Al Azhar University – Gaza
Faculty of Science
Department of Mathematics

Principles of

STATISTICS

ALI H. ABUZAID
Associate Professor of Statistics

2020

TABLE OF CONTENTS

CHAPTER 1
INTRODUCTION TO STATISTICS
(POPULATION AND SAMPLES)

The word ***statistics*** comes from:

Latin phrase "*statisticum collegium"* = lecture about state affairs.

Italian word *statista* = "*statesman*" or "*politician*" (compare to *status*).

German *Statistik* = originally designating the analysis of data about the state.

Statistics is the science and art of collecting, summarizing, and analyzing data that are subject to random variation. Whether you apply statistics to biological or other processes, it is the art of decision making in the face of uncertainty.

Literature is rich with various definition of Statistics, few of them are listed below:

- Statistics is the grammar of Science. (K. Pearson)
- Statistics is the key Technology of the current day. (P.C. Mahalanobis)
- Statistical Thinking will one day be as necessary for efficient citizenship, as the ability to read and write". (H.G.Wells)
- There are three kinds of lies: lies, damned lies and statistics." (Benjamin Disraeli)

Data are general term for numerical information that has been obtained on a set of objects. The objects can be anything, e.g., people, animals.

"Data" is the plural of "Datum".

1.1 Why Study Statistics?

- Essential for people going into research or graduate study in a specialized area.
- Effective presentation of researcher findings in papers, in reports for publication, and at professional meetings.
- Helpful to those who are preparing, or may be called upon to evaluate, research proposals.
- A knowledge of statistics is essential for persons who wish to keep their education up to date.
- Important to review and understand the writings in scientific journals, which use statistical terminology and methodology.
- An understanding of statistics can help *anyone* to discriminate between fact and fancy in everyday life in reading newspapers and watching television, and in making daily comparisons and evaluations.
- Finally, a course in statistics should help one know when, and for what purpose, a statistician should be consulted.

1.2 What statistics has to offer

- Help in developing concrete objectives and data gaining methods that meet the objectives
- Appropriate experimental and study design:
  - Source of bias
  - Measurement issues
  - Efficiency/power
  - Maximizing use of a given number of subjects
  - Interpretability of findings
  - Reproducibility of analyses

- Increase the likelihood that the sample will yield estimates of adequate precision to make experiments conclusive/affect medical practice.
- More efficient use of the data.
- Formulate analysis plans without making inappropriate assumptions.
- Estimate sample size.

When analyzing data, your goal is simple:

You wish to make the strongest possible conclusions from limited amounts of data.

How does one achieve this goal?

Limitation and misuse of statistics

- Statistical methods cannot be applied to all kinds of phenomenon and cannot answer all the quires.
- Statistical methods are subject to certain degree of error.
- Statistical statements are true on an average i.e. true for a group of individuals and may not be true for an individual.

- ***Doctor comforting his patient:***

"You have a serious disease. Of 10 people who get this disease, only one survives. But do not worry. It is lucky you came to me because I have recently had 9 patients with this disease and they all died of it. (Gambler's Fallacy)

- About 25% of biological research is faulty because of incorrect conclusions drawn from confounded experimental designs and misuse of statistical methods.
- Statistics is not your worst nightmare or the answer to all of your problems.
- Statistical significance does not equal clinical or theoretical significance.

How to use statistics properly?

- Develop an underlying question of interest.
- Generate a hypothesis.
- Design a study.
- Collect the relevant data.
- Analyze the collected data.

Branches of Statistics:

There are two types of statistics:

1- ***Descriptive statistics****:* deal with the enumeration, organization, and graphical representation of data.

2- ***Inferential statistics****:* a set of mathematical methods that employ probability theory for inferring the properties of a population from the analysis of the properties of a data sample drawn from it. .

Sources of Data:

- ***Routinely kept records*****:** Hospital medical records, for example, contain enormous amounts of information on patients.

- ***Surveys*****:**
  - Methods of survey:
  - Personal interview
  - Telephone interview
  - Questionnaires

- ***Experiments*****:** for example, the effect of specific medication on specific disease.

- ***Observations:*** attained by naked eyes or through video camera, etc.

- ***External sources*****:** published reports, commercially available data banks, or the research literature.

POPULATION AND SAMPLES

Population: a large set or collection of items that have something in common.

- A population consists of all elements, individuals, items, or objects whose characteristics are being studied. In medicine, population generally refers to patients or other living organisms.
- The population being studied is also called the target population.

Sample: A portion of the population selected for study.

- Suppose our population consist of the weights of all the elementary school children enrolled in a certain county school system. If we collect for analysis the weight of only a part of our population of weights, that is, we have a sample.

The purpose of sampling is to examine some portion of the population and to extend the knowledge obtained from the sample to the population at large.

Representative sample:
A representative sample contains the characteristics of the population as closely as possible

1.3 Why Sampling

It may not be practical or feasible to analyze the entire population. Thus, with sampling we have:

- ↓ Less costs
- ↓ Less field time
- ↑ More accuracy i.e. Can do a better job of data collection.
- ↘ When it's impossible to study the whole population
  Some testing is inherently destructive: We can't drain all the blood from a person and count every white cell.

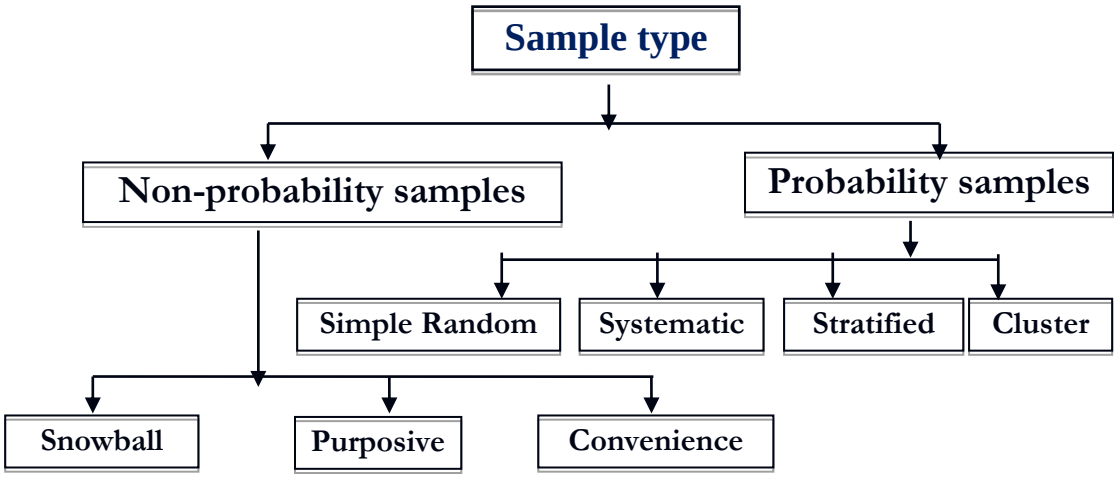
The language of sampling

- **population:** the entire collection of things of interest.
- **population parameter:** a number that results from measuring all the units in the population.
- **sampling frame:** the specific data from which the sample is drawn.
- **unit of analysis:** the type of object of interest (persons with condition *X*, animals, genes/cells)
- **sample:** a subset of some of the units in the population.
- **statistic:** is a number that results from measuring all the units in the sample.

Example:
To find out the average age of all students at Gaza universities in 2011:

- **Population** = all students at Gaza universities in 2011**.**
- **Sampling frame** = all students registered at any of Gaza universities start from January 1st , 2011.
- **Unit of analysis** = a student
- **Sample** = 300 students
- **Statistic** = the average age of the 300 students in the sample
- **Parameter** = the true average age of all students at Gaza universities in 2011.

1.4 Sampling Techniques



1.4.1 Non-probability samples

- Probability of being chosen is unknown.
- Cheaper- but unable to be generalised.
- potential for bias.

There are three main types of Non-probability samples:

I. Convenience sample (ease of access)

- It is the process of taking those members of the accessible population who are easily available.
- It is widely used in clinical research because of its obvious advantages in cost and logistics.

Example*:* A doctor conducting research might use patients volunteers to constitute a sample.

II. *Snowball sample:* (friend of friend….etc.)

Example: A researcher measure the blood pressure for certain student, followed by his father, mother and siblings.

III. Purposive sample: (judgemental)

It involves hand-picking from the accessible population those individuals judged most appropriate for the study.

Example: A researcher selecting four or five students from the faculty of medicine at Al Azhar university to assess the satisfaction of medicine student at Gaza Strip.

1.4.1 Probability Samples

- Subjects of the sample are chosen based on known probabilities.

- Guarantees that every element in the population of interest has the same probability of being chosen for the sample as all other elements in the population; "random" selection.

Random samples

- Each subject has a known probability of being selected.
- Allows application of statistical sampling theory to *Generalise* and *Hypotheses Testing*.
- Probability samples are the best compare to the Non-probability samples, but we need to ensure
    1. Representativeness,
    2. Precision.

Advantages of probability (random) sampling methods

- The population of interest is clear (because it must be identified before sampling from it.)
- Possible sources of bias are removed, such as self-selection and interviewer selection effects.
- The general size of the sampling error can be estimated.

I. Simple Random Sample (SRS):

- Every individual or item from the frame has an equal chance of being selected.
- Samples obtained from table of random numbers (Appendix 1) or computer random number generators.

Selecting a random sample involves three steps:

1. Define the population.
2. Enumerate it.
3. Use a random number table to select the sample.

To use a random number table, first randomly select a starting position and then **move in any direction** to select the numbers.

Example: evaluate the prevalence of tooth decay among the 850 children attending a school.

1. List of children attending the school,
2. Children are numerated from 1 to 850,
3. Sample size = 10 children,
4. Random sampling of 10 numbers between 1 and 850.

Characteristics of Simple Random Sample

- Easy.
- Rarely used.
- Availability of full list of population entire.
- Requires homogeneous population.

- The basic of other sampling methods.

II. Systematic Sample:

- Decide on sample size: *n*.
- Divide population 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.

Advantage: The sample usually will be easier to identify than it would be if simple random sampling were used.

Example: Selecting every 100th listing in a telephone book after the first randomly selected.

Example: If you want to select a sample of 50 children from the 850 children attending a school to evaluate the prevalence of tooth decay among them.

Solution
1. Divide the 850 children by 50 (sample size) = **17**.
2. Every 17th children is sampled.
3. Select a number randomly between 1 and 17 first, and we then select every 17th children.
4. Suppose we randomly select the number **10** from a random number table.
5. Then, the systematic sample consists of children with ID numbers 10, 27, 44, 61, 78, and so on; each subsequent number is determined by adding 17 to the last ID number.

III. Cluster sample:

- Population is divided into several "clusters," each representative of the population.
- A simple random sample of clusters is selected.
- All items in the selected clusters can be used, or items can be chosen from a cluster using another probability sampling technique.
- Known as Multi-stage sample.

Example: In conducting a survey of school children in a large city, we could first randomly select 5 schools and then include all the children from each selected school. This technique is more economical than the random selection of persons throughout the city.

IV. Stratified Samples

- The population is first divided into groups of elements called *strata*.
- Each element in the population belongs to one and only one stratum.
- Best results are obtained when the elements within each stratum are as much alike as possible (i.e. *homogeneous group*).
- A simple random sample is taken from each stratum.
- Formulas are available for combining the stratum sample results into one population parameter estimate.

Procedure for selecting of stratified sampling
- Firstly, the population is divided into at least two distinct strata or groups.
- Then a random sample of a certain size is drawn from each stratum.
- The groups or strata are often sampled in proportion to their actual percentage of occurrence in the overall population.
- Combine the results of all strata.

Advantage: If strata are homogeneous, this method is as "precise" as simple random sampling but with a smaller total sample size.

Example: The basis for forming the strata might be sex, location, age, industry type, etc.

Advantages and Disadvantages

- ***Simple random sample and systematic sample***
  a. Simple to use.
  b. May not be a good representation of the population's.
- ***Stratified sample***
  a. Ensures representation of individuals across the entire population.
- ***Cluster sample***
  a. More cost effective.
  b. Less efficient (need larger sample to acquire the same level of precision).

1.5 Errors in sample

There are two main sources of error in sampling:

1. **Systematic error (or bias)**
   - Inaccurate response (information bias).
   - Selection bias.

2. Random Sampling Error (random error)
   - Variability.
   - Sampling method.
   - Sample size.

1.6 Bias and sample size

- Any trend in the collection, analysis, interpretation, publication or review of data that can lead to conclusions that are systematically different from the truth, is known as bias. (Las, 2001)

- A systematic error in design or conduct of a study. (Szklo et al, 2000)

Sample size: The number of elements in the sample is called the sample size.

Statistically, there are five main factors affect the size of a sample as follows:

1- ***Population size:*** Total number of items in the population – ***Only*** important if the sample size is greater than 5% of the population in which case the sample size reduces.

2- ***The population proportion:*** the proportion of items in the population displaying the attributes that you are seeking.

3- ***Margin of error or precision:*** a measure of the possible difference between the sample estimate and the actual population value.

4- ***Variability in the population:*** the standard deviation is the most usual measure and often needs to be estimated.

5- ***Confidence level:*** how certain you want to be that the population figure is within the sample estimate and its associated precision.

6- Other factors:

- Time and money constraints influence sample size.
- The lower your sampling error must be, the larger your sample must be.
- The more diverse your population is, the larger your sample must be.
- The more complex your analysis, the larger your sample must be.
- The stronger your expected relationships, the smaller your sample can be.

CHAPTER 2
ORGANIZATION AND PRESENTATION OF DATA

- Data are collected on the specific characteristics of each subject, and groups are formed to be compared. These characteristics are called ***variables***, because they can change from each subject.

Variable: a characteristic under study that assumes different values for different elements.

Some *examples* of variables include

- diastolic blood pressure,
- the heights of adult's males,
- the weights of preschool children,
- pain score of patients postoperatively, etc…

Variable is obtained because it is:

- A result of interest – (dependent variable)
- It explains the dependent variable – (risk factor, independent variable).
- The value of dependent variable will depend on independent or exposure variable.

Data set: A collection of observation on a variable

- A typical data set is often represented with a matrix of information. Each row represents an individual or unit, while each column represents a variable

No	Age	Sex	Height	Weight
1	12	m	145	40
2	10	f	134	32

Parameter: Descriptive measure derived from a population.

- Usually we don't know the value of the parameter; consequently, we are estimating it from the sample.
- Usually it's impossible to be calculated.
- ***For example*** if we want to know the mean age of all Palestinian bladder cancer cases. This mean will be calculated from information collected from all bladder cancer cases in Palestine.

2.1 Variables Types:

Any survey or experiment yields a list of observations. These need to be organized and summarized in a logical fashion so that we may recognize the outcome clearly. Tables, graphs and numerical methods are popularly used to organize, summarize and description of data.

- It does not matter whether the observations are made on people, animals or objects
- What does matter is the ***kind of observation*** made and how the characteristic observed.

- These features determine:

  – **types of**
    - tables,
    - graphs,
    - summary statistics that best communicate the observations to someone else.

  – **types of test statistics**
    - to use in making conclusions about the observations.

Several ways to classify the variables

They may be defined as:

1. Qualitative (categorical) variables.
2. Quantitative variables.

Qualitative or categorical variables: are those variables that yield observations on which individuals can be categorized according to some characteristic.

***Examples*;** sex, marital status, and education level. (Non-numeric in nature)

Quantitative variables: are those variables that yield observations that can be measured.

Examples; weight, height, and serum cholesterol. (numeric values)

- Measurements made on quantitative variables convey information regarding amount.

Quantitative variables are either:

Discrete*:* only take values from some discrete set of possible values (whole are integers).

Examples: number of patients admitted to the hospital,
the number of children in one family,
the number of time you visit a doctor.

Continuous: Values from a continuous range of possible values, although the recorded measurements are rounded.

Examples: weight, height, hemoglobin levels, etc..

2.2 Scales of measurement

Another way to classify the variables is to assign number to the objects or events according to a set of rules. These rules are the scales of measurement.

- They are commonly broken down into four types:
  1. Nominal
  2. Ordinal
  3. Interval (numerical)
  4. Ratio (numerical)

1. Nominal scale:

The simplest level of measurement, where data values are fitted into categories and there is no ordering, (it makes no sense to state that M > F).

Examples:
– *Outcomes of a medical treatment:* occurring or not occurring.
– *Surgical procedure:* types of procedures.
– *Presence of possible risk factors*.

2. Ordinal Measurement:

Data may be arranged in some order, but actual differences between data values either cannot be determined or are meaningless.

Examples:
- *the degree of pain* (severe, moderate, mild, none),
- *class rank:* 1st, 2nd and 3rd ,
- *the age group:* (baby, infant, child, adult).

3. Interval Measurement

Data can be arranged in some order, (like the ordinal level) but it has addition property that meaningful intervals (differences) between data values can be computed.
Interval-level data ***have no absolute zero point*** or starting point. Consequently, ***differences*** are meaningful but ***ratios*** of data are not.

Examples:
- Difference between 100°C and 90°C is the same as the difference between 50°C and 40°C.
- It is not correct to say 20°C is twice as hot as 10°C or 100°C is not twice as hot as 50°C, because 0°C does not represent the point at which there is no heat, but it is the freezing point of pure water.
- Day times are also interval measurements, since the time 00:00, does not signify “no time”.
- An IQ “Intelligent Quotient" of zero value would not mean there is no intelligence at all, but serious intellectual or perceptual problem in using materials of the test.

4. Ratio Measurement

It has the same properties as an interval scale; but it ***has an absolute zero***. Thus, meaningful ratios do exist.

Examples:
- weight in grams or pounds,
- time in seconds or days,
- blood pressure in millimeters of mercury and
- pulse rate of 120 is twice as fast as a pulse rate of 60.

2.3 Data summary

Generally, we want to show the data in summarized form, to ease the communication with it in order to use proper statistical tests.

2.3.1 Summary of categorical data

We can obtain frequencies of categorical data and summarize them in a table or graphs.

1. The Frequency Table

Considerable information can be obtained from large masses of statistical data by grouping the data into classes and determining the number of observations that fall in each of the classes. Such an arrangement is called a ***frequency distribution*** or ***frequency table***.

- Frequency table may be the most convenient way of summarizing or displaying data.
- Two types of frequency distributions will be considered, ***categorical or qualitative*** frequency distributions, and ***grouped*** frequency distributions.

Example: Suppose we have 21 students from 4 different cities in Palestine.

Rafah	*Gaza*	*Gaza*
Nablus	*Rafah*	*Rafah*
Hebron	*Nablus*	*Nablus*
Nablus	*Hebron*	*Hebron*
Hebron	*Nablus*	*Nablus*
Gaza	*Hebron*	*Hebron*
Gaza	*Gaza*	*Gaza*

- List of cities show us an idea of the frequency of each city, but it is not clear.
- If we ordered them, the idea is more clearer.

Hebron	*Nablus*	*Gaza*	*Rafah*
Hebron	*Nablus*	*Gaza*	*Rafah*
Hebron	*Nablus*	*Gaza*	*Rafah*
Hebron	*Nablus*	*Gaza*	
Hebron	*Nablus*	*Gaza*	
Hebron		*Gaza*	

- The order of the list gives a clear idea of the frequency of data. We can show the results in a frequency distribution.

Frequency distribution of cities, n=21

City	***n***
Gaza	6
Nablus	6
Hebron	6
Rafah	3
Total	**21**

The first column shows city name and the second column shows how many times each city is repeated.

With a table of frequency distribution, we can see clearly and quickly, what city is more frequent. With 21 students the problems is not big, but in other studies the sample will be larger with hundreds or thousands.

- It is useful to show the frequency of each category, expressed as percentage of the total frequency.
- It is called distribution of relative frequencies.

$$\textbf{\textit{Relative frequency}} = \frac{\text{Frequency of category}}{\text{Total of frequency}} \times 100$$

Relative frequency distribution of cities.

City	*n*	%
Gaza	6	28.57
Nablus	6	28.57
Hebron	6	28.57
Rafah	3	14.29
Total	**21**	**100.00**

2. Grouped Frequency Distribution

For the continuous data set, a grouped frequency distribution is obtained by constructing class intervals, and then listing the corresponding number of values (frequency count) in each interval.

How to construct a frequency table?

Example: Following Table contains 63 systolic blood pressure readings. Construct the Frequency Table for the given data.

Systolic Blood Pressure of Non-Smokers

92	112	122	128	134	144	162
94	112	122	128	134	146	170
96	114	122	128	134	152	172
98	114	122	128	134	152	
100	118	124	130	134	154	
104	118	124	130	138	154	
106	118	128	130	140	154	
108	118	128	132	140	154	
108	118	128	132	142	156	
108	120	128	134	144	162	

Divide the range into a number of equal and nonoverlapping segments called **class intervals.**

1- Obtain the minimum and the maximum values to determine the range:
***R* = 172-92 = 80** mm**.**
2- Decide the number of intervals. (*Note:* Number of intervals is between 5 and 15).
Let the number of classes, $k = 5$.

