ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/339499419

Lecture notes on Biostatistics.

Book · February 2020

CITATIONS
0

READS
125,365

1 author:



Hamze ALI Abdillahi
Medical lecturer.
22 PUBLICATIONS **3** CITATIONS

SEE PROFILE

All content following this page was uploaded by Hamze ALI Abdillahi on 26 February 2020.

The user has requested enhancement of the downloaded file.

Dr-Hamze ALI ABDILLAHI

GOLLIS UNIVERSITY -ERIGAVO

Basic biostatistics

Introduction

- **Statistics:**

A field of study concerned with the collection, organization and summarization of data, and the drawing of inferences about a body of data when only part of the data are observed.

- **Biostatistics:** An application of statistical method to biological phenomena.

The science of assembling and interpreting numerical data (Bland 2000)

The discipline concerned with the treatment of numerical data derived from groups of individuals (Armitage et al.,2001)

Uses of Biostatistics

- •Hospital utility statistics
- •Resource allocation
- •Vaccination uptake
- •Magnitudes of a disease/condition
- •Assessing risk factors
- Disease frequency
- •Making diagnosis and choosing an appropriate treatment (implicit/probability).

Statistics can be used to:

1. Draw conclusions
2. Make predictions about what will happen in other subjects

Examples

1) **At Hargeisa general hospital, 5% of the patients were diagnosed with DM last year**
2. **Kat chewers are 3 times more likely to have MI than non-chewers**
3. **Antibiotics reduce the duration of viral throat infections by 1-2 days**

Medical research vs. Clinical Practice

- Data are collected from individual subjects
- Aim is to be able to make some general statements about a wider set of subjects that have been studied

- Data are collected from individual subjects
- Interested in the particular subjects

General steps in a research process

What does Biostatistics cover?

1. Planning
2. Design
3. Data collection
4. Data Processing
5. Data Presentation
6. Data Analysis
7. Interpretation
8. Publication

Population & Sample

- **Population:** is a complete set of items or subjects which can be studied
- **Target population:** A collection of items that have something in common for which we wish to draw conclusions at a particular time.
- **Study Population:** The specific population from which data are collected.
- **Sample:** A subset of the study population.

(A smaller part of that population)

Generalizability:

is a two-stage procedure: we want to generalize conclusions from the sample to the study population and then from the study population to the target population.

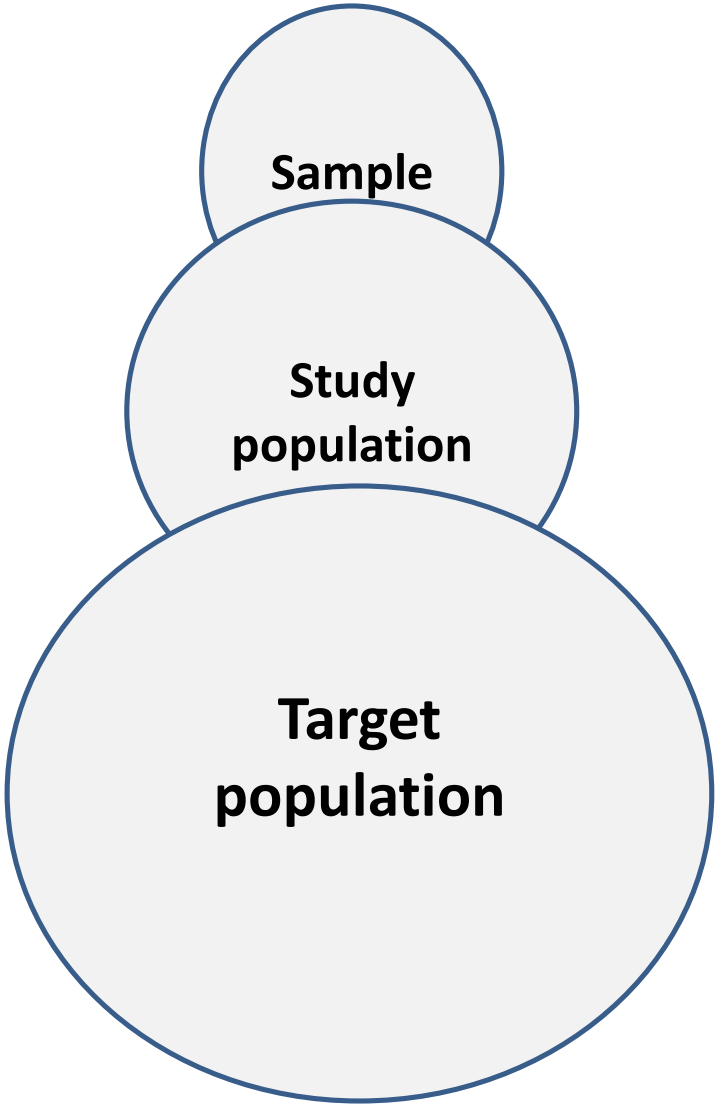
example

In a study of the prevalence of Kat chewing among secondary students in Somalia a random sample of Secondary students in Hargeisa were taken.

Target Population: All secondary students in Somalia

Study population: All secondary students in Somaliland

Sample: secondary students in Hargeisa


Sample
Study population
Target population

Parameter & Statistic

Parameter:

A descriptive measure computed from the data of a population. (Quantity calculated from population). E.g. mean serum glucose of the population is 100mg/dl

Statistic:

A descriptive measure computed from the data of a sample. (Quantity calculated from the sample). E.g. mean serum glucose of the sample is 110mg/dl

Scales of measurement (types of data)

- Clearly not all measurements are the same.
- Measuring an individuals weight is qualitatively different from measuring their response to some treatment on a three category of scale, “improved”, “stable”, “not improved”.
- Measuring scales are different according to the degree of precision involved.

Types of scales of measurement.

There are four types of scales of measurement:-

A. <u>QUALITATIVE DATA</u>:

1. Nominal scale: (can not be ordered)

uses names, labels, or symbols to assign each measurement to one of a limited number of categories that cannot be ordered.

Examples:

Blood type (A/B/AB/O) sex (Male/female) race (Somali/ Oromo) marital status (married/not married/ divorced). If there are only two possible categories the data is said to be **Dichotomous (e.g. Sex, male/female.**

2. Ordinal **scale** (categories can be placed in order): assigns each measurement to one of a limited number of categories that are ranked in terms of a graded order.

Examples:

- A questionnaire may ask respondents how happy they are with quality of services provided at the hospital, the choices can be: very happy, quite happy, unhappy, vey unhappy.

- **Degree of malnutrition**

  = mild, moderate, severe

- **Socio-economic status**

= upper, middle, lower

B. QUANTITATIVE DATA: (Numerical data).

Continuous data:

- **Interval scale**
- **Ratio scale**
- **Discrete (numbers)**

3. Interval scale (equally spaced intervals): assigns each measurement to one of an unlimited number of categories that are equally spaced. It has no true zero point.

Example:

body temperature measured on Celsius or Fahrenheit, heart rate measured per second. Thus the difference of interval between 5kg and 10kg is same as that between 20kg and 25kg.

These kind of measurement can be converted into dichotomous nominal scale e.g. afebrile (oral temp < 37) febrile (>37) also can be ordered (ordinal scale).

4.**Ratio scale:** measurement begins at a true zero point and the scale has equal space. Ratio data is similar to interval scales but it is the ratio of two measurements and also have a true zero.

Examples: Height per weight, blood pressure.

5. Discrete data: (numbers)

All values are clearly separated from each other, although numbers are used.

Examples: number of surgery operations performed in one month. Number of newly diagnosed psychiatric patients last year.

Variables

- ***Variable***: **A characteristic which takes different values in different persons, places, or things.**
- **Qualitative variable: The notion of magnitude is absent or implicit.**
- ***Quantitative variable***: **Variable that has magnitude.**
- **Discrete variable: It can only have a finite number of values in any given interval.**
- **Continuous variable: It can have an infinite number of possible values in any given interval.**

Data

The term DATA refers to (Items of information)

Systems for collecting data

1. **Regular system** (routine data collecting system): Registration of events as they become available.
2. **Ad hoc system** (non-routine): A form of survey to collect information that is not available on a regular basis.

Examples;

1. **Routine system:**

- **Census**: enumeration of all individuals in a country on a fixed day.
- **Vital registrations**: birth, deaths, marriage, divorce, ete.
- **Disease notification**: international notification, like cholera, national notification like polio, cholera, hepatitis = notification is from district level to national level to international level.
- **Disease registry:** TB, cancer, stroke, birth defects
- **Medical records:** schools, colleges, industries
- **Hospital records**
- **Environmental health records**

2. Non-routine

1. **Disease surveillance:** Polio, malaria, AIDS= it is important for control, prevention and eradication.
2. **Surveys**: nutritional status by interviewing examination or postal enquiry based.
3. **Social schemes:** medical insurance, sickness absenteeism, disability benefits, welfare schemes
4. **Economic data**: Consumption of goods, export and import, drugs, employment = helps panning commission for formulation of health policies
5. **Demographic data**: population movement, major epidemics

source of data

1.Primary data: collected from the items or individual respondents directly for the purpose of certain study**.**

2.Secondary data: which had been collected by certain people or agency, and statistically treated and the information contained in it is used for other purpose.

Biostatistics

methods of summarizing and displaying data

Summarizing and displaying data

- **Frequency tables**
- Histograms and polygons and Ogive
- Stem-and-leaf plots
- Box and Whisker plot
- Scatter plot
- Proportions or percentages
- Bar charts
- Pei charts

Measures of central tendency

- Mean
- Mode
- median

Measures of dispersion

- Variance
- Standard deviation
- Standard error
- Range
- Quartiles
- Coefficient of variation (CV)

Biostatistics

Presenting qualitative data

Charts and tables used to present qualitative data

1. **Pie charts**
2. **Bar charts** (**simple and clustered bar charts**)
3. Relative frequency (percentage) table

These **two charts** are used for presentation of qualitative data.

Pie charts

Pie charts are typically used to present the **relative frequency** of qualitative data.

In most cases the data are **nominal**, but **ordinal** data can also be displayed in a pie chart.

The complete circle represents the total number of measurements.

Partition into slices - one for each category.

The size of a slice is proportional to the relative frequency of that category.

Determine the angle of each slice by multiplying **the relative frequency** by 360 degree. (Recall a circle spans 360)

Steps to create a pie-chart

- Construct a frequency table
- Calculate relative frequency % (percentage)
- Change the percentages into degrees, where: degree = Percentage X 360^{o}.
- Draw a circle and divide it accordingly

For single variable:

For example in a class of 40 students, 15 are boys and 25 are girls. (See the pie chart)

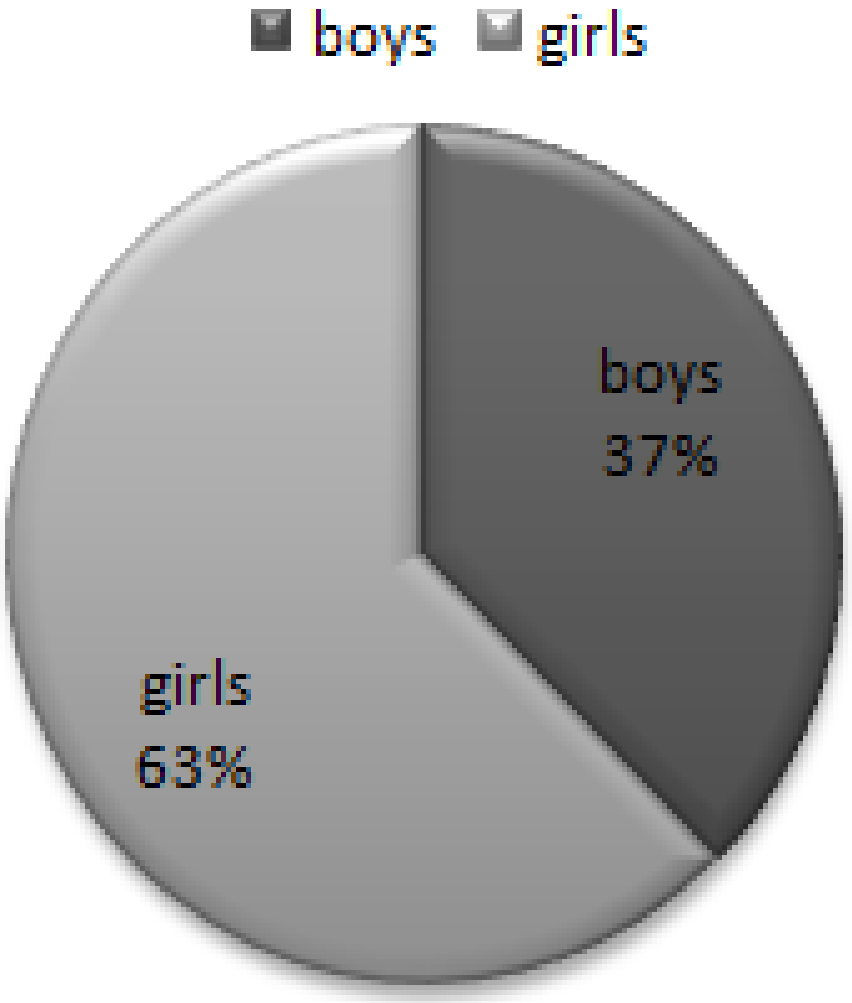
Frequency: number of times that something occurs.

