King Saud University
College of Science
Department of Statistics & OR

STAT – 145 BIOSTATISTICS

**Summer Semester
1431/1432**

Lectures' Notes

Prof. Abdullah Al-Shiha

Outline of the course

Course Code & No:	STAT 145
Course Name	Biostatistics
Credits:	2 (2+0)
Pre-requisite:	None

Instructor:

Professor Abdullah Al-Shiha

Office: 2B43 – Building 4

Office Hours: Saturday & Monday 5:00-6:00pm (might be changed later)

Textbook:

Biostatistics: Basic Concepts and Methodology for the Heath Sciences by Wayne W. Daniel., 9th edition.

The textbook is available at the university bookstore (ground floor – the floor under SAMBA bank). The textbook costs 70 Riyals for students.

Topics	No of Weeks	Contact hours	Pages
Introduction to biostatistics, types of data, frequency tables, and graphical representation.	1	2	2-13 19-28 + Notes
Descriptive statistics: Measures of central tendency: mean , median, mode. Measures of dispersion: range, variance, standard deviation, coefficient of variation.	3	6	38-41 43-46 + Notes
Basic probability, conditional probability, concept of independence, sensitivity, specificity, Bayes' Theorem.	3	6	65-83 + Notes
Discrete probability distributions, cumulative probability distribution, Binomial and Poisson distributions and their mean and variance.	2	4	93-112 + Notes
Continuous probability distributions, Normal distribution, Standard normal distribution, t-distribution.	2	4	114-128 135-141 143-157 172-173 + Notes
Sampling distributions: mean and proportion for one and two samples			
Statistical inference: Point and interval estimation, Type of errors, Concept of P-value, testing hypothesis about one and two samples means and proportions including paired data – different cases under normality.	3	6	162-175 185-187 215-234 250-255 + Notes

Leftover subjects and Review	1	2	

Marking scheme:

40 marks = 2 mid-term tests carrying 20 marks each

10 marks = Home woks and Quizzes

50 marks = Final exam

Notes:

1. This course is a coordinated course taught by several professors. The exams are common for all sections. Dates of the exams will be announced in two weeks in-sha-Allah. Exam dates will not be changed once they have been fixed.
2. Attendance is important. Latecomers will be marked absent. Students missing more than 25% of the lectures will be deprived of sitting the final exam.

References:

3. Bernard Rosner. Fundamentals of Biostatistics.
4. Pagano, Gauvreau. Principles of Biostatistics, 2nd edition.

CHAPTER 1: Getting Acquainted with Biostatistics

1.1 Introduction:

The course "Biostatistics" (STAT-145) is about information; how it is obtained, how it is analyzed, and how it is interpreted.

The objective of the course is to learn:

(1)   How to organize, summarize, and describe data.
        (Descriptive Statistics)

(2)   How to reach decisions about a large body of data by examine only a small part of the data.
        (Inferential Statistics)

1.2 Some Basic Concepts:

Data:

Data is the raw material of statistics. There are two types of data:

(1)   Quantitative data
        (numbers: weights, ages, …).

(2)   Qualitative data
        (words or attributes: nationalities, occupations, …).

Statistics:

Statistics is the field of study concerned with:

(1)   The collection, organization, summarization, and analysis of data. (Descriptive Statistics)

(2)   The drawing of inferences and conclusions about a body of data (population) when only a part of the data (sample) is observed. (Inferential Statistics)

Biostatistics:

When the data is obtained from the biological sciences and medicine, we use the term "biostatistics".

Sources of Data:
1. Routinely kept records.
2. Surveys.
3. Experiments.
4. External sources. (published reports, data bank, …)

Population:
- A population is the largest collection of entities (elements or individuals) in which we are interested at a particular time and about which we want to draw some conclusions.
- When we take a measurement of some variable on each of the entities in a population, we generate a population of values of that variable.
- Example: If we are interested in the weights of students enrolled in the college of engineering at KSU, then our population consists of the weights of all of these students, and our variable of interest is the weight.

Population Size (N):
The number of elements in the population is called the population size and is denoted by N.

Sample:
- A sample is a part of a population.
- From the population, we select various elements on which we collect our data. This part of the population on which we collect data is called the sample.
- Example: Suppose that we are interested in studying the characteristics of the weights of the students enrolled in the college of engineering at KSU. If we randomly select 50 students among the students of the college of engineering at KSU and measure their weights, then the weights of these 50 students form our sample.

Sample Size (n):
The number of elements in the sample is called the sample

size and is denoted by *n*.

Variables:
The characteristic to be measured on the elements is called variable. The value of the variable varies from element to element.

Example of Variables:

(1) No. of patients   (2) Height

(3) Sex   (4) Educational Level

Types of Variables:
(1) Quantitative Variables:
A quantitative variable is a characteristic that can be measured. The values of a quantitative variable are numbers indicating how much or how many of something.

Examples:

(i) Family Size   (ii) No. of patients

(iii) Weight   (iv) height

Types of Quantitative Variables:

(a) Discrete Variables:

There are jumps or gaps between the values.

Examples: - Family size (x = 1, 2, 3, …)

- Number of patients (x = 0, 1, 2, 3, …)

(b) Continuous Variables:

There are no gaps between the values.

A continuous variable can have any value within a certain interval of values.

Examples: - Height ($140 < x < 190$)

- Blood sugar level ($10 < x < 15$)

(2) Qualitative Variables:
The values of a qualitative variable are words or attributes indicating to which category an element belong.

Examples:

- Blood type   - Nationality
- Students Grades   - Educational level

Types of Qualitative Variables:

(a) Nominal Qualitative Variables:

A nominal variable classifies the observations into various mutually exclusive and collectively non-ranked categories. The values of a nominal variable are names or attributes that can not be ordered or sorted or ranked.

Examples: - Blood type (O, AB, A, B)
            - Nationality (Saudi, Egyptian, British, …)
            - Sex (male, female)

(b) Ordinal Qualitative Variables:

An ordinal variable classifies the observations into various mutually exclusive and collectively ranked categories. The values of an ordinal variable are categories that can be ordered, sorted, or ranked by some criterion.

Examples: - Educational level (elementary, intermediate, …)
            - Students grade (A, B, C, D, F)
            - Military rank
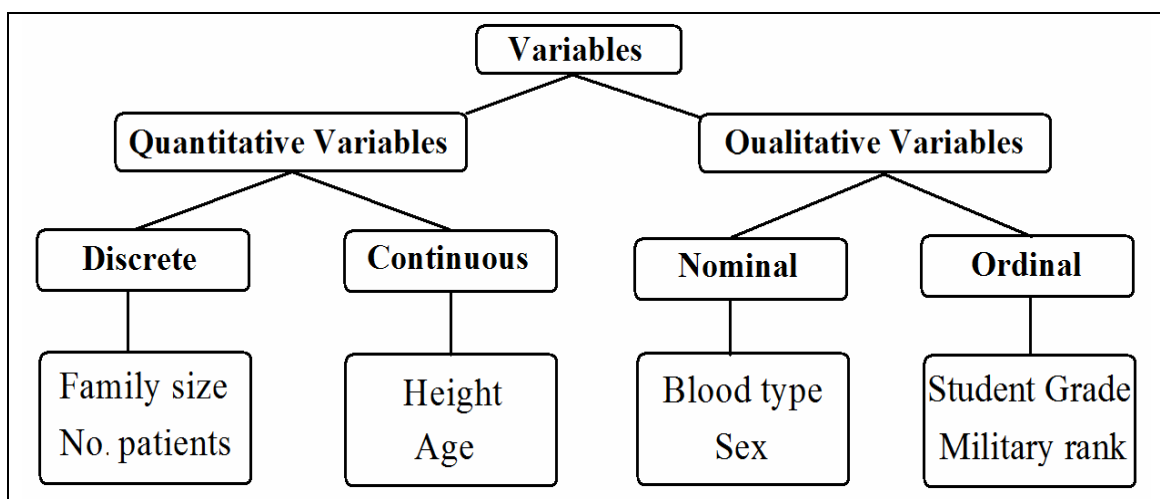
1.4 Sampling and Statistical Inference:

There are several types of sampling techniques, some of which are:

(1) Simple Random Sampling:

If a sample of size (n) is selected from a population of size (N) in such a way that each element in the population has the same chance to be selected, the sample is called a simple random sample.

(2) Stratified Random Sampling:

In this type of sampling, the elements of the population are classified into several homogenous groups (strata). From each group, an independent simple random sample is drawn. The sample resulting from combining these samples is called a stratified random Sample.

CHAPTER 2: Strategies for Understanding the Meaning of Data:

2.1 Introduction:

In this chapter, we learn several techniques for organizing and summarizing data so that we may more easily determine what information they contain. Summarization techniques involve:
- frequency distributions
- descriptive measures

2.2 The Ordered Array:

A first step in organizing data is the preparation of an ordered array.

An ordered array is a listing of the values in order of magnitude from the smallest to the largest value.

Example:

The following values represent a list of ages of subjects who participate in a study on smoking cessation:

55 46 58 54 52 69 40 65 53 58

The ordered array is:

40 46 52 53 54 55 58 58 65 69

2.3 Grouped Data: The Frequency Distribution:

To group a set of observations, we select a suitable set of contiguous, non-overlapping intervals such that each value in the set of observations can be placed in one, and only one, of the intervals. These intervals are called "class intervals".

Example:

The following table gives the <u>hemoglobin level</u> (g/dl) of a sample of 50 men.

17.0	17.7	15.9	15.2	16.2	17.1	15.7	17.3	**13.5**	16.3
14.6	15.8	15.3	16.4	13.7	16.2	16.4	16.1	17.0	15.9
14.0	16.2	16.4	14.9	17.8	16.1	15.5	**18.3**	15.8	16.7
15.9	15.3	13.9	16.8	15.9	16.3	17.4	15.0	17.5	16.1
14.2	16.1	15.7	15.1	17.4	16.5	14.4	16.3	17.3	15.8

We wish to summarize these data using the following class

intervals:

13.0 – 13.9 ,   14.0 – 14.9 ,   15.0 – 15.9 ,
16.0 – 16.9 ,   17.0 – 17.9 ,   18.0 – 18.9

Solution:

Variable = X = hemoglobin level (continuous, quantitative)

Sample size = n = 50

Max= 18.3

Min= 13.5

Class Interval	Tally	Frequency
13.0 – 13.9	|||	3
14.0 – 14.9	卌	5
15.0 – 15.9	卌 卌 卌	15
16.0 – 16.9	卌 卌 卌	–	16
17.0 – 17.9	卌 卌	10
18.0 – 18.9			1

The grouped frequency distribution for the hemoglobin level of the 50 men is:

Class Interval (Hemoglobin level)	Frequency (no. of men)
13.0 – 13.9	3
14.0 – 14.9	5
15.0 – 15.9	15
16.0 – 16.9	16
17.0 – 17.9	10
18.0 – 18.9	1
Total	n=50

Notes:

1. Minimum value $\in$ first interval.
2. Maximum value $\in$ last interval.
3. The intervals are not overlapped.
4. Each value belongs to one, and only one, interval.
5. Total of the frequencies = the sample size = n

Mid-Points of Class Intervals:

- Mid-point $= \dfrac{\text{upper limit} + \text{lower limit}}{2}$

True Class Intervals:

- d = gap between class intervals
- d = lower limit − upper limit of the preceding class interval
- true upper limit = upper limit +d/2
- true lower limit = lower limit − d/2

Class Interval	True Class Interval	Mid-point	Frequency
13.0 – **13.9**	12.95 - 13.95	13.45	3
14.0 – 14.9	13.95 - 14.95	14.45	5
15.0 – 15.9	14.95 - 15.95	15.45	15
16.0 – 16.9	15.95 - 16.95	16.45	16
17.0 – 17.9	16.95 - 17.95	17.45	10
18.0 – 18.9	17.95 – 18.95	18.45	1

For example:

Mid-point of the 1st interval = (13.0+13.9)/2 = 13.45

:

Mid-point of the last interval = (18.0+18.9)/2 = 18.45

Note:

(1) Mid-point of a class interval is considered as a typical (approximated) value for all values in that class interval.

For example: approximately we may say that:

there are 3 observations with the value of 13.45

there are 5 observations with the value of 14.45

:

there are 1 observation with the value of 18.45

(2) There are no gaps between true class intervals. The end-point (true upper limit) of each true class interval equals to the start-point (true lower limit) of the following true class interval.

Cumulative frequency:

Cumulative frequency of the 1st class interval = frequency.
Cumulative frequency of a class interval
= frequency + cumulative frequency of the preceding class interval

Relative frequency and Percentage frequency:

Relative frequency = frequency/n
Percentage frequency = Relative frequency × 100%

Class Interval	Frequency	Cumulative Frequency	Relative Frequency	Cumulative Relative Frequency	Percentage Frequency	Cumulative Percentage Frequency
13.0 – 13.9	3	3	0.06	0.06	6%	6%
14.0 – 14.9	5	8	0.10	0.16	10%	16%
15.0 – 15.9	15	23	0.30	0.46	30%	46%
16.0 – 16.9	16	39	0.32	0.78	32%	78%
17.0 – 17.9	10	49	0.20	0.98	20%	98%
18.0 – 18.9	1	50	0.02	1.00	2%	100%

From frequencies:
The number of people whose hemoglobin levels are between 17.0 and 17.9 = 10

From cumulative frequencies:
The number of people whose hemoglobin levels are less than or equal to 15.9 = 23
The number of people whose hemoglobin levels are less than or equal to 17.9 = 49

From percentage frequencies:
The percentage of people whose hemoglobin levels are between 17.0 and 17.9 = 20%

From cumulative percentage frequencies:
The percentage of people whose hemoglobin levels are less than or equal to 14.9 = 16%
The percentage of people whose hemoglobin levels are less than or equal to 16.9 = 78%

Displaying Grouped Frequency Distributions:

For representing frequency (or relative frequency or percentage frequency) distributions, we may use one of the following graphs:

- The Histogram
- The Frequency Polygon

Example:

Consider the following frequency distribution of the ages of 100 women.