3- Determine the size (length or width) of the class interval (w) by dividing the range (R) by the number of class intervals required or (k), i.e. $\boldsymbol{w \geq R/k} = 80/5 = 16$.
for easiness and for comparison purposes we will use 20.
4- Start the first class interval with the smallest value *or less*. This value is called as the *lower class limit*.

The frequency table is given in the first two columns of the following table.

3. Relative frequency

The relative frequency for a particular class is found by **dividing** the class frequency by the total of all frequencies (sample size) i.e., *f/n*.

4. Cumulative relative frequency (cumulative percentage)

It gives the percentage of individuals having a measurement less than or equal to the upper boundary of the class interval.

Frequency Table for systolic blood pressure of Nonsmokers

Class Interval (Systolic Blood Pressure)	Frequency	Relative Frequency (%)	Relative frequency	Cumulative Relative Frequency (%)
90-109	10	16	0.16	16
110-129	24	38	0.38	54
130-149	18	29	0.29	83
150-169	9	14	0.14	97
170-189	2	3	0.03	100
Total	**63**	**100**	**1**	**-**

Class boundaries may be used instead of class limits. **Class boundaries** are points that demarcate the true upper limit of one class and the true lower limit of the next. Class boundaries can be easily obtained by subtracting from the lower limit and adding to the upper limit ***one-half of the smallest unit used*** to record the data.

In the systolic blood pressure, the class boundaries will be (89.5-109.5, 109.5 -129.5, 129.5 - 149.5, 149.5 -169.5, 169.5 -189.5) .

3.4 Graphing of Data

The second way of displaying data is by use of graphs. Graphs give the user a nice overview of the essential features of the data. It is essential that each graph be *self-explanatory* that is, have

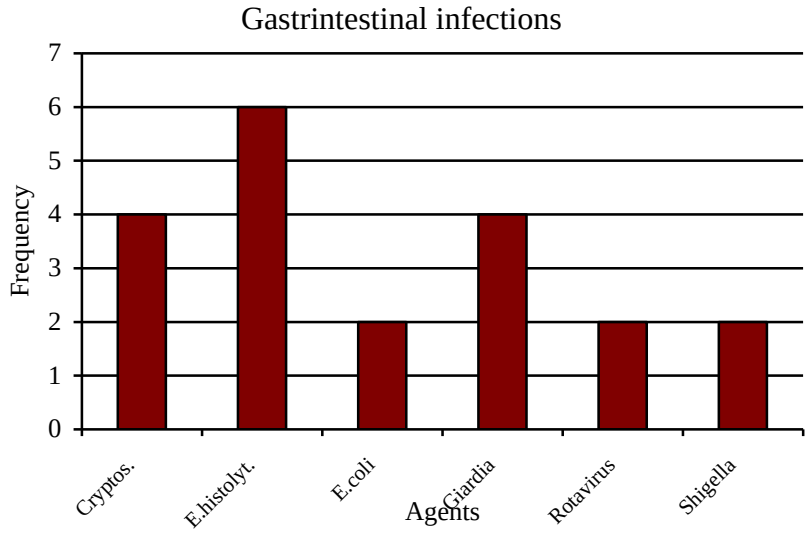
- A descriptive title,
- Labeled axes,
- An indication of the units of observation.

- An effective graph should not attempt to present so much information that it is difficult to comprehend. The type of chart to use depend on the nature of data.
  Categorical data
  1. Bar chart, 2. Pie chart.

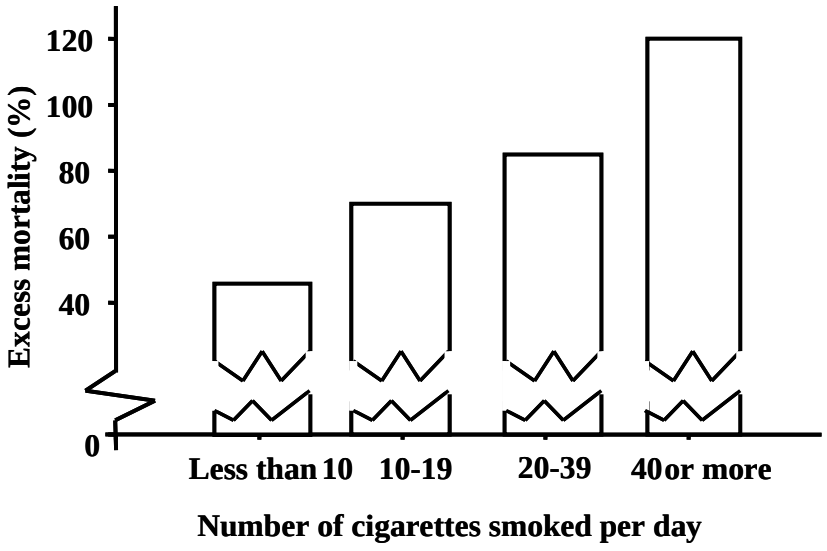
  Quantitative data
  1. Histogram, 2. Polygon, 3. Ogive, 4. Boxplot.

1. Bar chart

- It is used with categorical or numerical discrete data.
- The frequency or relative frequency of a categorical variable can be shown easily in a bar chart.
- Each bar represent one category and its height is the frequency or the relative frequency.
- Bars should be separated.
- It is very important that *Y* axis begin with 0.

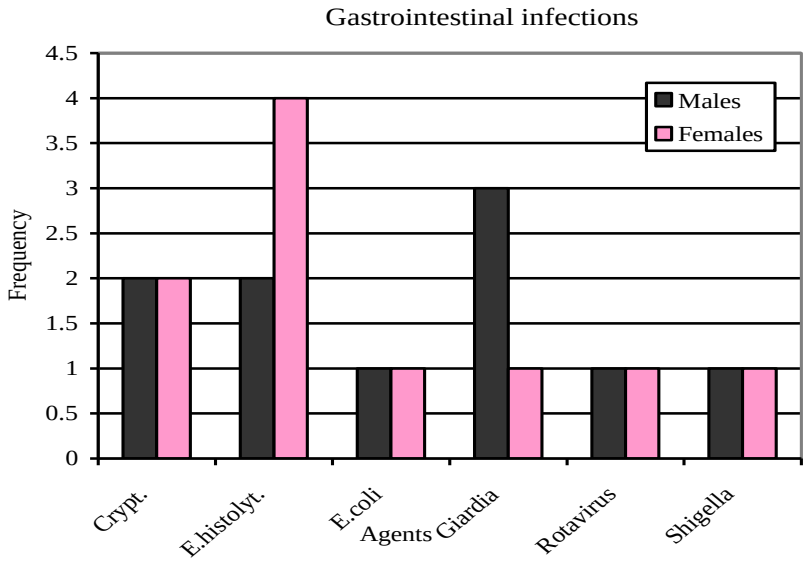


- When you use a change in scale, warn the viewer by using the squiggle or broken bars on the changed axis as shown in the following figure. Sometimes, if a single bar is unusually long, the bar length is compressed with a squiggle in the bar itself.



- If we have a nominal categorical variable, divided into two categories, can show data with a grouped bar chart. It allows an easy comparison between groups.

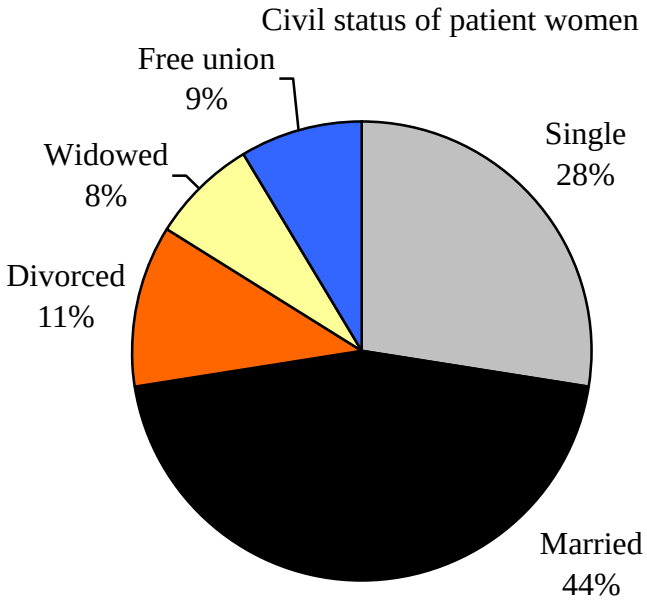
- Gastrointestinal infection category is divided by gender. We can easily see that *E.histolytic* infected, predominated females and in the *Gaza* infected predominated males.



2. Pie chart

- It is an alternative graph to present one categorical variable in the form of a circle.
- Each slice of pie correspond at frequency or relative frequency of categories of variable.
- If we want to make comparisons, we need to build more than one pie chart.
- The angle of any slice is obtained by the following formula:

$$\textbf{\textit{Angle}} = \frac{\text{Frequency of category}}{\text{Total of frequency}} \times 360^{\circ}$$



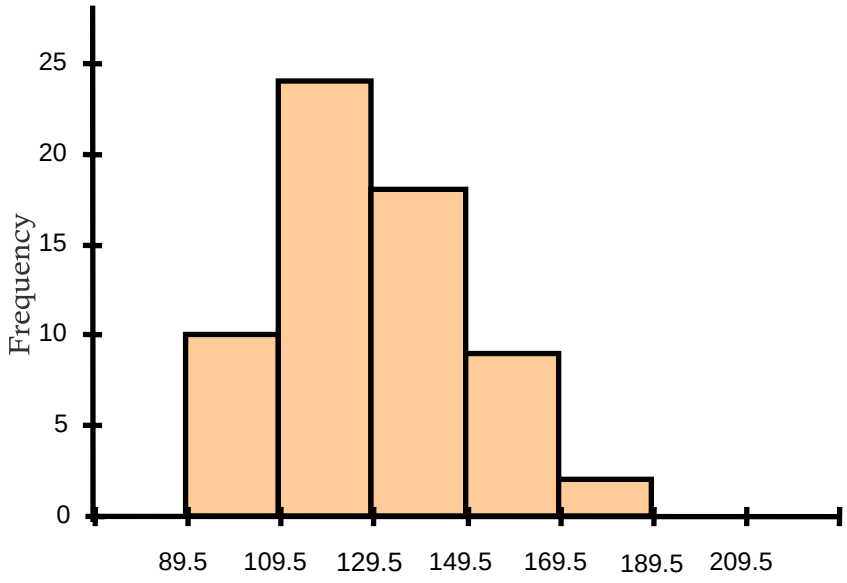
3. Distribution of frequency chart: (Histogram)

- It is useful to quantitative variables.
- There are no spaces between bars.
- The area bar, not its high, represent its frequency.
- ***X*** axis should be continuous.
- ***Y*** axis should begin in 0.
- Width represent the interval for each group.

To make a histogram

- Make frequency table that shows class intervals and class frequencies.
- Determine the ***class boundaries*** for each class interval.

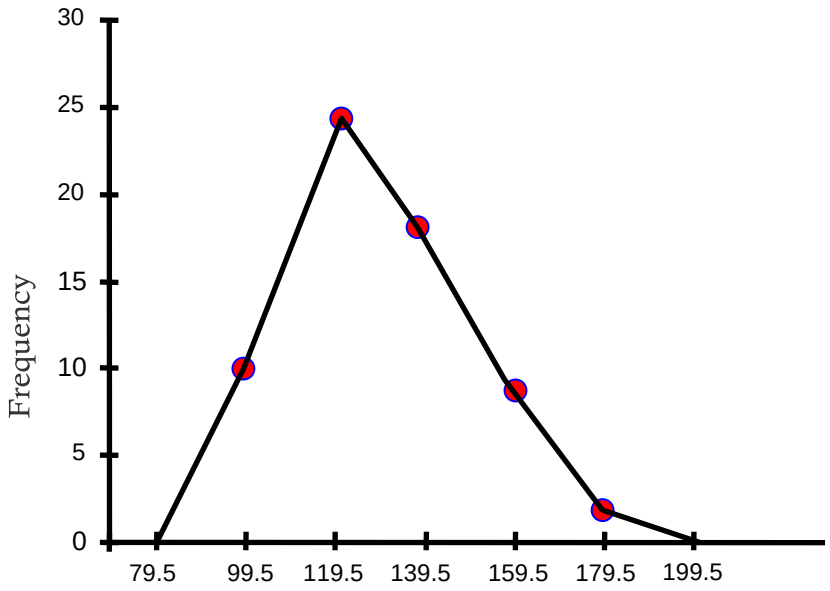
Class interval (Systolic Blood Pressure)	**Class boundaries**	***f*** **(frequency)**
90-109	89.5-109.5	10
110-129	109.5-129.5	24
130-149	129.5-149.5	18
150-169	149.5-169.5	9
170-189	169.5-189.5	2
Total		**63**



Systolic blood pressure (mm Hg) for Non-Smoker

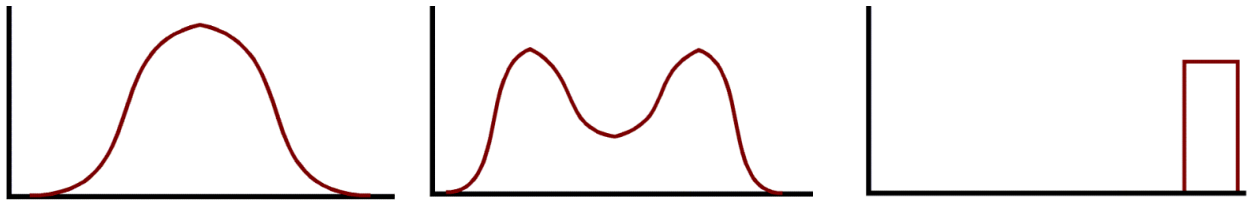
4. Frequency Polygon

The second commonly used graph is the frequency polygon, which uses the same axes as the histogram. It is constructed by making a point over the ***class midpoint*** at the same height of the class frequency. The coordinates of these dots are (class midpoint, class frequency). These points are then **connected** with straight lines.

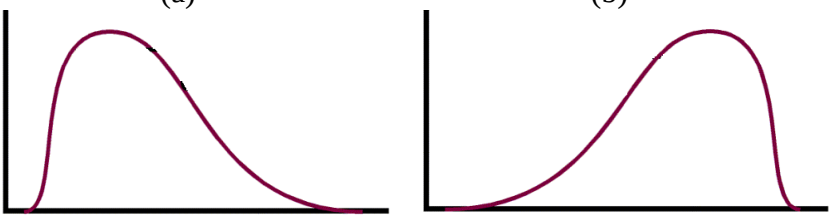


Systolic blood pressure (mm Hg) for Non-Smoker

Frequency polygons may take on a number of different shapes
(a) Bell-shaped" symmetrical distribution.
(b) Bimodal (having two peaks) distribution.
(c) Rectangular distribution in which each class interval is equally represented.
(d) Asymmetrical positively (right) skewed distribution, since its tail is in the positive direction.
(e) Asymmetrical negatively (left) skewed distribution, since its tail is in the negative direction.
(f) Other shapes, like (L) or (J) shapes.



(a) (b) (c)



(d) (e)

5. Cumulative Frequency Polygons (Ogive)

Ogive can be used to determine how many scores are above or below certain level.

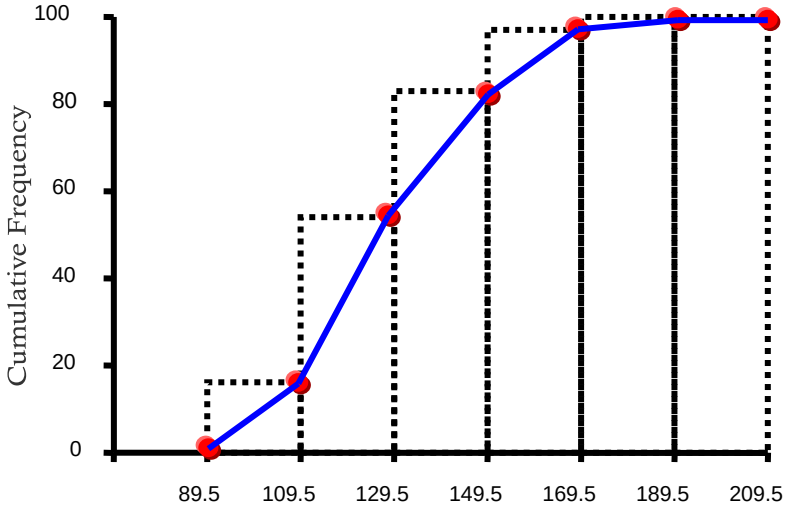
To construct an ogive:
Make a frequency table showing class boundaries and cumulative frequencies.

Class boundaries (Systolic Blood Pressure*)	Cumulative Relative Frequency (%)	
	Nonsmokers	Smokers
89.5-109.5	16	14
109.5-129.5	54	55
129.5-149.5	83	82
149.5-169.5	97	90
169.5-189.5	100	95
189.5-209.5	100	100

Use the same horizontal scale as that for a histogram, whereas the vertical scale indicates cumulative frequency or cumulative relative frequency.

For each class interval, make a dot over the *upper class boundary* at the height of the cumulative class frequency.

- The coordinates of the dots are (upper class boundary, cumulative class frequency). Connect these dots with line segments.
- By convention, an ogive begins on the horizontal axis at the lower class boundary of the first class interval.

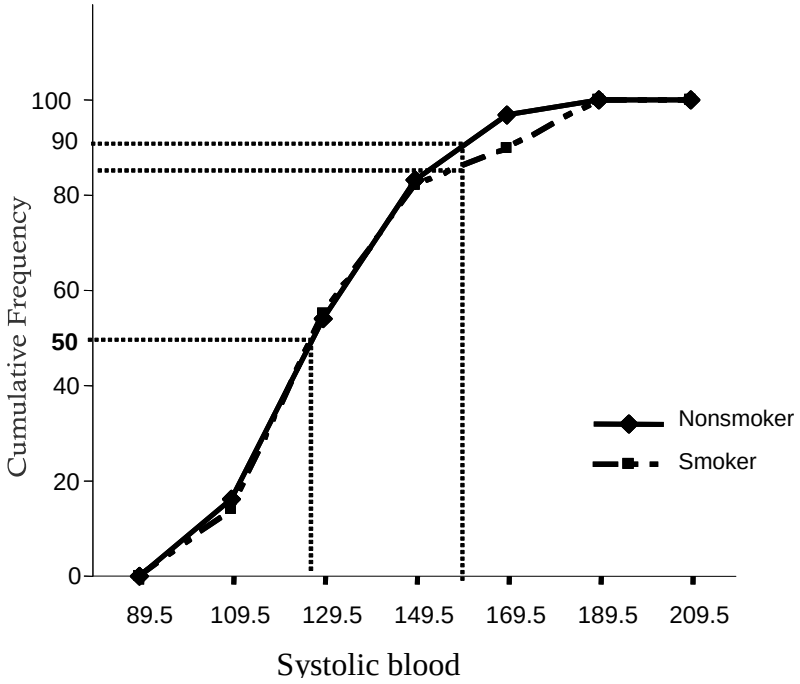


Systolic blood pressure for non smokers (mmHg)

Ogive are useful in comparing two sets of data, as, for example, data on smokers and non-smokers individuals.

In the Figure below we can see that:

- 90% of the nonsmokers and 86% of the smokers had systolic blood pressures below 160 mmHg.
- 50% of both nonsmokers and smokers had systolic blood pressures below 127 mmHg.



6. Boxplot:

Another exploratory tool, which uses quartiles (will be explained in the following chapter) of a set of measurements to depict the shape and range of the distribution.

The box is composed of a center line, which represents the median measurement. The upper and lower borders of the box represent the 'quartiles' of the distribution. Thus the middle 50% of the distribution of measurements falls in the range of the box.

A skewed distribution might appear as a box with an off-centered median. The lines emanating from the box extend an equal distance from the median, and serve to identify outliers.

The boxplot outliers are noted as exceptional cases in the distribution and may either result from truly abnormal values or be the result of a skewed distribution.