Relative frequency = frequency divide by sum of all frequencies

$$\textit{Relative frequency} = \frac{\textit{Frequency}}{\textit{Sum of all frequencies}}$$

Gender	Frequency	Relative frequency %
Boys	15	$\frac{15}{40} * 100 = 37.5\%$
Girls	25	$\frac{25}{40} * 100 = 62.5\%$
	40	

frequency distribution for gender



Angle computations:

Since a circle has 360 degrees, the degree measure of the sector for the category will be:

0.375*360 = 135

0.625*360 = 225

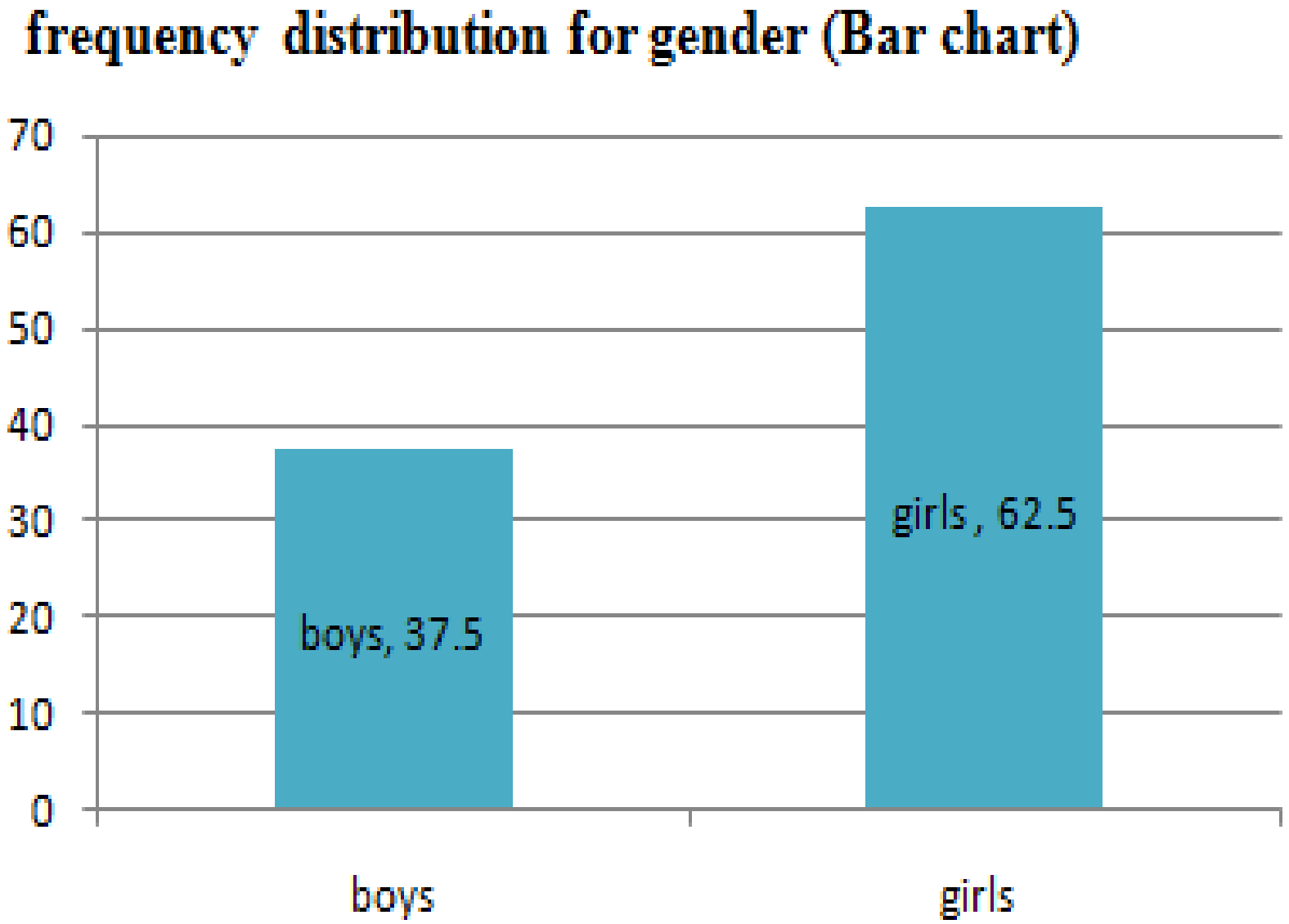
Total = 360

Bar Chart (Bar Graph):

- **Place categories on the horizontal axis.**
- **Place frequency (or relative frequency)** on the vertical axis.
- **Construct vertical bars of equal width,** one for each category.

Its **height** is proportional to the frequency (or **relative frequency**) of the category.

Simple bar chart



Two variables (cross tabulation)

Cross tabulation or cross tabs are often used in presenting the counts of two qualitative variables.

Suppose the variables of interest are :

- Gender and
- wearing spectacles.

The are presented in this table.

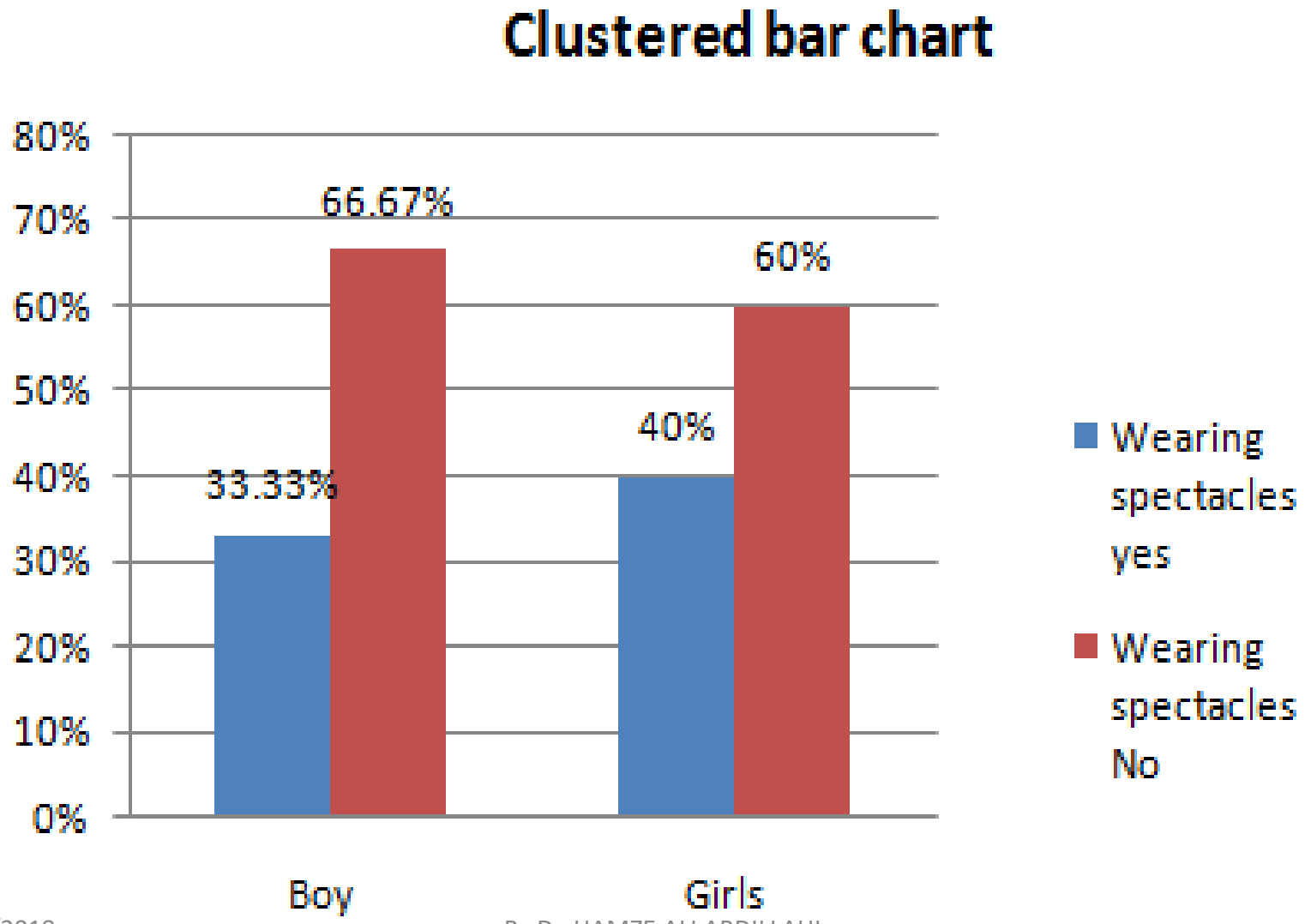
	Wearing spectacles		Total
	yes	No	
Boy	5	10	15
Girls	10	15	25
Total	15	25	40

Two variables (qualitative)
We cross tabulation

	Wearing spectacles		Total
	yes	**No**	
Boy	5	10	15
Girls	10	15	25
Total	15	25	40

	Wearing spectacles		Total
	yes	**No**	
Boy	33.33%	66.67%	100%
Girls	40%	60%	100%
Total	37.50%	62.50%	100%

Table showing the percentage of Gender and wearing spectacles.



Crosstabs and clustered bar chart

Expressed in percentage. 33.33% of the boys and 40% of the girls wear spectacles

Calculate the percentages

Smoking	Lung cancer		Total
	YES	NO	
YES	70	100	
NO	3	70	

BIOSTATISTICS

Methods of Displaying and Summarizing quantitative data

Summarizing and displaying data

- **Frequency tables**
- Histograms and polygons and Ogive
- Stem-and-leaf plots
- Box and Whisker plot
- Scatter plot

Measures of central tendency

- Mean
- Mode
- median

Measures of dispersion

- Variance
- Standard deviation
- Standard error
- Range
- Quartiles
- Coefficient of variation (CV)

Frequencies and frequency distribution tables:

Frequency distribution: is a table showing a listing of all observed values of the variable being studied and how many times each value is observed.

The number of times that something occurs is known as its ***frequency***.

The **notation** *fx* is used to denote the *frequency* or number of times the value *x* occurs.

The *relative frequency* is just the frequency divided by the sample size *n*.

Table: obtaining frequency, cumulative frequency and percentage

Age	***Frequency***	**Cumulative frequency**	**Relative Frequency %**	**Cumulative relative frequency %**
13	1	1	3	3
14	7	8	23	26
15	5	13	17	43
16	6	19	20	63
17	6	25	20	83
18	2	27	7	90
19	3	**30**	10	**100**
Total	**30**		**100**	

Computing Relative frequency

Frequency: number of times that something occurs.

Relative frequency = frequency divide by sum of all frequencies

$$\textit{Relative frequency} = \frac{\textit{Frequency}}{\textit{Sum of all frequencies}}$$

Cumulative frequency: frequencies are added up.

- **For example** 1/30*100= 3% and 7/30*100 =23%
- **Cumulative relative frequency:** sums of all relative frequencies below and including each category

Steps in constructing the **frequency distribution table** for **quantitative data:-**

1. Data are first divided into a number of intervals.
2. Then the number of data points falling within each interval is presented as the frequency or count for that interval.
3. Tally the data in the tally column and obtain the class frequencies.

Smoothing class intervals to obtain **Δ** = (class boundaries)

$$\Delta = \frac{\textit{(Upper limit of first class - lower limit of second class)}}{2}$$

- **Subtract Δ** from the first class limits to get the lower class boundaries
- **Add Δ** to the upper class limits to get the upper class boundaries

Sturge's rule: $K = 1+3.322(\log n)$

$$C = \frac{R}{K}$$

Where K = number of class intervals n = number of observations

C = class width

R (range) = minimum value – maximum value.

The beginning and end of each interval are called *boundaries* or *class interval* and the point midway between any two boundaries is called the *class mark* or *midpoint*.