True Class Interval (age)	Frequency (No. of women)	Cumulative Frequency	Mid-points
14.5 - 19.5	8	8	17
19.5 - 24.5	16	24	22
24.5 - 29.5	32	56	27
29.5 - 34.5	28	84	32
34.5 - 39.5	12	96	37
39.5 - 44.5	4	100	42
Total	$n=100$		

Width of the interval:

W = true upper limit − true lower limit = 19.5 − 14.5 = 5

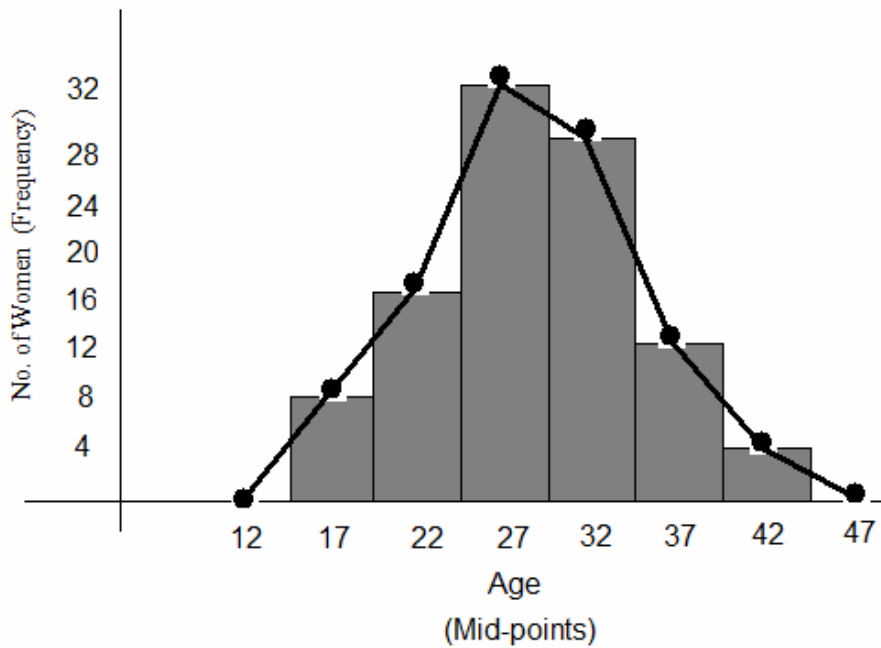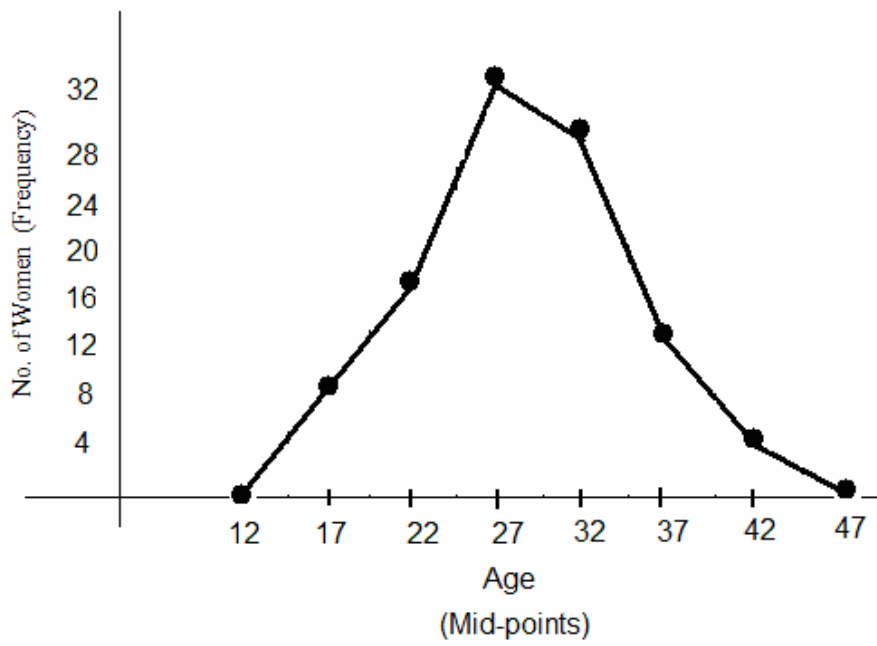(1) Histogram:

Organizing and Displaying Data using Histogram:
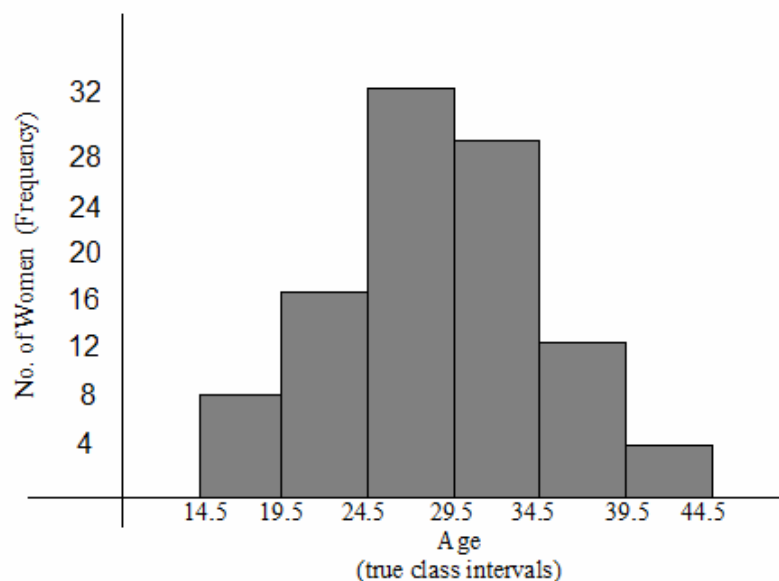
(2) The Frequency Polygon:
Organizing and Displaying Data using Polygon:

Polygon (Open)
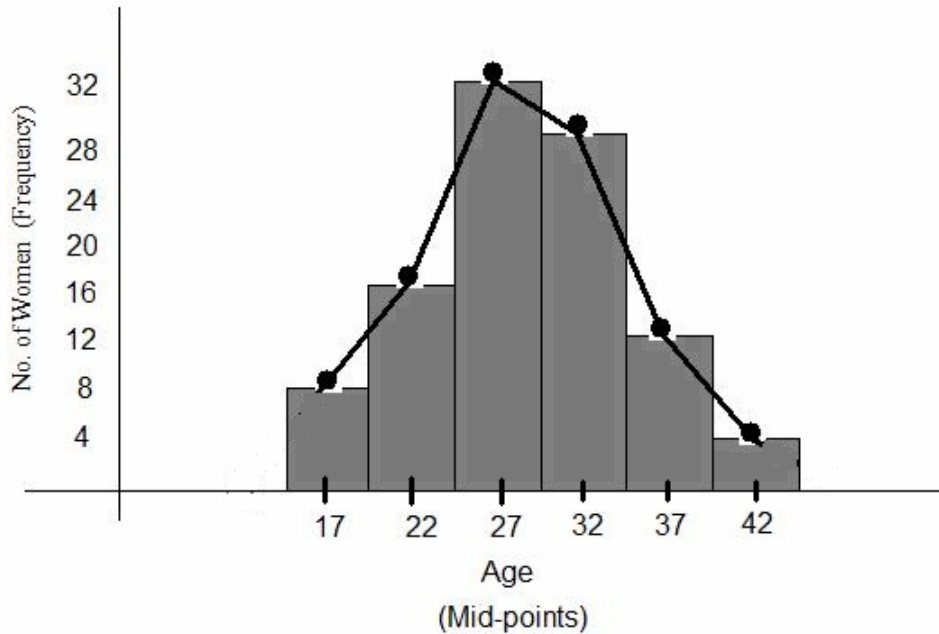
Polygon (Closed)

2.4 Descriptive Statistics: Measures of Central Tendency:

(Measures of location)

In the last section we summarize the data using frequency distributions (tables and figures). In this section, we will introduce the concept of summarization of the data by means of a single number called "a descriptive measure".

A descriptive measure computed from the values of a sample is called a "statistic".

A descriptive measure computed from the values of a population is called a "parameter".

For the variable of interest there are:

(1) "N" population values.

(2) "n" sample of values.

- Let $X_1, X_2, \ldots, X_N$ be the population values (in general, they are unknown) of the variable of interest.
  The population size $= N$

- Let $x_1, x_2, \ldots, x_n$ be the sample values (these values are known).
  The sample size $= n$.

(i) A **parameter** is a measure (or number) obtained from the population values: $X_1, X_2, \ldots, X_N$.
  - Values of the parameters are unknown in general.
  - We are interested to know true values of the parameters.

(ii) A **statistic** is a measure (or number) obtained from the sample values: $x_1, x_2, \ldots, x_n$.
  - Values of statistics are known in general.
  - Since parameters are unknown, statistics are used to approximate (estimate) parameters.

Measures of Central Tendency:
(or measures of location):

The most commonly used measures of central tendency are: the mean – the median – the mode.

- The values of a variable often tend to be concentrated around the center of the data.
- The center of the data can be determined by the measures of central tendency.
- A measure of central tendency is considered to be a typical (or a representative) value of the set of data as a whole.

Mean:
(1) The Population mean (μ):

If $X_1, X_2, \ldots, X_N$ are the population values, then the population mean is:

$$\mu = \frac{X_1 + X_2 + \cdots + X_N}{N} = \frac{\sum\limits_{i=1}^{N} X_i}{N} \qquad \text{(unit)}$$

- The population mean μ is a parameter (it is usually unknown, and we are interested to know its value)

(2) The Sample mean ($\bar{x}$):

If $x_1, x_2, \ldots, x_n$ are the sample values, then the sample mean is:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum\limits_{i=1}^{n} x_i}{n} \qquad \text{(unit)}$$

- The sample mean $\bar{x}$ is a statistic (it is known – we can calculate it from the sample).
- The sample mean $\bar{x}$ is used to approximate (estimate) the population mean μ.

Example:
Suppose that we have a population of 5 population values:

$$X_1 = 41, \ X_2 = 30, \ X_3 = 35, \ X_4 = 22, \ X_5 = 27. \ (N=5)$$

Suppose that we randomly select a sample of size 3, and the sample values we obtained are:

$$x_1 = 30, \ x_2 = 35, \ x_3 = 27. \ (n=3)$$

Then:

The population mean is:

$$\mu = \frac{41 + 30 + 35 + 22 + 27}{5} = \frac{155}{5} = 31 \qquad \text{(unit)}$$

The sample mean is:

$$\bar{x} = \frac{30 + 35 + 27}{3} = \frac{92}{3} = 30.67 \qquad \text{(unit)}$$

Notice that $\bar{x} = 30.67$ is approximately equals to $\mu = 31$.

Note: The unit of the mean is the same as the unit of the data.

Advantages and disadvantages of the mean:

Advantages:

- Simplicity: The mean is easily understood and easy to compute.
- Uniqueness: There is one and only one mean for a given set of data.
- The mean takes into account all values of the data.

Disadvantages:

- Extreme values have an influence on the mean. Therefore, the mean may be distorted by extreme values.
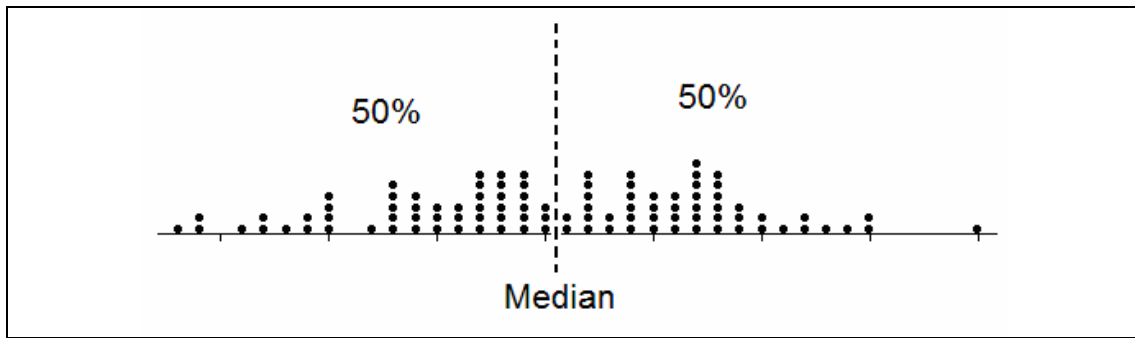
  For example:

Sample	Data						mean
A	2	4	5	7	7	10	5.83
B	2	4	5	7	7	100	20.83

- The mean can only be found for quantitative variables.

Median:

The median of a finite set of numbers is that value which divides the **ordered array** into two equal parts. The numbers in the first part are less than or equal to the median and the numbers in the second part are greater than or equal to the

median.

Notice that:

    50% (or less) of the data is $\leq$ Median

    50% (or less) of the data is $\geq$ Median

Calculating the Median:

    Let $x_1, x_2, \ldots, x_n$ be the sample values. The sample size (n) can be odd or even.

- First we order the sample to obtain the ordered array.
- Suppose that the ordered array is:

$$y_1, y_2, \ldots, y_n$$

- We compute the rank of the middle value (s):

$$rank = \frac{n+1}{2}$$

- If the sample size (n) is an odd number, there is only one value in the middle, and the rank will be an integer:

$$rank = \frac{n+1}{2} = m \quad \text{(m is integer)}$$

The median is the middle value of the **ordered** observations, which is:

$$\text{Median} = y_m.$$

Ordered set → (smallest to largest)	y_1	y_2	...	y_m middle value	...	y_n
Rank (or order) →	1	2	...	m	...	n

- If the sample size (n) is an even number, there are two values in the middle, and the rank will be an integer plus 0.5:

$$rank = \frac{n+1}{2} = m + 0.5$$

Therefore, the ranks of the middle values are (m) and (m+1). The median is the mean (average) of the two middle values of the **ordered** observations:

$$Median = \frac{y_m + y_{m+1}}{2}.$$

Ordered set →	y_1	y_2	...	y_m middle value	y_{m+1} middle value	...	y_n
Rank (or order) →	1	2	...	m	m+1	...	n

Example (odd number):

Find the median for the sample values: 10, 54, 21, 38, 53.

Solution:

$n = 5$ (odd number)

There is only one value in the middle.

The rank of the middle value is:

$$rank = \frac{n+1}{2} = \frac{5+1}{2} = 3. \quad (m=3)$$

Ordered set →	10	21	**38** (middle value)	53	54
Rank (or order) →	1	2	**3** (m)	4	5

The median =38 (unit)

Example (even number):

Find the median for the sample values: 10, 35, 41, 16, 20, 32

Solution:

$n = 6$ (even number)

There are two values in the middle.

The rank is:

$$rank = \frac{n+1}{2} = \frac{6+1}{2} = 3.5 = 3 + 0.5 = m+0.5 \quad (m=3)$$

Therefore, the ranks of the middle values are:

.$m = 3$ and $m+1 = 4$

Ordered set →	10	16	**20**	**32**	35	41
Rank (or order) →	1	2	**3** **(m)**	**4** **(m+1)**	5	6

The middle values are 20 and 32.

The median $= = \frac{20+32}{2} = \frac{52}{2} = 26$ (unit)

Note: The unit of the median is the same as the unit of the data.

Advantages and disadvantages of the median:

Advantages:

- Simplicity: The median is easily understood and easy to compute.
- Uniqueness: There is only one median for a given set of data.
- The median is not as drastically affected by extreme values as is the mean. (i.e., the median is not affected too much by extreme values).

For example:

Sample	Data						median
A	9	4	5	9	2	10	7
B	9	4	5	9	2	100	7

Disadvantages:

- The median does not take into account all values of the sample.
- In general, the median can only be found for quantitative variables. However, in some cases, the median can be found for ordinal qualitative variables.

Mode:

The mode of a set of values is that value which occurs most frequently (i.e., with the highest frequency).

- If all values are different or have the same frequencies, there will be <u>no</u> mode.
- A set of data may have more than one mode.

Example:

Data set	Type	Mode(s)
26, 25, 25, 34	Quantitative	25
3, 7, 12, 6, 19	Quantitative	No mode
3, 3, 7, 7, 12, 12, 6, 6, 19, 19	Quantitative	No mode
3, 3, 12, 6, 8, 8	Quantitative	3 and 8
B C A B B B C B B	Qualitative	B
B C A B A B C A C	Qualitative	No mode
B C A B B C B C C	Qualitative	B and C

Note: The unit of the mode is the same as the unit of the data.

Advantages and disadvantages of the mode:

Advantages:

- Simplicity: the mode is easily understood and easy to compute..
- The mode is not as drastically affected by extreme values as is the mean. (i.e., the mode is not affected too much by extreme values).

For example:

Sample	Data						Mode
A	7	4	5	7	2	10	7
B	7	4	5	7	2	100	7

- The mode may be found for both quantitative and qualitative variables.

Disadvantages:

- The mode is not a "good" measure of location, because it depends on a few values of the data.
- The mode does not take into account all values of the sample.
- There might be no mode for a data set.
- There might be more than one mode for a data set.

2.6 Descriptive Statistics: Measures of Dispersion (Measures of Variation):

The dispersion (variation) of a set of observations refers to the variety that they exhibit. A measure of dispersion conveys information regarding the amount of variability present in a set of data. There are several measures of dispersion, some of which are: Range, Variance, Standard Deviation, and Coefficient of Variation.

The variation or dispersion in a set of values refers to how spread out the values is from each other.