Follow these steps in order to construct a box plot:

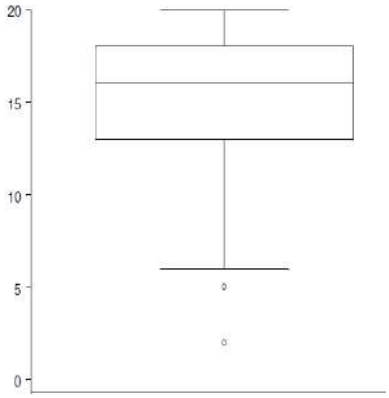
1. Calculate the median *m*, (will be explained in the following chapter).
2. Calculate the first and third quartiles Q_1 and Q_3.
3. Compute the inter-quartile range IQR = $Q_3 - Q_1$.
4. Find the lower fence LF = $Q_1 - 1.5 \times$ IQR.
5. Find the upper fence UF = $Q_3 + 1.5 \times$ IQR.
6. Find the lower adjacent value LAV = smallest value in the data that is greater or equal to the lower fence
7. Find the upper adjacent value UAV = largest value in the data that is smaller or equal to the upper fence
8. Any value outside the LAV or UAV is called an outlier and should receive extra attention.

Example: Consider the following depression scale scores, The Koopmans (1981) data set of depression scores:

2, 5, 6, 8, 8, 9, 9, 10, 11, 11, 11, 13, 13, 14, 14, 14, 14, 14, 14, 15, 15, 16, 16, 16, 16, 16, 16, 16, 16, 17, 17, 17, 18, 18, 18, 19, 19, 19, 19, 19, 19, 19, 19, 20, 20.

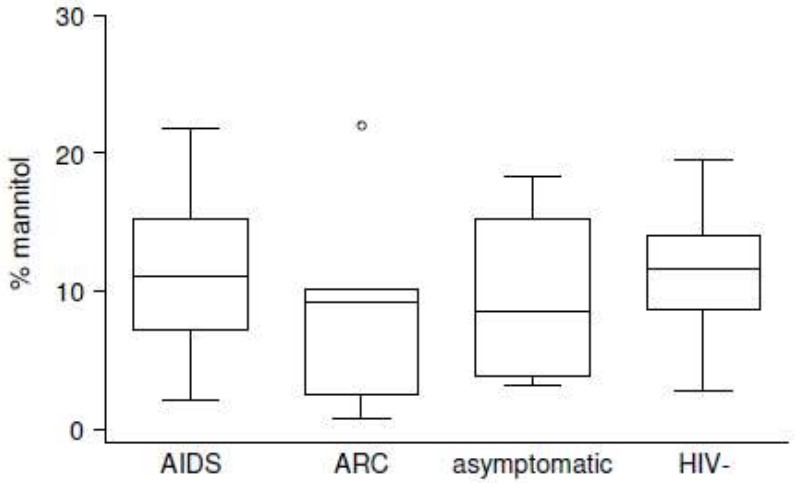
The Box Plot for Koopmans data is constructed as follows:

1. Calculate the median (m): Since the number of observations is 45 (odd number) the median is the $[(45 + 1)/2]^{th}$ i.e., the 23^{d} observation. That is, $m = 16$.
2. Calculate the first and third quartile Q_1 and Q_3 Split the data set into two equal parts (including the median in both of them), that is, split into the first and last 23 observations. Then Q_1 is the median of the first 23 observations (the 12^{th} observation), and Q_3 is the median of the last 23 observations (the 34^{th} observation). Thus, $Q_1 = 13$ and $Q_3 = 18$.
3. Compute the inter-quartile range IQR = $Q_3 - Q_1$, IQR = 18 − 13 = 5.
4. Find the lower fence LF = $Q_1 - 1.5\times$IQR. LF = $Q_1 - 1.5\times$IQR = 13−1.5(5) = 5.5.
5. Find the upper fence UF = $Q_3 + 1.5\times$IQR. UF = $Q_3 + 1.5\times$IQR = 18+1.5(5) = 25.5.
6. Find the lower adjacent value. LAV = smallest value in data > 5.5, LAV = 6.
7. Find the upper adjacent value. UAV = largest value in data < 25.5, UAV = 20.
8. Since 2 and 5 are lower than the LAV , these observations are outliers and must be investigated further. The boxplot of the Koopmans data set is given below:



Box plot of Koopmans depression scores

Boxplot is a very good procedure to compare several groups as shown in the following figure:



EXERCISES

1. Consider the following variables

Education level, Weight, Gender, Age, Smoking status, Blood glucose, Blood type, Speed, Religion, Satisfactory levels.

a. Classify each variable as to whether it is qualitative or quantitative.
b. Which of the quantitative variables are discrete? Which continuous?
c. Name an appropriate type of graph for presenting each variable.

2. The following are weight losses (in kilogram) of 25 individuals who enrolled in a five-week weight-control program:

9, 7, 10, 11, 10, 2, 3, 11, 5, 4, 8, 10, 9, 12, 5, 4, 11, 8, 3, 6, 9, 7, 4, 8, 9.

a. Construct a frequency table with five equal intervals of 3 units each.
b. Construct a histogram of weight losses.
c. Construct a frequency polygon and describe the shape of the frequency distribution.
d. What might be a possible interpretation of the particular shape of this distribution?
e. What was the most common weight losses.
f. Construct an ogive for the given data.
g. Construct a boxplot of the given data.

3. Prepare a suitable frequency distribution table for the following data:
21,11,3,42,49,7,33,10,25,45,33,36,14,19,41,35,32,26,38,14,13,12,14,35,23,23,24.

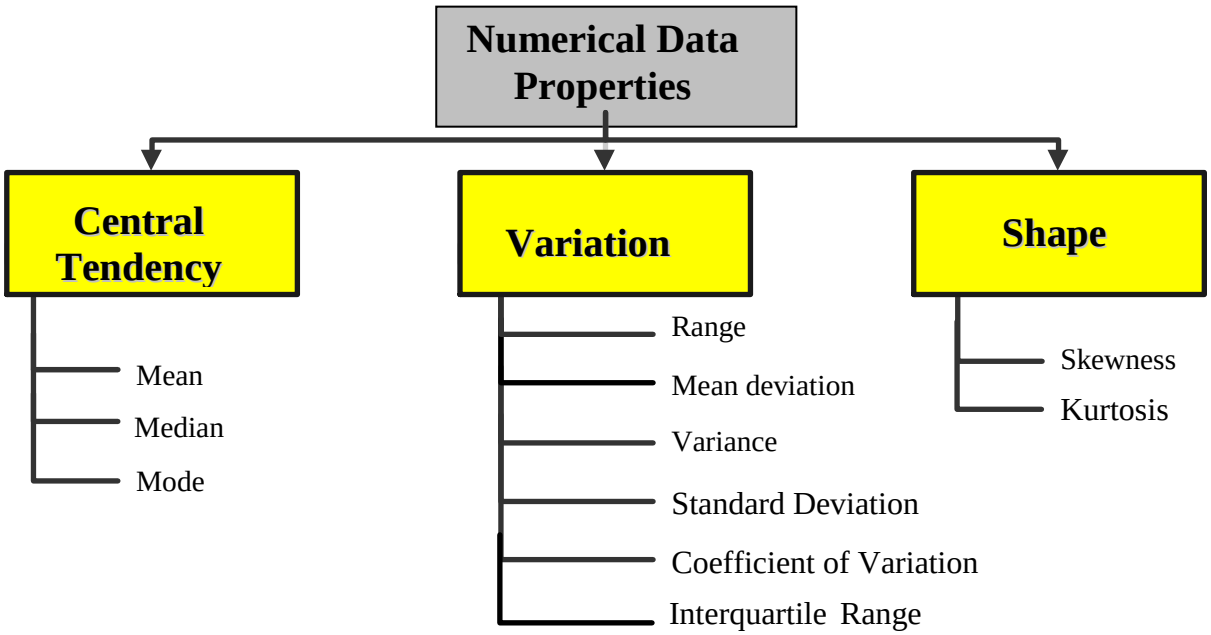
3. The following table represents the distribution of age for 95 persons selected at random.

Interval	**10-19**	**20-29**	**30-39**	**40-49**	**50-59**	**60-69**	**Total**
Frequency	2	10	30	33	15	5	**95**

a- Calculate the relative and the cumulative frequency of the age.
b- Construct a histogram of the age.
c- Using the previous draw plot, construct a frequency polygon and describe the shape of the frequency distribution.
d- Construct an ogive for the given data
e- Use the ogive to find the median of the age

CHAPTER 3
SUMMARIZING DATA

To summarize a set of data, there are three groups can be used to give an idea about the distribution of the data. The following chart summarize these summarize.



3.1 Central Tendency Measures:

Given a set of data, one regularly wishes to find a value about which the observations tend to cluster. The three most common values are the mean, the median, and the mode.
They are known as *measures of Central Tendency* -the tendency of a set of data to center around certain numerical values.

3.1.1 Mean

1- Arithmetic mean (Average):

The arithmetic mean is the most widely used measure of location. It requires the interval scale.
For ungrouped data, the sample mean is the sum of all the sample values divided by the number of sample values, *n*:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

For ungrouped data, the population mean is the sum of all the population values divided by the total number of population values:

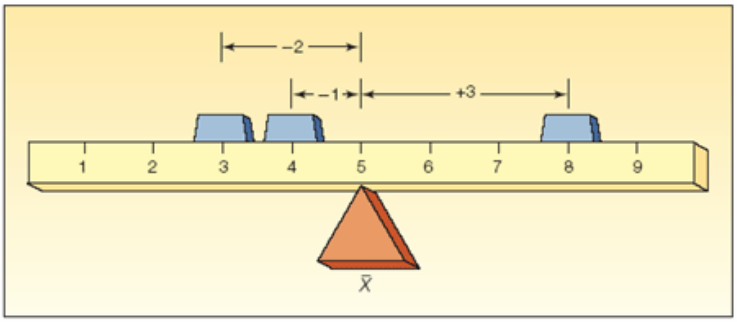
$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

where N is the population size.

Example: If we measure the height of 5 students: 167, 171, 163, 159, 175, mean will be $\bar{x} = \frac{167+171+163+159+175}{5} = 167$ cm.

Properties of the Arithmetic Mean

- Every set of interval-level or ratio-level data has a mean.
- All the values are included in computing the mean.
- A set of data has a unique mean.
- The mean is affected by unusually large or small data values (outliers).
- The arithmetic mean is the only measure of central tendency where the sum of the deviations of each value from the mean is zero, $\sum(x_i - \mu) = 0$.



2-Weighted Mean

The weighted mean of a set of numbers $x_1, x_2, ..., x_n$, with corresponding weights $w_1, w_2, ..., w_n$, is computed from the following formula:

$$\bar{x}_w = \frac{w_1 x_1 + w_2 x_2 + ... + w_n x_n}{w_1 + w_2 + ... + w_n}$$

Example: Al-Quds Hospital at Gaza pays its hourly employees \$16.50, \$19.00, or \$25.00 per day. There are 26 daily employees, 14 of which are paid at the \$16.50 rate, 10 at the \$19.00 rate, and 2 at the \$25.00 rate. What is the mean hourly rate paid the 26 employees?

$$\bar{x}_w = \frac{14(\$16.50) + 10(\$19.00) + 2(\$25.00)}{14 + 10 + 2} = \frac{\$471.00}{26} = \$18.1154$$

3- Mean of Grouped Data

In a grouped distribution, we use the middle point x_i of each interval as *x* value.

$$\bar{x} = \frac{\sum x_i f_i}{\sum f_i}$$

Example: find the mean of the age for the following data

Interval (age)	Middle point (x_i)	Frequency (f_i)
1-3	2	18
4-6	5	27
7-9	8	34
10-12	11	22
13-15	14	13
Total		**114**

$$\bar{x} = \frac{\sum x_i f_i}{\sum f_i} = \frac{(2\times18) + (5\times27) + (8\times34) + (11\times22) + (14\times13)}{18 + 27 + 34 + 22 + 13} = \frac{867}{114} = 7.61 \text{ years.}$$

3.1.2 The Median

1- Median of ungrouped data:

The Median is the midpoint of the values after they have been ordered from the smallest to the largest. There are as many values above the median as below it in the data array.
If the number of observations is even, then the median is the arithmetic mean of the two middle observations in the data set.

Properties of the Median

- It can be computed for ratio-level, interval-level, and ordinal-level data.
- There is a unique median for each data set.
- It is not affected by extremely large or small values. Therefore, it is a valuable measure of central tendency when such values occur.

Example
Suppose we have the observations 7, 4, 3, 5, 6, 8, 10, 1. Find the median of this data set.

Solution: First, we arrange the data set in ascending order

1 3 4 **5 6** 7 8 10.

Since the number of the observations n = 8, then by Definition the median is the average of the 4th (n/2 = 8/2 = 4th) and the 5th i.e. Median = (5+6)/2 = 5.5

Advantage of the median over the mean:

- It may be determined even if the values of all observations are not known, (i.e. 3,4,5,6, x_1, x_2, x_n).
- Extreme values in data set do not affect the median as strongly as they do the mean.

2. Median of grouped data:

In a grouped distribution, the following steps are followed:

- ***Step*** 1: Form the cumulative frequency (F)
- ***Step*** 2: Find the value of $N/2$ where $N = \sum f$
- ***Step*** 3: Find F value that the first exceeds $N/2$, which identifies the median class M.
- ***Step*** 4: Calculate the median using the following formula

$$Median = L_M + \left[\frac{\frac{N}{2} - F_{M-1}}{f_M} \right] c_M$$

where;

- L_M lower bound of the median class
- F_{M-1} cumulative frequency of class immediately prior to the median class
- f_M actual frequency of median class
- c_M median class width.

Example: Estimate the median for the age in the following data set

Age	20-25	25-30	30-35	35-40	40-45	45-50
frequency	2	14	29	43	33	9

Solution:
Step 1

Age	(*f*)	(F)
20-25	2	2
25-30	14	16
30-35	29	45
35-40	43	88
40-45	33	121
45-50	9	130

***Step* 2:** $N/2$=130/2 = 65
***Step* 3:** Median class is 35-40
***Step* 4:** L_M =35; F_{M-1}=45; $f_M = 43$, c_M =5.

$$Median = L_M + \left[\frac{\frac{N}{2} - F_{M-1}}{f_M}\right] c_M = 35 + \left[\frac{65-45}{43}\right] 5 = 37.33 \text{ years}$$

Median also known as the second quartile Q_2. (Quartiles will be considered in (3.2.6).

Percentiles are numerical values of the variable that divide a set of ordered data into 100 equal parts; each set of data has 99 percentiles.

The procedure for determining the value of any k^{th} percentile involve three basic steps.

1. The data must be ordered.
2. The position for the percentile is founded by first calculating the value of $\frac{k(n+1)}{100}$.
3. The percentile it self is obtained by finding the corresponding vlaue in the ordered data.

Example
Suppose we have the observations 7, 4, 3, 5, 6, 8, 10, 1. Find the 30th and 50th percentiles.

Solution: First, we arrange the data set in ascending order

1 3 4 5 6 7 8 10.

Since the number of the observations n = 8, then by Definition:

1. The position of the 30th percentile is $\frac{30(8+1)}{100} = 2.7$ which is between the second and third values. Thus, the 30th percentile is the average of the 2nd and 3rd values (3+4)/2=3.5.

2. The position of the 50th percentile is $\frac{50(8+1)}{100} = 4.5$. Thus, the 50th percentile is the average of the 4th and the 5th values (5+6)/2 = 5.5 median.

Percentiles of grouped data

The k^{th} percentile (P_k) of grouped data is obtained by:

$$P_k = L_P + \left[\frac{\frac{kN}{100} - F_{p-1}}{f_p} \right] c_p$$

where L_p, F_{p-1}, f_p and c_p have the same meaning of the terms in the median formula, but with respect to the k^{th} percentile.

Example: Estimate the 40th percentile of the age in the previous example:
Solution:
Step 1

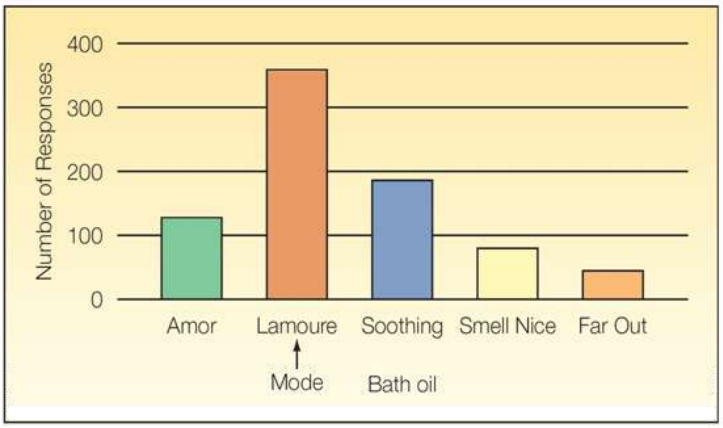
Age	(*f*)	(F)
20-25	2	2
25-30	14	16
30-35	29	45
35-40	43	88
40-45	33	121
45-50	9	130

***Step* 2:** $kN/100 = 5200/100 = 52$
***Step* 3:** k^{th} percentile class is 35-40
***Step* 4:** $L_p = 35$; $F_{p-1} = 45$; $f_p = 43$; $c_p = 5$.

$$P_{40} = L_P + \left[\frac{\frac{40N}{100} - F_{p-1}}{f_p} \right] c_p = 35 + \left[\frac{52-45}{43} \right] 5 = 35.81 \text{ years}$$

3.1.3 The Mode

The mode is the observation that occurs most frequently. i.e., is repeated most often in the data set.



Number of respondents favoring various bath oils

1- The mode of ungrouped data:

Example: For a given sample n =16:

33, 35, 36, 37, 38, 38, 38, 39, 39, 39, 39, 40, 40, 41, 41, 45.

The mode = 39

- It corresponds to the highest point on the frequency distribution.

Multimodal distribution: A data set may have several modes. In this case it is called multimodal distribution.

Example: The data set has two modes: 1 and 4. This distribution is called bimodal distribution.

9	6	2	0
10	6	4	0
11	7	4	1
11	8	4	1
12	9	5	1

- When data sets contain two or many modes, they are difficult to interpret and compare.

- Too often, there is no modal value because the data set contains no values that occur more than once. Other times, every value is the mode because every value occurs the same number of times. Clearly, the mode is a useless measure in these cases.

2- The mode of grouped data:

In a grouped distribution, the following steps are followed:

- ***Step*** 1: Determine the model class (class with the largest frequency).
- ***Step*** 2 Calculate D_1 = Difference between the largest frequency and frequency immediately preceding it.
- ***Step*** 3 Calculate D_2 = Difference between the largest frequency and the frequency immediately following it.
- ***Step*** 4 Calculate the mode using the following formula

$$Mode = L + \left[\frac{D_1}{D_1 + D_2}\right] C$$

- L = Lower bound of the modal class
- C = Model class width
- D_1 and D_2 are described in ***Step*** 2 and ***Step*** 3.

Example: Estimate the mode for the Age in the following data set

Age	20-25	25-30	30-35	35-40	40-45	45-50
frequency	2	14	29	43	33	9

Solution:

Step 1

Age	Number (f)
20-25	2
25-30	14
30-35	29
35-40	43
40-45	33
45-50	9

***Step* 2:** D_1 = 43-29=14

***Step* 3:** D_2 = 43-33=10

***Step* 4:** L =35; C = 40-35 =5

$$Mode = L + \left[\frac{D_1}{D_1 + D_2}\right] C = 35 + \left[\frac{14}{14+10}\right] 5 = 37.92 \text{ years}$$

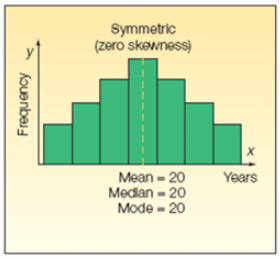
Advantages of Mode

- it is more appropriate average to use in situations where it is useful to know the most common value.
- easy to be understood, not difficult to be calculated.
- it is not affected by extreme values.

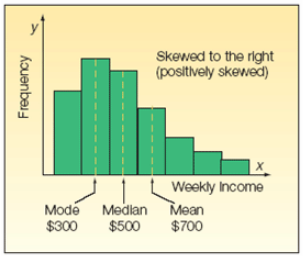
Disadvantages of Mode

- it ignores dispersion around the mode value and it does not take all the values into account.
- it is unsuitable for further statistical analysis.
- although it ignores extreme values, it is thought to be too much affected by the most popular class when a distribution is significantly skewed.

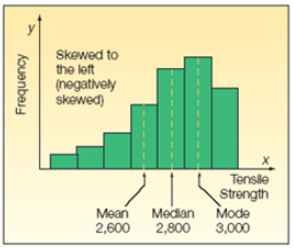
3.1.4 The Relative Positions of the Mean, Median and the Mode



zero skewness
mode = median = mean



positive skewness
mode < median < mean



negative skewness
mode > median > mean

Which measure of central tendency should you use?

