequency relative to the total of the frequencies, divide the cumulative frequency by the total number of measurements. Change the resulting decimal to a percent. The percentage of the cumulative frequency in the interval 76-80 is 18 dv 40 = .45= 45%. This figure may also be found by adding the percentage of frequency of all groups from the lowest up to and including the given interval.

Scores	Midpoint	Frequency	% of Frequency	Cumulative Frequency	% of Cumulative Frequency
99-100	98	3	7.5	40	100.0
9195	93	4	10.0	37	92.5
8690	88	8	20.0	33	82.5
8185	83	7	17.5	25	62.5
7680	78	6	15.0	18	45.0
7175	73	5	12.5	12	30.0
6670	68	3	7.5	7	17.5
6165	63	2	5.0	4	10.0
5660	58	2	5.0	2	5.0

***Exercises*:**

1.) Ask the students in each of your classes which of the following colors they prefer—red, blue, yellow, green, brown, or purple. Construct a frequency distribution to display the results of your survey
a.) How many people are included in the sample?

b.) What percent of the people surveyed prefer yellow? red? purple?

c.) What is the ratio of people who prefer green to those who prefer blue?

d.) What is the most popular color?

e.) What is the least popular color?

f.) If the number of people who prefer red were decreased by 2, what would be the percent of decrease?

2.) Tally the following scores in a frequency distribution. Do not use grouping.

84 98 92 88 91 91 85 80 84 93
92 80 91 84 87 85 84 80 87 95

3.) Make a frequency distribution of the following scores obtained by a basketball team.

72 104 95 93 96 76 105 100
88 62 79 78 87 78 89 81
110 68 96 106 80 87 86 84
102 84 96 88 82 83 92 87
87 85 108 90 94 98 78 80

a.) Use ten intervals and display the midpoint of each interval.

b.) Calculate the percentage of frequency of each interval.

c.) Find the cumulative frequency for each interval.

d.) Calculate the percentage of each cumulative frequency relative to the total of the frequencies.

Dot Diagrams

Many people find it easier to obtain information from pictures than from written material. Statisticians display mathematical relationships with diagrams and graphs. From these pictures numerical data can be summarized clearly and easily.

When the data of a frequency distribution have not been grouped in intervals, they can be represented on a dot *diagram*. A dot diagram illustrates the pattern of a distribution. It clearly shows whether the data are spread out evenly or if they tend to cluster about any point.

To construct a dot diagram list the measurements, from lowest to highest, horizontally across the bottom of the graph. On the left side vertically list the frequencies or number of times that the measurements occur. For each time a measurement occurs place a dot in the column above the measurement.

***Example*:**

Construct a dot diagram to represent the following distribution of daily temperature highs in twenty-four cities of the United States.

67 68 69 70 70 71 71 71
72 72 72 74 74 74 74 76
76 76 76 80 80 80 84 85

***Solution*:**

(figure available in print form)

***Exercises*:**

1.) Twenty workers were rated on a scale of 1 to 10 for efficiency. Construct a dot diagram to represent the following ratings: 7, 8, 9, 4, 5, 5, 7, 10, 6, 8, 7, 7. 5, 6, 9, 6.
2.) Draw a dot diagram to represent the following scores received on a spelling test: 98, 100, 78, 75, 68, 62, 75, 80, 82, 94, 80, 72, 75, 85, 85, 80, 70, 82, 78, 78, 72, 70, 90, 65.
3.) The distribution of heights of fifteen children is given below. Show the distribution on a dot diagram.

Height in Inches	Frequency
56	2
58	3
60	7
62	2
64	1

Histograms

A frequency distribution can be represented graphically on a *histogram*. A histogram is a bar graph on which the bars are adjacent to each other with no space between them. To construct a histogram, arrange the data in equal intervals. Represent the frequencies along the vertical axis and the scores along the horizontal axis. The true limits of any interval extend one half unit beyond the endpoints established for the interval and are represented in this manner on the horizontal axis. For example, the true limits of the interval 76-80 are 75.5 and 80.5. To get the proper perspective, the vertical axis should be approximately three-fourths as long as the horizontal axis.