- The dispersion (variation) is small when the values are close together.
- There is no dispersion (no variation) if the values are the same.
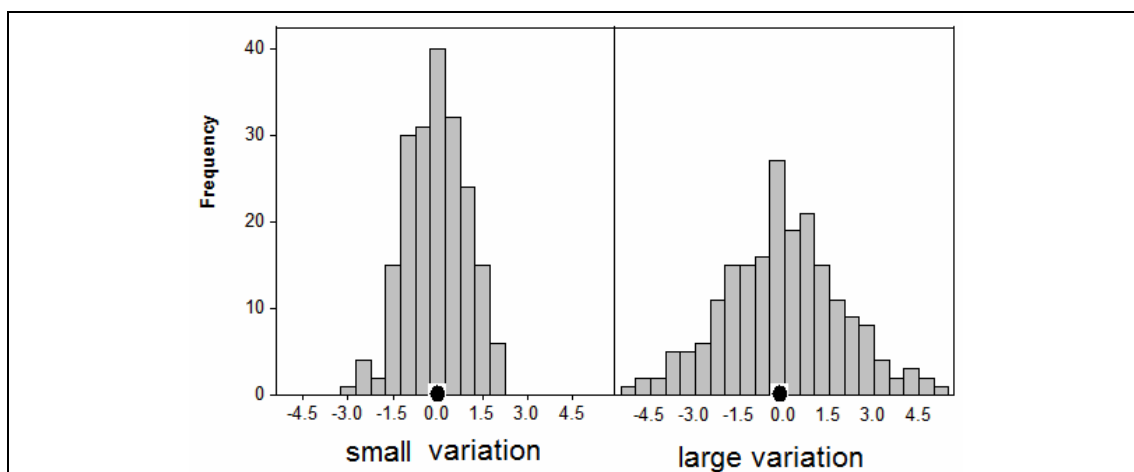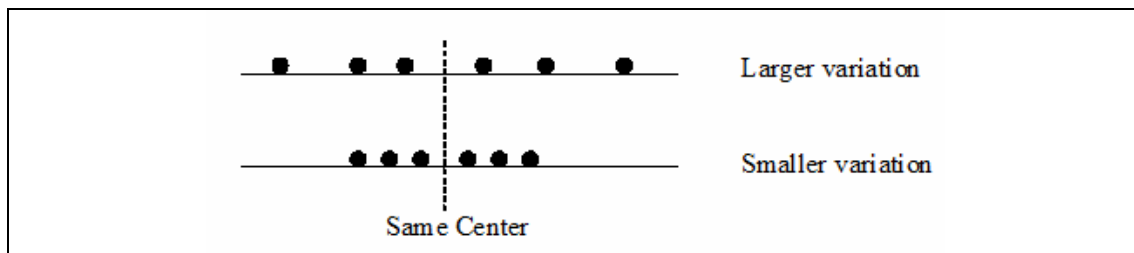
The Range:

The Range is the difference between the largest value (Max) and the smallest value (Min).

$$\text{Range } (R) = \text{Max} - \text{Min}$$

Example:

Find the range for the sample values: 26, 25, 35, 27, 29, 29.

Solution:

.max $= 35$

.min $= 25$

Range $(R) = 35 - 25 = 10$ (unit)

Notes:

1. The unit of the range is the same as the unit of the data.

2. The usefulness of the range is limited. The range is a poor measure of the dispersion because it only takes into account two of the values; however, it plays a significant role in many applications.

The Variance:

The variance is one of the most important measures of dispersion.

The variance is a measure that uses the mean as a point of reference.

- The variance of the data is small when the observations are close to the mean.
- The variance of the data is large when the observations are spread out from the mean.
- The variance of the data is zero (no variation) when all observations have the same value (concentrated at the mean).

Deviations of sample values from the sample mean:

Let $x_1, x_2, \ldots, x_n$ be the sample values, and $\bar{x}$ be the sample mean.

The deviation of the value x_i from the sample mean $\bar{x}$ is:
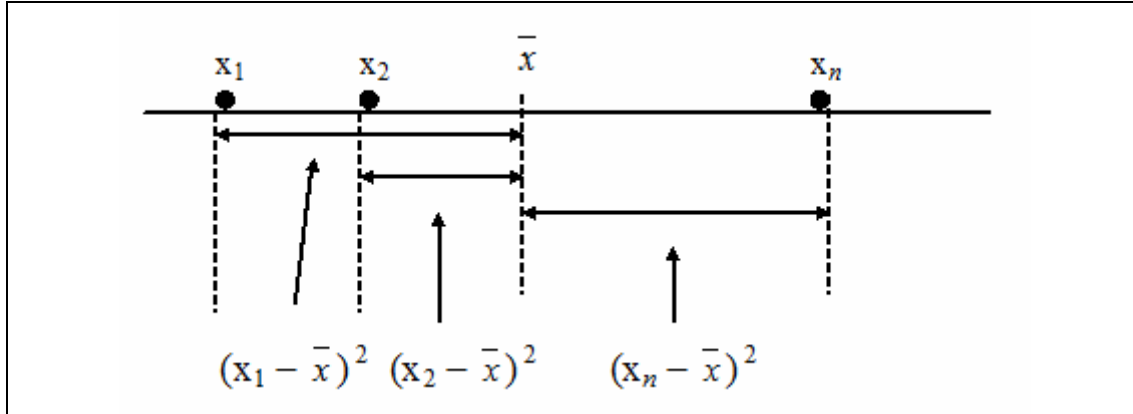
$$x_i - \bar{x}$$

The squared deviation is:

$$(x_i - \bar{x})^2$$

The sum of squared deviations is:

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

The following graph shows the squared deviations of the values from their mean:

(1)   The Population Variance σ^2 :

(Variance computed from the population)

Let $X_1, X_2, \ldots, X_N$ be the population values. The population variance (σ^2) is defined by:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}$$

$$= \frac{(X_1 - \mu)^2 + (X_2 - \mu)^2 + \cdots + (X_N - \mu)^2}{N} \qquad (unit)^2$$

where, $\mu = \dfrac{\sum_{i=1}^{N} X_i}{N}$ is the population mean, and (N) is the population size.

 Notes:
 - σ^2 is a parameter because it is obtained from the population values (it is unknown in general).
 - $\sigma^2 \geq 0$

(2)   The Sample Variance S^2:

(Variance computed from the sample)

Let $x_1, x_2, \ldots, x_n$ be the sample values. The sample variance (S^2) is defined by:

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

$$= \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1} \quad (unit)^2$$

where $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$ is the sample mean, and (n) is the sample size.

Notes:
- S^2 is a statistic because it is obtained from the sample values (it is known).
- S^2 is used to approximate (estimate) σ^2.
- $S^2 \geq 0$
- $S^2 = 0 \quad \Leftrightarrow \quad$ all observation have the same value
  $\Leftrightarrow \quad$ there is no dispersion (no variation)

Example:
We want to compute the sample variance of the following sample values: 10, 21, 33, 53, 54.
Solution:
$n=5$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum_{i=1}^{5} x_i}{5} = \frac{10 + 21 + 33 + 53 + 54}{5} = \frac{171}{5} = 34.2$$

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{5}(x_i - 34.2)^2}{5-1}$$

$$S^2 = \frac{(10 - 34.2)^2 + (21 - 34.2)^2 + (33 - 34.2)^2 + (53 - 34.2)^2 + (54 - 34.2)^2}{4}$$

$$= \frac{1506.8}{4} = 376.7 \quad (unit)^2$$

Another Method for calculating sample variance:

x_i	$(x_i - \bar{x}) = (x_i - 34.2)$	$(x_i - \bar{x})^2 = (x_i - 34.2)^2$
10	-24.2	585.64
21	-13.2	174.24

x_i	$(x_i - \bar{x}) = (x_i - 34.2)$	$(x_i - \bar{x})^2 = (x_i - 34.2)^2$
33	-1.2	1.44
53	18.8	353.44
54	19.8	392.04
$\sum\limits_{i=1}^{5} x_i = 171$	$\sum\limits_{i=1}^{5} (x_i - \bar{x}) = 0$	$\sum (x_i - \bar{x})^2 = 1506.8$

$$\bar{x} = \frac{\sum\limits_{i=1}^{5} x_i}{5} == \frac{171}{5} = 34.2 \quad \text{and} \quad S^2 = \frac{1506.8}{4} == 376.7$$

Standard Deviation:

The variance represents squared units, therefore, is not appropriate measure of dispersion when we wish to express the concept of dispersion in terms of the original unit.

- The standard deviation is another measure of dispersion.
- The standard deviation is the square root of the variance.
- The standard deviation is expressed in the original unit of the data.

(1) Population standard deviation is: $\sigma = \sqrt{\sigma^2}$ (unit)

(2) Sample standard deviation is: $S = \sqrt{S^2}$ (unit)

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

Example:

For the previous example, the sample standard deviation is
$$S = \sqrt{S^2} = \sqrt{376.7} = 19.41 \quad \text{(unit)}$$

Coefficient of Variation (C.V.):

- The variance and the standard deviation are useful as measures of variation of the values of a single variable for a single population.
- If we want to compare the variation of two variables we cannot use the variance or the standard deviation because:

1. The variables might have different units.
2. The variables might have different means.

- We need a measure of <u>the relative variation</u> that will not depend on either the units or on how large the values are. This measure is the coefficient of variation (C.V.).
- The coefficient of variation is defined by:

$$\text{C.V.} = \frac{S}{\bar{x}} \times 100\%$$

- The C.V. is free of unit (unit-less).
- To compare the variability of two sets of data (i.e., to determine which set is more variable), we need to calculate the following quantities:

	Mean	Standard deviation	C.V.
1st data set	$\bar{x}_1$	S_1	$C.V_1 = \frac{S_1}{\bar{x}_1}100\%$
2nd data set	$\bar{x}_2$	S_2	$C.V_2 = \frac{S_2}{\bar{x}_2}100\%$

- The data set with the larger value of CV has larger variation.
- The relative variability of the 1st data set is larger than the relative variability of the 2nd data set if $C.V_1 > C.V_2$ (and vice versa).

Example:
Suppose we have two data sets:
1st data set: $\bar{x}_1 = 66$ kg, $S_1 = 4.5$ kg

$$\Rightarrow C.V_1 = \frac{4.5}{66} * 100\% = 6.8\%$$

2nd data set: $\bar{x}_2 = 36$ kg, $S_2 = 4.5$ kg

$$\Rightarrow C.V_2 = \frac{4.5}{36} * 100\% = 12.5\%$$

Since $C.V_2 > C.V_1$, the relative variability of the 2nd data set is larger than the relative variability of the 1st data set.

If we use the standard deviation to compare the variability of the two data sets, we will wrongly conclude that the two data sets have the same variability because the standard deviation of both sets is 4.5 kg.

Chapter 3: Probability The Basis o Statistical Inference

3.1 Introduction
3.2 Probability
3.3 Elementary Properties of Probability
3.4 Calculating the Probability of an Event

General Definitions and Concepts:

Probability:

Probability is a measure (or number) used to measure the chance of the occurrence of some event. This number is between 0 and 1.

An Experiment:

An experiment is some procedure (or process) that we do.

Sample Space:

The sample space of an experiment is the set of all possible outcomes of an experiment. Also, it is called the universal set, and is denoted by Ω.

An Event:

Any subset of the sample space Ω is called an event.
- $\phi \subseteq \Omega$ is an event   (impossible event)
- $\Omega \subseteq \Omega$ is an event   (sure event)

Example:

Experiment: Selecting a ball from a box containing 6 balls numbered from 1 to 6 and observing the number on the selected ball.

This experiment has 6 possible outcomes.

The sample space is: $\Omega = \{1, 2, 3, 4, 5, 6\}$.

Consider the following events:

$E_1 = $ getting an even number $= \{2, 4, 6\} \subseteq \Omega$

$E_2 =$ getting a number less than $4 = \{1, 2, 3\} \subseteq \Omega$

$E_3 =$ getting 1 or $3 = \{1, 3\} \subseteq \Omega$

$E_4 =$ getting an odd number $= \{1, 3, 5\} \subseteq \Omega$

$E_5 =$ getting a negative number $= \{\ \} = \phi \subseteq \Omega$

$E_6 =$ getting a number less than $10 = \{1, 2, 3, 4, 5, 6\} = \Omega \subseteq \Omega$

Notation: $n(\Omega) =$ no. of outcomes (elements) in Ω

$n(E) =$ no. of outcomes (elements) in the event E

Equally Likely Outcomes:

The outcomes of an experiment are equally likely if the outcomes have the same chance of occurrence.

Probability of An Event:

If the experiment has $n(\Omega)$ equally likely outcomes, then the probability of the event E is denoted by $P(E)$ and is defined by:

$$P(E) = \frac{n(E)}{n(\Omega)} = \frac{no.\ of\ outcomes\ in\ E}{no.\ of\ outcomes\ in\ \Omega}$$

Example:

In the ball experiment in the previous example, suppose the ball is selected at random. Determine the probabilities of the following events:

$E_1 =$ getting an even number

$E_2 =$ getting a number less than 4

$E_3 =$ getting 1 or 3

Solution:

$\Omega = \{1, 2, 3, 4, 5, 6\}$; $n(\Omega) = 6$

$E_1 = \{2, 4, 6\}$; $n(E_1) = 3$

$E_2 = \{1, 2, 3\}$; $n(E_2) = 3$

$E_3 = \{1, 3\}$; $n(E_3) = 2$
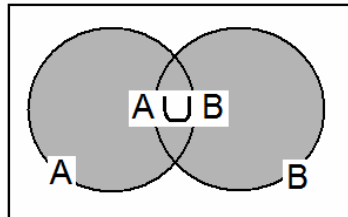
The outcomes are equally likely.

$$\therefore P(E_1) = \frac{3}{6}, \qquad P(E_2) = \frac{3}{6}, \qquad P(E_3) = \frac{2}{6},$$

Some Operations on Events:
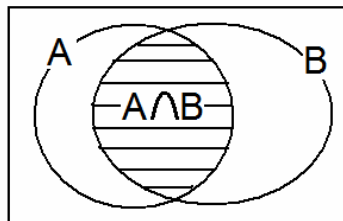   Let A and B be two events defined on the sample space Ω.

Union of Two events: $(A \cup B)$ or $(A + B)$
        The event $A \cup B$ consists of all outcomes in A **or** in B **or** in both A and B. The event $A \cup B$ occurs if A occurs, **or** B occurs, **or** both A and B occur.
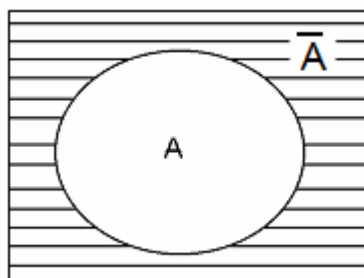

Intersection of Two Events: $(A \cap B)$
        The event $A \cap B$ Consists of all outcomes in both A **and** B. The event $A \cap B$ Occurs if both A **and** B occur.

Complement of an Event: $(\overline{A})$ or (A^c) or (A')
        The complement of the even A is denoted by $\overline{A}$. The even $\overline{A}$ consists of all outcomes of Ω but are not in A. The even $\overline{A}$ occurs if A does not.
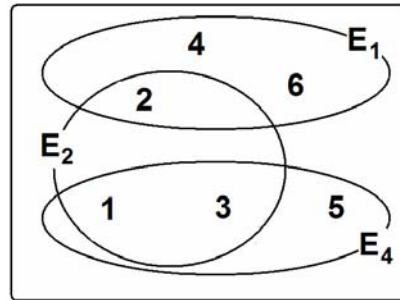

Example:
        Experiment: Selecting a ball from a box containing 6 balls numbered 1, 2, 3, 4, 5, and 6 randomly.
Define the following events:
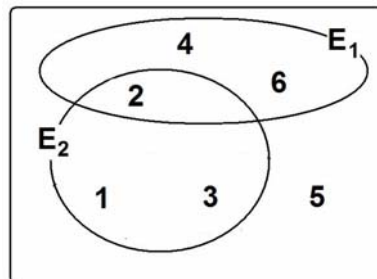 $E_1 = \{2, 4, 6\}$ = getting an even number.

$E_2 = \{1, 2, 3\} =$ getting a number < 4.
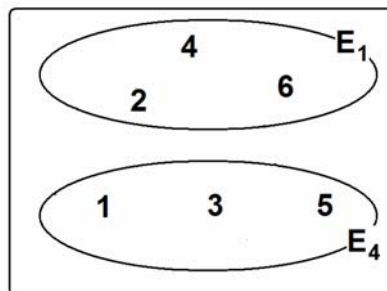
$E_4 = \{1, 3, 5\} =$ getting an odd number.