There are two general criteria for choosing between the measures of central tendency:

1. *Scale of measurement*
   - For nominal scale data, you can only use the Mode.
   - For ordinal scale data, you can only use Median or Mode; Median is more informative.
   - For interval or ratio scale data, you can use any one of the three measures.

2. *Shape of the distribution*
   - Mean is more informative, if you don't have a skewed distribution.
   - If you have skewed distribution, you use the median in place of mean.

3.2 Measures of Variation (dispersion)

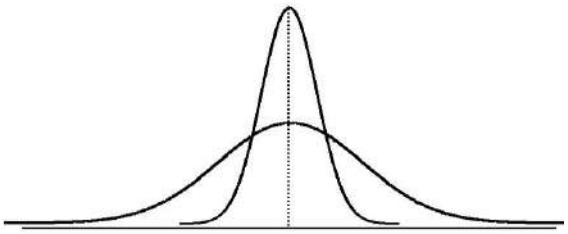
Measures of variation give information on the spread or variability of the data values.

Why Study Dispersion?

- A measure of location, such as the mean or the median, only describes the center of the data. It is valuable from that standpoint, but it does not tell us anything about the spread of the data.

Example: If your nature guide told you that the river ahead averaged 3 feet in depth, would you want to wade across on foot without additional information?
Probably not, you would want to know something about the variation in the depth.

- A second reason for studying the dispersion is to compare the spread in two or more distributions.



Two frequency distributions with equal means but different amounts of variation

3.2.1 Range

The range is the difference between the highest (maximum) and lowest (minimum) observation:

$$\text{Range} = x_{max} - x_{min}$$

The range can be computed quickly, but it is **not very useful** since it considers only the extremes and does not take into consideration the volume of the observations.

Advantages of Range

- it is easy to be found and to be understood.

Disadvantages of Range

- affected by extreme values.
- is not used in further advanced statistical work.

Example: Find the range of the following data set:

51.2, 53.5, 55.6, 61.4, 65.0, 74.2

Range of this distribution is 74.2 - 51.2 = 23 kg.

However, the extreme values of this distribution are far center of distribution, it unclear the fact that the most data are between 53.5 and 65 kg.

3.2.2 Mean deviation:

It is the deviation of all observation from the mean.

$$\text{Mean deviation: } = \frac{|x_1 - \bar{x}| + |x_2 - \bar{x}| + ... + |x_n - \bar{x}|}{n}$$

where $\bar{x}$ is the sample mean.

Example: Consider the percentage of graduates of a medical school who passed their National Boards with honors during a five-years period. Calculate the Mean deviation.

	Year of Graduation				
	2004	**2005**	**2006**	**2007**	**2008**
Percent of honors graduate (x_i)	4	6	5	8	7

The mean is $\bar{x} = 6$,

$$\text{Mean deviation: } = \frac{|4-6| + |6-6| + |5-6| + |8-6| + |7-6|}{5} = \frac{2+0+1+2+1}{5} = \frac{6}{5} = 1.2$$

3.2.3 Variance

- The Variance is a measure which uses the mean as a point of reference.
- The Variance is less when all value are close to the mean while it is more when the values are spread out from the mean.
- Variance is the average (approximately) of squared deviations of values from the mean

$$\text{The population variance, } \sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}. \qquad \text{Sample variance, } s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n\text{-}1}$$

- It is quite difficult to interpret the value of variance, because the unit is squared compared to the observations' unit. Thus, another measure of variability is defined, which is ***the standard deviation***.

3.2.4 Standard deviation

- It is the square root of the variance.
- Most commonly used measure of variation.
- Shows variation about the mean.
- Has the same units as the original data

Population standard deviation $\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$, Sample standard deviation $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$

Example: Consider the Annual percentage of Medical School National Board Honorees 2004-2008. Calculate the variance and the standard deviation.

The variance and standard deviation easily can be obtained as shown in the following Table

Year of Graduation	**Percent of honors graduate** (x_i)	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
2004	4	-2	4
2005	6	0	0
2006	5	-1	1
2007	8	2	4
2008	7	1	1
Total		**0**	**10**

Since the given data is a sample, then the Variance is $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{10}{4} = 2.5$

The standard deviation is $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{10}{4}} = \sqrt{2.5} = 1.58$.

The value 1.58 indicates, that on the average, observations fall 1.58 units from the mean.

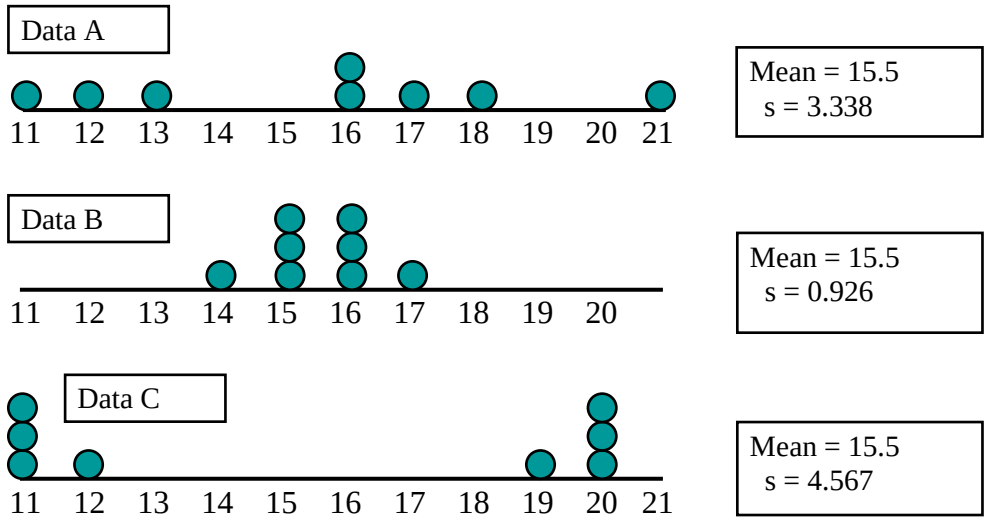
Advantages of variance

- it takes all values into account, therefore it can be regarded as truly representative of the data.
- it is suitable for further statistical analysis**.**

Disadvantages of variance

- it is more difficult to be understood than some other measures of dispersion.

Review: Comment on the following figure with the given statistics.



Standard deviation for grouped frequency distribution

$$s = \sqrt{\frac{\sum_{i=1}^{c} f_i x_i^2 - \frac{\left(\sum_{i=1}^{c} f_i x_i\right)^2}{n}}{n-1}}$$

Example: The Table below shows the temperature of a sample of 50 cities taken at the same time on a certain day. Determine the mean, and standard deviation of the sample.

(Temperature C°)	10-14	15-19	20-24	25-29	30-34
Number of cities	10	12	18	6	4

Solution:

Temperature	f	Cumulative	Mid point	x^2	fx	fx^2
10-14	10	10	12	144	120	1440
15-19	12	22	17	289	204	3468
20-24	18	40	22	484	396	8712
25-29	6	46	27	729	162	4374
30-34	4	50	32	1024	128	4096
	50				**1010**	**22090**

$$\bar{x} = \frac{1010}{50} = 20.2$$

$$s = \sqrt{\frac{\sum_{i=1}^{c} f_i x_i^2 - \frac{\left(\sum_{i=1}^{c} f_i x_i\right)^2}{n}}{n-1}} = \sqrt{\frac{22090 - \frac{(1010)^2}{50}}{49}} = \sqrt{\frac{22090 - \frac{(1010)^2}{50}}{49}} = \sqrt{\frac{1688}{49}} = \sqrt{34.45} = 5.87$$

3.2.5 Coefficient of Variation (CV)

- One important application of the mean and the standard deviation is the coefficient of variation.

- Measures relative variation and always in percentage (%)
- Can be used to compare two or more sets of data measured in different units, because CV itself is unitless.

$$\textbf{Coefficient of Variation} = CV = \left(\frac{S}{\overline{X}}\right).100\%$$

Example: Suppose that each day laboratory technician *A* completes 40 analyses with a standard deviation of 5. Technician *B* completes 160 analyses per day with a standard deviation of 15. Which employee shows less variability?

At first glance, it appears that technician *B* has three times more variation in the output rate than technician *A*. But *B* completes analyses at a rate 4 times faster than *A*. Taking all this information into account, we compute the coefficient of variation for both technicians:

For technician *A*: CV=5/40 × 100% = 12.5%

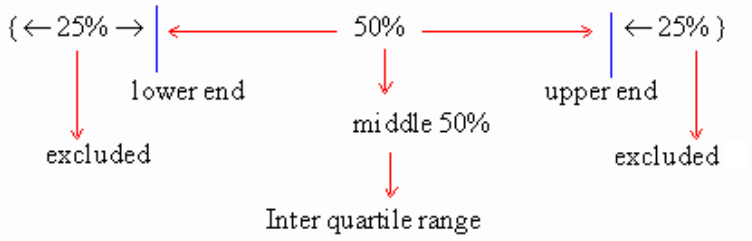
For technician *B*: CV=15/160 × 100% = 9.4%.

So, we find that, technician *B* who has more absolute variation in output than technician *A*, has less relative variation.

3.2.6 Interquartiles range (*IQR*)

If we concentrate on two extreme values - as in the case of range - we do not get any idea about the scatter of the data within the range.

If we discard these two values the limited range thus available might be more informative. For this reason the concept of interquartile range is developed. It is the range which includes middle 50% of the distribution. Here 1/4 (one quarter) of the lower end and 1/4 (one quarter) of the upper end of the observations are excluded.



Now the lower quartile (Q_1) is the 25th percentile and the upper quartile (Q_3) is the 75th percentile. It is interesting to note that the 50th percentile is the middle quartile (Q_2) which is in fact what you have studied under the title ' Median ". Thus symbolically

Interquartile Range, *IQR* = $Q_3 - Q_1$

For continuous series of data, the formulas for calculating Q_1 and Q_3 are same as the median Q_2

$$Median = Q_2 = L_2 + \left[\frac{\frac{2N}{4} - F_2}{f_2}\right] c_2$$

$$Q_1 = L_1 + \left[\frac{\frac{N}{4} - F_1}{f_1}\right] c_1 \text{ and } Q_3 = L_3 + \left[\frac{\frac{3N}{4} - F_3}{f_3}\right] c_3$$

Example: Estimate the range and the interquartile range for the weight in the following data set

Weight	32-36	37-41	42-46	47-51	52-56	57-61
Frequency	4	7	10	7	18	4

Solution:

Step 1

Weight	**Class boundaries**	**Number (*f*)**	**F**
32-36	31.5- 36.5	4	4
37-41	36.5- 41.5	7	11
42-46	41.5- 46.5	10	21
47-51	46.5- 51.5	7	28
52-56	51.5- 56.5	18	46
57-61	56.5- 61.5	4	50

To calculate the median

***Step* 2:** $N/2$=50/2 = 25

***Step* 3:** Median class is 46.5- 51.5

***Step* 4:** L_2 = 46.5 ; F_2 = 21 ; c_2 =5.

$$Median = Q_2 = L_2 + \left[\frac{\frac{2N}{4} - F_2}{f_2}\right] c_2 = 46.5 + \left[\frac{25-21}{7}\right] 5 = 49.36 \text{ k.g.}$$

To calculate the first quartile Q_1

***Step* 2:** $N/4$=50/2 = 12.5

***Step* 3:** the first quartile class is 41.5- 46.5.

***Step* 4:** L_1 = 41.5 ; F_1 = 11 ; c_2 =5.

$$Q_1 = L_1 + \left[\frac{\frac{N}{4} - F_1}{f_1}\right] c_1 = 41.5 + \left[\frac{12.5-11}{10}\right] 5 = 42.25 \text{ k.g.}$$

To calculate the third quartile Q_3

***Step* 2:** $3N/4$= 150/4 = 37.5

***Step* 3:** the third quartile class is 51.5- 56.5

***Step* 4:** L_3 = 51.5 ; F_3 = 18 ; c_2 =5.

$$Q_3 = L_3 + \left[\frac{\frac{3N}{4} - F_3}{f_3}\right] c_3 = 51.5 + \left[\frac{37.5 - 28}{18}\right] 5 = 54.14 \text{ k.g.}$$

Thus, **the range** = max – min = 59 – 34 = 25.

Interquartile Range = Q_3 - Q_1 = 54.14 - 42.25 = 11.89.

EXCERCISES

1. Find the mean, median, mode, range, 30^{th} and 70^{th} percentiles, interquartile range, variance and standard deviation for the data:

8, 5, 1, 5, 2, 3.

2. Using the sample: 3, 4, 6, 1, 10, 6.
   a. Find the median, mean, and interquartile range,.
   b. Compute the variance and standard deviation.

3. Determine the range, interquartile range, mean, median, 20^{th} and 90^{th} percentiles, mode, variance and standard deviation for the following sample of the blood pressure.

58, 82, 56, 58, 66, 102, 92, 68, 60, 78.

4. Assuming that the data in the previouse question is a population, find the mean, variance and the standard deviation

5. The following are weight losses (in kilogram) of 25 individuals who enrolled in a five-week weight-control program:

9, 7, 10, 11, 10, 2, 3, 11, 5, 4, 8, 10, 9, 12, 5, 4, 11, 8, 3, 6, 9, 7, 4, 8, 9.

   a. Calculate the coefficient of variation for the weight losses.
   b. Calculate $\bar{x} - s$ and $\bar{x} + s$
   c. Calculate $\bar{x} - 2s$ and $\bar{x} + 2s$
   d. Calculate $\bar{x} - 3s$ and $\bar{x} + 3s$
   e. What is the percentage of the weight losses observations fall within each of the three intervals you calculated in (b), (c) and (d).

6. The following table represents the distribution of age for 95 persons selected at random.

Interval	**10-19**	**20-29**	**30-39**	**40-49**	**50-59**	**60-69**	**Total**
Frequency	2	10	30	33	15	5	**95**

Calculate the range, mean, median, interquartile range, mode, 15^{th} and 75^{th} percentiles , variance and standard deviation.

7. Refer to the Systolic Blood Pressure data, which is discussed in Chapter 2, pg 15. Calculate the range, mean, median, interquartile range, mode, 38^{th} and 99^{th} percentiles, variance and standard deviation

CHAPTER 4
PROBABILITY

4.1 Combinatorial Methods :

Combinatorial mathematics is the mathematics of how we combine objects into arrangements. The use of mathematical probability is dependent on our understanding and application of several counting principles. In this topic, we will explore those counting principles we will use in finding probabilities.

4.1.1 Permutations

Permutation is an arrangement of objects in different orders. There are basically two types of permutations:

- Repetition is allowed
- No repetition

a) Permutations with Repetition

If you have *n* things to choose from, and you choose *r* of them, then the permutations are:

$$n \times n \times ...(r\ times) = n^r$$

Because there are *n* possibilities for the first choice. Then, there are *n* possibilities for the second choice, and so on.

Example: If there are 10 numbers to choose from (0,1,..9) and you choose 3 of them with repetition. Then, we have $10 \times 10 \times 10 = 10^3 = 1000$ permutations.

b) Permutations without Repetition

Example: List all permutations of the letters ABCD (without repetition)

Solution:

ABCD	BACD	CABD	DABC
ABDC	BADC	CADB	DACB
ACBD	BCAD	CBAD	DBAC
ACDB	BCDA	CBDA	DBCA
ADBC	BDAC	CDAB	DCAB
ADCB	BDCA	CDBA	DCBA

Now, if you didn't actually need a listing of all the permutations, you could use the formula for the number of permutations of *n* objects taken *r* at a time is

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

A factorial, "!" is a rule that defines an operation on some integer, which is to multiply together the integer each preceding integer down to the integer 1.

$$n! = n \times (n-1) \times (n-2) \times \ldots \times 3 \times 2 \times 1$$

Example: $3! = 3 \times 2 \times 1 = 6$ and $5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$

Since here are 4 objects and you're taking 4 at a time.

$$_4P_4 = \frac{4!}{(4-4)!} = \frac{4!}{0!} = \frac{24}{1} = 24$$

This also gives us another definition of permutations. A permutation when you include all *n* objects is *n*!

Example: If A={p,q,r,s} find the number of permutation for 3 elements.

Solution: $$^4P_3 = \frac{4!}{1!} = \frac{4.3.2.1}{1} = 24$$

The two key things to notice about permutations are that there is no repetition of objects allowed and that order is important.

4.1.2 Combinations

A combination is an arrangement of objects were the order is not important.

Example: List all combinations of the letters ABCD in groups of 3.

ABC, ABD, ACD, BCD

The number of combinations of *n* objects taken *r* at a time is obtained by the following formula

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

Since here are 4 objects and you're taking 3 at a time.

$$_4C_3 = \frac{4!}{(4-3)!\,3!} = 4$$

The difference between the two is whether or not order is important. If you have a problem where you can repeat objects, then you must use the Fundamental Counting Principle.

Example: A team comprising of 4 people are to be selected randomly from 6 women and 6 men. Find the number of possible selections if

(i) no conditions are imposed
(ii) the team must have more men than women
(iii) the team must have at most two men

Solution:

(i) ${}^{12}C_4 = 495$

(ii) ${}^{6}C_3 \times {}^{6}C_1 = 120$

(iii) ${}^{6}C_2 \times {}^{6}C_2 + {}^{6}C_1 \times {}^{6}C_3 + {}^{6}C_0 \times {}^{6}C_4 = 360$

4.2 Probability:

Probability refers to the study of randomness and uncertainty.
Probability is a numerical measure of chance or likelihood for the occurrence of an event.

4.1.1 Probability terminology:

- **Experiment**: a repeatable procedure with a well-defined set of possible **outcomes.**
- **Sample space:** is the set of all possible outcomes.
- An **event:** is a subset of outcomes.

Example: Consider an experiment of rolling a 6-sided die

Sample Space, *S* :{1, 2, 3, 4, 5, 6}.
Events, E_k:

E_1: odd number is rolled.→ Equivalently, {1,3,5}.
E_2: number less than four is rolled.→ Equivalently, {1, 2, 3}.
E_3: prime number is rolled.→ Equivalently, {2, 3, 5}.

Example: A three-digit winning lottery number is selected.

Sample Space: {000,001,002,003, . . . ,997,998,999}.There are 1000 simple events.

Probabilities for Simple Event: Probability any specific three-digit number is a winner is 1/1000. Assume all three-digit numbers are equally likely.

Event A = last digit is a 9 = {009,019, . . . ,999}.
Since one out of ten numbers in set, ***P*(A) = 1/10**.

Event B = three digits are all the same = {000, 111, 222, 333, 444, 555, 666, 777, 888, 999}.
Since event B contains 10 events, ***P*(B) = 10/1000 = 1/100**.

Example: An experiment consists of tossing three coins. Find the sample space if

1- We are interested in the observed face of each coin.

S={(H,H,H), (H,H,T), (H,T,H),(H,T,T), (T,H,H), (T,H,T), (T,T,H),(T,T,T)}

2- We are interested in the total number of heads obtained.

0,1,2 or 3

Example: How many sample points are in the sample space when a pair of dice is thrown once?

S={(1,1),(1,2),...,(1,6),
(2,1),(2,2),…,(2,6),
.
(6,1),(6,2),…,(6,6)}

$n(S)=36$

4.1.2 Probability

- The probability of an event is the proportion of times the event should occur when the experiment is run a large number of times.
- Suppose *S* is a sample space *in which all outcomes are assumed to be equally likely*, and *E* is an event. Then the **probability of *E***, denoted by ***P*(*E*)**, is

$$P(E)=\frac{\text{the number of outcomes in } E}{\text{the number of outcomes in } S}$$

The probability number or value ranges from 0 to 1.

There are three states of expectation in any event:

- ***certainty*** $P(E) = 1$, P(all of us will die someday) =1.
- ***impossibility*** $P(E) = 0$, P(a human will live for 1000 years) = 0.
- ***uncertainty*** $0 < P(E) < 1$, P(A human will live around 70 years)

  - The sum of the probabilities of all the outcomes of an experiment must total 1.
  - P(E does not occur) = P(E') = 1 – P(E).

Example : An experiment consists of tossing two dice.

1-What is the probability that the sum of the two number equal to 4?
E={(1,3),(3,1),(2,2)}, P(E) = 3/36 =1/12.

2- What is the probability that the sum of the two number equal to 7?
E={(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)}, P(E) =1/6.

Example : On February 1, 2003, the Space Shuttle Columbia exploded. This was the second disaster in 113 space missions for NASA. On the basis of this information, what is the probability that a future mission is successfully completed?

Solution:

Probability of a successful mission = $\frac{Number\ of\ successful\ flights}{Total\ number\ of\ flights} = \frac{111}{113} = 0.98$

Mutually exclusive events: Two events that cannot both happen

Example: event A = "Male" and B ="Pregnant" are two mutually exclusive events.
(as no males can be pregnant).