***Example*:**

Illustrate the following set of measurements on a histogram:

72 82 56 73 87 89 72 86 88 76
86 69 84 85 62 97 70 78 84 93
70 60 91 76 83 94 65 72 92 81
98 78 88 76 96 89 90 83 74 80

***Solution*:**

Scores	Frequency
96-100	3
91-95	3
86-90	4
81-85	6
76-80	8
71-75	5
66-70	3
61-65	2
56-60	

(figure available in print form)

***Exercises*:**

1.) Construct a histogram for the following scores earned by a group of high school students on a Scholastic Aptitude Examination.

Score	Number of Students
400-449	20
450-499	35
500-549	50
550-599	50
600-649	40
650-699	20
700-749	10

2.) The weights of 40 football players are as follows:

210 181 192 164 170 186 205 194

178 161 175 195 172 188 196 182
206 188 165 202 178 163 190 198
187 198 174 172 183 208 185 162
203 172 196 184 185 176 197 184

a.) Construct a frequency distribution for the given data.

b.) Make a histogram for the given data.

Frequency Polygon

A frequency polygon is a line graph which can be used to represent the frequency of a set of numbers. It is formed by connecting a series of points. The abscissa of each point is the midpoint of the interval in which the point lies. The ordinate of each point is the frequency for the interval. The polygon is closed at each end by drawing a line from the endpoints to the horizontal axis at the midpoint of the next interval.

***Example*:**

Illustrate the following data on a frequency polygon:

Scores	Midpoint	Frequency
96-100	98	3
9195	93	3
8690	88	4
8185	83	6
7680	78	8
7175	73	5
6670	68	3
6165	63	2
5660	58	2

***Solution*:**

(figure available in print form)

***Exercises*:**

1.) The following table shows the weekly wages earned by workers in a local hospital:

Number of People	11	11	15	18	13	12	10
Weekly Wage	$140	$200	$180	$160	$190	$150	$170

a.) Draw a histogram for the given data.

b.) Construct a frequency polygon for the given data.

2.) A baseball team made the following number of hits in a recent game:

Inning	1	2	3	4	5	6	7	8	9
Number of Hits	1	4	2	3	3	5	3	2	1

a.) Draw a dot diagram for the given data.

b.) Make a histogram for the given data.

c.) Construct a frequency polygon for the given data.

3.) The students in an English class received the following scores on a test:

60	95	85	100	81
56	87	80	62	75
73	64	69	86	93
82	77	91	58	69
76	94	72	88	78

a.) Make a frequency distribution for the given scores.

b.) Draw a histogram to represent the scores.

c.) Construct a frequency polygon for the given data.

Cumulative Frequency Polygon

Another method of graphical representation is the *cumulative frequency polygon*. The cumulative frequency polygon is a line graph which is used to picture cumulative frequencies of a set of numbers. The abscissa of each point is the upper limit of an interval in a frequency distribution. The ordinate of each point is the corresponding cumulative frequency. The graph starts at a frequency of zero for a group below the lowest interval in the distribution.

***Exam ple*:**

Construct a cumulative frequency polygon to represent the following scores obtained by 40 students on a mathematics test.

86	82	56	73	87	89	72	86	88	76
72	69	84	85	62	97	70	78	84	93
70	60	91	76	83	94	65	72	92	81
98	78	88	76	96	89	90	83	74	80

***Solution*:**

Make a frequency distribution for the scores, then draw the graph.

Scores	Frequency	Cumulative Frequency	% of Cumulative Frequency
96-100	3	40	100.0
9195	4	37	92.5

8690	8	33	82.5
8185	7	25	62.5
7680	6	18	45.0
7175	5	12	30.0
6670	3	7	17.5
6165	2	4	10.0
5660	2	2	5.0

(figure available in print form)

For some purposes the cumulative frequency polygon is very valuable. On the right side of the polygon is a scale of percent that parallels the scale of cumulative frequency. On the percent scale you read 25 corresponding to an abscissa of 72. This means that 25% of the scores were 72 or lower. The figure 72 is called the 25th *percentile*. The nth percentile is that score below which n percent of the scores in the distribution will fall.

To find the score that corresponds to a percentile on the graph, draw a horizontal line through the desired percent to intersect the cumulative frequency polygon. From the point of intersection draw a vertical line to the x-axis. The score at the point of intersection of the vertical line and the x-axis corresponds to the required percentile.

The fiftieth percentile is the *median* or middle score in a set of measurements. The 25th percentile is called the *lower quartile*, and the 75th percentile is the *upper quartile*.