(1) $E_1 \cup E_2 = \{1, 2, 3, 4, 6\}$

     = getting an even number **or** a number less than 4.

$$P(E_1 \cup E_2) = \frac{n(E_1 \cup E_2)}{n(\Omega)} = \frac{5}{6}$$

(2) $E_1 \cup E_4 = \{1, 2, 3, 4, 5, 6\} = \Omega$
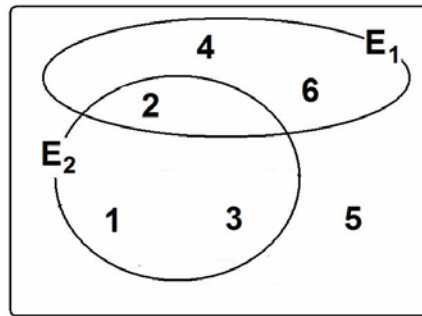
     = getting an even number **or** an odd number.

$$P(E_1 \cup E_4) = \frac{n(E_1 \cup E_4)}{n(\Omega)} = \frac{6}{6} = 1$$

Note: $E_1 \cup E_4 = \Omega$. E_1 and E_4 are called exhaustive events. The union of these events gives the whole sample space.
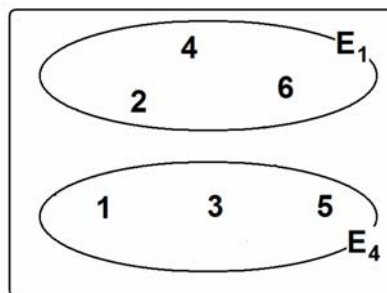
(3) $E_1 \cap E_2 = \{2\}$ = getting an even number **and** a number less than 4.

$$P(E_1 \cap E_2) = \frac{n(E_1 \cap E_2)}{n(\Omega)} = \frac{1}{6}$$

(4) $E_1 \cap E_4 = \phi$ = getting an even number **and** an odd number.

$$P(E_1 \cap E_4) = \frac{n(E_1 \cap E_4)}{n(\Omega)} = \frac{n(\phi)}{6} = \frac{0}{6} = 0$$

Note: $E_1 \cap E_4 = \phi$. E_1 and E_4 are called disjoint (or mutually exclusive) events. These kinds of events can not occurred simultaneously (together in the same time).

(5)   The complement of E_1

   $\overline{E_1}$ = <u>not</u> getting an even number = $\overline{\{2,4,6\}}$ = $\{1, 3, 5\}$
      = getting an odd number.
      = E_4

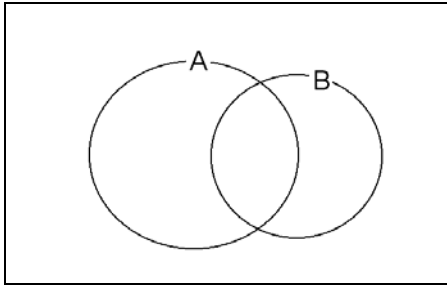
<u>Mutually exclusive (disjoint) Events:</u>
   The events A and B are disjoint (or mutually exclusive) if:
$$A \cap B = \phi.$$
For this case, it is impossible that both events occur simultaneously (i.e., together in the same time). In this case:
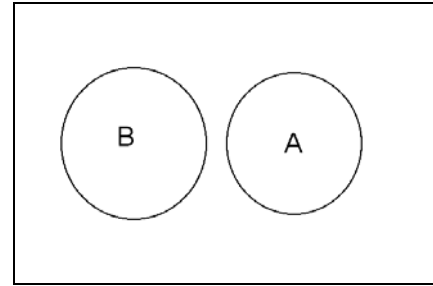   (i) $P(A \cap B) = 0$
   (ii) $P(A \cup B) = P(A) + P(B)$
If $A \cap B \neq \phi$, then A and B are not mutually exclusive (not disjoint).

$A \cap B \neq \phi$

A and *B* are not mutually exclusive
(It is possible that both events occur in the same time)

$A \cap B = \phi$

A and *B* are mutually exclusive (disjoint)
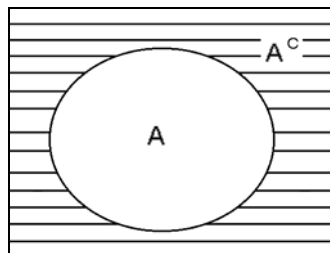(It is impossible that both events occur in the same time)

Exhaustive Events:

The events $A_1, A_2, \ldots, A_n$ are exhaustive events if:

$$A_1 \cup A_2 \cup \ldots \cup A_n = \Omega.$$

For this case, $P(A_1 \cup A_2 \cup \ldots \cup A_n) = P(\Omega) = 1$

Note:

1. $A \cup \overline{A} = \Omega$ (*A* and $\overline{A}$ are exhaustive events)
2. $A \cap \overline{A} = \phi$ (*A* and $\overline{A}$ are mutually exclusive (disjoint) events)
3. $n(\overline{A}) = n(\Omega) - n(A)$
4. $P(\overline{A}) = 1 - P(A)$

General Probability Rules:

1. $0 \leq P(A) \leq 1$
2. $P(\Omega) = 1$
3. $P(\phi) = 0$
4. $P(\overline{A}) = 1 - P(A)$

<u>The Addition Rule:</u>

For any two events *A* and *B:*
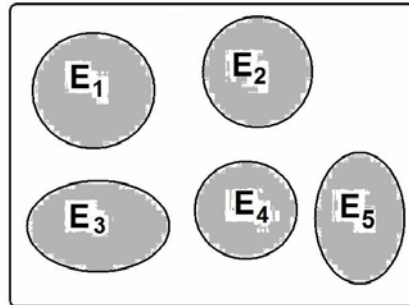$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Special Cases:

1. For mutually exclusive (disjoint) events *A* and *B*
$$P(A \cup B) = P(A) + P(B)$$

2. For mutually exclusive (disjoint) events $E_1, E_2, \ldots, E_n$:
$$P(E_1 \cup E_2 \cup \ldots \cup E_n) = P(E_1) + P(E_2) + \cdots + P(E_n)$$

Note:

If the events $A_1, A_2, \ldots, A_n$ are exhaustive and mutually exclusive (disjoint) events, then:
$$P(A_1 \cup A_2 \cup \ldots \cup A_n) = P(A_1) + P(A_2) + \cdots + P(A_n) = P(\Omega) = 1$$

<u>Marginal Probability:</u>

Given some variable that can be broken down into (m) categories designated by A_1, A_2, $\cdots$, A_m and another jointly occurring variable that is broken down into (n) categories designated by B_1, B_2, $\cdots$, B_n.

	B_1	B_2	…	B_n	Total
A_1	$n(A_1 \cap B_1)$	$n(A_1 \cap B_2)$	…	$n(A_1 \cap B_n)$	$n(A_1)$
A_2	$n(A_2 \cap B_1)$	$n(A_2 \cap B_2)$	…	$n(A_2 \cap B_n)$	$n(A_2)$
. . .	. . .	. . .	. . .	. . .	. . .
A_m	$n(A_m \cap B_1)$	$n(A_m \cap B_2)$	…	$n(A_m \cap B_n)$	$n(A_m)$
Total	$n(B_1)$	$n(B_2)$	…	$n(B_n)$	$n(\Omega)$

(This table contains the number of elements in each event)

	B_1	B_2	...	B_n	Marginal Probability
A_1	$P(A_1 \cap B_1)$	$P(A_1 \cap B_2)$	...	$P(A_1 \cap B_n)$	$P(A_1)$
A_2	$P(A_2 \cap B_1)$	$P(A_2 \cap B_2)$	...	$P(A_2 \cap B_n)$	$P(A_2)$
. . .	. . .	. . .	. . .	. . .	. . .
A_m	$P(A_m \cap B_1)$	$P(A_m \cap B_2)$	...	$P(A_m \cap B_n)$	$P(A_m)$
Marginal Probability	$P(B_1)$	$P(B_2)$	...	$P(B_n)$	1.00

(This table contains the probability of each event)

The marginal probability of A_i, $P(A_i)$, is equal to the sum of the joint probabilities of A_i with all categories of B. That is:

$$P(A_i) = P(A_i \cap B_1) + P(A_i \cap B_2) + \ldots + P(A_i \cap B_n)$$

$$= \sum_{j=1}^{n} P(A_i \cap B_j)$$

For example,

$$P(A_2) = P(A_2 \cap B_1) + P(A_2 \cap B_2) + \ldots + P(A_2 \cap B_n)$$

$$= \sum_{j=1}^{n} P(A_2 \cap B_j)$$

We define the marginal probability of B_j, $P(B_j)$, in a similar way.

Example:

Table of number of elements in each event:

	B_1	B_2	B_3	Total
A_1	50	30	70	150
A_2	20	70	10	100
A_3	30	100	120	250
Total	100	200	200	500

Table of probabilities of each event:

	B_1	B_2	B_3	Marginal Probability
A_1	0.1	0.06	0.14	0.3
A_2	0.04	0.14	0.02	0.2
A_3	0.06	0.2	0.24	0.5
Marginal Probability	0.2	0.4	0.4	1

For example:
$$P(A_2) = P(A_2 \cap B_1) + P(A_2 \cap B_2) + P(A_2 \cap B_n)$$
$$= 0.04 + 0.14 + 0.02$$
$$= 0.2$$

Applications:

Example:

630 patients are classified as follows:

Blood Type	O (E_1)	A (E_2)	B (E_3)	AB (E_4)	Total
No. of patients	284	258	63	25	630

- Experiment: Selecting a patient at random and observe his/her blood type.
- This experiment has 630 equally likely outcomes $n(\Omega) = 630$

Define the events:

$E_1 =$ The blood type of the selected patient is "O"

$E_2 =$ The blood type of the selected patient is "A"

$E_3 =$ The blood type of the selected patient is "B"

$E_4 =$ The blood type of the selected patient is "AB"

Number of elements in each event:

$n(E_1) = 284,$ $n(E_2) = 258,$

$n(E_3) = 63,$ $n(E_4) = 25.$

Probabilities of the events:

$$P(E_1) = \frac{284}{630} = 0.4508, \qquad P(E_2) = \frac{258}{630} = 0.4095,$$

$$P(E_3) = \frac{63}{630} = 0.1, \qquad P(E_4) = \frac{25}{630} = 0.0397,$$

Some operations on the events:

1. $E_2 \cap E_4 =$ the blood type of the selected patients is *"A"* **and** *"AB"*.

   $E_2 \cap E_4 = \phi$ (disjoint events / mutually exclusive events)

   $P(E_2 \cap E_4) = P(\phi) = 0$

2. $E_2 \cup E_4 =$ the blood type of the selected patients is *"A"* **or** *"AB"*

$$P(E_2 \cup E_4) = \begin{cases} \dfrac{n(E_2 \cup E_4)}{n(\Omega)} = \dfrac{258+25}{630} = \dfrac{283}{630} = 0.4492 \\ or \\ P(E_2) + P(E_4) = \dfrac{258}{630} + \dfrac{25}{630} = \dfrac{283}{630} = 0.4492 \end{cases}$$

$$(\text{since } E_2 \cap E_4 = \phi)$$

3. $\overline{E_1}$ = the blood type of the selected patients is not "O".

$$n(\overline{E_1}) = n(\Omega) - n(E_1) = 630 - 284 = 346$$

$$P(\overline{E_1}) = \frac{n(\overline{E_1})}{n(\Omega)} = \frac{346}{630} = 0.5492$$

another solution:

$$P(E_1^C) = 1 - P(E_1) = 1 - 0.4508 = 0.5492$$

Notes:

1. E_1, E_2, E_3, E_4 are mutually disjoint, $E_i \cap E_j = \phi \ (i \neq j)$.

2. E_1, E_2, E_3, E_4 are exhaustive events, $E_1 \cup E_2 \cup E_3 \cup E_4 = \Omega$.

Example:

339 physicians are classified based on their ages and smoking habits as follows.

Smoking Habit

	Daily (B_1)	Occasionally (B_2)	Not at all (B_3)	Total
20 - 29 (A_1)	31	9	7	47
30 - 39 (A_2)	110	30	49	189
40 - 49 (A_3)	29	21	29	79
50+ (A_4)	6	0	18	24
Total	176	60	103	339

Age

Experiment: Selecting a physician at random

The number of elements of the sample space is $n(\Omega) = 339$.

The outcomes of the experiment are equally likely.

Some events:

- A_3 = the selected physician is aged 40 - 49

$$P(A_3) = \frac{n(A_3)}{n(\Omega)} = \frac{79}{339} = 0.2330$$

- B_2 = the selected physician smokes occasionally

$$P(B_2) = \frac{n(B_2)}{n(\Omega)} = \frac{60}{339} = 0.1770$$

- $A_3 \cap B_2$ = the selected physician is aged 40-49 **and** smokes occasionally.

$$P(A_3 \cap B_2) = \frac{n(A_3 \cap B_2)}{n(\Omega)} = \frac{21}{339} = 0.06195$$

- $A_3 \cup B_2$ = the selected physician is aged 40-49 **or** smokes occasionally (**or** both)

$$P(A_3 \cup B_2) = P(A_3) + P(B_2) - P(A_3 \cap B_2)$$
$$= \frac{79}{339} + \frac{60}{339} - \frac{21}{339}$$
$$= 0.233 + 0.177 - 0.06195 = 0.3481$$

- $\overline{A_4}$ = the selected physician is **not** 50 years or older.
  $$= A_1 \cup A_2 \cup A_3$$

$$P(\overline{A_4}) = 1 - P(A_4)$$
$$= 1 - \frac{n(A_4)}{n(\Omega)} = 1 - \frac{24}{339} = 0.9292$$

- $A_2 \cup A_3$ = the selected physician is aged 30-39 **or** is aged 40-49
  = the selected physician is aged 30-49

$$\begin{cases} P(A_2 \cup A_3) = \dfrac{n(A_2 \cup A_3)}{n(\Omega)} = \dfrac{189 + 79}{339} = \dfrac{268}{339} = 0.7906 \\[2mm] or \\[2mm] P(A_2 \cup A_3) = P(A_2) + P(A_3) = \dfrac{189}{339} + \dfrac{79}{339} = 0.7906 \end{cases}$$

$$\text{(Since } A_2 \cap A_3 = \phi \text{)}$$

Example:

Suppose that there is a population of pregnant women with:

- 10% of the pregnant women delivered prematurely.
- 25% of the pregnant women used some sort of medication.

- 5% of the pregnant women delivered prematurely <u>and</u> used some sort of medication.

Experiment: Selecting a woman randomly from this population.
Define the events:

- D = The selected woman delivered prematurely.
- M = The selected women used medication.
- $D \cap M$ = The selected woman delivered prematurely <u>and</u> used some sort of medication.

Percentages:
$$\%(D) = 10\% \qquad \%(M) = 25\% \qquad \%(D \cap M) = 5\%$$

The complement events:

$\bar{D}$ = The selected woman did not deliver prematurely.
$\bar{M}$ = The selected women did not use medication.