Not mutually exclusive: If two or more events occur at one time.

Example: event A = "Female" and B ="Pregnant" are not mutually exclusive events

Independent events: two events are said to be independent if the occurrence (or not) of one of the events will in no way affect the occurrence (or not) of other.

Alternatively, two events that are defined on two physically different experiments are said to be independent.

Example: event A = "Ahmed's blood pressure" and B ="Hassan's weight"

4.2 Probability Rules

4.2.1 Rule of Addition

1. Mutually Exclusive events

If E and F are mutually exclusive events, then

$P(E \cup F) = P(E) + P(F)$

(i.e. the probability that either E or F occurs is the probability that E occurs, plus the probability that F occurs)

2. General Addition Principle

If E and F are not mutually exclusive events, then

$P(E \cup F) = P(E) + P(F) - P(E \cap F)$

Example: What is the probability that on one roll of a die, I get a 3 or a 4.

- The events are mutually exclusive so the probability is:
- P(3 or 4) = P(3) + P(4) = (1/6) + (1/6) = (2/6) = 0.333.

Example: Consider the example on share prices at the end of a month compared with the price at the beginning of the month. Let the possible outcomes be :
A, a rise in price of more than 10%; P(A) = 0.3
B, a rise in price of less than 10%; P(B) = 0.4
C, no change; P(C) = 0. and D, a fall in price; P(D) = 0.
What is the probability that the share price will rise ?

Solution:

P (the share price will rise) = P(A or B) = P(A) + P(B)
= 0.3 + 0.4 = 0.7.

Example: What is the probability that a randomly selected person from the sample either smokes or drinks coffee.

	Coffee	**No Coffee**	**Total**
Smoker	60	40	**100**
Non-Smoker	115	85	**200**
Total	**175**	**125**	**300**

Event A: A person smokes
Event B: A person drinks coffee
These are not mutually exclusive events because some people smoke and drink coffee.

P(A or B) = P(A) + P(B) – P(A and B)
P(A or B) = (100/300) + (175/300) – (60/300) = (215/300) = 0.7167.

4.2.2 Rule of multiplication

If A and B are two events, then the probability of P(A and B), i.e. probability that A and B occur can be calculated as below:

Probabilities under conditions of statistical independence

$$P(A \text{ and } B) = P(A) \times P(B)$$

***Example*:** At a hospital, there are 4 boys and 6 girls. If we choose two children without replacement, what is the probability that both are boys?

Probability of the first child being boy = 4/10
Thus the probability of both children are boys is equal to 4/10 × 3/9 = 2/15

Example: A family has 3 boys and 2 girls. A second family has 1 boy and 3 girls. If one child is chosen from each family, find the probability that children are

(a) both are boys.

$$\frac{3}{5}\times\frac{1}{4}=\frac{3}{20}$$

(b) one boy and one girl.

$$\left(\frac{3}{5}\times\frac{3}{4}\right)+\left(\frac{2}{5}\times\frac{1}{4}\right)=\frac{11}{20}$$

Example: A survey by the American Automobile association (AAA) revealed 60 percent of its members made airline reservations last year. Two members are selected at random. What is the probability both made airline reservations last year?

Solution:
The probability the first member made an airline reservation last year is .60, $P(R_1)$ = .60
The probability that the second member selected made a reservation is also .60, so $P(R_2)$ = .60.
Since the number of AAA members is very large, you may assume that R_1 and R_2 are independent. Thus, $P(R_1 \text{ and } R_2) = P(R_1)P(R_2)$ = (.60)(.60) = .36

4.2.3 Conditional Probability

The probability of an event B occurring conditional (or given) that event A occurs (or has occurred) is defined as: $P(B|A)=\frac{P(A\cap B)}{P(A)}$

and consequently, $P(A\cap B)=P(B|A).P(A)$

Example: Consider the following events:
A= "A person in the U.S. is alive at age 60"
B = "A person in the U.S. will live to the age of 65"
Compute the probability of the event B|A= "A 60 year-old person in the U.S. will live to the age of 65."

From life tables collected on the U.S. population, it is known that out of 100,000 individuals born, in 1988, 85,331 have reached 60 years of age while 79,123 have reached 65 years of age. Given the large *n* we can consider these proportions as reasonably accurate estimates of P(A) and P(B).

That is, P(A) = P("Lives to 60") = 0.85
P(B) = P("Lives to 65") = 0.79

Also, notice that P(A ∩ B) = P("Lives to 60"and "Lives to 65")
= P("Lives to 65") = P(B) = 0.79.

Finally, $P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{0.79}{0.85} = 0.93$

That is, a person has 79% chance of reaching 65 at birth, but a 60-year-old has 93% chance to reach the same age. The reason of course is that all situations where an individual would have died prior to having reached 60 years of age (i.e., the elements of S that are incompatible with A) have been excluded from the calculations (by the division with P(A)).

Example: The following data are characteristics of the voting-age population regarding the Health insurance issue in the United States. Number of persons is measured in thousands.

	Supported	**Did not Support**	**Total**
Males	53,312	35,245	**88,557**
Females	60,554	36,573	**97,127**
Total	**113,866**	**71,818**	**185,684**

For a randomly selected person from the population, let A be the event that the person selected supported, and B be the event that the person selected is a male. Find each of the following:

1. $P(A)$
2. $P(\overline{A})$
3. $P(A \mid B)$

Solution

1. $P(A) = {}^{113866}\!/_{185684} = 0.613$
2. $P(\overline{A}) = 1 - P(A) = 1 - 0.613 = 0.387$
3. $P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{{}^{53312}\!/_{185684}}{{}^{88557}\!/_{185684}} = 0.602$

Steps for Finding Probabilities

Step 1: List each separate random circumstance involved in the problem.
Step 2: List the possible outcomes for each random circumstance.
Step 3: Assign whatever probabilities you can with the knowledge you have.
Step 4: Specify the event for which you want to determine the probability.
Step 5: Determine which of the probabilities from step 3 and which probability rules can be combined to find the probability of interest.

Example: At a university, 22.9% of the boys and 4.5% of the girls admitted they visit a doctor at least once a month during the previous year. The population consisted of 50.9% girls and 49.1% boys. What is Probability that a randomly selected student will be a male who also visited a doctor?

Solution:
Event A = male, P(A) = 0.491.
Event B = weekly visitor, P(B|A) = 0.229.
P(male and visited a doctor once a month)= P(A and B) = P(A)P(B|A) = (0.491)(0.229) = 0.1124
About 11% of all university students are males and weekly visiting a doctor .

Example: In a cafeteria, 80% of the customers order chips and 60% order buns. If 20% of those ordering buns do not want chips, find the probability that two customers chosen at random,
(i) both order chips but not buns.
(ii) exactly one of them orders a bun only.

Solution:

$P(C) = 0.8, \quad P(B) = 0.6, \quad P(C'|B) = 0.2$

(i) P(both order chips but not buns) = $P(C \cap B') \times P(C \cap B')$

Now, $P(C'|B) = \dfrac{P(C' \cap B)}{P(B)} = 0.2$, Thus, $P(C' \cap B) = 0.2 \times P(B) = 0.2 \times 0.6 = 0.12$

$P(C' \cap B) = P(B) - P(C \cap B) = 0.12$, $P(C \cap B) = 0.6\text{-}0.12 = 0.48$

$P(C \cap B') = P(C) - P(B \cap C) = 0.8 - 0.48 = 0.32$

P(both order chips but not buns) = $P(C \cap B') \times P(C \cap B') = 0.32 \times 0.32 = 0.1024$

(ii) P(exactly one of them orders a bun only) **,**

Let the event of ordering bun only = D, where $P(D) = 0.12$ **, How?!**

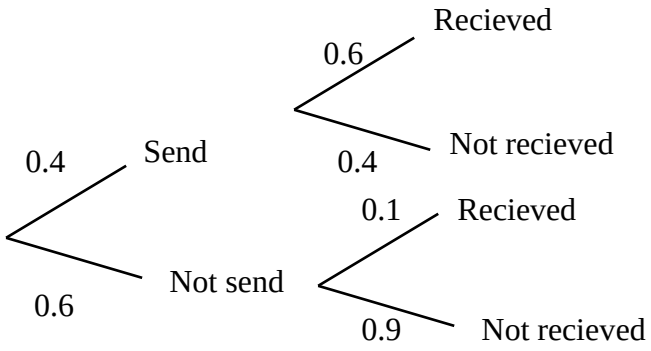
= P(1st D and 2nd D')+P(1st D' and 2nd D)

$= P(D \cap D') + P(D' \cap D) = (0.12 \times 0.88) + (0.88 \times 0.12) = 0.2112.$

Example: A scientist believes that the probability is 0.4 that creatures from Mars are trying to communicate with us by sending high-frequency signals. By using sophisticated equipment, the

scientist hopes to pick up these signals. The manufacturer of the equipment claims that if creatures are indeed sending signals, the probability that the equipment will detect them is 0.6. However if creatures are not sending signals, the probability that the equipment will seem to detect such a signal is 0.1.

a) Represent the above information in a tree diagram.



b) If the scientist uses the equipment, find the probability that:

i) the equipment detects the signals sent by the creatures.

$$P(S \cap D) = P(S) \times P(D) = 0.4 \times 0.6 = 0.24.$$

ii) the equipment doesn't detect any signal.

$P(D') = P(S \cap D') + P(S' \cap D') =$ Pr = $(0.4 \times 0.4)+(0.6 \times 0.9)=0.7$

iii) the equipment detects any signal.

does $P(D) = 1 - P(D') = = 1 - 0.7 = 0.3$

Or , $P(D) = P(S \cap D) + P(S' \cap D) = = (0.4 \times 0.6) + (0.6 \times 0.1) = 0.3.$

iv) the creatures send signals, given that the equipment detects the signals.

$$P(S|D) = \frac{P(S \cap D)}{P(D)} = \frac{0.4 \times 0.6}{0.3} = 0.8$$

v) the creatures send signals, given that the equipment doesn't detect any signal.

$$P(S|D') = \frac{P(S \cap D')}{P(D')} = \frac{0.4 \times 0.4}{0.7} = 0.229$$

EXCERCISES

1- If A={α,β,μ,λ}, how many words of that can be built with length 3, repetition is allowed?

2- In how many different ways can four members drawn for the offices of president, vice president, treasure, and secretary from among the 24 members of a club?

3- In how many different ways can a person gathering data for a market research organization select three of the 20 households living in a certain apartment complex?

4- There are ten teaching assistants available for grading papers in a particular course. The exam consists of four questions, and the professor wishes to select different assistant to grade each question (one assistant per question). In how may ways can assistant be chosen to grade the exam?

5- An experiment has four possible outcomes, A, B, C and D, which are mutually exclusive. Explain why the following assignments of probabilities are not permissible:

a. $P(A) = 0.12, P(B) = 0.63, P(C) = 0.45, P(D) = -0.2;$

b. $P(A) = \frac{9}{120}, P(B) = \frac{45}{120}, P(C) = \frac{27}{120}, P(D) = \frac{46}{120}.$

6- A dice is loaded in such a way that each odd number is twice as likely to occur as each even number. Find $P(G)$, where G is the event that a number greater than 3 occurs on a single roll of dice.

7- What is the total number of possible 5-letter arrangements of the letters W, H, I, T, E if each letter is used only once in each arrangement?

8- How many different number-plates for cars can be made if each number-plate contains four of the digits 0 to 9 followed by a letter A to Z, assuming that

(a) no repetition of digits is allowed?
(b) repetition of digits is allowed?

9- A couple is planning to have three children. Find the following probabilities by listing all the possibilities:

a. two boys and one girl
b. at least one boy.
c. no girls.
d. at most two girls
e. two boys followed by a girl.
f. how does (e) differ from (a).

10- In an experiment involving a toxic substance, the probability that a white mouse will be alive for 10 hours is 7/10, and the probability that a black mouse will be alive for 10 hours is 9/10. Find the probability that, at the end of 10 hours,

a. both mice will be alive.
b. Only the black mouse will be alive.
c. Only the white mouse will be alive.
d. At least one mouse will be alive.

11- Assume that the engine component of a spacecraft consists of two engines in parallel. If the main engine is 95% reliable, the backup is 80% reliable, and the engine component as a whole is 99% reliable, what is the probability that

(i) Both engines will be operable; i.e $P(M \cap B)$?

(ii) The main engine will fail but the backup will be operable, i.e $P(M' \cap B)$?

(iii) The engine component will fail, i.e $P(M \cup B)'$?

12- A golfer has 12 golf shirts in his closet. Suppose 9 of these shirts are white and the others blue. He gets dressed in the dark, so he just grabs a shirt and puts it on. He plays golf two days in a row and does not do laundry.
What is the likelihood both shirts selected are white?

13- A bag contains 3 green balls and 2 white balls, two balls are drawn together, what is the probability that:
1. Both are green.
2. One is green and one is white.

14- In a certain school class, consisting of 60 girls and 40 boys, it is observed that 24 girls and 16 boys wear eyeglasses. If a student is picked at random from this class, find the following probabilities:

a) the picked student is wearing eyeglasses.
b) the picked student is wearing eyeglasses and being a boy.
c) the picked student is wearing eyeglasses and not being a boy.
d) are the two events; wearing glasses and being a boy independent?

16- A dice is loaded in such a way that each odd number is twice as likely to occur as each even number. Find P(G), where G is the event that a number greater than 3 occurs on a single roll of dice.

CHAPTER 5
THE BINOMIAL AND NORMAL DISTRIBUTION

5.1 Probability Distributions:

A probability distribution shows all possible values of a random variable along with their respective probabilities**.**

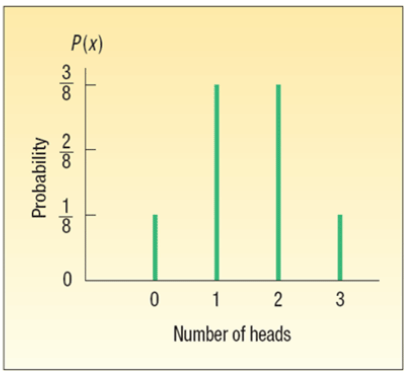
Experiment: Toss a coin three times. Observe the number of heads.

Possible Result	Coin Toss			Number of Heads
	First	Second	Third	
1	T	T	T	0
2	T	T	H	1
3	T	H	T	1
4	T	H	H	2
5	H	T	T	1
6	H	T	H	2
7	H	H	T	2
8	H	H	H	3

What is the probability distribution for the number of heads?

Probability Distribution of Number of Heads Observed in 3 Tosses of a Coin

Number of Heads, x	Probability of Outcome, $P(x)$
0	$\frac{1}{8} = .125$
1	$\frac{3}{8} = .375$
2	$\frac{3}{8} = .375$
3	$\frac{1}{8} = .125$
Total	$\frac{8}{8} = 1.000$



Characteristics of a probability distribution:

1. The probability of a particular outcome is between 0 and 1 inclusive.
2. The outcomes are mutually exclusive events.
3. The list is exhaustive. So the sum of the probabilities of the various event is equal to 1.

Example: A balanced coin is tossed three times. Find the probability distribution for the total number of heads.

Solution:

Example: Check whether the function given by

$$f(x) = \frac{x+2}{25}, \text{ for } x = 1, 2, 3, 4, 5$$

Can serve as the probability distribution of a discrete random variable.

Solution:

Example: A box contains 4 white and three red balls, if X denotes the number of red balls in three draws with replacement. Find the probability distribution of x.

Solution:

Types of Random probability distributions

A discrete probability distribution: It is the probability distribution of a discrete random variable, for example: Binomial Distribution, Poisson Distribution, etc.

A continuous probability distribution: It is the probability distribution of a continuous random variable, for example: The normal distribution, the *t* distribution, the chi-square distribution, the F distribution, etc.

5.2 Binomail Distribution

A binomial experiment has the following conditions:
1. There are *n* repeated trials.
2. Each trial has only two possible outcomes-success or failure, girl or boy, sick or well, dead or alive, at risk or not at risk, infected–not infected, or simply positive–negative etc
3. The probabilities of the two outcomes remains constant from trial to trial.

The probability of success denoted by *p*.
The probability of a failure, $q = (1 - p)$.

4. The outcome of each trial is independent of the outcomes of any other trial; that is, the outcome of one trial has no effect on the outcome of any other trial.

$$P(r \text{ success}) = \binom{n}{r} p^r q^{n-r} = \frac{n!}{r!(n-r)!} p^r q^{n-r}$$

"!" represents the factorial function.
$n! = (n)(n - 1)(n - 2)(n - 3) \ldots (1)$.

For example: $4! = (4)(3)(2)(1) = 24$

By definition, $0! = 1$

Example: In a three-child family if the probability of having a boy is 0.5. What is the probability of having
a- two girls and one boy,
b- 3 boys,
c- and at least one boy?

Solution:

Let a family having a boy is a success with, $p = 0.5$, $n = 3$.

a- $P(r = 1) = \binom{3}{1}(0.5)^1(0.5)^2 = \frac{3!}{1!\,2!}(0.5)^3 = 3(0.125) = 0.375$.

b- $P(r = 3) = \binom{3}{3}(0.5)^3(0.5)^0 = \frac{3!}{3!}(0.5)^3 = 0.125$.

by considering the girl is the success, $n = 3$, $r = 0$

$P(r = 0) = \binom{3}{0}(0.5)^0(0.5)^3 = \frac{3!}{3!}(0.5)^3 = 0.125$.

c- $P(\text{at least one boy}) = P(r = 1) + P(r = 2) + P(r = 3) = 1 - P(r = 0)$

$$= 1 - \binom{3}{0}(0.5)^0(0.5)^3 = 1 - 0.125 = 0.875.$$

Example: Ten individuals are treated surgically. For each individual there is a 70% chance of successful surgery. Among these 10 people, the number of successful surgeries follows a binomial distribution with n =10, and p =0.7.
What is the probability of exactly 5 successful surgeries?

Solution:

$$P(r=5)=\binom{10}{5}(0.7)^5(0.3)^5=\frac{10!}{5!\,5!}(0.7)^5(0.3)^5=(252)(0.168)(0.0024)=0.102$$

Example: Let r = number of patients who will experience nausea following treatment with Phe-Mycin, n = 4 and p=0.1. Find the probability that 2 of the 4 patients treated will experience nausea.

Solution:

$$P(r=2)=\binom{4}{2}(0.1)^2(0.9)^2=\frac{4!}{2!\,2!}(0.1)^2(0.9)^2=0.0486$$

Mean & Variance of the Binomial Distribution

The mean of the binomial distribution is found by:

$$\mu = np$$

The variance of the binomial distribution is found by:

$$\sigma^2 = np(1-p)$$

Example: Obtain the expected value (mean) and the standard deviation of the of successful surgeries in the previous example.

Solution:

The expected value, $\mu = np$ = (10) (0.7) = 7.

The variance, $\sigma^2 = np(1-p)$ = (10) (0.7) (0.3) = 2.1.

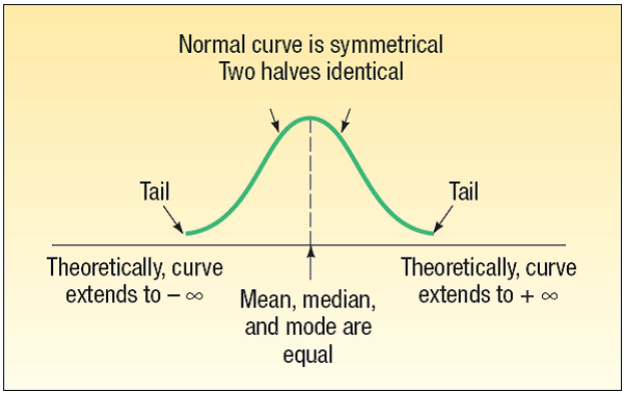
Thus, the standard deviation, $\sigma = \sqrt{np(1-p)} = \sqrt{2.1} = 1.45$

5.3 Normal Probability Distribution:

- It is the most important distribution for describing a continuous random variable.
- It has been used in a wide variety of applications:
  - Heights and weights of people.
  - Test scores.
  - Scientific measurements.
  - Amounts of rainfall.
- It is widely used in statistical inference
- The Normal Probability Density Function is given by:

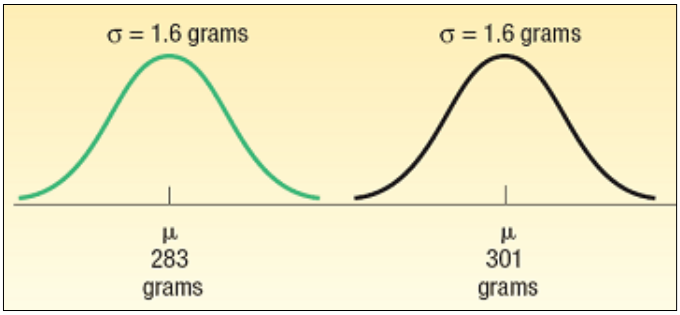
$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-(x-\mu)^2/2\sigma^2}$$

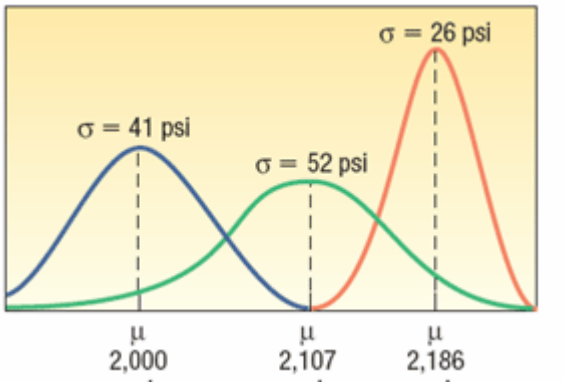
μ = mean
σ = standard deviation
π = 3.14159
e = 2.71828



Characteristics of a Normal Probability Distribution

- It is **bell-shaped** and has a single peak at the center of the distribution.
- The arithmetic mean, median, and mode are equal.
- It is **symmetrical** about the mean.
- The location of a normal distribution is determined by the mean, μ, the dispersion or spread of the distribution is determined by the standard deviation, σ .
- There are an infinite number of normal distributions By varying the parameters σ and μ, we obtain different normal distributions.