For example: table: Body Mass Index Data for a Sample of 120 U.S. Adults: **Ordered Array**

18.3	21.9	23.0	24.3	25.4	26.6	27.5	28.8	30.9	34.4
19.2	21.9	23.1	24.3	25.6	26.9	27.5	28.8	30.9	34.9
19.8	21.9	23.1	24.5	25.7	27.1	27.6	28.9	31.0	35.0
20.2	22.3	23.3	24.6	25.7	27.3	28.2	29.3	31.1	35.5
20.7	22.3	23.4	24.6	25.8	27.3	28.3	29.5	31.3	35.8
20.8	22.3	23.5	24.7	25.8	27.3	28.3	29.8	31.6	35.9
21.1	22.4	24.0	24.7	25.9	27.3	28.3	30.0	31.6	36.6
21.1	22.5	24.0	24.8	25.9	27.4	28.4	30.1	32.6	37.1
21.1	22.7	24.0	24.8	26.2	27.4	28.6	30.2	32.8	37.5
21.3	22.7	24.1	25.0	26.5	27.4	28.7	30.3	33.2	37.8
21.3	22.8	24.1	25.4	26.5	27.4	28.7	30.8	33.6	38.2
21.5	22.9	24.2	25.4	26.5	27.4	28.8	30.8	34.2	38.8

Usually, for a data set of **100 to 150** observations, the number chosen ranges from about **5 to 10**.

In our example, the **range** of the data is **38.8 – 18.3 = 20.5.** Suppose we divide the data set into **seven** intervals. Then, we have **20.5 ÷ 7 = 2.93,** which rounds to **3.0.** So the intervals have a width of 3.

These seven intervals are as follows:

- **18.0 – 20.9**
- **21.0 – 23.9**
- **24.0 – 26.9**
- **27.0 – 29.9**
- **30.0 – 32.9**
- **33.0 – 35.9**
- **36.0 – 38.9**

Frequency Distribution table

Class Interval for BMI levels	*Frequency (f)*	*Cumulative Frequency (cf)*	*Relative Frequency (%)*	*Cumulative Relative Frequency (%)*
18.0 – 20.9	**6**	**6**	**5.00**	**5.00**
21.0 – 23.9	**24**	**30**	**20.00**	**25.00**
24.0 – 26.9	**32**	**62**	**26.67**	**51.67**
27.0 – 29.9	**28**	**90**	**23.33**	**75**
30.0 – 32.9	**15**	**105**	**12.50**	**87.50**
33.0 – 35.9	**9**	**114**	**7.50**	**95.00**
36.0 – 38.9	**6**	**120**	**5.00**	**100.00**
Total	**120**		**100.00**	**100.00**

Graphs for displaying quantitative data include:

- **Histogram**
- **Frequency Polygon and Ogive**
- **Stem-and-leaf plot**
- **Box and Whisker plot (** used when we are constructing **quartiles**)
- **Scatter plot** (used in correlation and regression analysis

Histogram & frequency polygons:

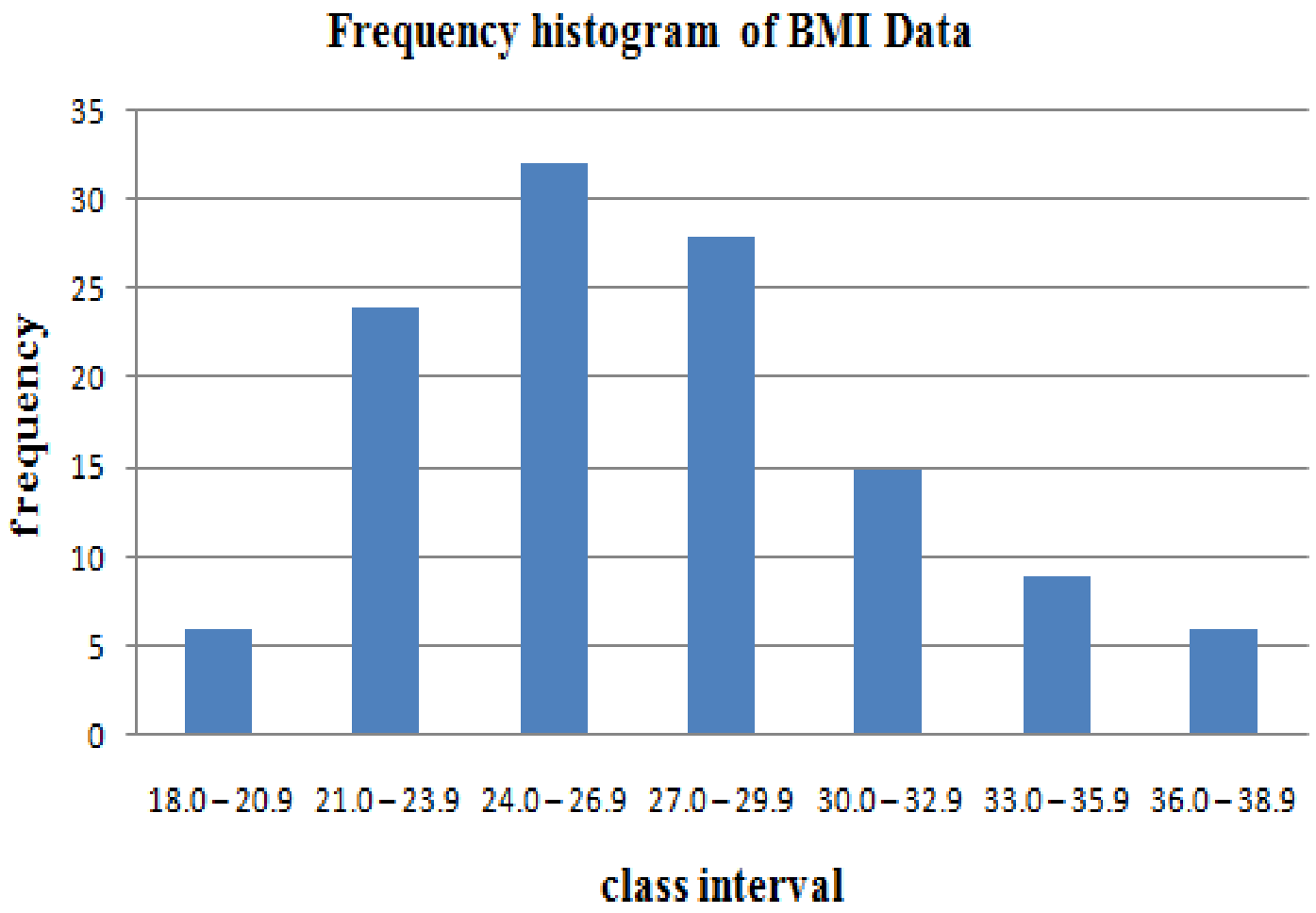
Frequency distributions are often displayed with **a histogram**, which looks like a bar chart but there is no space between bars. The heights of the bars represent either the number or percent of observations within each interval.

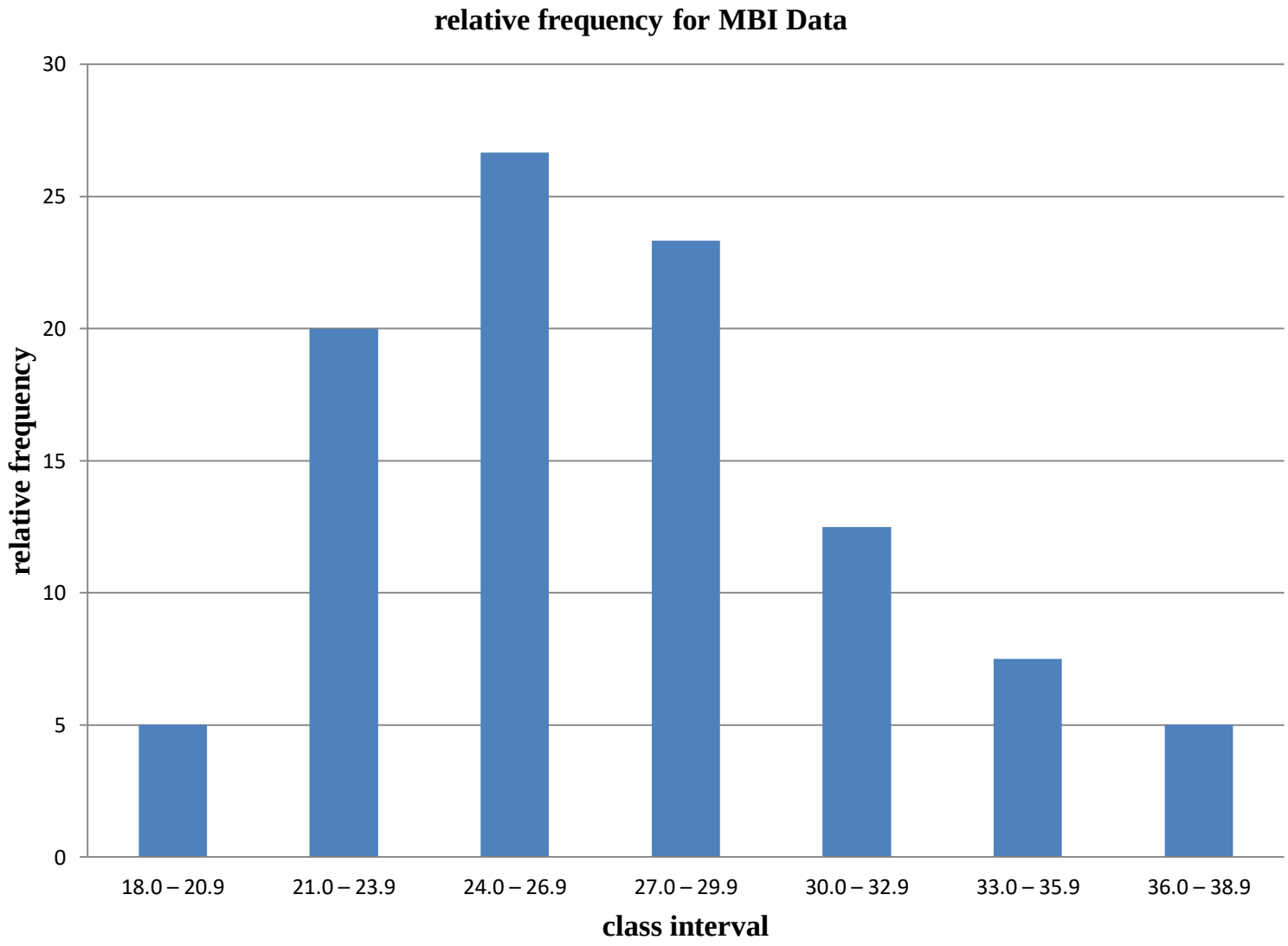
Frequency polygons, which are essentially a line that connects the middle of each of the bars of the histogram, are also used extensively.