A Two-way table: (Percentages given by a two-way table):

	M	$\bar{M}$	Total
D	**5**	?	**10**
$\bar{D}$	?	?	?
Total	**25**	?	**100**

	M	$\bar{M}$	Total
D	**5**	5	**10**
$\bar{D}$	20	70	90
Total	**25**	75	**100**

The probabilities of the given events are:
$$P(D) = \frac{\%(D)}{100\%} = \frac{10\%}{100\%} = 0.1$$
$$P(M) = \frac{\%(M)}{100\%} = \frac{25\%}{100\%} = 0.25$$
$$P(D \cap M) = \frac{\%(D \cap M)}{100\%} = \frac{5\%}{100\%} = 0.05$$

Calculating probabilities of some events:
$D \cup M$ = the selected woman delivered prematurely <u>or</u> used medication.
$$P(D \cup M) = P(D) + (M) - P(D \cap M)$$
$$= 0.1 + 0.25 - 0.05 = 0.3 \qquad \text{(by the rule)}$$

$\overline{M}$ = The selected woman did <u>not</u> use medication

$$P(\overline{M}) = 1 - P(M) = 1 - 0.25 = 0.75 \quad \text{(by the rule)}$$

$$P(\overline{M}) = \frac{75}{100} = 0.75 \qquad \text{(from the table)}$$

$\overline{D}$ = The selected woman did <u>not</u> deliver prematurely

$$P(\overline{D}) = 1 - P(D) = 1 - 0.10 = 0.90 \quad \text{(by the rule)}$$

$$P(\overline{D}) = \frac{90}{100} = 0.90 \qquad \text{(from the table)}$$

$\overline{D} \cap \overline{M}$ = the selected woman did not deliver prematurely <u>and</u> did not use medication.

$$P(\overline{D} \cap \overline{M}) = \frac{70}{100} = 0.70 \qquad \text{(from the table)}$$

$\overline{D} \cap M$ = the selected woman did not deliver prematurely <u>and</u> used medication.

$$P(\overline{D} \cap M) = \frac{20}{100} = 0.20 \qquad \text{(from the table)}$$

$D \cap \overline{M}$ = the selected woman delivered prematurely <u>and</u> did not use medication.

$$P(D \cap \overline{M}) = \frac{5}{100} = 0.05 \qquad \text{(from the table)}$$

$D \cup \overline{M}$ = the selected woman delivered prematurely <u>or</u> did not use medication.

$$P(D \cup \overline{M}) = P(D) + (\overline{M}) - P(D \cap \overline{M})$$
$$= 0.1 + 0.75 - 0.05 = 0.8 \qquad \text{(by the rule)}$$

$\overline{D} \cup M$ = the selected woman did not deliver prematurely <u>or</u> used medication.

$$P(\overline{D} \cup M) = P(\overline{D}) + (M) - P(\overline{D} \cap M)$$
$$= 0.9 + 0.25 - 0.20 = 0.95 \qquad \text{(by the rule)}$$

$\overline{D} \cup \overline{M}$ = the selected woman did not deliver prematurely <u>or</u> did not use medication.
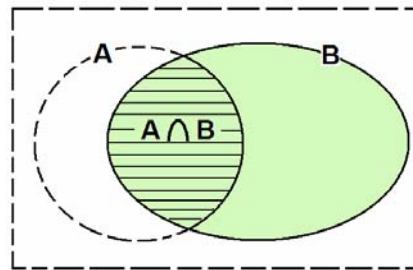
$$P(\overline{D} \cup \overline{M}) = P(\overline{D}) + (\overline{M}) - P(\overline{D} \cap \overline{M})$$
$$= 0.9 + 0.75 - 0.70 = 0.95 \qquad \text{(by the rule)}$$

Conditional Probability:

- The conditional probability of the event *A* when we know that the event *B* has already occurred is defined by:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \qquad ; P(B) \neq 0$$

- P(A | B) = The conditional probability of *A* given *B*.

Notes:

(1) $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)} = \dfrac{n(A \cap B)/n(\Omega)}{n(B)/n(\Omega)} = \dfrac{n(A \cap B)}{n(B)}$

(2) $P(B \mid A) = \dfrac{P(A \cap B)}{P(A)}$

(3) For calculating $P(A \mid B)$, we may use any one of the following:

(i) $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)}$

(ii) $P(A \mid B) = \dfrac{n(A \cap B)}{n(B)}$

(iii) Using the restricted table directly.

Multiplication Rules of Probability:

For any two events A and B, we have:
$$P(A \cap B) = P(B)P(A \mid B)$$
$$P(A \cap B) = P(A)P(B \mid A)$$

Example:

Smoking Habit

	Daily (B_1)	Occasionally (B_2)	Not at all (B_3)	Total
20-29 (A_1)	31	9	7	47
30-39 (A_2)	110	30	49	189
40-49 (A_3)	29	21	29	79
50+ (A_4)	6	0	18	24
Total	176	60	103	339

Age

Consider the following event:

$(B_1 \mid A_2)$ = the selected physician smokes daily given that his

age is between 30 and 39

- $P(B_1) = \dfrac{n(B_1)}{n(\Omega)} = \dfrac{176}{339} = 0.519$

- $P(B_1 \mid A_2) = \dfrac{P(B_1 \cap A_2)}{P(A_2)}$

$$= \dfrac{0.324484}{0.557522} = 0.5820$$

$$\begin{cases} P(B_1 \cap A_2) = \dfrac{n(B_1 \cap A_2)}{n(\Omega)} = \dfrac{110}{339} = 0.324484 \\[4mm] P(A_2) = \dfrac{n(A_2)}{n(\Omega)} = \dfrac{189}{339} = 0.557522 \end{cases}$$

another solution:

$$P(B_1 \mid A_2) = \dfrac{n(B_1 \cap A_2)}{n(A_2)} = \dfrac{110}{189} = 0.5820$$

Notice that:

$$P(B_1) = 0.519$$
$$P(B_1 \mid A_2) = 0.5820$$
$$P(B_1 \mid A_2) > P(B_1) \; !! \; \dots \; P(B_1) \neq P(B_1 \mid A_2)$$

What does this mean?
We will answer this question after talking about the concept of independent events.

Example: (Multiplication Rule of Probability)

A training health program consists of two consecutive parts. To pass this program, the trainee must pass both parts of the program. From the past experience, it is known that 90% of the trainees pass the first part, and 80% of those who pass the first part pass the second part. If you are admitted to this program, what is the probability that you will pass the program? What is the percentage of trainees who pass the program?

Solution:

Define the following events:

A = the event of passing the first part

B = the event of passing the second part

A∩B = the event of passing the first part and the second Part

= the event of passing both parts

= the event of passing the program

Therefore, the probability of passing the program is P(A∩B).

From the given information:

The probability of passing the first part is:

$$P(A) = 0.9 \qquad (\frac{90\%}{100\%} = 0.9)$$

The probability of passing the second part given that the trainee has already passed the first part is:

$$P(B|A) = 0.8 \qquad (\frac{80\%}{100\%} = 0.8)$$

Now, we use the multiplication rule to find P(A∩B) as follows:

$$P(A∩B) = P(A)\ P(B|A) = (0.9)(0.8) = 0.72$$

We can conclude that 72% of the trainees pass the program.

Independent Events

There are 3 cases:

- $P(A|B) > P(A)$

    (knowing B increases the probability of occurrence of A)

- $P(A|B) < P(A)$

    (knowing B decreases the probability of occurrence of A)

- $P(A|B) = P(A)$

    (knowing B has no effect on the probability of occurrence of A). In this case A is independent of B.

Independent Events:

- Two events A and B are independent if one of the following conditions is satisfied:

$$(i)\ \ P(A|B) = P(A)$$
$$\Leftrightarrow (ii)\ \ P(B|A) = P(B)$$
$$\Leftrightarrow (iii)\ \ P(B∩A) = P(A)P(B)$$

Note: The third condition is the multiplication rule of independent events.

Example:

Suppose that A and B are two events such that:

$$P(A) = 0.5 , \quad P(B)=0.6, \quad P(A \cap B)=0.2.$$

Theses two events are not independent (they are dependent) because:

$P(A) P(B) = 0.5 \times 0.6 = 0.3$

$P(A \cap B)=0.2.$

$P(A \cap B) \neq P(A) P(B)$

Also, $P(A) = 0.5 \neq P(A|B) = \dfrac{P(A \cap B)}{P(B)} = \dfrac{0.2}{0.6} = 0.3333$.

Also, $P(B) = 0.6 \neq P(B|A) = \dfrac{P(A \cap B)}{P(A)} = \dfrac{0.2}{0.5} = 0.4$.

For this example, we may calculate probabilities of all events. We can use a two-way table of the probabilities as follows:

	B	$\overline{B}$	Total
A	**0.2**	?	**0.5**
$\overline{A}$	?	?	?
Total	**0.6**	?	**1.00**

We complete the table:

	B	$\overline{B}$	Total
A	**0.2**	0.3	**0.5**
$\overline{A}$	0.4	0.1	0.5
Total	**0.6**	0.4	**1.00**

$P(\overline{A}) = 0.5$

$P(\overline{B}) = 0.4$

$P(A \cap \overline{B}) = 0.3$

$P(\overline{A} \cap B) = 0.4$

$P(\overline{A} \cap \overline{B}) = 0.1$

$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.5 + 0.6 - 0.2 = 0.9$

$P(A \cup \overline{B}) = P(A) + P(\overline{B}) - P(A \cap \overline{B}) = 0.5 + 0.4 - 0.3 = 0.6$

$P(\overline{A} \cup B) = exercise$

$P(\overline{A} \cup \overline{B}) = exercise$

Note: The Addition Rule for Independent Events:

If the events *A* and *B* are independent, then

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad \text{(Addition rule)}$$
$$= P(A) + P(B) - P(A)P(B)$$

Example: (Reading Assignment)

Suppose that a dental clinic has 12 nurses classified as follows:

Nurse	1	2	3	4	5	6	7	8	9	10	11	12
Has children	Yes	No	No	No	No	Yes	No	No	Yes	No	No	No
Works at night	No	No	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	Yes

The experiment is to randomly choose one of these nurses. Consider the following events:

C = the chosen nurse has children

N = the chosen nurse works night shift

a) Find The probabilities of the following events:
   1. the chosen nurse has children.
   2. the chosen nurse works night shift.
   3. the chosen nurse has children and works night shift.
   4. the chosen nurse has children and does not work night shift.
b) Find the probability of choosing a nurse who woks at night given that she has children.
c) Are the events C and N independent? Why?
d) Are the events C and N disjoint? Why?
e) Sketch the events C and N with their probabilities using Venn diagram.

Solution:

We can classify the nurses as follows:

	N (Night shift)	$\overline{N}$ (No night shift)	total
C (Has Children)	2	1	3
$\overline{C}$ (No Children)	6	3	9
total	8	4	12

a) The experiment has n(Ω) = 12 equally likely outcomes.

P(The chosen nurse has children) = P(C) = $\dfrac{n(C)}{n(\Omega)} = \dfrac{3}{12} = 0.25$

P(The chosen nurse works night shift) = P(N) = $\dfrac{n(N)}{n(\Omega)} = \dfrac{8}{12} = 0.6667$

P(The chosen nurse has children and works night shift)

= P(C∩N) = $\dfrac{n(C \cap N)}{n(\Omega)} = \dfrac{2}{12} = 0.16667$

P(The chosen nurse has children and does not work night shift)

$$= \ P(C \cap \bar{N}) = \frac{n(C \cap \bar{N})}{n(\Omega)} = \frac{1}{12} = 0.0833$$

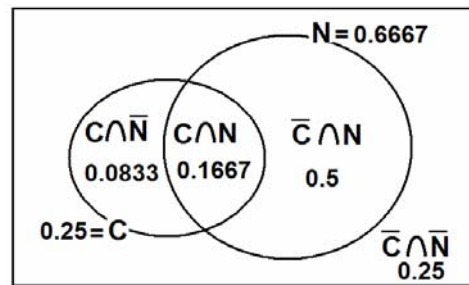b) The probability of choosing a nurse who woks at night given that she has children:

$$P(N \mid C) = \frac{P(C \cap N)}{P(C)} = \frac{2/12}{0.25} = 0.6667$$

c) The events C and N are independent because $P(N \mid C) = P(N)$.

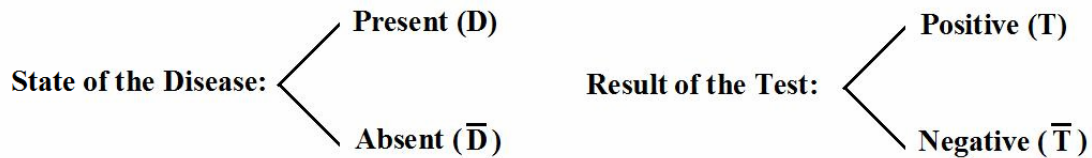d) The events C and N not are disjoint because $C \cap N \neq \phi$.

   (Note: $n(C \cap N) = 2$)

e) Venn diagram

3.5 Bayes' Theorem, Screening Tests, Sensitivity, Specificity, and Predictive Value Positive and Negative: (pp.79-83)

There are two states regarding the disease and two states regarding the result of the screening test:

State of the Disease:
- Present (D)
- Absent ($\overline{D}$)

Result of the Test:
- Positive (T)
- Negative ($\overline{T}$)

We define the following events of interest:

D : the individual has the disease (presence of the disease)

$\overline{D}$: the individual does not have the disease (absence of The disease)

T : the individual has a positive screening test result

$\overline{T}$: the individual has a negative screening test result

- There are 4 possible situations:

		True status of the disease	
		+ve (D: Present)	−ve ($\overline{D}$:Absent)
Result of	+ve (T)	Correct diagnosing	false positive result
the test	−ve ($\overline{T}$)	false negative result	Correct diagnosing

Definitions of False Results:

There are two false results:

1. **A false positive** result:

This result happens when a test indicates a positive status when the true status is negative. Its probability is:

$P(T \mid \overline{D})$ = P(positive result | absence of the disease)

2. **A false negative** result:

This result happens when a test indicates a negative status when the true status is positive. Its probability is:

$P(\overline{T} \mid D)$ = P(negative result | presence of the disease)

Definitions of the Sensitivity and Specificity of the test:

1. The Sensitivity:

The sensitivity of a test is the probability of a positive test result given the presence of the disease.

$P(T \mid D) = $ P(positive result of the test | presence of the disease)

2. The specificity:

The specificity of a test is the probability of a negative test result given the absence of the disease.

$P(\overline{T} \mid \overline{D}) = $ P(negative result of the test | absence of the disease)

To clarify these concepts, suppose we have a sample of (n) subjects who are cross-classified according to Disease Status and Screening Test Result as follows:

Test Result	Disease		Total
	Present (D)	Absent ($\overline{D}$)	
Positive (T)	a	b	a + b = n(T)
Negative ($\overline{T}$)	c	d	c + d = n($\overline{T}$)
Total	a + c = n(D)	b + d = n($\overline{D}$)	n

For example, there are (a) subjects who have the disease and whose screening test result was positive.

From this table we may compute the following conditional probabilities:

1. The probability of false positive result:

$$P(T \mid \overline{D}) = \frac{n(T \cap \overline{D})}{n(\overline{D})} = \frac{b}{b+d}$$

2. The probability of false negative result:

$$P(\overline{T} \mid D) = \frac{n(\overline{T} \cap D)}{n(D)} = \frac{c}{a+c}$$

3. The sensitivity of the screening test:

$$P(T \mid D) = \frac{n(T \cap D)}{n(D)} = \frac{a}{a+c}$$

4. The specificity of the screening test:

$$P(\overline{T} \mid \overline{D}) = \frac{n(\overline{T} \cap \overline{D})}{n(\overline{D})} = \frac{d}{b+d}$$

Definitions of the Predictive Value Positive and Predictive Value Negative of a Screening Test:

1. The predictive value positive of a screening test:
The predictive value positive is the probability that a subject has the disease, given that the subject has a positive screening test result:

$P(D\,|\,T)$ = P(the subject has the disease | positive result)
        = P(presence of the disease | positive result)

2. The predictive value negative of a screening test:
The predictive value negative is the probability that a subject does not have the disease, given that the subject has a negative screening test result:

$P(\overline{D}\,|\,\overline{T})$ = P(the subject does not have the disease | negative result)
        = P(absence of the disease | negative result)

Calculating the Predictive Value Positive and Predictive Value Negative:
(How to calculate $P(D\,|\,T)$ and $P(\overline{D}\,|\,\overline{T})$):
    We calculate these conditional probabilities using the knowledge of:
    1. The sensitivity of the test = $P(T\,|\,D)$
    2. The specificity of the test = $P(\overline{T}\,|\,\overline{D})$
    3. The probability of the relevant disease in the general population, P(D). (It is usually obtained from another independent study)

Calculating the Predictive Value Positive, $P(D\,|\,T)$:
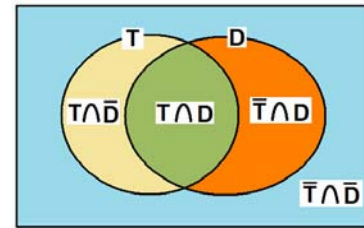$$P(D\,|\,T) = \frac{P(T \cap D)}{P(T)}$$

But we know that:

$P(T) = P(T \cap D) + P(T \cap \overline{D})$

$P(T \cap D) = P(T|D)P(D)$ (multiplication rule)

$P(T \cap \overline{D}) = P(T|\overline{D})P(\overline{D})$ (multiplication rule)

$P(T) = P(T|D)P(D) + P(T|\overline{D})P(\overline{D})$

Therefore, we reach the following version of Bayes' Theorem:

$$P(D|T) = \frac{P(T|D)\,P(D)}{P(T|D)\,P(D) + P(T|\overline{D})\,P(\overline{D})} \quad \cdots\cdots\cdots\cdots (1)$$

Note:

$P(T|D)$ = sensitivity.