***Exercises*:**

1.) During one week a dealer sold the following number of cars: Monday 12, Tuesday 15, Wednesday 5, Thursday 6, Friday 10, Saturday 12.

a.) Construct a histogram to represent the given data.

b.) Make a frequency polygon to represent the given data.

c.) Draw a cumulative frequency polygon to represent the given data.

2.) The heights in inches of 50 high school students are:

60 68 74 79 62 75 60 65 61 64
71 72 63 66 71 60 60 73 63 65
73 68 76 75 62 76 72 70 69 62
78 71 68 62 74 69 67 70 61 63
72 67 71 68 62 60 70 69 65 64

a.) Group the data into a frequency table.

b.) Construct a histogram to represent the data.

c.) Construct a cumulative frequency polygon.

d.) Find the median height. Find the upper and lower quartiles.

e.) Determine the 8Oth percentile.

3.) Forty students have the following IQ scores:

120 100 115 126 82 108 114 95

150 92 140 88 98 116 134 138
98 87 110 92 106 96 126 102
80 82 100 128 110 100 118 84
88 98 94 85 124 90 80 112

a.) Group the data into a frequency table.

b.) Construct a cumulative frequency polygon.

c.) Determine the median IQ score and the 7Oth percentile.

Measures of Central Tendency

When statisticians study a group of measurements, they try to determine which measure is most representative of the group. The score about which most of the other scores tend to cluster is a *measure of central tendency*. Three measures of central tendency are the mode, the median and the mean.

The *mode* of a set of numbers is the element that appears most frequently in the set. There can be more than one mode in a set of numbers. A set that has two modes is *bimodal*, and one that has three modes is *trimodal*. If no element of a set appears more often than any other element, the set has no mode. The mode is an important measure for business people. It tells them what items are most popular with consumers.

***Example 1*:**

Find the mode of the following set of numbers: 34, 26, 30, 34, 28, 32, 32, 34, 33, 31, 33, 30.

***Solution*:**

Element	Frequency
26	1
28	1
30	2
31	1
32	2
33	2
34	3

The number 34 occurs most frequently, hence 34 is the mode of the set.

***Example 2*:**

Find the mode of the following set of numbers: 13 17, 14, 20, 18

***Solution*:**

Element	Frequency
13	1
17	1

14	1
20	1
18	1

No number appears more than any other number in the set. The set has no mode.

***Example 3*:**

Find the mode of the following set of numbers: 1, 2, 2, 3, 4, 4, 5

***Solution*:**

Element	Frequency
1	1
2	2
3	14
4	2
5	1

The numbers 2 and 4 each appear twice. The set has two modes: 2 and 4.

Another measure of central tendency is the *median*. When the elements of a set of numbers have been arranged in ascending order, the number in the middle of the set is the median of the set. The median divides the set of data into two equal parts. On a cumulative frequency polygon the median is the 50th percentile. To determine which element of a set is the middle number, use the following formula:

Middle Number = (Total Number of Elements + 1); = 2

If the set contains an even number of elements, the median is the average of the two middle numbers.

***Example 1*:**

The weights of nine children are as follows: 99, 98, 73, 81, 79, 86, 90, 94, 71. Find the median weight.

***Solution*:**

Arrange the weights in order from lowest to highest: 71, 73, 79, 81, 86, 90, 94, 98, 99 (9 + 1) dv 2 = 10 dv 2 = 5 The fifth number of the set is the middle number. The median weight is 86.

***Example 2*:**

Ten students received the following scores on an examination: 96, 68, 78, 82, 87, 74, 80, 70, 86, 84. Find the median score.

***Solution*:**

Arrange the scores in ascending order: 68, 70, 74, 78, 80, 82, 84, 86, 87, 96.

(10 + 1) Ö 2 = 11 Ö 2 = 5.5

The two middle numbers of the set are the fifth and sixth numbers: 80 and 82.

(80 + 82) Ö 2 = 162 Ö 2 = 81

The median score is 81.

A third, and most widely used, measure of central tendency is the *arithmetic mean*. The arithmetic mean is the average of a set of numbers. It is usually denoted by the symbol x. To calculate the arithmetic mean of a set of numbers, add the members of the set and divide the sum by the number of items in the set.

***Example*:**

Find the arithmetic mean of the following set of numbers: 25, 15, 20, 20, 10.

***Solution*:**

(25 + 15 + 20 + 30 + 10) Ö 5 = 100 Ö 5 = 20

The arithmetic mean of the set is 20.