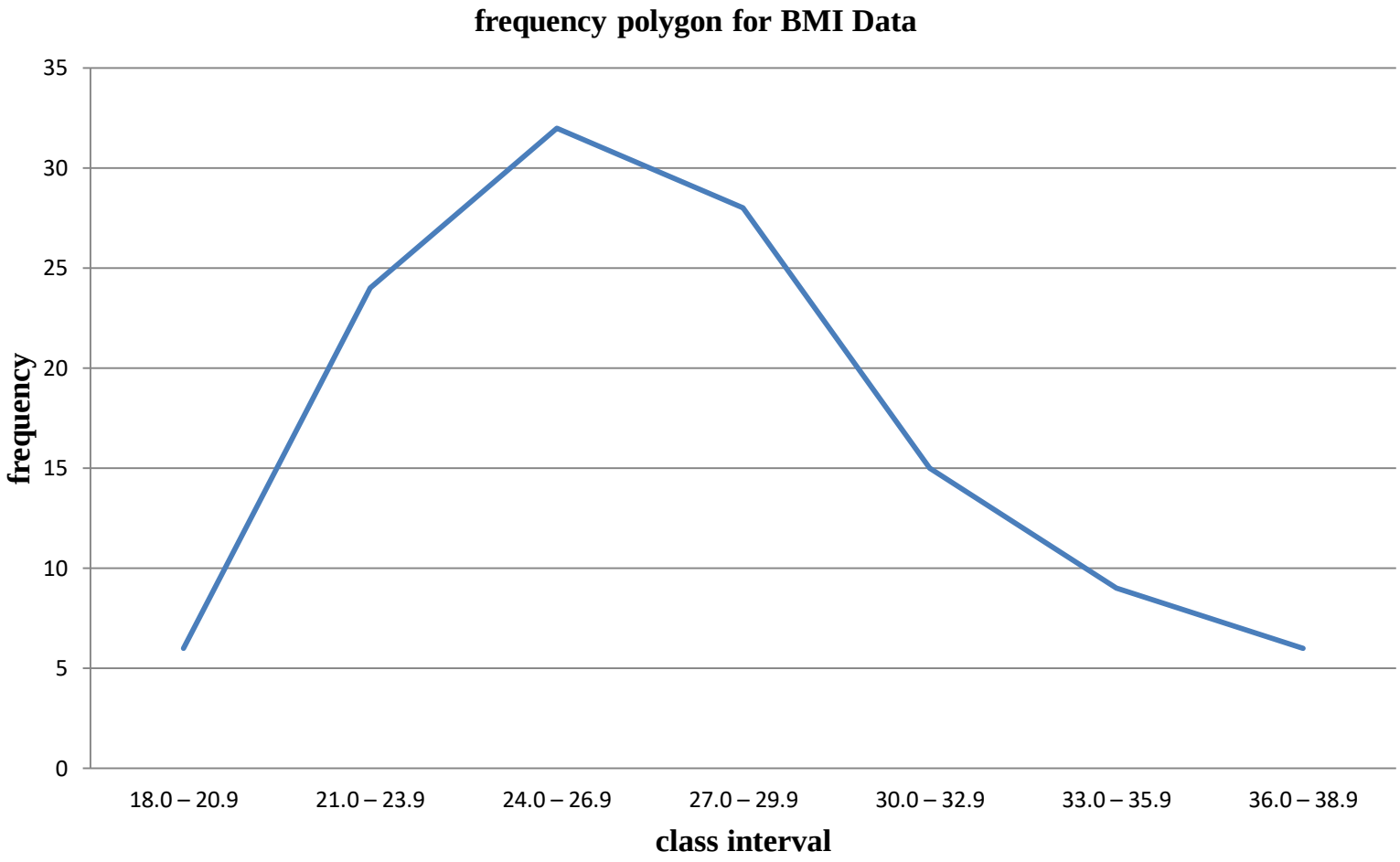
To construct a histogram

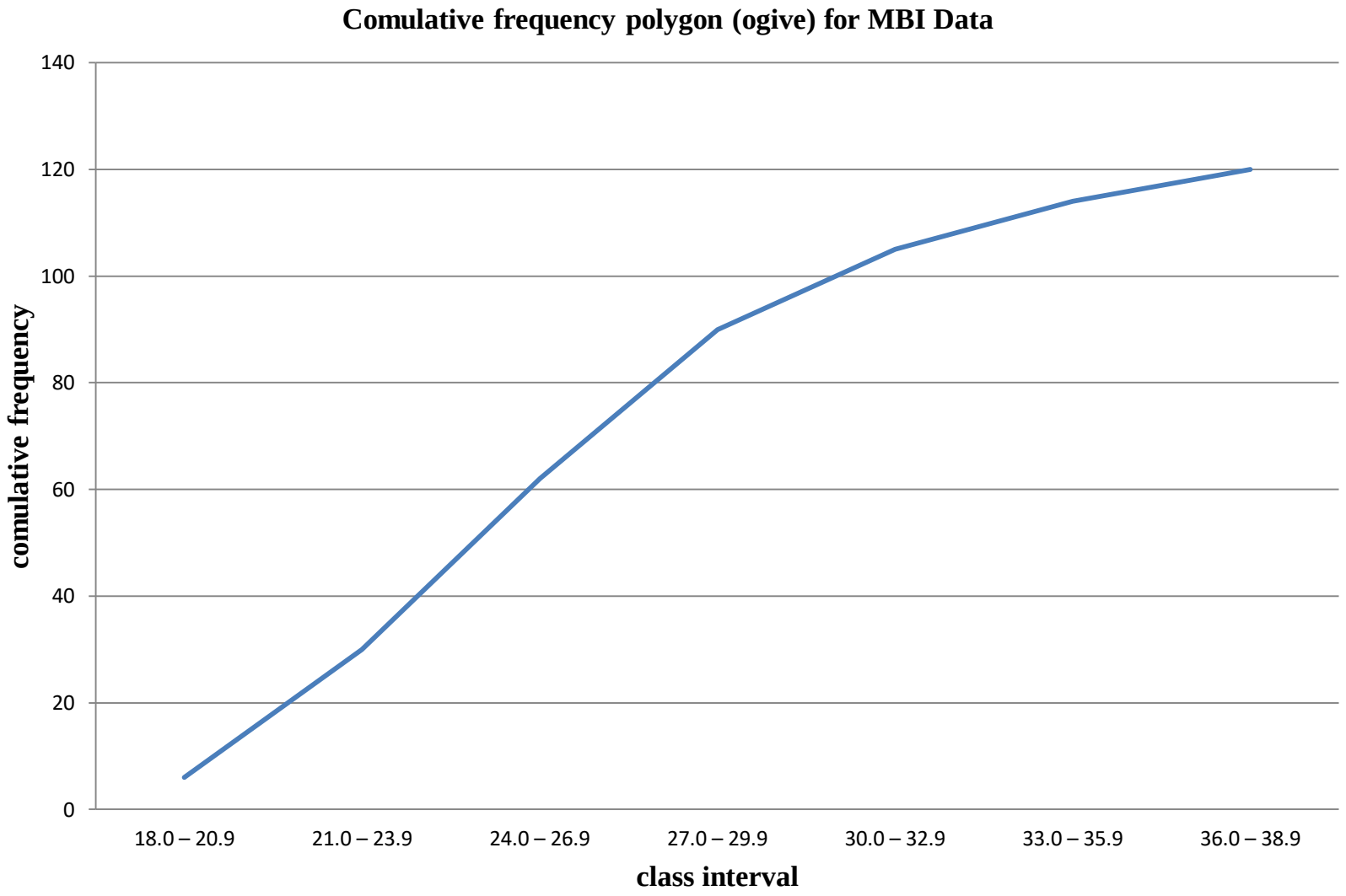
- Draw the interval boundaries on a horizontal line and the frequencies on a vertical line.
- Non-overlapping intervals that cover all of the data values must be used.
- Bars are then drawn over the intervals in such a way that the areas of the bars are all proportional in the same way to their interval frequencies.

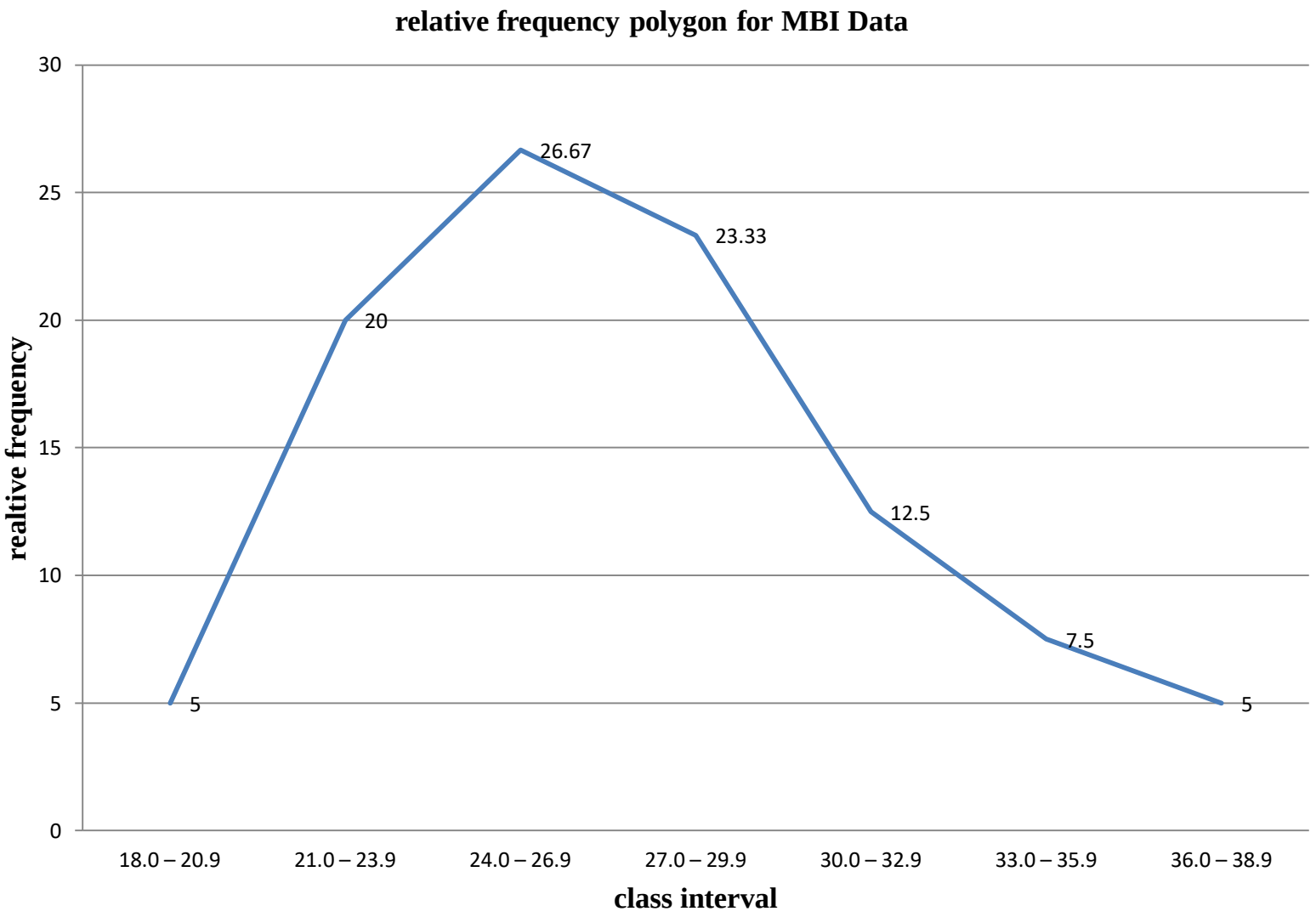
Using the above data we can contract histogram and polygon using **Excel.**


Frequency histogram of BMI Data
frequency
35
30
25
20
15
10
5
0
18.0 – 20.9
21.0 – 23.9
24.0 – 26.9
27.0 – 29.9
30.0 – 32.9
33.0 – 35.9
36.0 – 38.9
class interval


relative frequency for MBI Data
30
25
20
15
10
5
0
relative frequency
18.0 – 20.9
21.0 – 23.9
24.0 – 26.9
27.0 – 29.9
30.0 – 32.9
33.0 – 35.9
36.0 – 38.9
class interval


frequency polygon for BMI Data
35
30
25
20
15
10
5
0
frequency
18.0 – 20.9
21.0 – 23.9
24.0 – 26.9
27.0 – 29.9
30.0 – 32.9
33.0 – 35.9
36.0 – 38.9
class interval