The Standard Normal Probability Distribution

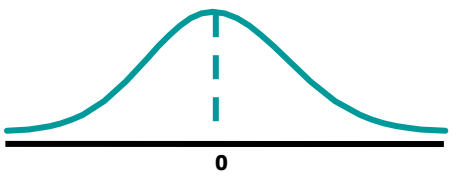
- The standard normal distribution is a normal distribution with a mean of 0 and a standard deviation of 1.
- It is also called the *Z* distribution.
- A *z*-value is the distance between a selected value, designated *X*, and the population mean μ, divided by the population standard deviation, σ.
- The formula of the standardized value is: $Z = \dfrac{X - \mu}{\sigma}$
- By standardising any normally distributed random variable, we can use just the table namely, ***Areas Under the Normal Curve Or Areas of a Standard Normal Distribution,*** Such tables are usually found in the Appendix of most of statistics books.

Example: Assume that the Intelligence Quiz (IQ) of a given population is normally distributed with μ =100 and σ =15.

i. what is the proportion of persons having IQs between 100 and 120.
ii. what is the proportion of persons has IQs greater than 120?
iii. what is the proportion of persons with IQs between 80 and 120?
iv. what is the proportion of persons with IQs between 95 and 125?
v. what is the Z value of the normal curve that marks the upper 10% of the area?
vi. what is the 90th percentile of IQ scores?

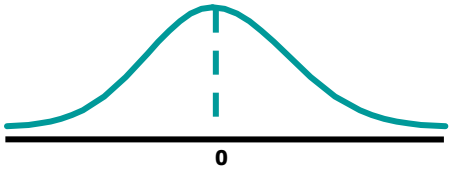
Solution:

a. The Z corresponding to x =100 is $Z = \dfrac{100-100}{15} = 0$ and Z corresponding to x =120 is $Z = \dfrac{120-100}{15} = 1.33$.

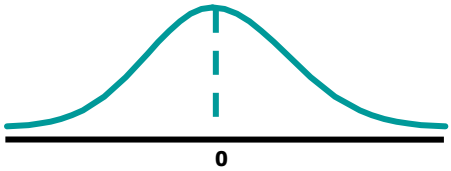


By using Table B to find the area for a Z of 1.33, you'll find the answer to be 0.4082.
Therefore the proportion of persons having IQs between 100 and 120 is 0.4082, about 41%.

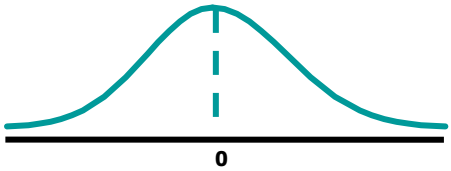
ii. Since the area to the right of Z = 0 is 0.5, and the area between Z= 0 and 1.33 is 0.4082, by subtraction you will obtain the area beyond a Z of 1.33, namely, 0.5 - 0.4082 = 0.0918.
So the answer is that about 9% have IQs over 120.



iii. The Z corresponding to x =80 is $Z = \frac{80-100}{15} = -1.33$. Thus the area under the standard normal curve between -1.33 and 1.33.
Using the symmetry argument, simply double the area between Z = 0 and 1.33, 2(0.4082) = 0.8164.



iv. The Z corresponding to x =95 is $Z = \frac{95-100}{15} = -0.33$ and Z corresponding to x =125 is $Z = \frac{125-100}{15} = 1.67$.

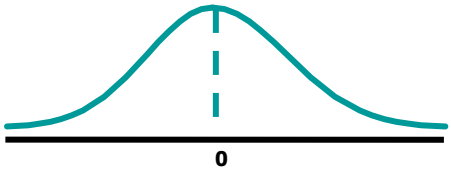


The required area is divided into two sub-areas , $A = A_1 + A_2$.
The area A_1 is between Z = 0 and - 0.33 which is the same as the area between Z = 0 and 0.33.
In Table B, The area of A_1 is 0 >1293.
The area A_2 is between Z= 0 and 1.67 is 0.4525. Thus $A = A_1 + A_2$ = 0.1293 + 0.4525 = 0.5818.

v. The desired normal deviate is that value corresponding to 0.40 of the area.
In Table B the value is found to be approximately Z=1.28.

vi. First we need to find the Z score of the 90th percentile to be 1.28. But the 1.28 is not in term of IQs. Thus, by applying the standardization formula then you have

$Z = \frac{100-100}{15} = 0$. Therefore, x = 1.28 (15) +100 = 119.2



EXCERCISES

1. Verify that $f(x) = \frac{2x}{k(k+1)}$ for $x = 1,2,3,\ldots, k.$ can serve as the probability distribution of a random variable with given range.

2. A software company Mysoftcom is concerned about a low retention rate for employees. On the basis of past experience, management has seen a turnover of 10% of the hourly employees annually. Thus, for any hourly employees chosen at random, management estimates a probability of 0.1 that the person will not be with the company next year. Choosing 3 hourly employees at random, what is the probability that 1 of them will leave the company this year?

3. The probability that a patient recovers from SARS is 0.4. If 15 people are known to have contracted this disease, what is the probability that
   (i) at least 13 survive
   (ii) at most 2 die

4. Find the areas under the normal curve that lie between the given values of Z.
   a. Z = 0 and Z = 2.37.
   b. Z = 0 and Z = -1.94.
   c. Z = - 1.85 and Z = 1.85.
   d. Z = - 0.76 and Z = 3.09.
   e. Z = - 2.77 and Z = - 0.96.

2. What Z scores correspond to the following areas under the normal curve?
   a. area of 0.05 to the right of +Z.
   b. area of 0.01 to the left of –Z.
   c. area of 0.9 between $\pm Z$.
   d. area of 0.95 between $\pm Z$.

3. If the heights of male youngsters are normally distributed with a mean of 60 inches. and standard deviation of 10 inches, what percentage of the boys' heights (in inches) would we expect to be
   a. between 45 and 75?
   b. Between 30 and 90?
   c. Less than 50?
   d. 45 or more?
   e. 75 or more?
   f. Between 50 and 75?

4. Assume that the age at onset of disease ***X*** is distributed normally with a mean of 50 years and a standard deviation of 12 years. What is the probability that an individual afflicted with X had developed it before age 35?

5. Describe the normal distribution.

6. why do statisticians prefer to work with the standard normal distribution rather than the normal distribution?

7. If a clinical variable $\boldsymbol{X}$ is normally distributed with a standard deviation of 8, what is the mean value of $\boldsymbol{X}$, if 0.3413 of the values of $\boldsymbol{X}$ is greater than 58?

8. The probability that a patient recovers from a rare blood disease is 0.4. If 10 people are known to have contracted this disease, what is the probability that
   a. exactly 3 survive,
   b. at least 8 survive, and
   c. from 2 to 5 survive?

9. A biologist is studying a new hybrid tomato. It is known that the seeds of this hybrid tomato have probability 0.70 of germinating. The biologist plants 10 seeds.

   a. What is the probability that exactly 8 seeds will germinate?
   b. What is the probability that at least 8 seeds will germinate?
   c. What is the probability that at most 8 seeds will germinate?

CHAPTER 6
CORRELATION AND REGRESSION

6.1 Introduction

In analyzing data for the health sciences, we find that it is frequently desirable to investigate the relationship between two or more variables.

The nature and strength of the relationships between variables, may be examined by correlation and regression analysis .

Examples of two related variables:

- Blood pressure and age.
- Height and weight.
- The concentration of an injected drug and heart rate.
- The consumption level of some nutrient and weight gain.

Correlation Analysis is a statistical technique concerned with measuring the strength of the relationship between variables.

Regression Analysis is used to predict or estimate the value of one variable corresponding to a given value of another variable.

Scatter Diagram: is a chart that portrays the relationship between the two quantitative variables.

Features of scatter diagram:

- It is the usual first step in correlations analysis.
- One variable is called independent (*X*) and the second is called dependent (*Y*).
- Points are not joined.

In the analysis of relationships it is important to classify variables into two different types:

The Dependent Variable is the variable being predicted or estimated.
The Independent Variable provides the basis for estimation, it is the predictor variable.

6.2 Correlation Coefficient:

Correlation coefficient of variables ***X*** and ***Y*** shows how strongly the values of these variables are related to each other.

Characteristics of correlation coefficient:

- It is denoted by r , where $r \in [-1, 1]$.
- If the correlation coefficient is positive (*a direct relationship*), then both variables are simultaneously increasing (or simultaneously decreasing).
- If the correlation coefficient is negative (*an inverse relationship*), then when one variable increases the other decreases, and reciprocally.
- The correlation coefficient measures the degree of linear association between two variables.

There is a strong relationship if $r \in [0.8, 1]$ or $r \in [-1, -0.8]$,
a moderate relationship if $r \in (0.5, 0.8)$ or $r \in (-0.8, -0.5)$,
a weak relationship if $r \in [-0.5, 0.5]$.

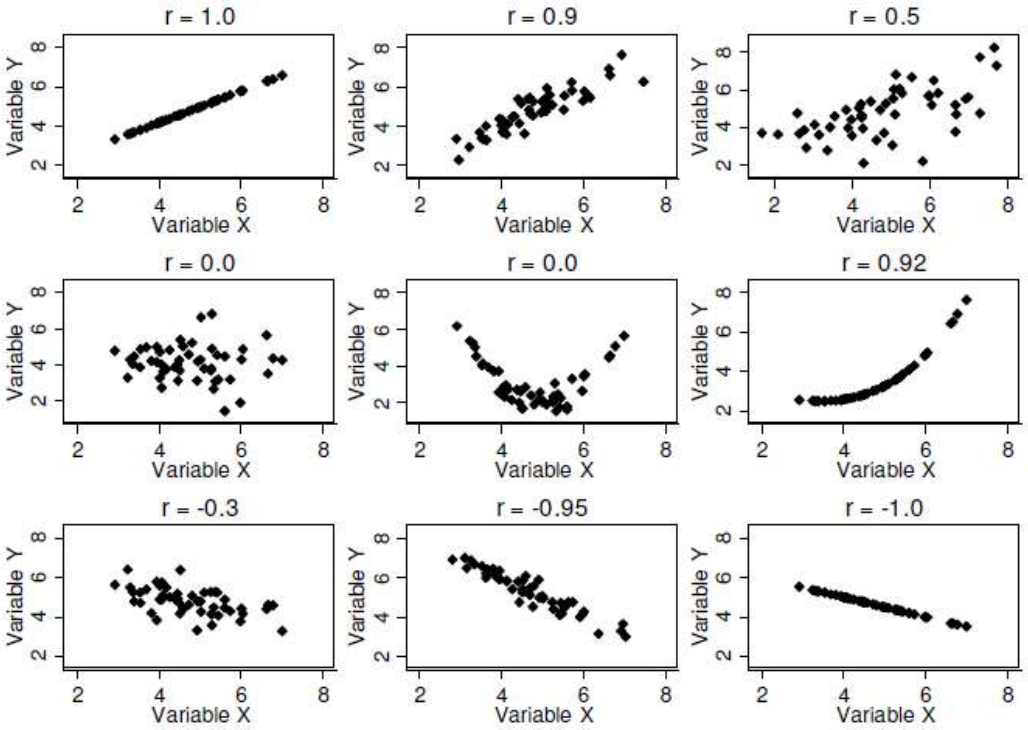


Figure 6.1 scatter diagrams with different correlation coefficient.

6.2.1 Simple Correlation coefficient (*r*)

It is also called Pearson's correlation, it measures the nature and strength between two variables of the quantitative type.

The simple correlation coefficient is obtained using the following formula:

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left(\sum x^2 - \frac{(\sum x)^2}{n}\right)\cdot\left(\sum y^2 - \frac{(\sum y)^2}{n}\right)}}$$

where n is the sample size, x is the independent variable and y is the dependent variable.

***Example*:** A sample of 6 children was selected, their age in years and gained weight in the last year in pounds (1 pound = 0.453 k.g.) was recorded as shown in the following table . Find the correlation between age and weight.

No	Age (years)	Gained weight
1	7	12
2	6	8
3	8	12
4	5	10
5	6	11
6	9	13

These 2 variables are of the quantitative type, one variable (Age) is called the independent and denoted as (*X*) variable and the other (Gained weight) is called the dependent and denoted as (*Y*) variables. To find the relation between age and weight compute the simple correlation coefficient, the following table will ease the calculation of the correlation coefficient

No	Age (x)	Gained weight (y)	xy	x^2	y^2
1	7	12	84	49	144
2	6	8	48	36	64
3	8	12	96	64	144
4	5	10	50	25	100
5	6	11	66	36	121
6	9	13	117	81	169
Total	**41**	**66**	**461**	**291**	**742**

$$r = \frac{461 - \frac{41 \times 66}{6}}{\sqrt{\left(291 - \frac{(41)^2}{6}\right)\cdot\left(742 - \frac{(66)^2}{6}\right)}} = 0.759$$

r indicates a strong direct correlation.

Example: Relationship between Anxiety and Test Scores

No.	Anxiety (x)	Test score (y)	xy	y^2	x^2
1	10	2	20	4	100
2	8	3	24	9	64
3	2	9	18	81	4
4	1	7	7	49	1
5	5	6	30	36	25
6	6	5	30	25	36
Total	**32**	**32**	**129**	**204**	**230**

$$r = \frac{(6)(129) - (32)(32)}{\sqrt{\left(6(230) - 32^2\right)\left(6(204) - 32^2\right)}} = \frac{774 - 1024}{\sqrt{(356)(200)}} = -0.94.$$ Indirect strong correlation

6.2.2 Spearman Rank Correlation Coefficient (r_s)

It is a non-parametric measure of correlation used in the case of ordinal or qualitative (ratio or relative) variables.

- This procedure makes use of the two sets of ranks that may be assigned to the sample values of *x* and *y*.

Spearman Rank correlation coefficient could be computed in the following cases:

1. Both variables are quantitative.
2. Both variables are qualitative ordinal.
3. One variable is quantitative and the other is qualitative ordinal.

Procedure:

1. Rank the values of X from 1 to n , where n is the numbers of pairs of values of X and Y in the sample.
2. Rank the values of Y from 1 to n.
3. Compute the value of d_i for each pair of observations by subtracting the rank of Y_i from the rank of X_i
4. Square each d_i and compute $\sum d_i^2$ which is the sum of the squared values.
5. Apply the following formula

$$r_s = 1 - \frac{6\sum(di)^2}{n(n^2-1)}$$

The value of r_s denotes the magnitude and nature of association giving the same interpretation as simple r.

Example: In a study of the relationship between education level and health awareness, the following data was obtained. Find the relationship between them and comment.

No.	Education level (X)	Health awareness (Y)
1	preparatory.	25
2	primary.	10
3	university.	8
4	secondary	10
5	secondary	15
6	illiterate	50
7	university.	60

Solution:

No.	(X)	(Y)	Rank (X)	Rank (Y)	d_i	d_i^2
1	Preparatory	25	5	3	2	4
2	Primary	10	6	5.5	0.5	0.25
3	University	8	1.5	7	-5.5	30.25
4	secondary	10	3.5	5.5	-2	4
5	secondary	15	3.5	4	-0.5	0.25
6	illiterate	50	7	2	5	25
7	university	60	1.5	1	0.5	0.25
Total						64

$$r_s = 1 - \frac{6 \times 64}{7(48)} = -0.1$$

Comment: There is an indirect weak correlation between education level and health awareness.

Would you justify your findings?!

6.3 Regression Analyses

Characteristics of regression analysis:

- It uses a variable (*X*) to predict some outcome variable (*Y*) if there is a linear relationship between *X* and *Y*.
- It tells how values in *Y* change as a function of changes in values of *X*.
- Calculates the "best-fit" line for a certain set of data.
- The regression line makes the sum of the squares of the residuals smaller than for any other line

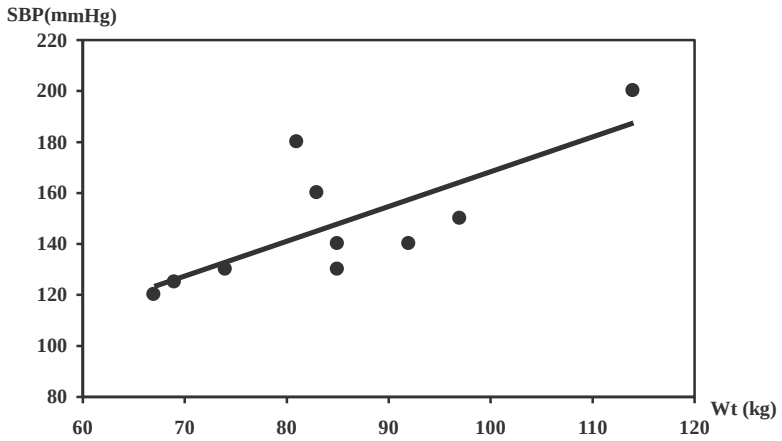
Least Square Method: a procedure that minimizes the vertical deviations of plotted points surrounding a straight line

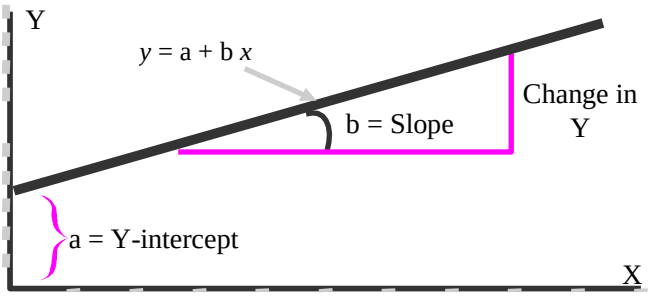
By using the least squares method, we are able to construct a best fitting straight line to the scatter diagram points and then formulate a regression equation in the form of:

$$\hat{y} = \hat{a} + \hat{b}x + \varepsilon$$

where $\hat{y}$ is the predicted values of dependent variable y, $\hat{a}$ is the intercept of the best fitted line with the y-axis, $\hat{b}$ is the slope of the best-fitted line, and ε is the residuals term.

Regression minimizes the residuals





The least square estimate of the simple linear regression parameters are given by

$$\hat{b} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} \quad \text{and} \quad \hat{a} = \bar{y} - b\bar{x}\,.$$

The estimated residuals can be obtained by using the following formula:

$$\varepsilon = y - \hat{y}$$

Example: A sample of 6 persons was selected the value of their age (X variable) and their gained weight is demonstrated in the following table. Find the regression equation and what is the predicted weight when age is 8.5 years. Obtain the error of predication when the age is 7 years.

no.	Age (X)	Gained weight (Y)
1	7	12
2	6	8
3	8	12
4	5	10
5	6	11
6	9	13

***Solution*:**

Construct the following table to ease the estimation of regression parameters

no.	Age (X)	Gained weight (Y)	xy	x^2
1	7	12	84	49
2	6	8	48	36
3	8	12	96	64
4	5	10	50	25
5	6	11	66	36
6	9	13	117	81
Total	**41**	**66**	**461**	**291**

The mean of X , is $\bar{x} = \frac{41}{6} = 6.83$.

The mean of Y , is $\bar{y} = \frac{66}{6} = 11$.