Sometimes an item appears more than once in a set of measures. To find the arithmetic mean of a set of measures when some items occur several times, multiply each item in the set by Its frequency and divide the sum of these products by the total number of items in the set.

***Example*:**

Find the arithmetic mean of the following numbers: 28, 24, 22, 24, 26, 26, 22, 24, 22, 28, 30, 24.

***Solution*:**

Item	*Frequency*	*Product*
22	3	66
24	4	96
26	2	52
28	2	56
30	1	30

Sum of Products = 66 + 96 + 52 + 56 + 30 = 300

Total Number of Items = 3 + 4 + 2 + 2 + 1 = 12

Sum of Products dv Total # of Items = 300 dv 12 = 25

The arithmetic mean is 25.

When the data have been arranged in intervals in a frequency distribution, the arithmetic mean is obtained in the following manner:

1.) Multiply the midpoint of each interval by the frequency of the interval.
2.) Find the sum of the products obtained in step 1.
3.) Divide the sum obtained in step 2 by the total number of items in the distribution.

The formula used to find the arithmetic mean is:

n

x = 1/ni å1 xifi

x ; arithmetic mean xi = midpoint of the interval
n = number of items in the distribution fi = frequency of the interval
= sum

***Example*:**

Find the arithmetic mean for the following distribution:

Scores	Midpoint	Frequency	xifi
96-100	98	3	294
91-95	93	4	372
86-90	88	8	684
81-85	83	7	581
76-80	78	6	468
71-75	73	5	365
66-70	68	3	204
61-65	63	2	126
56-60	58	2	116

Solution

n = 3 + 4 + 8+ 7+ 6+ 5+ 3+ 2 + 2 = 40

40

åxifi= 294 + 372 + 684 + 581 + 468 + 365+ i= 1 204 + 126 + 116 = 3210

40

x = 1/n å xifi 1/40 x 3210 = 3210/40 = 80.25

i = 1

The arithmetic mean of the distribution is 80.25.

***Exercises*:**

1.) Ten employees of a department store earn the following weekly wages: $200, $150, $160, $125, $160, $150, $180, $130, $170 $150

a.) Find the average weekly income.

b.) What is the median wage?

c.) Find the mode.

2.) Write mean, median, or mode to complete the sentence.

a.) 7, 13, 8, 5, 9, 12. The ____ is 9.

b.) 6, 2, 4, 7, 6, 3. The ____ is 6.
c.) 18, 10, 21, 17, 12. The ____ is 17.
d.) 8, 3, 9, 4, 10, 14. The ____ is 8.
e.) 13, 11, 8, 15, 9, 10. The ____ is 10.5.

3.) Find the mean, the median and the mode for each set of numbers.

a.) 72, 68, 56, 65, 72, 56, 68.

b.) 13, 19, 12, 18, 24, 10.

c.) 125, 132, 120, 118, 128, 126, 120.

d.) 8, 4, 6, 4, 10, 4, 10.

4.) Find the arithmetic mean of the following numbers:

Number	Frequency
32	4
36	2
38	6
40	8

5.) The salaries of thirty people are listed below.

$12,500 $23,900 $18,750 $24,000 $$14,000
$18,750 $11,570 $25,000 $ 9,200 $15,000
$24,000 $22,000 $20,500 $12,500 $17,300
$10,980 $15,550 $18,750 $18,000 $16,200
$ 8,750 $12,500 $10 980 $13,000 $19,850
$32,000 $13,000 $22,000 $35,000 $21,000

a.) Arrange the salaries in intervals and make a frequency table for the set of data.

b.) What is the mode of the salaries?

c.) What is the median salary?

d.) What is the mean salary?

Measures of Disperseion

Measures of central tendency very often present an incomplete picture of data. In order to evaluate more completely any group of scores it is necessary to measure the spread or dispersion of the data being studied. One way to indicate the spread of scores is by the range of scores. The *range* of a set of numbers is the difference between the highest and lowest numbers of the set. To find the range of a set of numbers, use the following formula:

Range = Highest Number—Lowest number

***Example*:**

What is the range of the following set of numbers? 3, 1, 6, 12, 9, 8, 10, 15

***Solution*:**

The highest number in the set is 15. The lowest number in the set is 1.

15Ð1 = 14. The range of the set is 14.