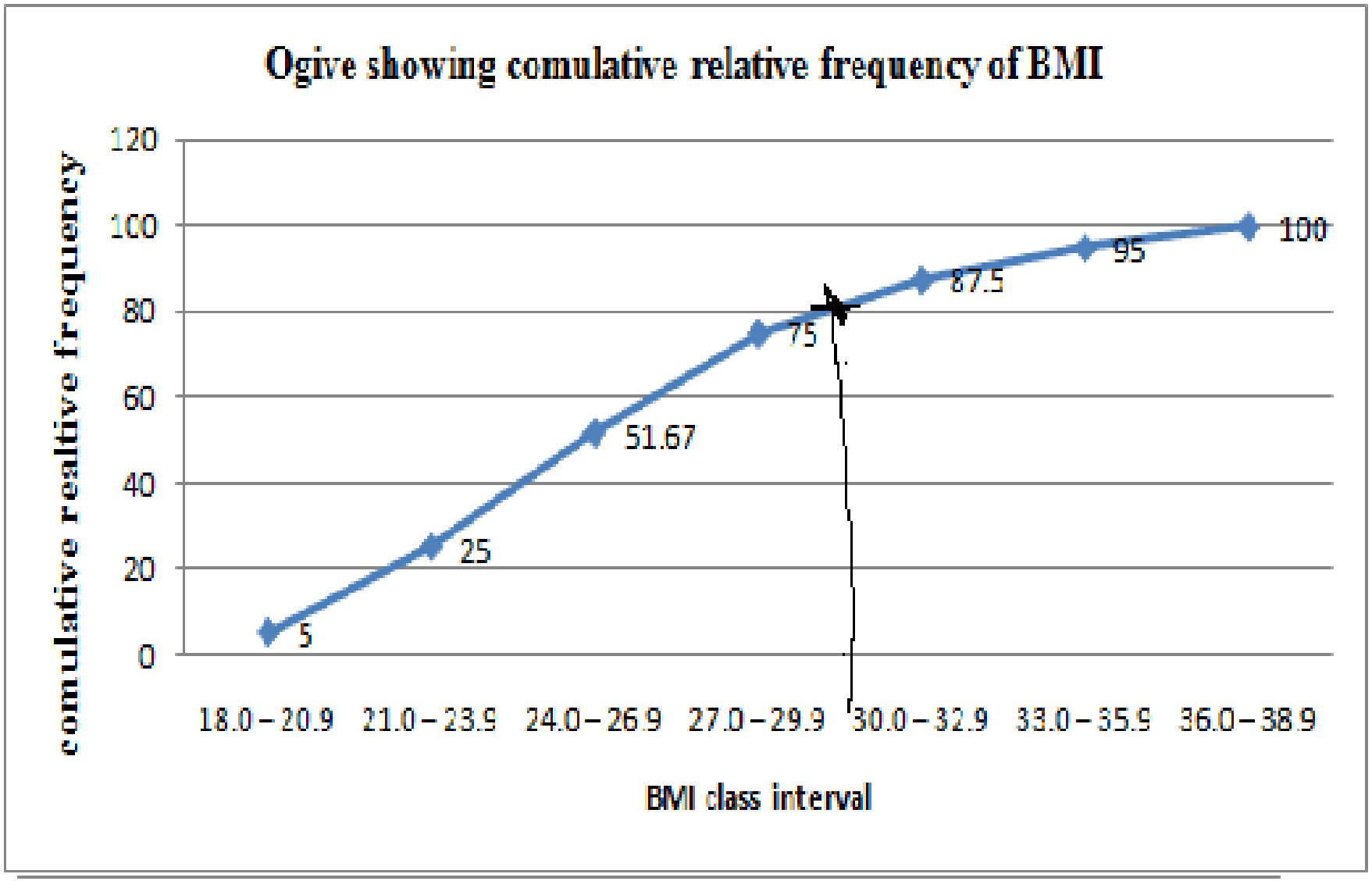

Comulative frequency polygon (ogive) for MBI Data
140
120
100
80
60
40
20
0
comulative frequency
18.0 – 20.9
21.0 – 23.9
24.0 – 26.9
27.0 – 29.9
30.0 – 32.9
33.0 – 35.9
36.0 – 38.9
class interval


relative frequency polygon for MBI Data
realtive frequency
30
25
20
15
10
5
0
26.67
23.33
20
12.5
7.5
5
5
18.0 – 20.9
21.0 – 23.9
24.0 – 26.9
27.0 – 29.9
30.0 – 32.9
33.0 – 35.9
36.0 – 38.9
class interval

Cumulative relative frequency using Ogive

Another way of representing of quantitative data is the Ogive which is the graphical presentation of the commutative relative frequency. Sometimes it may become necessary to know **the number of items whose values are more or less than a certain amount.** We can use Ogive to estimate the cumulative relative frequencies of other values.

For example 80% of the respondents have a BMI less than 30.


Ogive showing comulative relative frequency of BMI
comulative realtive frequency
120
100
80
60
40
20
0
5
25
51.67
75
87.5
95
100
18.0 – 20.9
21.0 – 23.9
24.0 – 26.9
27.0 – 29.9
30.0 – 32.9
33.0 – 35.9
36.0 – 38.9
BMI class interval

Stem-and-leaf plot

Example 4: HbA1c from diabetic patients (in %)

7.1 8.0 7.2 7.5 6.4

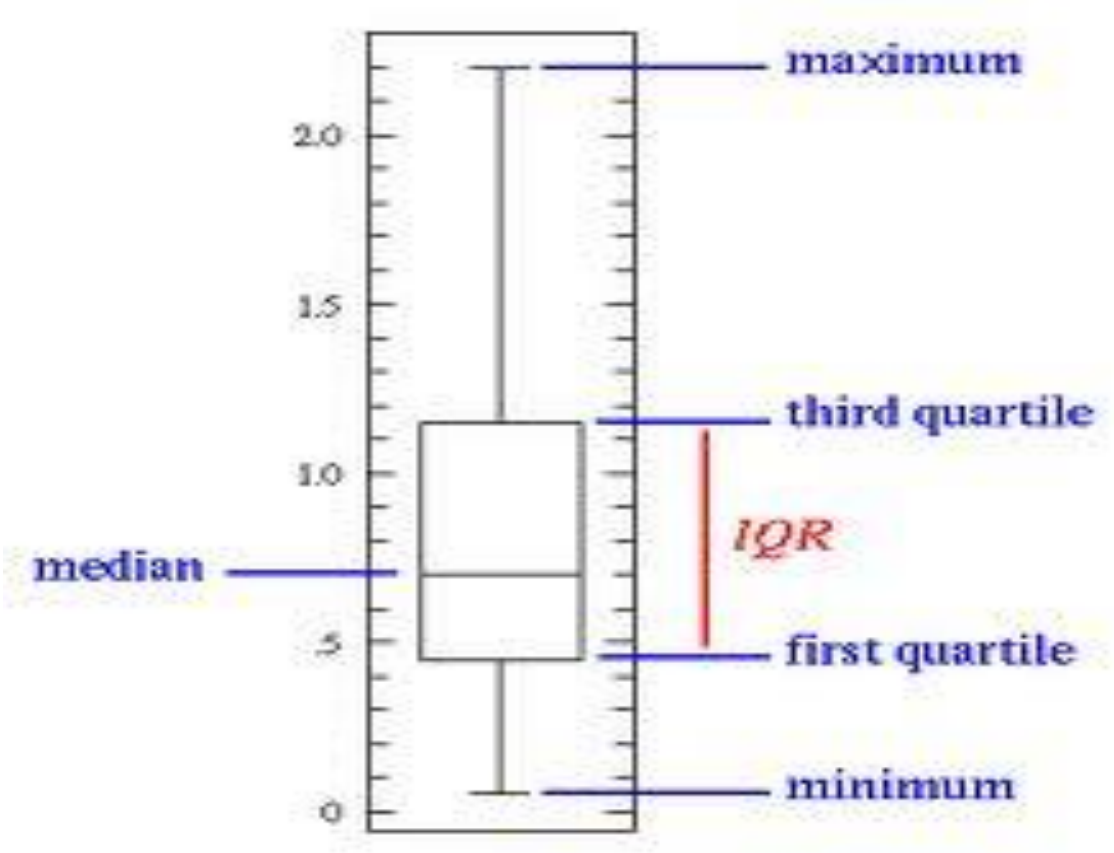
6.8 8.2 9.1 7.8 8.1

Stem	Leaf
6	4 8
7	1 2 5 8
8	0 1 2
9	1

Advantages of Stem-and-leaf plot:

- Orders the data, so that the maximum and minimum are evident
- Gaps in the data become evident
- All the data is displayed
- The shape of the data becomes clearer

Box and Whisker plot



Box and Whisker plot

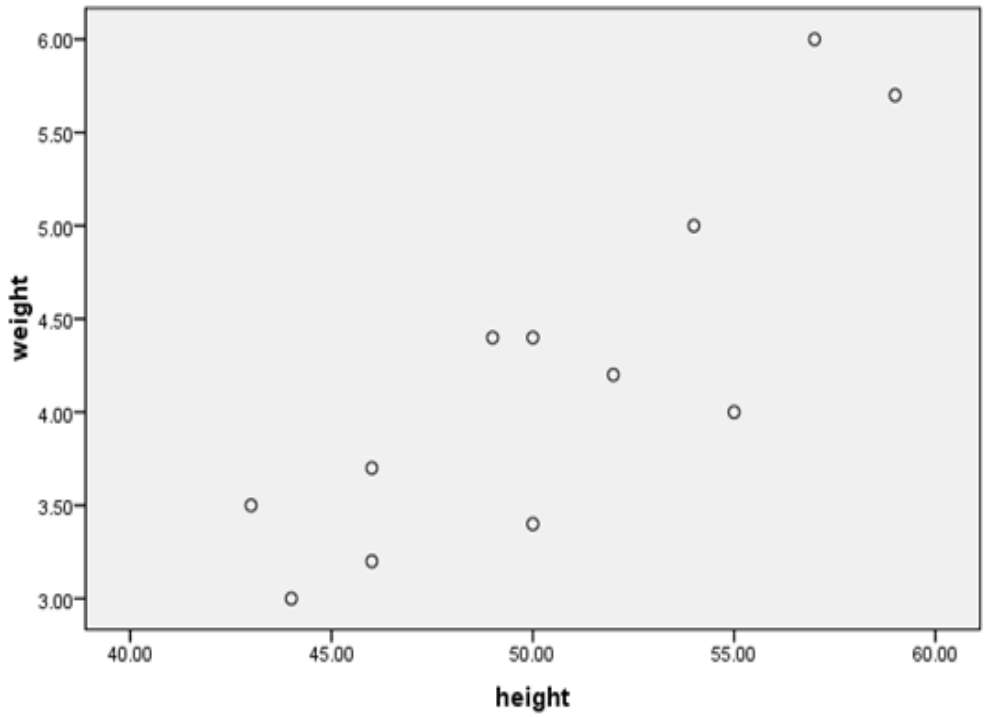
It is another way to display information when the objective is to illustrate certain locations in the distribution. A box plot is a good alternative or complement to a histogram and is usually better for showing several simultaneous comparisons.

It is useful for the detection of outliers.

It displays median, minimum, maximum first quartile (Q1) third quartile (Q3) and inter-quartile range (IQR).

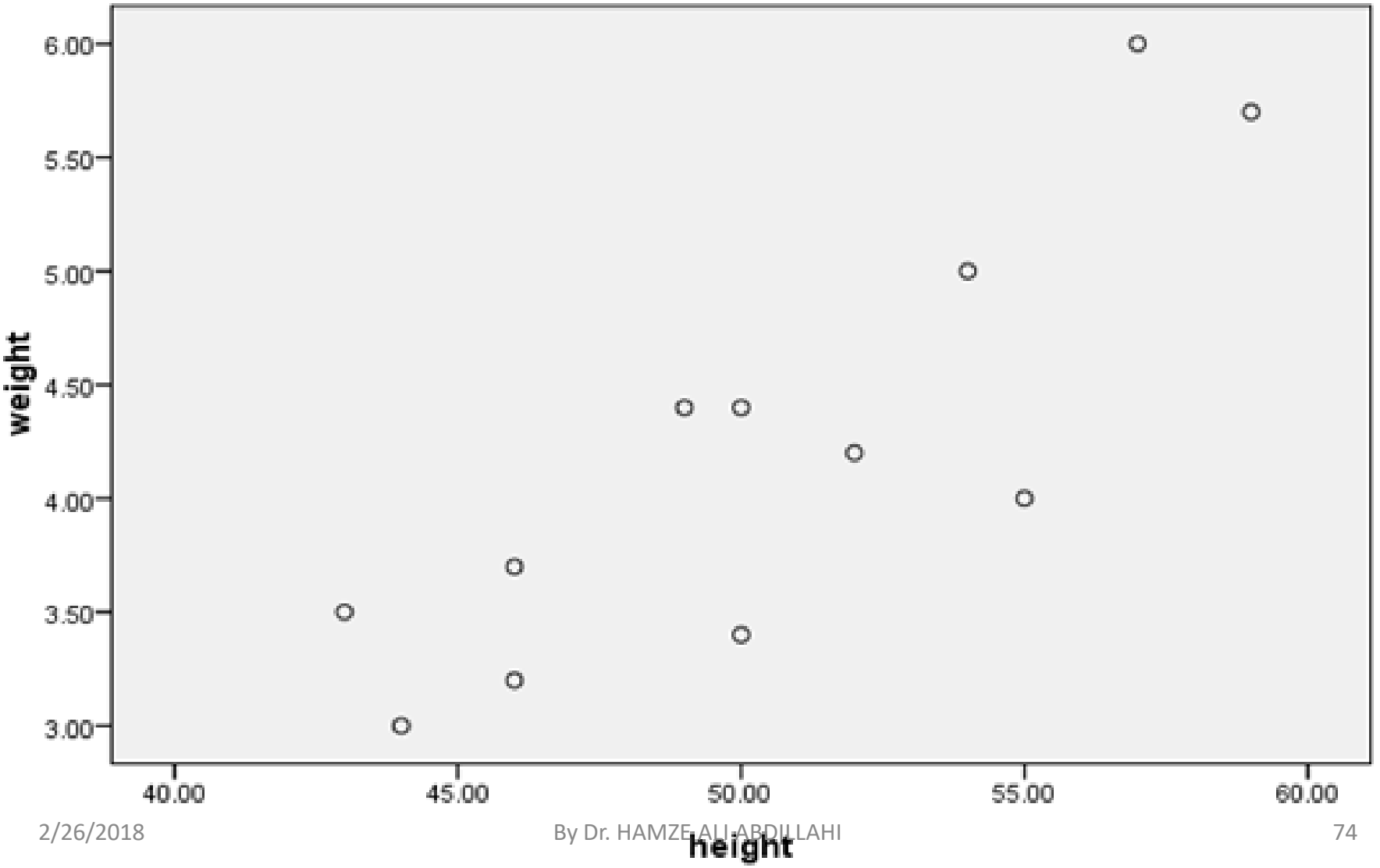
1. A box is drawn with the top of the box at the third quartile and the bottom at the first quartile.
2. The location of the mid-point of the distribution is indicated with a horizontal line in the box, which the median or the (Q2)
3. Finally, straight lines, or **whiskers**, are drawn from the centre of the top of the box to the largest observation and from the centre of the bottom of the box to the smallest observation

Scatter plot



To illustrate the relationship between two characteristics when both are quantitative variables we use bivariate plots (also called scatter plots or scatter diagrams).