$P(T|\overline{D}) = 1 - P(\overline{T}|\overline{D}) = 1 - $ specificity.

$P(D)$ = The probability of the relevant disease in the general population.

$P(\overline{D}) = 1 - P(D)$.

Calculating the Predictive Value Negative, $P(\overline{D}|\overline{T})$:

To obtain the predictive value negative of a screening test, we use the following statement of Bayes' theorem:

$$P(\overline{D}|\overline{T}) = \frac{P(\overline{T}|\overline{D})\,P(\overline{D})}{P(\overline{T}|\overline{D})\,P(\overline{D}) + P(\overline{T}|D)\,P(D)} \quad \cdots\cdots\cdots\cdots (2)$$

Note:

$P(\overline{T}|\overline{D})$ = specificity.

$P(\overline{T}|D) = 1 - P(T|D) = 1 - $ sensitivity.

Example:

A medical research team wished to evaluate a proposed screening test for Alzheimer's disease. The test was given to a random sample of 450 patients with Alzheimer's disease and an independent random sample of 500 patients without symptoms of the disease. The two samples were drawn from populations of subjects who were 65 years of age or older. The results are as follows:

| Test Result | Alzheimer Disease | | |
	Present (D)	Absent ($\overline{D}$)	Total
Positive (T)	436	5	441
Negative ($\overline{T}$)	14	495	509
Total	450	500	950

Based on another independent study, it is known that the percentage of patients with Alzheimer's disease (the rate of prevalence of the disease) is 11.3% out of all subjects who were 65 years of age or older.

Solution:

Using these data we estimate the following quantities:

1. The sensitivity of the test:

$$P(T \mid D) = \frac{n(T \cap D)}{n(D)} = \frac{436}{450} = 0.9689$$

2. The specificity of the test:

$$P(\overline{T} \mid \overline{D}) = \frac{n(\overline{T} \cap \overline{D})}{n(\overline{D})} = \frac{495}{500} = 0.99$$

3. The probability of the disease in the general population, P(D): The rate of disease in the relevant general population, P(D), cannot be computed from the sample data given in the table. However, it is given that the percentage of patients with Alzheimer's disease is 11.3% out of all subjects who were 65 years of age or older. Therefore P(D) can be computed to be:

$$P(D) = \frac{11.3\%}{100\%} = 0.113$$

4. The predictive value positive of the test:

We wish to estimate the probability that a subject who is positive on the test has Alzheimer disease. We use the Bayes' formula of Equation (1):

$$P(D \mid T) = \frac{P(T \mid D)\, P(D)}{P(T \mid D)\, P(D) \; + \; P(T \mid \overline{D})\, P(\overline{D})} .$$

From the tabulated data we compute:

$$P(T \mid D) = \frac{436}{450} = 0.9689 \quad \text{(From part no. 1)}$$

$$P(T \mid \overline{D}) = \frac{n(T \cap \overline{D})}{n(\overline{D})} = \frac{5}{500} = 0.01$$

Substituting of these results into Equation (1), we get:

$$P(D \mid T) = \frac{(0.9689)\, P(D)}{(0.9689)\, P(D) \ + \ (0.01)\, P(\overline{D})}$$

$$= \frac{(0.9689)\,(0.113)}{(0.9689)\,(0.113) \ + \ (0.01)\,(1 - 0.113)} = 0.93$$

As we see, in this case, the predictive value positive of the test is very high.

5. The predictive value negative of the test:

We wish to estimate the probability that a subject who is negative on the test does not have Alzheimer disease. We use the Bayes' formula of Equation (2):

$$P(\overline{D} \mid \overline{T}) = \frac{P(\overline{T} \mid \overline{D})\, P(\overline{D})}{P(\overline{T} \mid \overline{D})\, P(\overline{D}) \ + \ P(\overline{T} \mid D)\, P(D)}$$

To compute $P(\overline{D} \mid \overline{T})$, we first compute the following probabilities:

$$P(\overline{T} \mid \overline{D}) = \frac{495}{500} = 0.99 \quad \text{(From part no. 2)}$$

$$P(\overline{D}) = 1 - P(D) = 1 - 0.113 = 0.887$$

$$P(\overline{T} \mid D) = \frac{n(\overline{T} \cap D)}{n(D)} = \frac{14}{450} = 0.0311$$

Substitution in Equation (2) gives:

$$P(\overline{D} \mid \overline{T}) = \frac{P(\overline{T} \mid \overline{D})\, P(\overline{D})}{P(\overline{T} \mid \overline{D})\, P(\overline{D}) \ + \ P(\overline{T} \mid D)\, P(D)}$$

$$= \frac{(0.99)(0.887)}{(0.99)(0.887) + (0.0311)(0.113)}$$

$$= 0.996$$

As we see, the predictive value negative is also very high.

CHAPTER 4: Probabilistic Features of Certain Data Distribution (Probability Distributions)

4.1 Introduction:

The concept of random variables is very important in Statistics. Some events can be defined using random variables.

There are two types of random variables:

Random variables $\begin{cases} Discrete\ \ Random\ \ Variables \\ Continuous\ \ Random\ \ Variables \end{cases}$

4.2 Probability Distributions of Discrete Random Variables:

Definition:

The probability distribution of a discrete random variable is a table, graph, formula, or other device used to specify all possible values of the random variable along with their respective probabilities.

Examples of discrete r v.'s

- The no. of patients visiting KKUH in a week.
- The no. of times a person had a cold in last year.

Example:

Consider the following discrete random variable.

X = The number of times a Saudi person had a cold in January 2010.

Suppose we are able to count the no. of Saudis which $X = x$:

x (no. of colds a Saudi person had in January 2010)	Frequency of x (no. of Saudi people who had a cold x times in January 2010)
0	10,000,000
1	3,000,000
2	2,000,000
3	1,000,000
Total	$N = 16,000,000$

Note that the possible values of the random variable X are:
$$x = 0, 1, 2, 3$$
Experiment: Selecting a person at random
Define the event:

$(X = 0)$ = The event that the selected person had no cold.
$(X = 1)$ = The event that the selected person had 1 cold.
$(X = 2)$ = The event that the selected person had 2 colds.
$(X = 3)$ = The event that the selected person had 3 colds.

In general:

$(X = x)$ =The event that the selected person had x colds.

For this experiment, there are $n(\Omega) = 16,000,000$ equally likely outcomes.

The number of elements of the event (X = x) is:

n(X=x) = no. of Saudi people who had a cold x times
in January 2010.
= frequency of x.

The probability of the event (X = x) is:

$$P(X = x) = \frac{n(X = x)}{n(\Omega)} = \frac{n(X = x)}{16000000} \quad , \text{ for x=0, 1, 2, 3}$$

x	freq. of x $n(X = x)$	$P(X = x) = \dfrac{n(X = x)}{16000000}$ (Relative frequency)
0	10000000	0.6250
1	3000000	0.1875
2	2000000	0.1250
3	1000000	0.0625
Total	16000000	1.0000

Note:

$$P(X = x) = \frac{n(X = x)}{16000000} = Relative\ Frequency = \frac{frequency}{16000000}$$

The probability distribution of the discrete random variable X is given by the following table:

x	$P(X=x)=f(x)$
0	0.6250
1	0.1874
2	0.1250
3	0.0625
Total	1.0000

Notes:

- The probability distribution of any discrete random variable X must satisfy the following two properties:

   (1) $0 \leq P(X=x) \leq 1$

   (2) $\sum_{x} P(X=x) = 1$

- Using the probability distribution of a discrete r.v. we can find the probability of any event expressed in term of the r.v. X.

Example:

Consider the discrete r.v. X in the previous example.

x	$P(X=x)$
0	0.6250
1	0.1875
2	0.1250
3	0.0625
Total	1.0000

(1) $P(X \geq 2) = P(X=2) + P(X=3) = 0.1250 + 0.0625 = 0.1875$

(2) $P(X > 2) = P(X=3) = 0.0625$ [note: $P(X > 2) \neq P(X \geq 2)$]

(3) $P(1 \leq X < 3) = P(X=1) + P(X=2) = 0.1875 + 0.1250 = 0.3125$

(4) $P(X \leq 2) = P(X=0) + P(X=1) + P(X=2)$
$= 0.6250 + 0.1875 + 0.1250 = 0.9375$

   another solution:

$P(X \leq 2) = 1 - P(\overline{(X \leq 2)})$

$= 1 - P(X > 2) = 1 - P(X=3) = 1 - 0.625 = 0.9375$

(5) $P(-1 \leq X < 2) = P(X=0) + P(X=1)$
$= 0.6250 + 0.1875 = 0.8125$

(6) $P(-1.5 \le X < 1.3) = P(X = 0) + P(X = 1)$
$$= 0.6250 + 0.1875 = 0.8125$$

(7) $P(X = 3.5) = P(\phi) = 0$

(8) $P(X \le 10) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) = P(\Omega) = 1$

(9) The probability that the selected person had <u>at least</u> 2 cold:
$$P(X \ge 2) = P(X = 2) + P(X = 3) = 0.1875$$

(10) The probability that the selected person had <u>at most</u> 2 colds:
$$P(X \le 2) = 0.9375$$

(11) The probability that the selected person had <u>more than</u> 2 colds:
$$P(X > 2) = P(X = 3) = 0.0625$$

(12) The probability that the selected person had <u>less than</u> 2 colds:
$$P(X < 2) = P(X = 0) + P(X = 1) = 0.8125$$

Graphical Presentation:

The probability distribution of a discrete r. v. X can be graphically represented.

Example:

The probability distribution of the random variable in the previous example is:

x	$P(X = x)$
0	0.6250
1	0.1875
2	0.1250
3	0.0625

The graphical presentation of this probability distribution is given by the following figure:
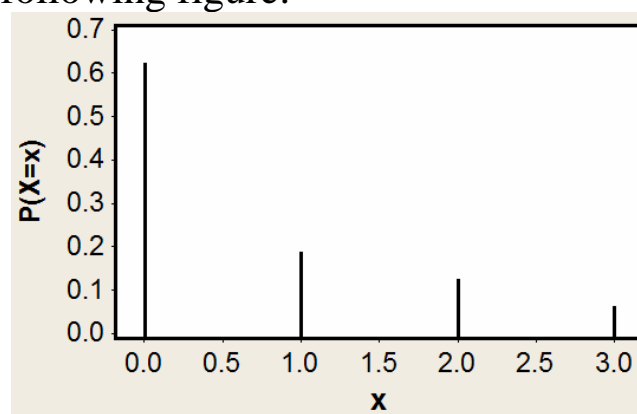
<u>Mean and Variance of a Discrete Random Variable</u>

Mean: The mean (or expected value) of a discrete random variable X is denoted by μ or μ_x. It is defined by:

$$\mu = \sum_x x\, P(X = x)$$

Variance: The variance of a discrete random variable X is denoted by σ^2 or σ_x^2. It is defined by:

$$\sigma^2 = \sum_x (x - \mu)^2\, P(X = x)$$

Example:

We wish to calculate the mean μ and the variance of the discrete r. v. X whose probability distribution is given by the following table:

x	$P(X = x)$
0	0.05
1	0.25
2	0.45
3	0.25

Solution:

x	$P(X = x)$	$x P(X = x)$	$(x - \mu)$	$(x - \mu)^2$	$(x - \mu)^2 P(X = x)$
0	0.05	0	-1.9	3.61	0.1805
1	0.25	0.25	-0.9	0.81	0.2025
2	0.45	0.9	0.1	0.01	0.0045
3	0.25	0.75	1.1	1.21	0.3025
Total		$\mu = $ $\sum x\, P(X = x)$ $= 1.9$			$\sigma^2 = $ $\sum (x - \mu)^2 P(X = x)$ $= 0.69$

$$\mu = \sum_x x\, P(X = x) = (0)(0.05) + (1)(0.25) + (2)(0.45) + (3)(0.25) = 1.9$$

$$\sigma^2 = \sum_x (x - 1.9)^2\, P(X = x)$$

$$= (0 - 1.9)^2 (0.05) + (1 - 1.9)^2 (0.25) + (2 - 1.9)^2 (0.45) + (3 - 1.9)^2 (0.25)$$

$$= 0.69$$

Cumulative Distributions:

The cumulative distribution function of a discrete r. v. X is defined by:

$$P(X \leq x) = \sum_{a \leq x} P(X = a)$$ (Sum over all values ≤x)

Example:

Calculate the cumulative distribution of the discrete r. v. X whose probability distribution is given by the following table:

x	$P(X = x)$
0	0.05
1	0.25
2	0.45
3	0.25

Use the cumulative distribution to find:

P(X≤2), P(X<2), P(X≤1.5), P(X<1.5), P(X>1), P(X≥1)

Solution:

The cumulative distribution of X is:

x	$P(X \leq x)$
0	0.05
1	0.30
2	0.75
3	1.0000

$P(X \leq 0) = P(X = 0)$
$P(X \leq 1) = P(X = 0) + P(X = 1)$
$P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2)$
$P(X \leq 3) = P(X = 0) + \cdots + P(X = 3)$

Using the cumulative distribution,

P(X≤2) = 0.75
P(X<2) = P(X≤1) = 0.30
P(X≤1.5) = P(X≤1) = 0.30
P(X<1.5) = P(X≤1) = 0.30
P(X>1) = 1- P($\overline{(X > 1)}$) = 1-P(X≤1) = 1- 0.30 = 0.70
P(X≥1) = 1- P($\overline{(X \geq 1)}$) = 1-P(X<1) = 1- P(X≤0)
              = 1- 0.05 = 0.95

Example: (Reading Assignment)

Given the following probability distribution of a discrete random variable X representing the number of defective teeth of the patient visiting a

certain dental clinic:

x	P(X = x)
1	0.25
2	0.35
3	0.20
4	0.15
5	K

a) Find the value of K.
b) Find the flowing probabilities:
   1. $P(X < 3)$
   2. $P(X \le 3)$
   3. $P(X < 6)$
   4. $P(X < 1)$
   5. $P(X = 3.5)$
c) Find the probability that the patient has at least 4 defective teeth.
d) Find the probability that the patient has at most 2 defective teeth.
e) Find the expected number of defective teeth (mean of X).
f) Find the variance of X.