Another way of indicating the dispersion of scores is in terms of their deviations from the mean. This method is known as *standard deviation* and tells how scores tend to scatter about the mean of a set of data. If the standard deviation is small the scores tend to cluster closely about the mean. If the standard deviation is large, there is a wide scattering of scores about the mean. Standard deviation is represented by the symbol s and may be computed by the formula:

Standard Deviation = s=

(figure available in print form)

where x is a score, x is the mean, n is the number of scores, and means “the sum of”.

Six steps are used to find standard deviation:

1.) List each score (x) in the set of data.
2.) Compute the mean (x) for the data.
3.) Subtract the mean from each score (xÐx). The result is the deviation of each score from the mean.
4.) Square the deviations.
5.) Find the average of the squares of the deviations by dividing the sum of the squares of the deviations by the number of scores in the distribution.
6.) Take the square root of this average. The result is the standard deviation.

***Example*:**

Compute the standard deviation for the scores: 2, 3, 4, 5, 6, 7, 8

Solution

(figure available in print form)

The standard deviation is a number that is used to compare scores in a distribution. If the mean of a group of test scores is 75, and the standard deviation is 10, a person who receives a score of 85 is one standard deviation above the mean. If the mean of another group of test scores is 80, and the standard deviation is 3, a person who receives a score of 83 is one standard deviation above the mean. This person has done equally well, with respect to the other class members, as the person who received 85 on the first test.

***Exercises*:**

1.) Compute the range for the following sets of scores:

a.) 24, 15, 19, 29, 24, 22

4 More Measure of Location Extension

- Measures of locations include measures of central tendency and measures of partition.
- Measures of central tendency include; mean (Arithmetic mean, Geometric mean, harmonic mean), median and mode, while measures of partition include; quartiles, deciles and percentiles.
- The arithmetic mean, usually referred to as mean, of a set of numbers is generally denoted by $\bar{x}$ and is defined for ungrouped data as

$$\bar{x} = \frac{\sum x}{\sum n} \tag{1}$$

Example 3.1. The mean of the following data; 2, 5, 6, 10, 7, 3, $\bar{x} = \frac{2+5+6+10+73}{6} = 5.5$

- While for grouped data it is defined as;

$$\bar{x} = \frac{\sum fx}{\sum f} \tag{2}$$

Example 3.2: The following data is the distribution of the moisture content of 50 dishes of a diet. Compute the mean $\bar{x} = \frac{\sum fx}{\sum f} = 15.5$

Table 1: Distribution of moisture content of 50 Dishes of a diet.

Class	Class mark	Frequency	fx	Cumulative frequency (F)
0 - 5	2.5	4	10	4
5 - 10	7.5	8	60	12
10 - 15	12.5	10	125	22
15 - 20	17.5	14	245	36
20 - 25	22.5	9	202.5	45
25 - 30	27.5	5	137.5	50
Total		50	780	

5 Computation from Group Data cont.

- **Geometric Mean:** The geometric mean is an analytic method of finding the average rate of growth or decline in the values of an item over a particular period of time. For ungrouped data it is given as;

$$G.M. = \sqrt[n]{(x_1 \cdot x_2 \ldots x_n)} = \left(\prod_{i=1}^{n}(x_i)\right)^{\frac{1}{n}} \tag{3}$$

While for grouped data it is given as;

$$G.M. = \sqrt[n]{(x_1^{f_1} \cdot x_2^{f_2} \ldots x_n^{f_n}} = \left(\prod_{i=1}^{n}(x_i)^{f_i}\right)^{\frac{1}{n}}, \quad \text{where } n = \sum f \tag{4}$$

where x_i is the class mark and f_i the corresponding frequency for class i and $\prod$ is the multiplication symbol.

- **Example 3.2:** Find the geometric mean of the numbers: 8, 8, 10, 11, 12, 20.
solution: $G.M. = \sqrt[6]{(8 \times 8 \times 10 \times 11 \times 12 \times 20)} = 10.91$

5.1 Computation from Group Data cont.

- **Harmonic Mean:(HM)** of positive observations $x_1, x_2, \ldots, x_n$ is the reciprocal of the mean of the reciprocal of the observations. It is used when dealing with average ratios and rates such as kilometer per hour, naira per liter, plants per stand, insect counts per plot e.t.c For ungrouped data, it is given as

$$H.M. = \frac{n}{\sum\left(\frac{1}{x_i}\right)} \tag{5}$$

While for grouped data it is given as;

$$H.M. = \frac{n}{\sum\left(\frac{f_i}{x_i}\right)} \tag{6}$$

Example 3.3: Find the harmonic mean of the numbers: 8, 8, 10, 11, 12, 20.
solution: $H.M. = 6/(1/8 + 1/8 + 1/101/11 + 1/12 + 1/20) = 10.45$