Scatter plot showing height and weight of newborn babies



Summation notation

Summation notation is simply way of saying that a collection of numbers is to be added.

Generally, some letter is used is to represent whatever is being measured; the letter X is the most common choice.

The notation $X1$ is used to indicate the first observation.

The next observation is X_2, and so on.... Generally, n is typically used to represent the total number of observations, and the observations themselves are represented by $X1, X2, \ldots, Xn$.

In symbols, adding the numbers $X1, X2, \ldots, Xn$ is denoted by

$$\sum_{i=1}^{n} xi$$

Where Xi = X_1 +X_2+· · ·+Xn,

Where $\sum$ is an upper case Greek sigma. The subscript *i* is the *index of summation* and the 1 and *n* that appear respectively below and above the symbol $\sum$ designate the range of the summation.

The i is where the X values start and the n is where the values end.

Sometimes, the sum extends over all *n* observations, in which case it is customary to omit the index of summation. That is, simply use the notation

$$\sum Xi = X1 + X2 + \cdots + Xn.$$

For example:

1.2, 2.2, 6.4, 3.8, 0.9.

$$\sum_{i=2}^{4} xi$$

Then the

$$= 2.2+6.4+3.8 = 12.4$$

And $\sum Xi = 1.2+2.2+6.4+3.8+0.9 = 14.5.$

Another common arithmetic operation is squaring each observed value and summing the results.

This is written as:

$$\sum X^2_i = X^2_1 + X^2_2 + \cdots + X^2_n$$

The adding of all the values and squaring them, is written as:

$$\sum (X_i)^2$$

For example

$$\sum X^2_i = 1.2^2 + 2.2^2 + 6.4^2 + 3.8^2 + 0.9^2 = 62.49$$

$$(\sum X_i)^2 = (1.2+2.2+6.4+3.8+0.9)^2 = 14.5^2 = 210.25$$

Let *c* be any constant. In some situations it helps to note that multiplying each value by *c* and adding the results is the same as first computing the sum and then multiplying by *c*. This is written as:

$$\sum cX_i = c\sum X_i$$

For example

$$\sum 60X_i = 60\sum Xi = 60\times 14.5 = 870.$$

Another common operation is to subtract a constant from each observed value, square each difference, and add the results. In summation notation, **this is written as:**

$$\sum (Xi - c)^2.$$

For example:

For example, suppose we want to subtract 2.9 from each value, square each of the results, and then sum these squared differences.

So $c = 2.9$, and

$$\sum(Xi - c)^2 = (1.2-2.9)^2 + (2.2-2.9)^2 + \cdots + (0.9-2.9)^2 = 20.44.$$

Basic Biostatistics

Measures of central tendency

Measures of central tendency

1. **Mean** - average (*arithmetic mean)*
2. **Median** - middle value
3. **Mode** - most frequently observed value(s).

Means, medians, and modes are methods of measuring the central tendency of a group of values- that is, the tendency for values in a group to gather around a central or average value which is typical of the group.

To avoid biased reporting central tendency must be addressed collectively, based on all the three measures mean, median, mode.

Formulas for Mean: (*arithmetic mean)*

$$\bar{X} = \frac{\sum X_i}{n}$$

$$\bar{X} = \frac{\sum Xi}{n}$$

Mean

$$\bar{X} = \frac{\sum Xi}{n}$$

The mean is the sum of all the values in a data set, divided by the number of values. The mean of a whole population is usually denoted by **μ**, (*called mu)* while the mean of a sample is usually denoted by $\bar{x}$ *called x-bar).*

To calculate the mean:

$$\bar{X} = \frac{\sum Xi}{n}$$

- ✓ Sum up all the values.
- ✓ Divide the sum by the number of values.

Mean is a simple **point-estimate** for the **population mean,** which is just the average of the data collected. The mean is very sensitive to outliers and the estimate can be biased in the presence of extreme values. Unlike the median and mode, where a change to an extreme value usually has no effect

Mean of the ungrouped data:

Example:

The results of **HbA1c** of patients with diabetes is; 4.0, 5.4, 4.6, 6.0.

Calculate the mean of the data?

$$\text{Mean} = \frac{\text{Sum of all data values}}{\text{Number of data values}}$$

Symbolically,

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

where $\bar{x}$ (read as 'x bar') is the mean of the set of x values,

$\sum x$ is the sum of all the x values, and

n is the number of x values.

Result

$$\text{Mean} = \frac{(4.0+5.4+4.6+6.0)}{4} = 20/4 = 5$$

The mean of the **HbA1c is = 5.** Remember that when writing the mean, it is good practice to refer to the unit of measured; in this case it is an HbA1c value of 5%.

Example 2

- Data set is 4, 7, 5, 9, 5.

Calculate the mean?

- Data set is 10, 12, 16,14.

Calculate the mean?

Result

$$M = \frac{4+7+5+9+5}{5} = 6$$

$$M = \frac{10+12+16+14}{4} = 13$$

Mean of the grouped data

In calculating the mean from grouped data, we assume that all values falling into a particular class interval are located at the mid-point of the interval. It is calculated as follow:

$$M = \frac{\sum mi\,fi}{\sum fi}$$

Example:

Age	fi	mi	mifi
15-19	11	17	187
20-24	36	22	792
25-29	28	27	756
30-34	13	32	416
35-39	7	37	259
40-44	3	42	126
45-49	2	47	94
Total	**100**		**2630**

Where

k = the number of class intervals

mi = the mid-point of the ith class interval

fi = the frequency of the ith class interval

Mean = 2630/100 = 26.3

Trimmed mean

It trims all but one or two values.

No specific amount of trimming is always best, but 20% trimming is often a good choice in the literature. This means that the smallest 20%, as well as the largest 20%, are trimmed and the average of the remaining data is computed. Although there are circumstances where this extreme amount of trimming can be beneficial, but sometimes this extreme amount of trimming can be detrimental.

Computation of trimmed mean:

- **first compute 0.2*n**
- **Round down to the nearest number.**
- **call this result g,**

The formula of 20% trimmed mean is given by:

$$\bar{X}_t = \frac{1}{n-2g}\left(X_{(g+1)} + \cdots + X_{(n-g)}\right)$$

Example

Data values are:

46,12,33,15,29,19,4,24,11,31,38,69,10

Calculate the trimmed mean?.

Ordered data:

4,10,11,12,15,19,24,29,31,33,38,46,69.

The number of values is n = 13, 0.2(n) = 0.2(13) = 2.6,

•Rounding this down to the nearest integer yields g = 2.

•That is, trim the two smallest values, 4 and 10, trim the two largest values, 46 and 69

•Average the numbers that remain yielding.

$$M\,t = \frac{1}{9}(11+12+15+19+24+29+31+33+38) = 23.56.$$

Median

It is the second measure, is the middle number of a set of numbers arranged in numerical order.

To calculate the median of the ungrouped data?

- First arrange the values in order of size and then find the middle value.
- If the number of observations, ***n,*** is even, Then location of the sample median is, ***m=n/2.*** Then the median is the two middle numbers divided by **2.** Or we can use the formula $m = (n+1)/2$ for both odd an even.
- If the number of observations, ***n,*** is odd, Then the location of the sample median is ***m = (n+1)/2.***

Finding the location of the median

Median **= (n+1)/2**

Example1

Median of the Ungrouped data

Find the median of (13, 3, 20, 22, and 25)

Ordered data: 3, 13, 20, 22, and 25. The **median** = n+1/2 = 5+1/2 = 3 so the location of the median is third data value which is = 20

Example 2

If there is an even number of values, use the mean of the two middle values. For example the values 3, 13, 13, 20, 22, 25: **median** = n+1/2 = 6+1/2 = 3.5, so the median lies between number 3 and 4.

Median = **(13 + 20)/2** = 16.5. It is the **point** that divides a distribution of scores into two equal halves

Median of the Grouped data

$$M = Lm + \left(\frac{n/2 - Fc}{Fm} \right) W$$

1. **Lm**= lower true class boundary of the interval containing the median.
2. **Fc** = cumulative frequency of the interval just above the median class interval.
3. **Fm** = frequency of the interval containing the median
4. W= class interval width.
5. **n** = total number of observations

Example:

$$M = Lm + \left(\frac{n/2 - Fc}{Fm} \right) W$$

Age	fi	Cum. F
5-14	5	5
15-24	10	15
25-34	20	**35**
35-44	**22**	57
45-54	13	70
55-64	5	75

- **n/2** = 75/2 = 37.5.
- **Median class interval** = 35-44.
- **Lm**=34.5.
- **Fc** = 35
- **W** = 10.
- **n** = 75.
- **Fm** =22.
- **Median** = 34.5 + (37.5-35)/22 x 10 = 35.64

The mean versus the median

- The mean is sensitive to outliers
- The median is not sensitive to outliers
- When the data are highly skewed, the median is usually preferred
- When the data are not skewed, the median and the mean will be very close

Mode

The last measure is the **mode**, which is the most frequent occurring number.

Example: 3, 13, 13, 20, 22, 25: the **mode** = 13. It is usually more informative to quote the mode accompanied by the percentage of times it happened; e.g, the **mode** is 13 with 33% of the occurrences. In medical research, mean and median are usually presented. A set can have more than one mode; if it has two, it is said to be bimodal.

Example

Data values:

Ordered data : 1,1,3,3,4,5, 60

The mean is : 77/7 = 11

$$\textbf{Median is} = \frac{(n+1)}{2} = \frac{7+1}{2} = 4 \textit{ (location)}$$

So the median is the fourth data value , **m = 3**

Mode = most frequent number in the data set

Which is = **1 & 3** , so the **mode is bimoda**l

Mode of the grouped data

$$\text{Mode} = Lo + \frac{D_1}{D_1 + D_2} * C_o$$

Lo = the lower boundary of the modal class

D1 = difference in frequency between modal class and the one before

D2 = difference in frequency between modal class and the one after

Co = the width of the modal class

Note , the modal class is the one that contains the highest frequency

Example

class	mi (midpoint)	fi	fc
9.5 – 13.5	11.5	3	3
13.5 – 17.5	15.5	4	7
17.5 – 21.5	19.5	8	15
21.5 – 25.5	23.5	3	18
25.5 – 29.5	27.5	2	20
Sum		**20**	

Calculate :

Mode , mean and median of the data.

Mode, the third class has the largest frequency = 8

So the class (17.5-21.5) is the modal class.

For the modal class , Lo = 17.5, D1 = (8-4) = 4

D2 = (8-3) 5 and Co = (21.5 -17.5) = 4

So the mode = 17.5 + (4/4+5)

Calculate the: mean and median

Result

- ✓ Mean = 378/20 = 18.9
- ✓ Median = 19

Measures of dispersion

1. **Range**
2. **Variation (SS)** the sum of squared deviation from the mean.
3. **Variance (S^2)**
4. **Standard deviation (S)**
5. **Standard error (SE)**
6. **Quartiles and inter quartile range (QR)**
7. **Coefficient of variation (CV)**

Range

Is the difference between the maximum and the minimum data values.

$R = X_L - X_S$, where X_L = is the largest value and X_S = is the smallest value.

It is the simplest measure and can be easily understood. It takes into account only two values which causes it to be a poor measure of dispersion. One application is in quality control charts, especially when small sample sizes are involved.

For example:

data set: 4, 5, 6 , 7, 14

The maximum value is 14 and minimum value is 4

So, the range is 14-4 = 10

Variation (SS) the sum of squared deviation from the mean

Variation (SS)

$$\text{Sum of squares} = \sum (X - \bar{X})^2 \quad \text{Or} \quad = \sum X^2 - \frac{(\sum X)^2}{n}$$

Variation is used in the construction of analysis of variance (ANOVA) tables which will be discussed later.

Variance (S^2)

The variance is the average of the squares of the deviations taken from the mean.

Variance is = ***Variation divided by (n-1).***

$$S^2 = \frac{(\sum x - m)^2}{(n-1)} \quad \text{or simply,} \quad S^2 = \frac{SS}{(n-1)}$$

Variance is used to account for the sample size used.