Solution:

a) $1 = \sum P(X = x) = 0.25 + 0.35 + 0.20 + 0.15 + K$

      $1 = 0.95 + K$

      $K = 1 - 0.95$

      $K = 0.05$

The probability distribution of X is:

x	P(X = x)
1	0.25
2	0.35
3	0.20
4	0.15
5	0.05
Total	1.00

b) Finding the probabilities:
   1. $P(X < 3) = P(X=1)+P(X=2) = 0.25+0.35 = 0.60$
   2. $P(X \le 3) = P(X=1)+P(X=2)+P(X=3) = 0.8$
   3. $P(X < 6) = P(X=1)+P(X=2)+P(X=3)+P(X=4)+P(X=5) = P(\Omega)=1$
   4. $P(X < 1) = P(\phi)=0$
   5. $P(X = 3.5) = P(\phi)=0$

c) The probability that the patient has at least 4 defective teeth
     $P(X \ge 4) = P(X=4)+P(X=5) = 0.15+0.05=0.2$

d) The probability that the patient has at most 2 defective teeth
     $P(X \le 2) = P(X=1)+P(X=2) = 0.25+0.35=0.6$

e) The expected number of defective teeth (mean of X)

x	P(X = x)	x P(X = x)
1	0.25	0.25
2	0.35	0.70
3	0.20	0.60
4	0.15	0.60
5	0.05	0.25
Total	$\sum P(X = x) = 1$	$\mu = \sum x P(X = x) = 2.4$

The expected number of defective teeth (mean of X) is

$\mu = \sum x P(X = x) = (1)(0.25) + (2)(0.35) + (3)(0.2) + (4)(0.15) + (5)(0.05) = 2.4$

f) The variance of X:

x	$P(X = x)$	$(x - \mu)$	$(x - \mu)^2$	$(x - \mu)^2 P(X = x)$
1	0.25	-1.4	1.96	0.49
2	0.35	-0.4	0.16	0.056
3	0.20	0.6	0.36	0.072
4	0.15	1.6	2.56	0.384
5	0.05	2.6	6.76	0.338
Total				$\sigma^2 = \sum (x - \mu)^2 P(X = x) = 1.34$

The variance is $\sigma^2 = \sum (x - \mu)^2 P(X = x) = 1.34$

Combinations:

Notation (n!):

$n!$ is read "n factorial". It defined by:

$$n! = n(n-1)(n-2)\cdots(2)(1) \qquad for \quad n \geq 1$$

$$0! = 1$$

Example: $5! = (5)(4)(3)(2)(1) = 120$

Combinations:

The number of different ways for selecting r objects from n distinct objects is denoted by $_nC_r$ or $\binom{n}{r}$ and is given by:

$$_nC_r = \frac{n!}{r!\,(n-r)!}; \qquad for \quad r = 0, 1, 2, \ldots, n$$
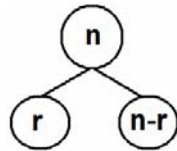
Notes:

1. $_nC_r$ is read as " n " choose " r ".

2. $_nC_n = 1$, $_nC_0 = 1$,

3. $_nC_r = \ _nC_{n-r}$ (for example: $_{10}C_3 = \ _{10}C_7$)

4. $_nC_r$ = number of unordered subsets of a set of (n) objects such that each subset contains (r) objects.

Example:

For n = 4 and r = 2:

$$_4C_2 = \ = \frac{4!}{2!\,(4-2)!} = \frac{4!}{2! \times 2!} = \frac{4 \times 3 \times 2 \times 1}{(2 \times 1) \times (2 \times 1)} = 6$$

$_4C_2 = \ 6$ = The number of different ways for selecting 2 objects from 4 distinct objects.

Example:

Suppose that we have the set {a, b, c, d} of (n=4) objects.
We wish to choose a subset of two objects. The possible subsets of this set with 2 elements in each subset are:

{a , b}, {a , c}, {a , d}, {b , d}, {b , c}, {c , d}

The number of these subsets is $_4C_2 = 6$.

4.3 Binomial Distribution:

- **Bernoulli Trial**: is an experiment with only two possible outcomes: S = success and F= failure (Boy or girl, Saudi or non-Saudi, sick or well, dead or alive).
- Binomial distribution is a discrete distribution.
- Binomial distribution is used to model an experiment for which:
  1. The experiment has a sequence of n Bernoulli trials.
  2. The probability of success is $P(S) = p$, and the probability of failure is $P(F) = 1 - p = q$.
  3. The probability of success $P(S) = p$ is constant for each trial.
  4. The trials are independent; that is the outcome of one trial has no effect on the outcome of any other trial.

In this type of experiment, we are interested in the discrete r. v. representing the number of successes in the n trials.

X = The number of successes in the n trials

The possible values of X (number of success in n trails) are:
$$x = 0, 1, 3, \ldots, n$$

The r.v. X has a binomial distribution with parameters n and p, and we write:
$$X \sim Binomial(n, p)$$

The probability distribution of X is given by:
$$P(X = x) = \begin{cases} {}_nC_x \; p^x \; q^{n-x} & for \; x = 0, 1, 2, \ldots, n \\ 0 & otherwise \end{cases}$$

Where: $${}_nC_x = \frac{n!}{x! \, (n-x)!}$$

We can write the probability distribution of X as a table as follows.

x	$P(X = x)$
0	${}_nC_0 \; p^0 \; q^{n-0} = q^n$
1	${}_nC_1 \; p^1 \; q^{n-1}$

x	$P(X = x)$
2	$_nC_2 \ p^2 \ q^{n-2}$
$\vdots$	$\vdots$
$n - 1$	$_nC_{n-1} \ p^{n-1} \ q^1$
n	$_nC_n \ p^n \ q^0 = p^n$
Total	1.00

Result: (Mean and Variance for normal distribution)
If $X \sim$ Binomial(n, p) , then

- The mean: $\mu = np$ (expected value)
- The variance: $\sigma^2 = npq$

Example:

Suppose that the probability that a Saudi man has high blood pressure is 0.15. Suppose that we randomly select a sample of 6 Saudi men.

(1) Find the probability distribution of the random variable (X) representing the number of men with high blood pressure in the sample.

(2) Find the expected number of men with high blood pressure in the sample (mean of X).

(3) Find the variance X.

(4) What is the probability that there will be exactly 2 men with high blood pressure?

(5) What is the probability that there will be at most 2 men with high blood pressure?

(6) What is the probability that there will be at lease 4 men with high blood pressure?

Solution:

We are interested in the following random variable:

X = The number of men with high blood pressure in the sample of 6 men.

Notes:

- Bernoulli trial: diagnosing whether a man has a high blood pressure or not. There are two outcomes for each trial:

S = Success: The man has high blood pressure
F = failure: The man does not have high blood pressure.
- Number of trials = 6 (we need to check 6 men)
- Probability of success: $P(S) = p = 0.15$
- Probability of failure: $P(F) = q = 1 - p = 0.85$
- Number of trials: $n = 6$
- The trials are independent because of the fact that the result of each man does not affect the result of any other man since the selection was made ate random.

The random variable X has a binomial distribution with parameters: n=6 and p=0.15, that is:
$$X \sim \text{Binomial } (n, p)$$
$$X \sim \text{Binomial } (6, 0.15)$$

The possible values of X are:
$$x = 0, 1, 3, 4, 5, 6$$

(1) The probability distribution of X is:

$$P(X = x) = \begin{cases} {}_6C_x \ (0.15)^x (0.85)^{6-x} & ; x = 0,1,2,3,4,5,6 \\ 0 & ; \quad otherwise \end{cases}$$

The probabilities of all values of X are:
$$P(X = 0) = {}_6C_0 \ (0.15)^0 (0.85)^6 = (1)(0.15)^0 (0.85)^6 = 0.37715$$
$$P(X = 1) = {}_6C_1 \ (0.15)^1 (0.85)^5 = (6)(0.15)(0.85)^5 = 0.39933$$
$$P(X = 2) = {}_6C_2 \ (0.15)^2 (0.85)^4 = (15)(0.15)^2 (0.85)^4 = 0.17618$$
$$P(X = 3) = {}_6C_3 \ (0.15)^3 (0.85)^3 = (20)(0.15)^3 (0.85)^3 = 0.04145$$
$$P(X = 4) = {}_6C_4 \ (0.15)^4 (0.85)^2 = (15)(0.15)^4 (0.85)^2 = 0.00549$$
$$P(X = 5) = {}_6C_5 \ (0.15)^5 (0.85)^1 = (6)(0.15)^5 (0.85)^1 = 0.00039$$
$$P(X = 6) = {}_6C_6 \ (0.15)^6 (0.85)^0 = (1)(0.15)^6 (1) = 0.00001$$

The probability distribution of X can by presented by the following table:

x	$P(X = x)$
0	0.37715
1	0.39933
2	0.17618
3	0.04145
4	0.00549
5	0.00039
6	0.00001

The probability distribution of X can by presented by the following graph:
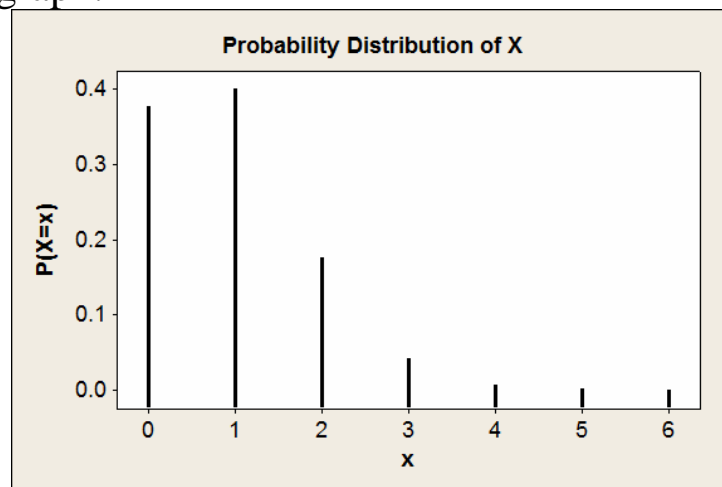
(2) The mean of the distribution (the expected number of men out of 6 with high blood pressure) is:
$$\mu = np = (6)(0.15) = 0.9$$

(3) The variance is:
$$\sigma^2 = npq = (6)(0.15)(0.85) = 0.765$$

(4) The probability that there will be exactly 2 men with high blood pressure is:
$$P(X = 2) = 0.17618$$

(5) The probability that there will be at most 2 men with high blood pressure is:
$$P(X \leq 2) = P(X=0) + P(X=1) + P(X=2)$$
$$= 0.37715 + 0.39933 + 0.17618$$
$$= 0.95266$$

(6) The probability that there will be at lease 4 men with high blood pressure is:

$$P(X \geq 4) = P(X=4) + P(X=5) + P(X=6)$$
$$= 0.00549 + 0.00039 + 0.00001$$
$$= 0.00589$$

Example: (Reading Assignment)

Suppose that 25% of the people in a certain population have low hemoglobin levels. The experiment is to choose 5 people at random from this population. Let the discrete random variable X be the number of people out of 5 with low hemoglobin levels.

  a) Find the probability distribution of X.
  b) Find the probability that at least 2 people have low hemoglobin levels.
  c) Find the probability that at most 3 people have low hemoglobin levels.
  d) Find the expected number of people with low hemoglobin levels out of the 5 people.
  e) Find the variance of the number of people with low hemoglobin levels out of the 5 people.

Solution:

X = the number of people out of 5 with low hemoglobin levels

The Bernoulli trail is the process of diagnosing the person

   Success = the person has low hemoglobin

   Failure = the person does not have low hemoglobin

$n = 5$ (no. of trials)

$p = 0.25$ (probability of success)

$q = 1 - p = 0.75$ (probability of failure)

a) X has a binomial distribution with parameter $n = 5$ and $p = 0.25$

$$X \sim Binomial(n, p)$$
$$X \sim Binomial(5, 0.25)$$

The possible values of X are:

$$x = 0, 1, 2, 3, 4, 5$$

The probability distribution is:

$$P(X = x) = \begin{cases} {}_nC_x \; p^x \; q^{n-x} ; & for \; x = 0, 1, 2, \ldots, n \\ 0 & ; \quad otherwise \end{cases}$$

$$P(X = x) = \begin{cases} {}_5C_x \; (0.25)^x \; (0.75)^{5-x} ; & for \; x = 0, 1, 2, 3, 4, 5 \\ 0 & ; \quad otherwise \end{cases}$$

x	P(X = x)
0	${}_5C_0 \times 0.25^0 \times 0.75^{5-0}$ $= 0.23730$

x	P(X = x)	
1	$_5C_1 \times 0.25^1 \times 0.75^{5-1}$	$= 0.39551$
2	$_5C_2 \times 0.25^2 \times 0.75^{5-2}$	$= 0.26367$
3	$_5C_3 \times 0.25^3 \times 0.75^{5-3}$	$= 0.08789$
4	$_5C_4 \times 0.25^4 \times 0.75^{5-4}$	$= 0.01465$
5	$_5C_5 \times 0.25^5 \times 0.75^{5-5}$	$= 0.00098$
Total	$\sum P(X = x) = 1$	

b) The probability that at least 2 people have low hemoglobin levels:

P(X≥2) = P(X=2)+P(X=3)+P(X=4)+P(X=5)

= 0.26367+ 0.08789+ 0.01465+ 0.00098

= 0. 0.36719

c) The probability that at most 3 people have low hemoglobin levels:

P(X≤3) = P(X=0)+P(X=1)+P(X=2)+P(X=3)

= 0.23730+ 0.39551+ 0.26367+ 0.08789

= 0.98437

d) The expected number of people with low hemoglobin levels out of the 5 people (the mean of X):

$$\mu = n\,p = 5 \times 0.25 = 1.25$$

e) The variance of the number of people with low hemoglobin levels out of the 5 people (the variance of X) is:

$$\sigma^2 = n\,pq = 5 \times 0.25 \times 0.75 = 0.9375$$

4.4 The Poisson Distribution:

- It is a discrete distribution.
- The Poisson distribution is used to model a discrete r. v. representing the number of occurrences of some random event in an interval of time or space (or some volume of matter).
- The possible values of X are:

  x= 0, 1, 2, 3, …

- The discrete r. v. *X* is said to have a Poisson distribution with parameter (average or mean) λ if the probability distribution of *X* is given by

$$P(X = x) = \begin{cases} \dfrac{e^{-\lambda} \lambda^x}{x!} & ; \quad for \quad x = 0, 1, 2, 3, \ldots \\ \\ 0 & ; \quad otherwise \end{cases}$$

where　e = 2.71828　(the natural number).

We write :

$$X \sim \text{Poisson } (\lambda)$$

Result: (Mean and Variance of Poisson distribution)

If $X \sim$ Poisson (λ), then:

- The mean (average) of X is : $\mu = \lambda$　　(Expected value)
- The variance of X is: $\sigma^2 = \lambda$

Example:

Some random quantities that can be modeled by Poisson distribution:

- No. of patients in a waiting room in an hours.
- No. of surgeries performed in a month.
- No. of rats in each house in a particular city.

Note:

- λ is the average (mean) of the distribution.
- If X = The number of patients seen in the emergency unit in a day, and if $X \sim$ Poisson (λ), then:
  1. The average (mean) of patients seen every day in the emergency unit = λ.
  2. The average (mean) of patients seen every month in the emergency unit = 30λ.
  3. The average (mean) of patients seen every year in the emergency unit = 365λ.
  4. The average (mean) of patients seen every hour in the emergency unit = $\lambda/24$.

Also, notice that:

(i) If $Y =$ The number of patients seen every month, then:

$Y \sim$ Poisson (λ^{*}), where $\lambda^{*}=30\lambda$

(ii) $W =$ The number of patients seen every year, then:

$W \sim$ Poisson (λ^{*}), where $\lambda^{*}=365\lambda$

(iii) $V =$ The number of patients seen every hour, then:

$V \sim$ Poisson (λ^{*}), where $\lambda^{*}=\dfrac{\lambda}{24}$

Example:

Suppose that the number of snake bites cases seen at KKUH in a year has a Poisson distribution with average 6 bite cases.

(1) What is the probability that in a year:

(i) The no. of snake bite cases will be 7?

(ii) The no. of snake bite cases will be less than 2?

(2) What is the probability that there will be 10 snake bite cases in 2 years?

(3) What is the probability that there will be no snake bite cases in a month?