5.2 Exercise

1 Find the harmonic mean of milk production (liter) per day from a data of 20 days for a particular cow from the following frequency distribution:

Milk(l/day)	10	12	14	16	18
Frequency	5	7	2	2	4

Answer: 12.71l

1 The following is the distribution of Fat (percentage) in 100 samples collected from different milk centres in villages.

Fat(%)	1-3	3-5	5-7	7-9	9-11
Samples	40	26	30	2	2

Compute Mean, Median, Mode, GM and H.M. of Fat content per sample.

2 . The following is the distribution of body weights of 100 calves at the 1st lactation

Body weight (kg)	30-40	40-50	50-60	60-70	70-80
Calves	12	26	34	20	8

Find Mean, Median, Mode, GM. and H.M. of body weight of calves.

3 Compute the Arithmetic Mean , median, mode, First and third Quartiles for the yield (bags) of paddy rice given in the following distribution.

Yield (bags)	less than 20	less than 25	less than 30	less than 35	less than 40	less than 45
Farms	6	18	30	34	16	14

6 Questions and Answers Section

- **Example 5**.

Determine what the key terms refer to in the following study.

As part of a study designed to test the safety of automobiles, the National Transportation Safety Board collected and reviewed data about the effects of an automobile crash on test dummies. Here is the criterion they used:

Speed at which Cars Crashed	Location of "drive" (i.e. dummies)
35 miles/hour	Front Seat

Table 1.1

Cars with dummies in the front seats were crashed into a wall at a speed of 35 miles per hour. We want to know the proportion of dummies in the driver's seat that would have had head injuries, if they had been actual drivers. We start with a simple random sample of 75 cars.

Solution 1.3

The **population** is all cars containing dummies in the front seat.

The **sample** is the 75 cars, selected by a simple random sample.

The **parameter** is the proportion of driver dummies (if they had been real people) who would have suffered head injuries in the population.

The **statistic** is proportion of driver dummies (if they had been real people) who would have suffered head injuries in the sample.

The **variable** X = the number of driver dummies (if they had been real people) who would have suffered head injuries.

The **data** are either: yes, had head injury, or no, did not.

- **Example 7**.

You go to the supermarket and purchase three cans of soup (19 ounces tomato bisque, 14.1 ounces lentil, and 19 ounces Italian wedding), two packages of nuts (walnuts and peanuts), four different kinds of vegetable (broccoli, cauliflower, spinach, and carrots), and two desserts (16 ounces pistachio ice cream and 32 ounces chocolate chip cookies).

Name data sets that are quantitative discrete, quantitative continuous, and qualitative.

Solution 1.7

One Possible Solution:

- The three cans of soup, two packages of nuts, four kinds of vegetables and two desserts are quantitative discrete data because you count them.
- The weights of the soups (19 ounces, 14.1 ounces, 19 ounces) are quantitative continuous data because you measure weights as precisely as possible.
- Types of soups, nuts, vegetables and desserts are qualitative data because they are categorical.

Try to identify additional data sets in this example.

- **Example 8.**

The data are the colors of backpacks. Again, you sample the same five students. One student has a red backpack, two students have black backpacks, one student has a green backpack, and one student has a gray backpack. The colors red, black, black, green, and gray are qualitative data.

1.8 The data are the colors of houses. You sample five houses. The colors of the houses are white, yellow, white, red and white. What type of data is this?

- **Exercise 9.**

Work collaboratively to determine the correct data type (quantitative or qualitative). Indicate whether quantitative data are continuous or discrete. Hint: Data that are discrete often start with the words "the number of."

a. the number of pairs of shoes you own
b. the type of car you drive
c. the distance it is from your home to the nearest grocery store
d. the number of classes you take per school year.
e. the type of calculator you use
f. weights of sumo wrestlers
g. number of correct answers on a quiz
h. IQ scores (This may cause some discussion.)

Solution 1.9

Items a, d, and g are quantitative discrete; items c, f, and h are quantitative continuous; items b and e are qualitative, or categorical.