A small data set, that has a bigger dispersion (the points are too far from each other) compared with a large data set, may show a smaller computed variation

This is due to the fact that only a small number of values are used in the small data set compared to a large one.

$$s^2 = \frac{(\sum x - m)^2}{(n-1)} \quad \text{or simply,} \quad s^2 = \frac{SS}{(n-1)}$$

Note:

that the variation is divided by (n-1) instead of n. when the variation is divided by n, the formula is said to be biased because it underreports the dispersion especially in small data set.

But when using a large data set it does not matter to use n as a denominator.

To calculate the variance:

1. **Calculate the mean of the distribution**
2. **Find the difference between each score and the mean:**
3. **Square each of these results**
4. **Sum these squared deviations (differences)**
5. **Add up the number of observed values, and subtract 1.** This is called the **variance.** (This is the average squared deviation from the mean).

Standard deviation (S)

It is the square root of variance. In variation, the unit of measurement is in the squared form. And when divided by (n-1) into variance the unit is still in squared form.

To bring back to the original unit of measurement, the square root of the variance of the variance must be obtained

The **standard deviation** (SD) quantifies variability or scatter. ***Standard deviation*** is a measure of precision of the population distribution.

Tells us what we could expect about individuals in the population

$$S = \sqrt{\frac{\sum(x-m)^2}{n-1}}$$

The standard deviation computed this way (with a denominator of N-1) is called the sample sd, in contrast to the population sd, which would have a denominator of N. **(N-1) known as degrees of freedom**. **Sd** is always reported alongside the mean value.

For example, the mean cholesterol is 5.2 ± 0.6 mmol/l.

- Sd parameter used in establishing data symmetry and normality that will be discussed later.
- Sd also used in quality control charts to monitor the process variation from time to time.

Steps in calculating SD

1. Find the mean .
2. Subtract this from every value in the group individually - this shows the deviation from the mean, for every value.
3. Work out the square *(x^2) of every deviation (that is, multiply each* deviation by itself (e.g. 5*5) - this produces a squared deviation for every value.
4. Add up all of the squared deviations.
5. Add up the number of observed values, and subtract 1.
6. Divide the sum of squared deviations by this number, to produce the *sample variance.*
7. *Work out the square root of the variance.*

Standard error of the mean (SEM)

SE quantifies the precision of the mean. It is a measure of precision of a sample statistic. Tells us how precise our estimate of the parameter is. It is a measure of how far your sample mean is likely to be from the true population mean.

Standard error (SE) $= SE = \frac{S}{\sqrt{n}}$

To calculate SE, sd divided by the square root of n, the sample size.

It is an indication of sample to sample variation.

For example, if we took a large number of samples of a particular size from a population and recorded the mean for each sample, we could calculate the sd of all their means- this is called SE. because it is based on a very large number of theoretical samples, it should be more precise and therefore smaller than sd.

It is used in hypothesis testing and the calculation of confidence intervals.

The difference between the SD and SEM

Students confuse about the difference between the standard deviation (SD) and the standard error of the mean **(SEM**

a) The **SD** quantifies scatter — how much the values vary from one another.
b) The **SEM** quantifies how accurately the true mean of the population.

The **SEM** gets smaller as your samples get larger. Because the mean of a large sample is likely to be closer to the true population mean than is the mean of a small sample.

Example

Data set = 4, 7, 5, 9, 5.

Calculate :

a) Mean
b) Maximum & minimum
c) Range
d) Variation
e) Variance
f) Standard deviation
g) Standard error

Result

Mean = 30/5 = 6

Maximum = 9, minimum = 4

Range = 9 – 4 = 5

$$S^2 = \frac{(\sum x - m)^2}{(n-1)} \quad \text{or simply,} \quad S^2 = \frac{SS}{(n-1)}$$

$$\underline{\text{Variation ss}} = (4-6)^2 + (7-6)^2 + (5-6)^2 + (9-6)^2 + (5-6)^2$$

$$= (-2)^2 + (-1)^2 + (-1)^2 + (3)^2 + (-1)^2 = 4+1+1+9+1 = 16$$

$$\underline{\text{Variance } S^2} = \frac{16}{(5-1)} = 4$$

$$\underline{S} = \sqrt{4} = 2$$

$$\underline{SE} = \frac{2}{\sqrt{5}} = 0.894$$

Problem

Data set

10, 12, 16, 14

Calculate:

a) ***Mean***
b) ***Maximum & minimum***
c) ***Range***
d) ***Variation***
e) ***Variance***
f) ***Standard deviation***
g) ***Standard error of the mean***

Result

a) Mean = 13

b) Maximum = 16

c) Minimum = 10

d) Range = 16 – 10 = 6

e) Variation, SS = 20

f) Variance , S^2 = 6.67

g) Standard deviation = 2.58

Measures of dispersion 2

- Quartiles & inter-quartile range
- Coefficient variation
- Detecting outliers

Quartiles

Values which divide the sorted data set into four equal parts, so that each part represents 25% of the data. Quartiles are divided by the 25^{th} percentile, 50^{th} percentile, and 75^{th} percentile. One quarter of the values are less than or equal to the 25^{th} percentile. The median is the 50^{th} percentile.

Quartiles

- Q1 = gives the cut-point for the lower 25% of the data set.
- Q2 = is the median.
- Q3 = gives the cut-point for the upper 25% of the data set

Used of Quartiles

1. Qs and IQR are used in the construction of the box plot.
2. This box plot can be used to detect **outlier**s in data set.
3. An **outlier** is said to be a number more than 1.5 IQRs below Q1 or above Q3.
4. Qs are reported with median

Finding the location of Quartiles

The location of the Quartiles

$$Q1 = (n+1)\frac{1}{4}$$

$$Q2 = (n+1)\frac{1}{2}$$

$$Q3 = (n+1)\frac{3}{4}$$

Inter-quartile range

$$Q3 - Q1$$

Example:

Data set, 10, 12, 16, and 14.

Calculate the:

- *Mean*
- *Median*
- *Standard deviation*
- *Quartiles*
- *CV %*

Mean = 13, **median** = 13, **Sd** = 2.58

Ordered data = 10, 12, 14, and 16.

The location of the first Q1 =

$$Q1 = \frac{(n+1)1}{4}$$

5/4 = 1.25 **location**, which is the mid point $\frac{(10+12)}{2} = 11$

The location of the Q3 $= 5\left(\frac{3}{4}\right)$ = 3.75 the location.

$$Q3 = \frac{(n+1)3}{4}$$

The Q3 is mid point $\frac{(14+16)}{2} = 15$

IQR = Q3 – Q1 = 15 – 11 = 4

Coefficient of variation (CV)

- Also known as ***relative variability***.
- It is the measure of normalised dispersion.
- It is the ratio between measure of spread and measure of location.
- It is expressed in percentage form.

$$CV = \frac{S}{M} * 100$$

Coefficient of variation (CV)

- A small value implies that the spread is small with respect to the location and there is high level of precision.
- It is often used for the evaluation of instrument reliability.
- Because it is a unit-less ratio, you can compare the CV of variables expressed in different units.

Example

Data set, 10, 12, 16, and 14.

Calculate the:

- ***Coefficient of variation***

Mean = 13, **Sd** = 2.58

$$CV = \frac{S}{M} * 100$$

$$CV = \frac{2.58}{13} * 100 = 19.85\%$$

Detecting outliers

- *Outliers* are values that are unusually large or small.
- A single outlier can grossly affect the sample mean and variance.
- The detection of outliers is important for a variety of reasons.
- Detecting an outlier can help recognize erroneously recorded results.

A simple approach to detecting outliers is to simply

1. **Look at the data**. Checking data entry.
2. **A classic outlier detection method**
3. **Inspect graphs of the data (box plot)**

A classic outlier detection method

- A classic outlier detection technique illustrates the problem of masking.
- This classic technique **declares** the value *X* an outlier if

$$\frac{|X - \bar{X}.|}{S} \geq 2$$

For example

Data values are:

2,2,2,2,2,3,3,3,3,3,4,4,4,4,4,1,000.

The sample mean is $\boldsymbol{X}$ = **65.94** the sample standard deviation is **S** = **249.1.**

$$\frac{|1000 - 65.94|}{249.1} = 3.75.$$

Since **3.75** is greater than 2, so the value 1,000 is declared an outlier

Another Example

Data values are:

2,2,3,3,3,4,4,4,100,000,100,000.

The sample mean is = 20,002.5, the sample standard deviation is s = 42,162.38,

$$\frac{|100{,}000 - 20{,}002.5|}{42{,}162.38} = 1.897$$

The box plot rule

***Box plot is another rule** of* outlier detection. It is based on the fundamental strategy of avoiding masking by replacing the mean and standard deviation with measures of location and dispersion that are relatively insensitive to outliers.

This rule is based on the lower and upper quartiles, as well as the inter-quartile range, which provide resistance to outliers.

The *box plot rule* declares the value X an outlier if

$$\boldsymbol{X < q1 - 1.5\,(q2 - q1)}$$

Or

$$\boldsymbol{X > q2 + 1.5(q2 - q1)}$$

For example:

Data values are:

1,2,3,4,5,6,7,8,9,10,11,12,13,14,100,500.

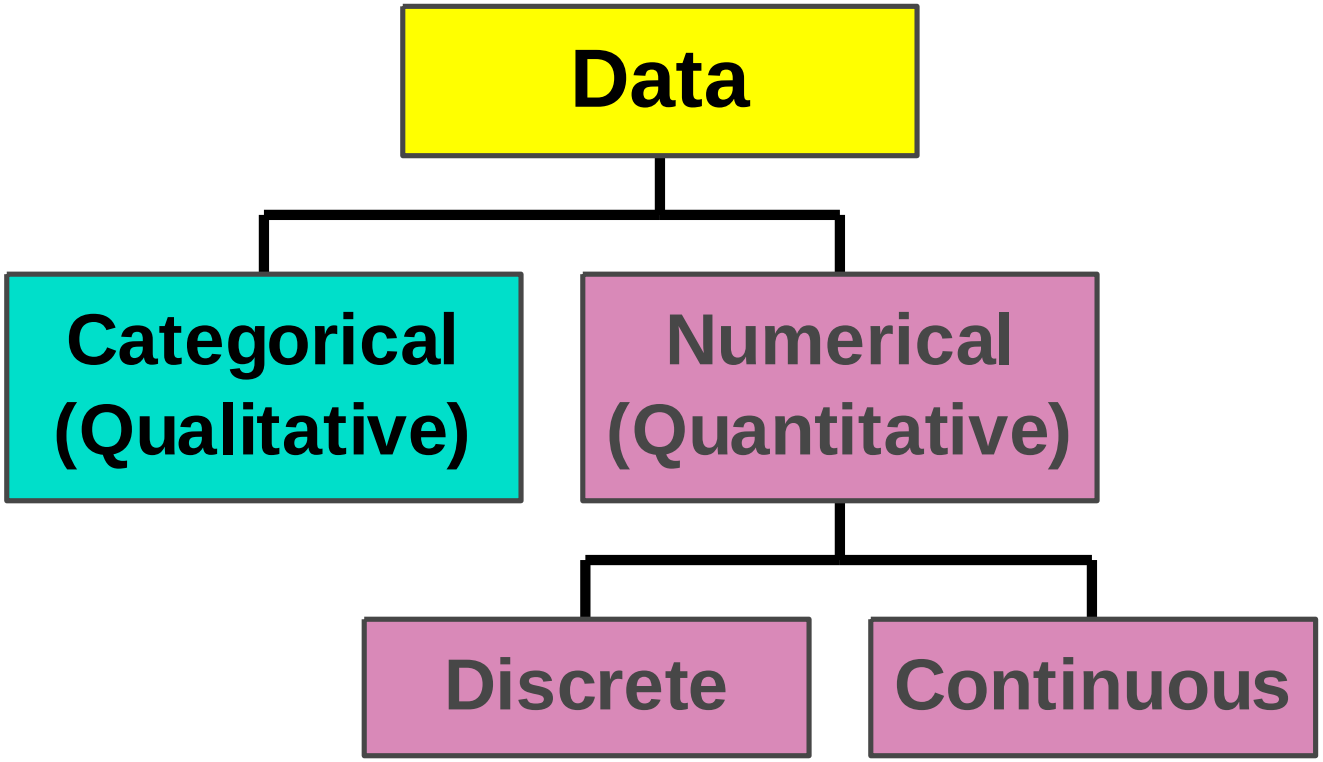
The lower quartile is $q1 = 4.417$, the upper quartile is $q2 = 12.583$.