The estimate of the slop parameter is

$$\hat{b} = \frac{461 - \frac{41 \times 66}{6}}{291 - \frac{(41)^2}{6}} = 0.92.$$

The estimate of intercept parameter is

$$\hat{a} = 11 - 0.92(6.83) = 4.675.$$

Thus, the regression equation is given by

$$\hat{y} = 4.675 + 0.92x + \varepsilon$$

The predicted weight when age is 8.5 years can be obtained as follows

$$\hat{y}_{(8.5)} = 4.675 + 0.92(8.5) = 12.50 \text{ pound.}$$

The predicted weight when age is 7 years can be obtained as follows

$$\hat{y}_{(7)} = 4.675 + 0.92(7) = 11.115 \ pound.$$

Therefore, the error of prediction is

$$\varepsilon = 12 - 11.115 = 0.885 pound.$$

we create a regression line by plotting two estimated values for *y* against their *x* component, then extending the line right and left.

Example: The following are the age (in years) and systolic blood pressure of 20 apparently healthy adults.

- Find the correlation between age and blood pressure using simple and Spearman's correlation coefficients, and comment.
- Find the regression equation.
- What is the predicted blood pressure for a man aging 25 years?

Age (x)	**B.P (y)**	**Age (x)**	**B.P (y)**
20	120	46	128
43	128	53	136
63	141	60	146
26	126	20	124
53	134	63	143
31	128	43	130
58	136	26	124
46	132	19	121
58	140	31	126
70	144	23	123

Solution:

Following summations can be obtained

$\sum x = 852$

$\sum y = 2630$

$\sum x^2 = 41678$

$\sum xy = 114486$

The estimate of the slop parameter is $\hat{b} = \dfrac{114486 - \dfrac{852 \times 2630}{20}}{41678 - \dfrac{852^2}{20}} = 0.4547$

The estimate of intercept parameter is $\hat{a} = \dfrac{2630}{20} - 0.4547(\dfrac{852}{20}) = 112.13$

Thus, the regression equation is given by

$$\hat{y} = 112.13 + 0.4547x + \varepsilon$$

6.4 Correlation and Regression

Correlation describes the strength of a **linear** relationship between two variables **Linear** means "**straight line**"

Regression tells us how to draw the straight line described by the correlation

6.4.1 Limitations of Correlation

a- linearity:
can't describe non-linear relationships.

b- truncation of range:
under estimate strength of relationship if you can not see full range of *x* value.

c- no proof of causation:
third variable problem: could be 3rd variable causing change in both variables.

6.5 Multiple Regression

Multiple regression analysis is a straightforward extension of simple regression analysis which allows more than one independent variable.

EXERCISES

1- Compute the correlation coefficient and regression equation for Blood glucose (x) and serum cholesterol (y)

$n = 100,\ \sum x = 15{,}214,\ \sum y = 21{,}696,\ \sum x^2 = 2{,}611{,}160,\ \sum xy = 3{,}371{,}580,\ \sum y^2 = 4{,}856{,}320$

2- Compute the correlation coefficient and regression equation for Ponderal index (x) and systolic blood pressure (y).

$n = 100,\ \sum x = 13{,}010,\ \sum y = 4{,}052,\ \sum x^2 = 1{,}736{,}990,\ \sum xy = 527{,}185,\ \sum y^2 = 164{,}521.$

3- In a study of systolic blood pressure (SBP) in relation to whole blood cadmium (Cd) and zinc (Zn) levels the following data were obtained

Cd (ppm/g ash)	68	63	56	48	96	70	66
Zn (ppm/g ash)	127	118	78	76	181	134	122
SBP (mmHg)	166	162	116	120	160	120	182

a- Make a scatter diagram of cadmium and systolic blood pressure, using the later as the dependent variable.
b- Judging from the diagram, would you be justified in using linear regression analysis to determine a line of best fit for cadmium and blood pressure? Why or why not?
c- Compute the correlation coefficient for cadmium and blood pressure.
d- Using zinc as the dependent variable, plot the scatter diagram of cadmium and zinc.
e- Does the diagram of (d) provide justification for using regression analysis to determine a line of best fit?
f- Calculate the equation of the line on the scatter diagram for d.

CHAPTER 7
SAMPLING DISTRIBUTION

The distribution of values of a statistic obtained from repeated samples of the same size from a given population is called the sampling distribution of that statistic.

The sampling distribution of the mean is the distribution of all possible sample means of sample size *n* taken from a population.

The mean of the variable $\overline{X}$ is always equal to the mean of the population from which the random samples are chosen.

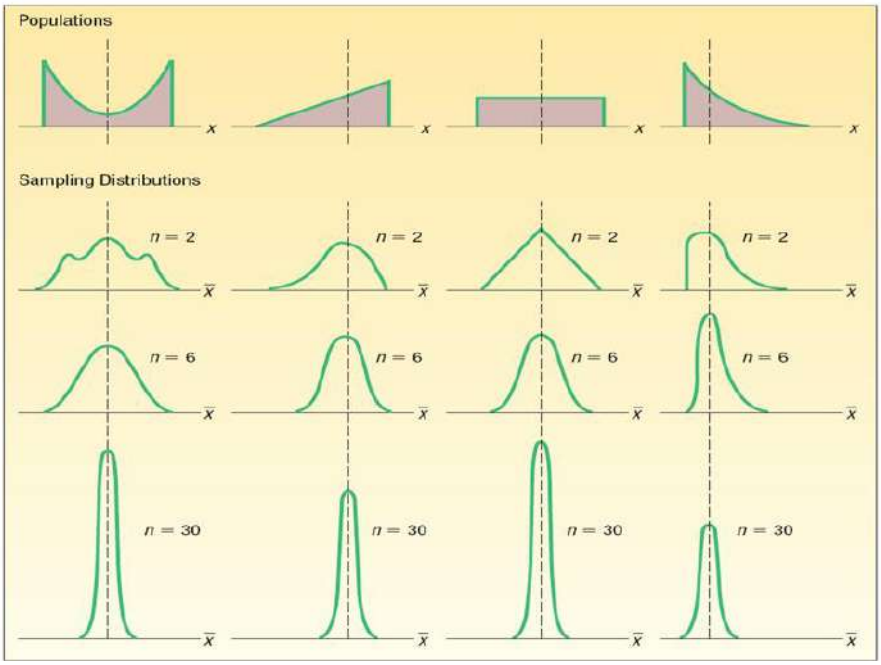
7.1 Central Limit Theorem

- For a population with a mean μ and a variance σ^2 the sampling distribution of the means of all possible samples of size *n* generated from the population will be approximately normally distributed.
- The mean of the sampling distribution equal to, $\mu_{\bar{x}} = \mu$ and the standard deviation equal to, $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.
- If the population is ***not normal***, it can safely be applied if the sample size exceeds **30 (*n*>30)**
- If the population is ***normal***, the sampling distribution is *also normal for any sample size*.

Standard Error of Mean

- It is the standard deviation of all possible sample means, $\overline{X}$
- It is less than population standard deviation
- It can be calculate using the following formula:

$$se(\bar{x}) = \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$



Results of the Central Limit Theorem for Several Populations

- If a population follows the normal distribution, the sampling distribution of the sample mean will also follow the normal distribution.
- To determine the probability a sample mean falls within a particular region, use:

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}.$$

Example: If the mean and standard deviation of serum iron values for healthy men are 120 and 15 micrograms per 100 ml, respectively, what is the probability that a random sample of 50 normal men will yield a mean between 115 and 125 micrograms per 100 ℓ?

Solution

Observe here that we do not know the distribution of serum iron values, but the sample size is **large**, so we can apply the Central Limit Theorem for the sample means.

In this example we aim to find the $P(115 \le \bar{x} \le 125)$

$$P\left(\frac{\bar{x}_1 - \mu}{\sigma/\sqrt{n}} \le z \le \frac{\bar{x}_2 - \mu}{\sigma/\sqrt{n}}\right) = P\left(\frac{115-120}{15/\sqrt{50}} \le z \le \frac{125-120}{15/\sqrt{50}}\right)$$

$$P(-2.36 \le z \le 2.36) = 2\,P(0 \le z \le 2.36) = 2(0.4909) = 0.9818.$$

Example: The serum cholesterol levels for all 20-74 year-old US males has mean μ = 211 mg/100 ml and the standard deviation is σ = 46 mg/100 ml. That is, each individual serum cholesterol level is distributed around μ = 211 mg/100 ml, with variability expressed by the standard deviation σ.
Let's say that we take a sample of size *n* = 25. What if $\bar{x} \ge$ 217 mg/100 ml?

Solution:

If μ = 217 mg/100 ml, then from the Central Limit theorem we have that

$$P(\bar{x} \ge 217) = P\left(\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \ge \frac{217-211}{46/\sqrt{25}}\right) = P(Z \ge 0.65) = 0.258$$

7.2 Using the Sampling Distribution of the Sample Mean (Sigma Unknown)

In some cases the population standard deviation σ is unknown. Without σ we are unable to calculate the normal deviate.

If the standard deviation σ is unknown, it may be estimated by the sample standard deviation s,

$$s = \sqrt{\frac{\sum(x - \bar{x})}{n-1}}.$$

The use of the sample standard deviation *s* instead of the population standard deviation σ. Leads to use a new distribution, which is the Student- *t* distribution.

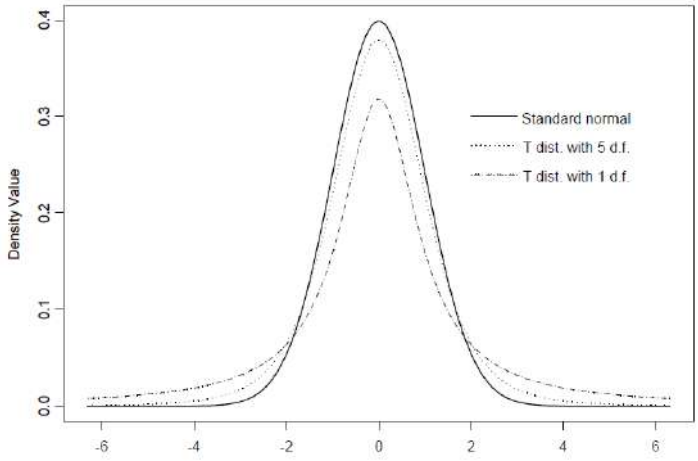
The equation for its t-score is

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

The *t* distribution is similar to the standard normal distribution in that it is:

- Unimodel
- Bell-shaped
- Symmetrical

Its curve has more variance than the normal distribution.



Areas under the curve, designated as α in Table C, are a function of a quantity called degrees of freedom (df)), where df= *n*-1.

When should the t -distribution be used?

Use it when the population standard deviation is not known and sample less than 30.

Example: A random sample of size 9 is drawn from a normal population with unknown variance and a mean of 20. Find the probability that the sample mean is greater than 24 if the sample standard deviation is 4.1436

Solution:

Given: μ = 20, *n* = 9, *s* = 4.1436.

$P(\bar{x} \geq 24) =$

Convert to T statistic:

$$P(\frac{\overline{X} - \mu}{s/\sqrt{n}} \geq \frac{24-20}{4.1436/\sqrt{9}})$$

$$P(t > 2.896) = 0.01$$

Example: Supposing you know that average Malaysian man exercises for 60 minutes a week and a random sample of 25 means you have drawn from the population has a mean of 65 minutes per week with a standard deviation of 10 minutes.
What is the probability of getting a random sample of this size with a mean less than or equals 65 if the actual population mean is 60?

Solution:

Given: μ = 60, *n* = 25, $\bar{x}$ = 65, *s* = 10.

Since the population standard deviation is not known, we use the t-distribution with T.

$$t = \frac{\overline{X} - \mu}{s/\sqrt{n}} = \frac{65-60}{10/\sqrt{25}} = 2.5$$

$$P(\overline{X} \leq 2.5) = P(T \leq 2.5) = ?$$

Referring to Table of t-Curve Tail areas with df = (n –1) = 24, we obtain P(t > 2.5) = 0.01.

Thus $P(\overline{X} \leq 2.5) = 1 - P(t > 2.5) = 0.99$

CHAPTER 8
TESTS OF SIGNIFICANCE

8.1 Introduction

While doing a particular research, one often proposes hypothesis, design an experiment and collect the data to be analyzed.
In order to reach a conclusion about the hypothesis:

• Data may support the research hypothesis or
• Data may not support the research hypothesis.

Since in an experiment, most often one looks at a sample of the population, thus there always exists a chance that our conclusion about the hypothesis may turn out to be wrong!

Thus the main objective of hypothesis testing is

To validate (to accept or reject) an estimation of a particular sample statistics whether it reflects the population parameter.

A statistical hypothesis, or just hypothesis, is a conjecture or claim concerning one or more population parameter.

Evidence from the sample that is inconsistent with the stated hypothesis leads to a rejection of the hypothesis, whereas evidence supporting the hypothesis leads to its acceptance.

Null hypothesis is a hypothesis which is tested for possible rejection under the assumption that it is true and is denoted by H_0 and always contains the equality "=" sign.

Alternative hypothesis is a complimentary hypothesis to null hypothesis and is denoted by H_a (or H_1). It never contains the equality "=" sign.

Logical Argument of Hypothesis:

- The rejection of H_0 leads to the acceptance of an alternative hypothesis, H_a.
- A test of hypothesis is a method of using sample data to decide whether H_0 should be rejected.
- A null hypothesis concerning ***a population parameter*** will always be stated so as to specify ***an exact value*** of the parameter, whereas the alternative hypothesis allows for the possibility of several values.

Hence, if the null hypothesis $H_0 : p = 0.5$ for a binomial population, the alternative hypothesis H_1 would be one of the following:

$$p > 0.5, \quad p < 0.5, \quad p \neq 0.5.$$

Example:

H_0: The mean weights at birth of children born to urban women is 3.5 kg.

H_1: The mean weights at birth of children born to urban women is not 3.5 kg

H_1: The mean weights at birth of children born to urban women is more than 3.5 kg

H_1; The mean weights at birth of children born to urban women is less than 3.5 kg

Example :
For each of the following assertions, state whether it is a legitimate statistical hypothesis and why:

(a) $H_0 : \sigma \geq 100$

Yes. It is an assertion about the value of a parameter.

(b) $H_0 : \tilde{x} = 45$

No. The sample median $\tilde{x}$ is not a parameter.

(c) $H_0 : s \leq 0.20$

No. The sample standard deviation s is not a parameter.

(d) $H : \sigma_1 / \sigma_2 < 1$

Yes. The assertion is that the standard deviation of population #2 exceeds that of population #1.

(e) $H_0 : \bar{x} - \bar{y} = 5$

No. $\bar{x}$ and $\bar{y}$ are statistics rather than parameters, so cannot appear in a hypothesis.

How to define hypothesis?

First read the problem carefully and determine the claim that you want to test.

- If the claim suggest a simple direction such as more than, less than, superior to, inferior to, and so on, then H_1 will be stated using the inequality symbol (< or >) corresponding to the suggested direction.

- If the claim suggest a compound direction (equality as well as direction) such as at least, equal to or greater, at most, no more than, and so on, then this entire compound direction ($\leq$ or $\geq$) is expressed as H_0, but using only the equality sign, and Ha is given by the opposite direction.

- If no direction whatsoever is suggested by the claim, then Ha is stated using the not equal symbol, $\neq$.

In summary:

- Null hypothesis H_0 :{=, $\leq$, $\geq$, (take =)}, then
- Alternative hypothesis H_1 : {$\neq$, <, >} chosen respectively.

Example:
(1) *The researcher wants to support the claim that μ_1 is not equal to 3.5 kg; therefore, the null and alternative hypotheses, in terms of these parameters, are*

$$H_0 : \mu_1 = 3.5 \text{ kg.}$$

(i.e., μ_1 = 3.5 kg; there is no difference between the mean weights of children born to urban and the claim)

$$H_1 : \mu_1 \neq 3.5 \text{ kg}$$

(i.e., $\mu_1 \neq 3.5$ kg; the mean weights of children born to urban women is not equal to (different from) 3.5 kg).

(2) *The researcher wants to support the claim that* μ_1 *is different from* μ_2*; therefore, the null and alternative hypotheses, in terms of these parameters, are*

$$H_0: \mu_1 = \mu_2$$

(i.e., there is no difference between the mean weights of children born to urban and rural women)

$$H_1: \mu_1 \neq \mu_2$$

(i.e., the mean weights of children born to urban women is different from that of the rural women)

8.2 TESTING A STATISTICAL HYPOTHESIS:

Important Concepts in Hypothesis Testing

A test statistic: is the sample statistic that is used in the hypothesis testing process. It is used for either rejecting or accepting the null hypothesis.

Example of test statistic for the mean, μ is either $Z = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ or $t = \dfrac{\overline{X} - \mu}{s/\sqrt{n}}$.

Critical region (w): is the subset of the sample space that corresponds to the rejection of null hypothesis based on level of significance α given.

Procedure for testing a statistical hypothesis:

1. Set up the null hypothesis and its alternative.
2. Choose a random sample $x_1, x_2, ..., x_n$ from a random variable X.
3. Check any assumptions of the test.
4. Find the value of the test statistic.
5. Refer the test statistic to a known distribution which it would follow if the null hypothesis were true.
6. If the value of the test statistic at the observed values $x_1, x_2, ..., x_n$ fall in critical region, then reject the null hypothesis, H_0 and accept (not reject) the alternative hypothesis H_1.
7. Conclude that the data are consistent or inconsistent with the null hypothesis.

Error Committed

The procedure of hypothesis testing can lead to two kinds of errors:

Type I error: consists of rejecting the null hypothesis H_0 when it is true.

P(Type I error) = P(Reject H_0 when it is true) = α

Type II error: involves accepting (not rejecting) H_0 when it is false.

P(Type II error) = β

The following table summarized the possible decisions in the significant tests:

	Accept H_0	***Reject*** H_0
H_0 ***is true***	Correct	Type I error
H_0 ***is false***	Type II error	Correct

Power of the test: It is the probability of rejecting H_0 when it is false.

Tests are classified into three different types:

(a) **Two-sided test:** In which we have $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$.
(b) **Right one-sided test**: In which we have $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$.
(c) **Left one-sided test**: In which we have $H_0 : \theta = \theta_0$ against $H_1 : \theta < \theta_0$.

P-value: is the lowest level of significance at which the null hypothesis could have been rejected.

The *P*-value of the two-sided test is twice the probability of the used statistics will be greater than or equal to its value calculated from the sample. Assuming that the null hypothesis is true, *P*-value is $2P(\overline{X} \geq \bar{x})$ or $2P(\overline{X} \leq \bar{x})$.

In the case of the one-sided test, then there is no need to multiply by 2.

- For the right-tailed test, the *P*-value is $P(\overline{X} \geq \bar{x})$.
- For the left-tailed test, the *P*-value is $P(\overline{X} \leq \bar{x})$.

If the P-value is less than the significance level α***, then we reject*** H_0**.**

The smaller the *P*-value, the more contradictory is the data to H_0.

8.3 Tests Concerning Means

Case 1: σ^2 *is known or* σ^2 *is unknown but* $n \geq 30$

Suppose we have a sample of size *n* taken from a population whose mean is μ and variance σ^2. We want to test whether this sample is taken from a population whose mean is μ_0. We know that the sample mean $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ if *n* is large.

For two-tailed hypothesis:
(i) $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$
(ii) $z = \dfrac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$
(iii) Critical region: Reject H_0 if $|z| \geq z_{\alpha/2}$

For right-tailed hypothesis:
(i) $H_0 : \mu = \mu_0$ vs $H_1 : \mu > \mu_0$

(ii) $z = \dfrac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$

(iii) Critical region: Reject H_0 if $z \geq z_\alpha$

For left-tailed hypothesis:

(i) $H_0 : \mu = \mu_0$ vs $H_1 : \mu < \mu_0$

(ii) $z = \dfrac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$

(ii) Critical region: Reject H_0 if $z \leq -z_\alpha$

Note : Replace** σ **with s if** σ **is unknown.

Example: In a heart study, the standard deviation σ=5.5, a random sample of *n* =100, has a mean age $\bar{x}$ =54.85, test the null hypothesis that the sample come from a population whose mean is 53 against alternative hypothesis the mean does not equal 53 years, at 0.05 level of significance.

Solution:

$H_0 : \mu = 53$ vs $H_1 : \mu \neq 53$

$\alpha = 0.05$

The test statistic

$$z = \frac{54.85 - 53}{5.5/\sqrt{100}} = \frac{1.85}{0.55} = 3.36$$

The critical region from the Z distribution, we find, for a two-tailed test where $\alpha/2 = 0.025$, the corresponding Z= ± 1.96.

Since the test statistic Z=3.36 falls within the critical region, thus, we reject the null hypothesis and not reject the alternative hypothesis that the sample comes from a population with a mean not equal to 53 yeas.