so $\boldsymbol{q2 + 1.5(q2 - q1) = 12.583 + 1.5(12.583 - 4.417) = 24.83.}$

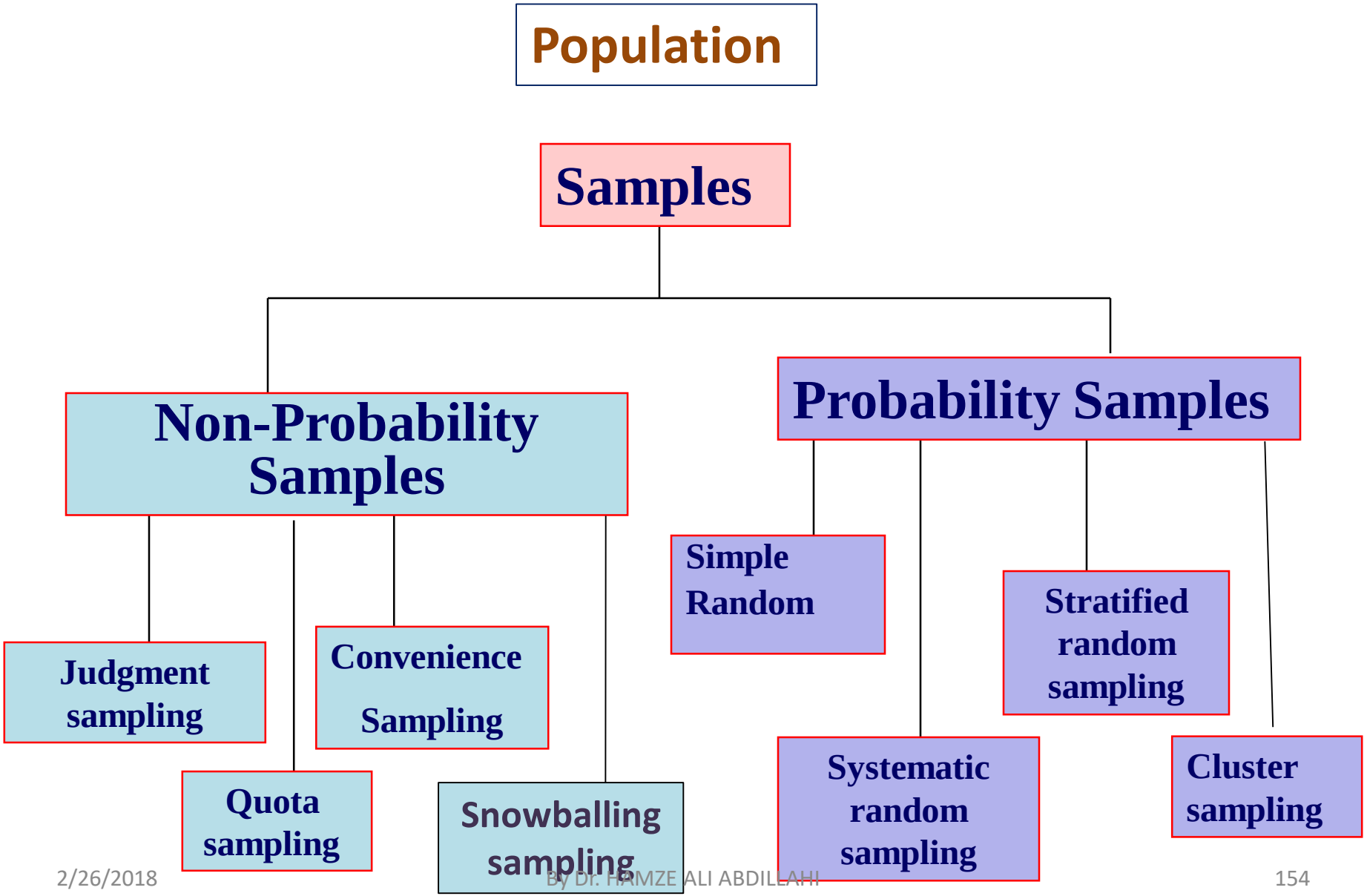
That is, any value greater than 24.83 is declared an outlier.

Hence, the values 100 and 500 are labeled outliers.

Types of Data



Types of Sampling Methods



Probability: means the chance of an occurrence. To compute the chance of occurrence, we need to know all the items in the population.

Sampling frame refers to complete list of all the items in the population.

Random means that every item in the population has an equal chance of being picked.

Why sampling?

Investigation entire population by a census

- ✓ is costly
- ✓ Time consuming
- ✓ Requires large manpower

Sampling is a more cost-effective and convenient means of collecting information.

Simple Random Sampling

- Every individual or item from the frame has an equal chance of being selected

Samples obtained from:

✓ table of random numbers or

✓computer random number generators.

- Advantages of SR
  - minimal knowledge of population needed
  - statistical estimation of error
  - Easy to analyze data
- Disadvantages
  - High cost; low frequency of use
  - Requires sampling frame
  - Does not use researchers' expertise
  - Larger risk of random error than stratified

Table of random numbers

6 8 4 20, 5 7 9 57, 4 1 82 5, 6 3 29 1,
5 8 2 10, 3 62 1 5, 4 07 8 5, 9 6 02 0,
3 6 25 3, 3 34 2 5, 4 77 8 9, 1 22 0 3,
9 8 56 2, 6 31 0 1, 7 84 2 4, 5 05 3 6

- Locate one row and one column in the table.
- Close the eyes and use pencil to choose any number.
- Say the number is 5821.
- Read the digits horizontally, can also be read vertically down.
- Split the digits into two-digit numbers : **example** 58, 21, 03 …
- Remove the repeat numbers and rearrange the selected numbers

Fore example in a class of 40 students, each students has a 1/40 (0.025) chance of being picked.

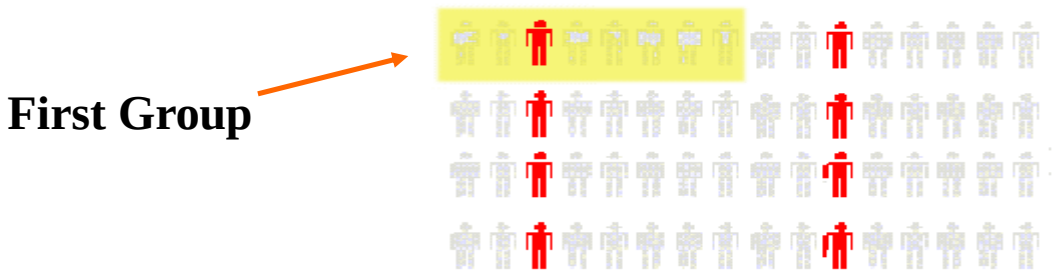
Systematic random sampling

- Decide on sample size: n
- Divide frame of N individuals into groups of k individuals: k=N/n
- Randomly select one individual from the 1st group
- Select every k-th individual thereafter.
- First number that is within the range 1 – 8 is 3
- Then the next number is 3+8 = 11 and third is 11 + 8 = 19 and so on…..

N = 64

n = 8

k = 8



- Advantages: Systematic Sampling
  - Moderate cost; moderate usage
  - statistical estimation of error
  - Simple to draw sample; easy to verify
- Disadvantages
  - Requires sampling frame
  - Potential for bias if there are underlying patterns to the sampling frame

Stratified Samples

- Population divided into two or more groups according to some common characteristic with similar groups in each strata.
- Simple random sample selected from each group
- The two or more samples are combined into one.

- Advantages
  - minimal knowledge of population needed
  - Allows calculation statistical estimation of error
  - Easy to analyze data
- Disadvantages
  - High cost
  - Requires sampling frame
  - Does not use researchers' expertise
  - Larger risk of random error than stratified
  - Unhelpful if there are no homogenous groups

For example:

we have 16 boys and 24 girls in a class, and we wand to stratify the class by gender.

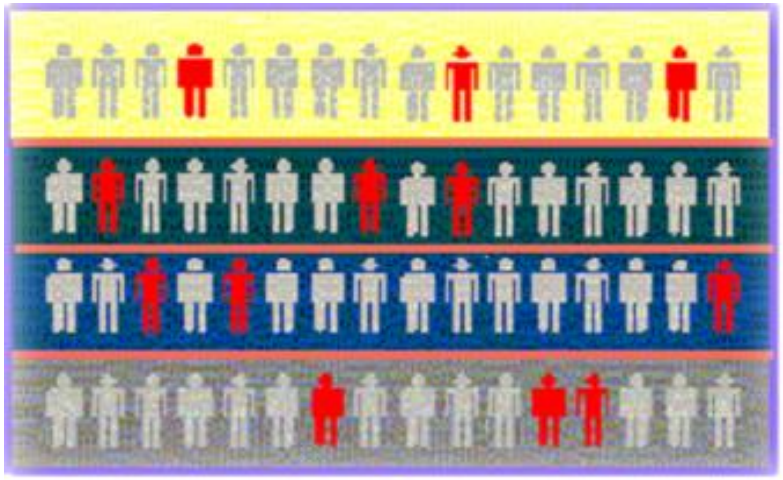
- First divide class list into two (boys and girls lists).
- We want select 5 from the sampling frame.
- Subjects from each stratum is usually proportionate to the population size within each stratum.

n = 5/40 *100 = 12.5% . The number of boys will be 16*12.5/100 = 2, we select two boys from sampling frame using simple random sampling.

The number of girls = 24 *12.5/100 = 3 we select 3 girls from the sampling frame using simple random sampling.

Cluster Samples

- Population divided into several “clusters,” each representative of the population
- Simple random sample selected from each
- The samples are combined into one



Population divided into 4 clusters.

Cluster sampling is useful when it is difficult or costly to develop a complete list of the population members or when the population elements are widely dispersed geographically.

Cluster sampling may increase sampling error due to similarities among cluster members.

- Advantages
  - Low cost
  - Requires list of all clusters
  - Can estimate characteristics of both cluster and population
- Disadvantages
  - Increase sampling error

Stratification vs. Clustering

Stratification

- Divide population into groups different from each other: sexes, races, ages
- Sample randomly from each group
- Less error compared to simple random
- More expensive to obtain stratification information before sampling

Clustering

- Divide population into comparable groups: schools, cities
- Randomly sample some of the groups
- More error compared to simple random
- Reduces costs to sample only some areas or organizations

Non-probability Samples

We use when the sampling frame is absent .

1. **Convenience sampling**
2. **Quota sampling**
3. **Judgment sampling**
4. **Snowballing sampling**

Convenience Sample

- ✓ Subjects are selected on basis of being readily available.
- ✓ Target population is defined and the required sample size is determined.
- ✓ Subjects are selected until we reach the required sample size.

- Advantages
  - Very low cost
  - Extensively used/understood
  - No need for list of population elements

- Disadvantages
  - Variability and bias cannot be measured or controlled- volunteer bias

Quota Sampling

1. Select demographic characteristics of interest (e.g. age, sex, ethnicity).
2. After selecting the target population into homogenous groups , the number of subjects in each group will not be the same.
3. So we find the percentage composition of each group in the population, similar to the first stage of stratified sampling method.
4. Then we choose the subjects using convenient procedure , on first-come-first serve basis

- Advantages
  - moderate cost
  - Very extensively used/understood
  - No need for list of population elements
  - Introduces some elements of stratification
  - Representative with regard to known characteristics
- Disadvantages
  - Variability and bias cannot be measured or controlled –volunteer bias

For example

In a study on perception of outpatients on services provided at a hospital, the patients may be sub-divided into various age groups .

Target population is (patients between 21 to 60 years old seeking services at the particular hospital. Age groups are (21,30) (31,40) (41,50) (51, 60) . The percentage of the patients taken from hospital records were 10%, 30%, 40%, 20% respectively. If the overall sample size is 50 , then the 50*10/100 = 5 patients will be choosing from the first group interval (21,30) …also 15, 20 and 10 from other groups respectively.

Judgment sampling

- ✓ Subjects chosen purposively on the basis of having particular features
- ✓ Used by specialists or authorities in a specific area.
- ✓ Most case studies are done in this manner.
- ✓ Sample size may not be large but an in-depth study of the cases is the main focus.
- ✓ Also used when choosing controls for epidemiological studies.
- ✓ Useful for rare characteristics

- Advantages
  - Moderate cost
  - Commonly used/understood
  - Sample will meet a specific objective
  - Useful for qualitative research
  - Useful for rare characteristics

- Disadvantages
  - Bias

Snowballing sampling

- ✓ Researchers move from one known case to another just by referrals.
- ✓ Used in rare events (sentinel events) .
- ✓ Enables researcher to reach groups that are otherwise hard to reach.

For example; when studying rare behaviors in the population such as drug abuse

- Advantages
  - low cost
  - Useful in specific circumstances
  - Useful for locating rare populations
- Disadvantages
  - Bias because sampling units not independent

View publication stats