Case 2: σ^2 *is unknown and n < 30*

Suppose we have a sample of size *n* taken from a normal population whose mean is μ and variance unknown. We want to test whether this sample is taken from a population whose mean is μ_0.

For two-tailed hypothesis:

(i) $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$

(ii) $T = \dfrac{\bar{x} - \mu_0}{s / \sqrt{n}}$

(iii) Critical region: Reject H_0 if $|T| \geq t_{\alpha/2, n-1}$

- *Use appropriate hypothesis statement and rejection criteria for upper-tail or lower-tail test.*

Example: A smog alert is issued when the amount of a particular pollutant in the air is found to be greater than 7 ppm. Samples collected from 16 stations give an $\bar{x}$ of 7.84 with an s of 2.01. Do these findings indicate that the smog alert criterion has been exceeded, or can the results be explained by chance?

Solution:

Since σ is unknown and the sample size is less than 30, the t test is used.

$H_0 : \mu = 7$ vs $H_1 : \mu > 7$

$\alpha = 0.05$

The test statistic

$$t = \frac{7.84 - 7}{2.01/\sqrt{16}} = \frac{0.84}{0.5} = 1.68$$

The critical region: since the $H_1 : \mu > 7$ indicates a one-tails test, we place all $\alpha = 0.05$ on the positive side. From t- distribution table we find that for 15 df $t_{0.05} = 1.753$.

Since the calculated t = 1.68 dose not fall in the critical region , we do not reject H_0, we conclude that the data were insufficient to indicate that the critical air pollution level of 7 ppm was exceeded.

Example: Suppose that it is known from experience that the standard deviation of the 8-cm diameter CDs made by a certain company is 0.16 cm. To check whether its production is under control on a given day, namely, to check whether the true average diameter of the CD is 8 cm, the employee selected a random sample of 25 pieces of CDs and finds that their mean diameter is $\bar{x} = 8.091$cm. Since the company stands to lose money when $\mu > 8$ and the customer loses out when $\mu < 8$, test the null hypothesis $\mu = 8$ against the alternative hypothesis $\mu \neq 8$ at $\alpha = 0.01$.

Solution:

Example: Suppose that 100 tires made by a certain manufacturer lasted on the average 21,819 miles with a standard deviation of 1,295 miles. Test the null hypothesis $\mu = 22{,}000$ miles against the alternative hypothesis $\mu < 22{,}000$ miles at the 0.05 level of significance.

Solution:

8.4 Hypothesis Tests About the Difference Between Two Populations Means (Comparing two means)

Large-sample test of hypothesis about $(\mu_1 - \mu_2)$

Assumptions:

The validity of the two-sample t-test depends on various assumptions being satisfied:

1. Each subject must be randomly selected from the population.
2. The random samples are selected in an independent manner from the two populations.
3. The populations from which the samples are selected both have approximately normal distributions.
4. The two samples come from populations with equal (or approximately equal) variances. (see the next slide)

According to the test direction, there are two types of t- test

One tailed test	**Two- tailed test**
$H_0 : \mu_1 - \mu_2 = D$	$H_0 : \mu_1 - \mu_2 = D$
$H_1 : \mu_1 - \mu_2 > D$ or $H_1 : \mu_1 - \mu_2 < D$	$H_1 : \mu_1 - \mu_2 \neq D$

Test Statistic:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - D}{\sigma_{(\bar{x}_1 - \bar{x}_2)}} \approx \frac{(\bar{x}_1 - \bar{x}_2) - D}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Rejection region

$z > z_\alpha$ or $z < -z_\alpha$

$H_1 : \mu_1 - \mu_2 > D$ or $H_1 : \mu_1 - \mu_2 < D$

Rejection region

$z > z_{\alpha/2}$ or $z < -z_{\alpha/2}$

$H_1 : \mu_1 - \mu_2 \neq D$

Example: In a study on pregnant women in their third trimester who delivered during Ramadan or the first two weeks of Shawwal, the birthweight of the baby (in kg) was measured for *independent random samples* of babies of *fasting* and *non-fasting women*. The results of the investigation are summarized in the Table below. Does this evidence indicate that the mean of the baby of a non-fasting mother is significantly **higher** than the mean of the baby of a fasting mother? Use a significance level of $\alpha = .01$.

Solution:

Non-fasting	**Fasting**
$n_1 = 75$	$n_2 = 64$
$\bar{x}_1 = 3$	$\bar{x}_2 = 2.95$
$s_1 = 0.11$	$s_2 = 0.09$

The researcher wants to test the hypotheses

$H_0 : \mu_1 - \mu_2 = 0$ (i.e., no difference between means)

$H_1 : \mu_1 - \mu_2 > D$ (i.e., the mean of the baby of a non-fasting mother is higher than the mean of the baby of a fasting mother)

where

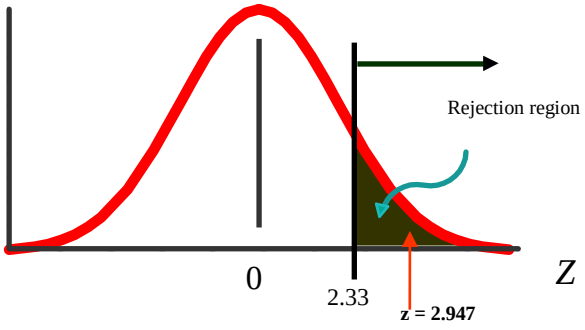
μ_1 = The mean of the baby of a non-fasting mother

μ_2 = The mean of the baby of a fasting mother

This one-tailed, large-sample test is based on a *z* statistic. Thus, we will reject H_0 if

$z > z_\alpha = z_{0.01} = 2.33$, the rejection region is given by $z > 2.33$

We compute the test statistic as follows:



Example: A company claims that its light bulbs are superior to those of its main competitor. If a study showed that a sample of 40 of its bulbs has a mean lifetime of 647 hours of continuous use with a standard deviation of 27 hours, while a sample of 40 bulbs made by main competitor had a mean lifetime of 638 hours of continuous use with a standard deviation of 31 hours. Does this substantiate the claim at the 0.05 level of significance?

Solution:

EXERCISES

1- What is the critical value for a test of significance in each of the following situations?
   a. One-tailed test, $\alpha = 0.05$, σ known, *n*=20.
   b. One-tailed test, $\alpha = 0.05$, σ unknown, *n*=10.
   c. Two-tailed test, $\alpha = 0.01$, σ unknown, *n*=14.
   d. Two-tailed test, $\alpha = 0.01$, σ known, *n*=25.
   e. Two-tailed test, $\alpha = 0.05$, σ unknown, *n*=35.

2- in which of the situations in Exercise (1) would you use (a) a Z test? (b) a t-test? Why?

3- For each of the following, state the null hypothesis (H_0) and alternative (H_1):
   a. Has the average community level of suspended particulates for the month of August exceeded 30 units per cubic meter?
   b. Does mean age of onest of a certain acute disease for schoolchildren differ from 11.5?
   c. A psychologist claims that the average IQ of a sample of 60 children is significantly above the normal IQ of 100.
   d. Is the average cross-sectional area of the lumen of coronary arteries for men, age 40 to 59, less than 31.5% of the total arterial cross section?
   e. Is the mean hemoglobin level of a group of high-attitude workers different from 16 g/cc?
   f. Does the average speed of 50 cars as checked by radar on a particular high way differ from 55 mph?

4- Determine the critical value that would be used to test a hypothesis under the conditions given in each of the following:
   a. $H_0 : \mu_0 = 220, \quad H_1 : \mu_0 \neq 220, \quad \alpha = 0.05, \quad n = 20, \quad \sigma$ known.
   b. $H_0 : \mu_0 = 15, \quad H_1 : \mu_0 > 15, \quad \alpha = 0.01, \quad n = 35, \quad \sigma$ unknown.
   c. $H_0 : \mu_0 = 70, \quad H_1 : \mu_0 \neq 70, \quad \alpha = 0.01, \quad n = 18, \quad \sigma$ known.
   d. $H_0 : \mu_0 = 120, \quad H_1 : \mu_0 \neq 120, \quad \alpha = 0.05, \quad n = 25, \quad \sigma$ unknown.
   e. $H_0 : \mu_0 = 100, \quad H_1 : \mu_0 < 1000, \quad \alpha = 0.01, \quad n = 16, \quad \sigma$ unknown.
   f. $H_0 : \mu_0 = 55, \quad H_1 : \mu_0 < 55, \quad \alpha = 0.05, \quad n = 49, \quad \sigma$ unknown.

5- Boys of a certain age have a mean weight of 85 Ib. A complaint was made that in municipal children's home the boys are underfed. As one bit of evidence, all 25 boys of the given age were weighted and found to have a mean weight of 80.94 Ib.

   a. If it is known in advance that the population standard deviation for weights of boys this age is 11.6 Ib, what you conclude regarding the complaint? Use $\alpha = 0.05$.

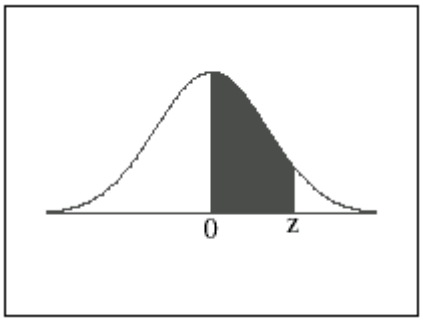
   b. Suppose that the population standard deviation is unknown. If the sample standard deviation is found to be 12.3 Ib, what conclusion regarding the complaint might you draw? Use $\alpha = 0.05$.

6- A company claims that its eyes lenses are superior to those of its main competitor. If a study showed that a sample of 40 of its eyes lenses has a mean lifetime of 647 days of continuous use with a standard deviation of 27 day, while a sample of 40 eyes lenses made by main competitor had a mean lifetime of 638 days of continuous use with a standard deviation of 31 days. Does this substantiate the claim at the 0.05 level of significance?

APPENDIX A: Random Number Table

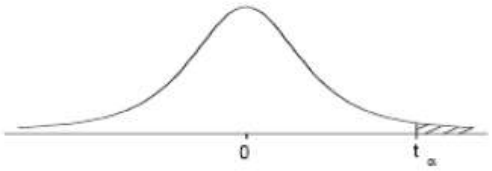
13962	70992	65172	28053	02190	83634	66012	70305	66761	88344
43905	46941	72300	11641	43548	30455	07686	31840	03261	89139
00504	48658	38051	59408	16508	82979	92002	63606	41078	86326
61274	57238	47267	35303	29066	02140	60867	39847	50968	96719
43753	21159	16239	50595	62509	61207	86816	29902	23395	72640
83503	51662	21636	68192	84294	38754	84755	34053	94582	29215
36807	71420	35804	44862	23577	79551	42003	58684	09271	68396
19110	55680	18792	41487	16614	83053	00812	16749	45347	88199
82615	86984	93290	87971	60022	35415	20852	02909	99476	45568
05621	26584	36493	63013	68181	57702	49510	75304	38724	15712
06936	37293	55875	71213	83025	46063	74665	12178	10741	58362
84981	60458	16194	92403	80951	80068	47076	23310	74899	87929
66354	88441	96191	04794	14714	64749	43097	83976	83281	72038
49602	94109	36460	62353	00721	66980	82554	90270	12312	56299
78430	72391	96973	70437	97803	78683	04670	70667	58912	21883
33331	51803	15934	75807	46561	80188	78984	29317	27971	16440
62843	84445	56652	91797	45284	25842	96246	73504	21631	81223
19528	15445	77764	33446	41204	70067	33354	70680	66664	75486
16737	01887	50934	43306	75190	86997	56561	79018	34273	25196
99389	06685	45945	62000	76228	60645	87750	46329	46544	95665
36160	38196	77705	28891	12106	56281	86222	66116	39626	06080
05505	45420	44016	79662	92069	27628	50002	32540	19848	27319
85962	19758	92795	00458	71289	05884	37963	23322	73243	98185
28763	04900	54460	22083	89279	43492	00066	40857	86568	49336
42222	40446	82240	79159	44168	38213	46839	26598	29983	67645
43626	40039	51492	36488	70280	24218	14596	04744	89336	35630
97761	43444	95895	24102	07006	71923	04800	32062	41425	66862
49275	44270	52512	03951	21651	53867	73531	70073	45542	22831
15797	75134	39856	73527	78417	36208	59510	76913	22499	68467
04497	24853	43879	07613	26400	17180	18880	66083	02196	10638
95468	87411	30647	88711	01765	57688	60665	57636	36070	37285
01420	74218	71047	14401	74537	14820	45248	78007	65911	38583
74633	40171	97092	79137	30698	97915	36305	42613	87251	75608
46662	99688	59576	04887	02310	35508	69481	30300	94047	57096
10853	10393	03013	90372	89639	65800	88532	71789	59964	50681
68583	01032	67938	29733	71176	35699	10551	15091	52947	20134
75818	78982	24258	93051	02081	83890	66944	99856	87950	13952
16395	16837	00538	57133	89398	78205	72122	99655	25294	20941
53892	15105	40963	69267	85534	00533	27130	90420	72584	84576
66009	26869	91829	65078	89616	49016	14200	97469	88307	92282
45292	93427	92326	70206	15847	14302	60043	30530	57149	08642
34033	45008	41621	79437	98745	84455	66769	94729	17975	50963
13364	09937	00535	88122	47278	90758	23542	35273	67912	97670
03343	62593	93332	09921	25306	57483	98115	33460	55304	43572
46145	24476	62507	19530	41257	97919	02290	40357	38408	50031
37703	51658	17420	30593	39637	64220	45486	03698	80220	12139
12622	98083	17689	59677	56603	93316	79858	52548	67367	72416
56043	00251	70085	28067	78135	53000	18138	40564	77086	49557
43401	35924	28308	55140	07515	53854	23023	70268	80435	24269
18053	53460	32125	81357	26935	67234	78460	47833	20496	35645

APPENDIX B: Areas Under the Standard Normal Curve



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4996	.4997
3.4	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4998
3.5	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998

APPENDIX C: Upper Critical Values of Student's *t* Distribution



	Tail Area α				
df	.10	.05	.025	.01	.005
1	3.0777	6.3138	12.706	31.821	63.657
2	1.8856	2.9200	4.3027	6.9646	9.9248
3	1.6377	2.3534	3.1824	4.5407	5.8409
4	1.5332	2.1318	2.7764	3.7469	4.6041
5	1.4759	2.0150	2.5706	3.3649	4.0321
6	1.4398	1.9432	2.4469	3.1427	3.7074
7	1.4149	1.8946	2.3646	2.9980	3.4995
8	1.3968	1.8595	2.3060	2.8965	3.3554
9	1.3830	1.8331	2.2622	2.8214	3.2498
10	1.3722	1.8125	2.2281	2.7638	3.1693
11	1.3634	1.7959	2.2010	2.7181	3.1058
12	1.3562	1.7823	2.1788	2.6810	3.0545
13	1.3502	1.7709	2.1604	2.6503	3.0123
14	1.3450	1.7613	2.1448	2.6245	2.9768
15	1.3406	1.7531	2.1314	2.6025	2.9467
16	1.3368	1.7459	2.1199	2.5835	2.9208
17	1.3334	1.7396	2.1098	2.5669	2.8982
18	1.3304	1.7341	2.1009	2.5524	2.8784
19	1.3277	1.7291	2.0930	2.5395	2.8609
20	1.3253	1.7247	2.0860	2.5280	2.8453
21	1.3232	1.7207	2.0796	2.5176	2.8314
22	1.3212	1.7171	2.0739	2.5083	2.8188
23	1.3195	1.7139	2.0687	2.4999	2.8073
24	1.3178	1.7109	2.0639	2.4922	2.7969
25	1.3163	1.7081	2.0595	2.4851	2.7874
26	1.3150	1.7056	2.0555	2.4786	2.7787
27	1.3137	1.7033	2.0518	2.4727	2.7707
28	1.3125	1.7011	2.0484	2.4671	2.7633
29	1.3114	1.6991	2.0452	2.4620	2.7564
30	1.3104	1.6973	2.0423	2.4573	2.7500
31	1.3095	1.6955	2.0395	2.4528	2.7440
32	1.3086	1.6939	2.0369	2.4487	2.7385
33	1.3077	1.6924	2.0345	2.4448	2.7333
34	1.3070	1.6909	2.0322	2.4411	2.7284
35	1.3062	1.6896	2.0301	2.4377	2.7238
36	1.3055	1.6883	2.0281	2.4345	2.7195
37	1.3049	1.6871	2.0262	2.4314	2.7154
38	1.3042	1.6860	2.0244	2.4286	2.7116
39	1.3036	1.6849	2.0227	2.4258	2.7079
40	1.3031	1.6839	2.0211	2.4233	2.7045
41	1.3025	1.6829	2.0195	2.4208	2.7012
42	1.3020	1.6820	2.0181	2.4185	2.6981
43	1.3016	1.6811	2.0167	2.4163	2.6951
44	1.3011	1.6802	2.0154	2.4141	2.6923
45	1.3006	1.6794	2.0141	2.4121	2.6896
46	1.3002	1.6787	2.0129	2.4102	2.6870
47	1.2998	1.6779	2.0117	2.4083	2.6846
48	1.2994	1.6772	2.0106	2.4066	2.6822
49	1.2991	1.6766	2.0096	2.4049	2.6800
50	1.2987	1.6759	2.0086	2.4033	2.6778

	Tail Area α				
df	.10	.05	.025	.01	.005
51	1.2984	1.6753	2.0076	2.4017	2.6757
52	1.2980	1.6747	2.0066	2.4002	2.6737
53	1.2977	1.6741	2.0057	2.3988	2.6718
54	1.2974	1.6736	2.0049	2.3974	2.6700
55	1.2971	1.6730	2.0040	2.3961	2.6682
56	1.2969	1.6725	2.0032	2.3948	2.6665
57	1.2966	1.6720	2.0025	2.3936	2.6649
58	1.2963	1.6716	2.0017	2.3924	2.6633
59	1.2961	1.6711	2.0010	2.3912	2.6618
60	1.2958	1.6706	2.0003	2.3901	2.6603
61	1.2956	1.6702	1.9996	2.3890	2.6589
62	1.2954	1.6698	1.9990	2.3880	2.6575
63	1.2951	1.6694	1.9983	2.3870	2.6561
64	1.2949	1.6690	1.9977	2.3860	2.6549
65	1.2947	1.6686	1.9971	2.3851	2.6536
66	1.2945	1.6683	1.9966	2.3842	2.6524
67	1.2943	1.6679	1.9960	2.3833	2.6512
68	1.2941	1.6676	1.9955	2.3824	2.6501
69	1.2939	1.6672	1.9949	2.3816	2.6490
70	1.2938	1.6669	1.9944	2.3808	2.6479
71	1.2936	1.6666	1.9939	2.3800	2.6469
72	1.2934	1.6663	1.9935	2.3793	2.6459
73	1.2933	1.6660	1.9930	2.3785	2.6449
74	1.2931	1.6657	1.9925	2.3778	2.6439
75	1.2929	1.6654	1.9921	2.3771	2.6430
76	1.2928	1.6652	1.9917	2.3764	2.6421
77	1.2926	1.6649	1.9913	2.3758	2.6412
78	1.2925	1.6646	1.9908	2.3751	2.6403
79	1.2924	1.6644	1.9905	2.3745	2.6395
80	1.2922	1.6641	1.9901	2.3739	2.6387
81	1.2921	1.6639	1.9897	2.3733	2.6379
82	1.2920	1.6636	1.9893	2.3727	2.6371
83	1.2918	1.6634	1.9890	2.3721	2.6364
84	1.2917	1.6632	1.9886	2.3716	2.6356
85	1.2916	1.6630	1.9883	2.3710	2.6349
86	1.2915	1.6628	1.9879	2.3705	2.6342
87	1.2914	1.6626	1.9876	2.3700	2.6335
88	1.2912	1.6624	1.9873	2.3695	2.6329
89	1.2911	1.6622	1.9870	2.3690	2.6322
90	1.2910	1.6620	1.9867	2.3685	2.6316
91	1.2909	1.6618	1.9864	2.3680	2.6309
92	1.2908	1.6616	1.9861	2.3676	2.6303
93	1.2907	1.6614	1.9858	2.3671	2.6297
94	1.2906	1.6612	1.9855	2.3667	2.6291
95	1.2905	1.6611	1.9853	2.3662	2.6286
96	1.2904	1.6609	1.9850	2.3658	2.6280
97	1.2903	1.6607	1.9847	2.3654	2.6275
98	1.2902	1.6606	1.9845	2.3650	2.6269
99	1.2902	1.6604	1.9842	2.3646	2.6264
100	1.2901	1.6602	1.9840	2.3642	2.6259
∞	1.2816	1.6449	1.9600	2.3263	2.5758