Solution:

(1) $X =$ no. of snake bite cases in a year.

$$X \sim \text{Poisson (6)} \qquad (\lambda=6)$$

$$P(X = x)=\frac{e^{-6}6^{x}}{x!} \ ; \ \ x = 0, 1, 2, \ldots$$

(i) $P(X = 7)=\dfrac{e^{-6}6^{7}}{7!} = 0.13768$

(ii) $P(X < 2)= P(X = 0)+ P(X = 1)$

$$= \frac{e^{-6}6^{0}}{0!} + \frac{e^{-6}6^{1}}{1!} = 0.00248 + 0.01487 = 0.01735$$

(2) $Y =$ no of snake bite cases in 2 years

$$Y \sim \text{Poisson}(12) \qquad (\lambda^{*} = 2\lambda = (2)(6)=12)$$

$$P(Y = y)= \frac{e^{-12}12^{y}}{y!} : \qquad y = 0 , 1 , 2\ldots$$

$$\therefore P(Y = 10)= \frac{e^{-12}12^{10}}{10!} = 0.1048$$

(3) $W =$ no. of snake bite cases in a month.

$$W \sim \text{Poisson (0.5)} \qquad\qquad (\lambda^{**} = \frac{\lambda}{12} = \frac{6}{12} = 0.5)$$

$$P(W = w) = \frac{e^{-0.5} \ 0.5^w}{w!} : \quad w = 0, 1, 2\ldots$$

$$P(W = 0) = \frac{e^{-0.5}(0.5)^0}{0!} = 0.6065$$

4.5 Continuous Probability Distributions:

For any continuous r. v. X, there exists a function $f(x)$, called the probability density function (pdf) of X, for which:

(1) The total area under the curve of $f(x)$ equals to 1.
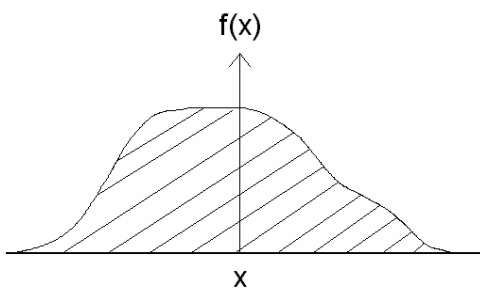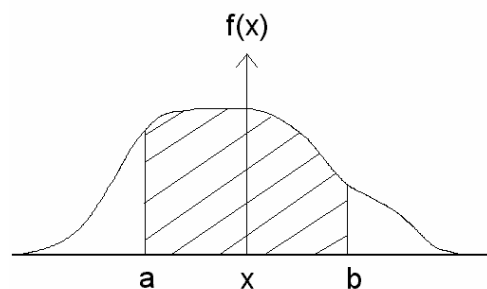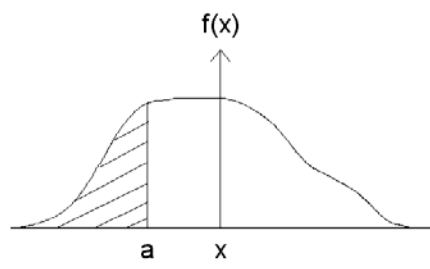

$$Total\ area = \int_{-\infty}^{\infty} f(x)\, dx = 1 \qquad\qquad P(a \leq X \leq b) = \int_{a}^{b} f(x)dx = area$$

(2) The probability hat X is between the points (a) and (b) equals to the area under the curve of f(x) which is bounded by the point a and b.

(3) In general, the probability of an interval event is given by the area under the curve of $f(x)$ and above that interval.
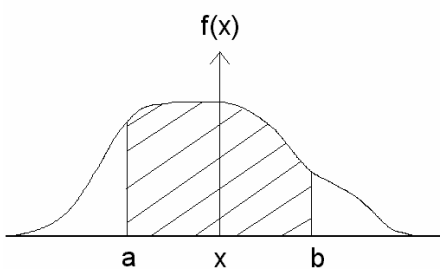
$$P(a \leq X \leq b) = \int_{a}^{b} f(x)dx = area \qquad P(X \leq a) = \int_{-\infty}^{a} f(x)dx = area \qquad P(X \geq b) = \int_{b}^{\infty} f(x)\, dx = area$$
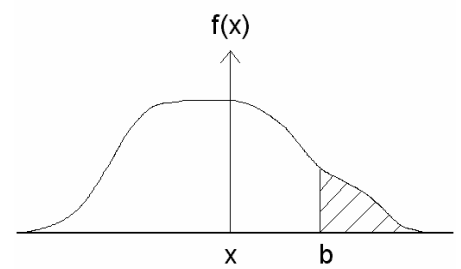
Note:

If X is continuous r.v. then:

    1. $P(X = a) = 0$ for any a.

    2. $P(X \leq a) = P(X < a)$

3. $P(X \geq b) = P(X > b)$
4. $P(a \leq X \leq b) = P(a \leq X < b) = P(a < X \leq b) = P(a < X < b)$
5. $P(X \leq x) =$ cumulative probability
6. $P(X \geq a) = 1 - P(X < a) = 1 - P(X \leq a)$
7. $P(a \leq X \leq b) = P(X \leq b) - P(X \leq a)$
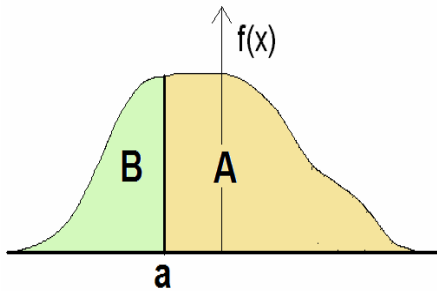
$P(X \geq a) = 1 - P(X \leq a)$
$A = 1 - B$
Total area $= 1$
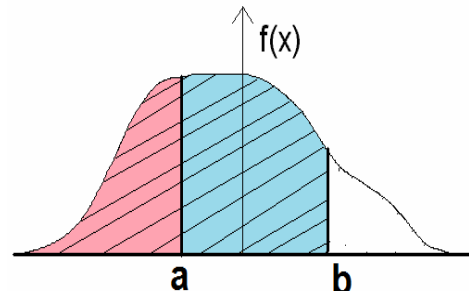
$P(a \leq X \leq b) = P(X \leq b) - P(X \leq a)$
$$\int_a^b f(x)dx = \int_{-\infty}^b f(x)dx - \int_{-\infty}^a f(x)dx$$

4.6 The Normal Distribution:

■ One of the most important continuous distributions.
■ Many measurable characteristics are normally or approximately normally distributed.
 (Examples: height, weight, …)
■ The probability density function of the normal distribution is given by:

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

where (e=2.71828) and (π=3.14159).
The parameters of the distribution are the mean (μ) and the standard deviation (σ).
■ The continuous r.v. X which has a normal distribution has several important characteristics:
   1. $-\infty < X < \infty$,
   2. The density function of X, $f(x)$, has a bell-Shaped curve:
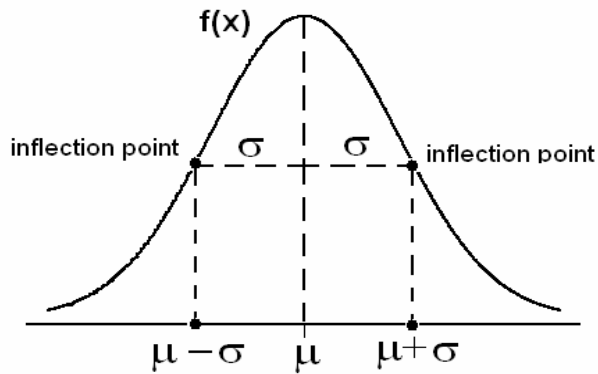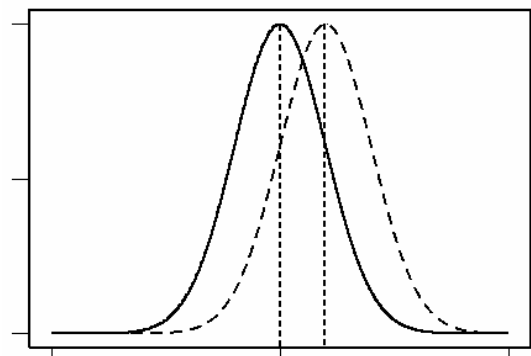
mean $= \mu$

standard deviation $= \sigma$

variance $= \sigma^2$

3. The highest point of the curve of f(x) at the mean μ.
   (Mode $= \mu$)

4. The curve of f(x) is symmetric about the mean μ.

   μ = mean = mode = median

5. The normal distribution depends on two parameters:

   mean $= \mu$ (determines the location)

   standard deviation $= \sigma$ (determines the shape)

6. If the r.v. X is normally distributed with mean μ and

   standard deviation σ (variance σ^2), we write:

   $X \sim \text{Normal}(\mu, \sigma^2)$ or $X \sim N(\mu, \sigma^2)$

7. The location of the normal distribution depends on μ. The
   shape of the normal distribution depends on σ.

Note: The location of the normal distribution depends on μ and its shape depends on σ.
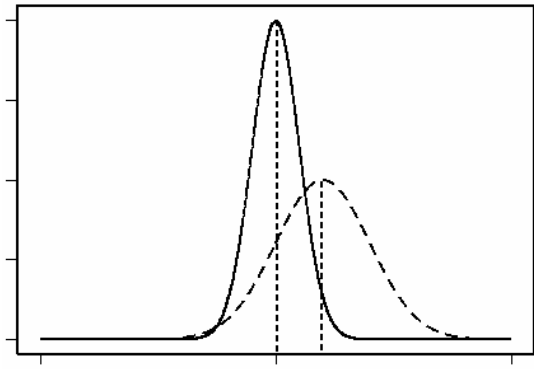Suppose we have two normal distributions:

_____ $N(\mu_1, \sigma_1)$

------------ $N(\mu_2, \sigma_2)$

$\mu_1 < \mu_2, \sigma_1 = \sigma_2$
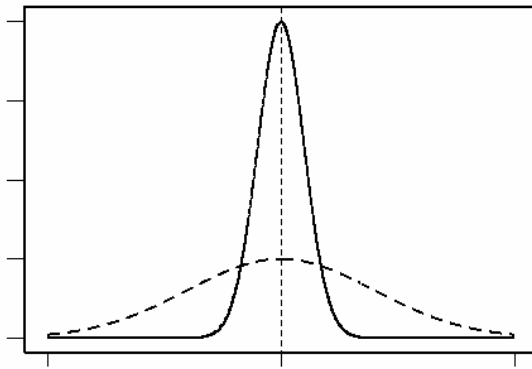
$\mu_1 = \mu_2,\ \sigma_1 < \sigma_2$ $\mu_1 < \mu_2,\ \sigma_1 < \sigma_2$
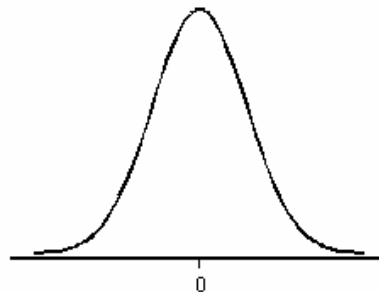
The Standard Normal Distribution:

The normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$ is called the standard normal distribution and is denoted by Normal (0,1) or N(0,1). The standard normal random variable is denoted by (Z), and we write:
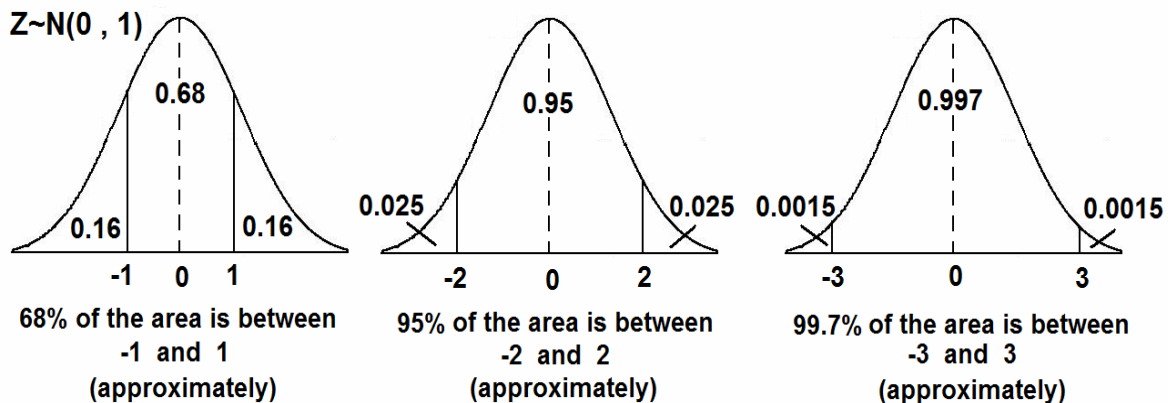
$$Z \sim N(0, 1)$$

The probability density function (pdf) of Z~N(0,1) is given by:

$$f(z) = n(z;0,1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

The standard normal distribution, Normal (0,1), is very important because probabilities of any normal distribution can be calculated from the probabilities of the standard normal distribution.

Z~N(0 , 1)

68% of the area is between -1 and 1 (approximately)

95% of the area is between -2 and 2 (approximately)

99.7% of the area is between -3 and 3 (approximately)

<u>Result:</u>

If $X \sim \text{Normal}\left(\mu, \sigma^2\right)$,   then $Z = \dfrac{X - \mu}{\sigma} \sim \text{Normal }(0,1)$.
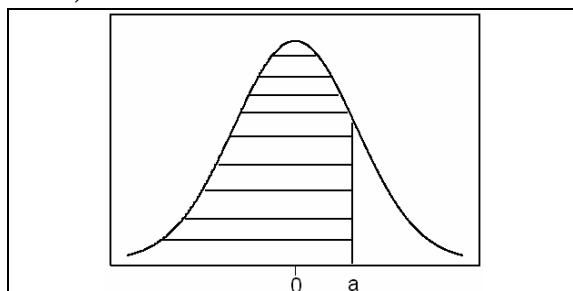
<u>Calculating Probabilities of Normal (0,1):</u>

Suppose $Z \sim \text{Normal }(0,1)$.

For the standard normal distribution $Z \sim N(0,1)$, there is a special table used to calculate probabilities of the form:
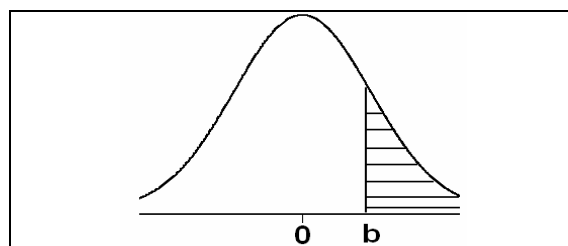
$$P(Z \leq a)$$

(i) $P(Z \leq a) = $ From the table

(ii) $P(Z \geq b) = 1 - P(Z \leq b)$

Where:

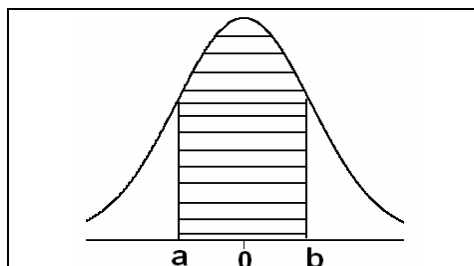$P(Z \leq b) = $ From the table

(iii) $P(a \leq Z \leq b) = P(Z \leq b) - P(z \leq a)$

Where:

$P(Z \leq b) = $ from the table
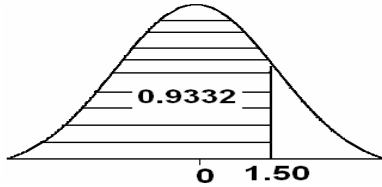
$P(z \leq a) = $ from the table

(iv) $P(Z = a) = 0$ for every a.

Example:

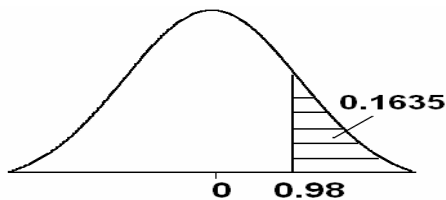Suppose that $Z \sim N(0,1)$

(1) $P(Z \leq 1.50) = 0.9332$

Z	0.00	0.01	...
:	⇓		
1.50 ⇒	0.9332		
:			

(2)

$P(Z \geq 0.98) = 1 - P(Z \leq 0.98)$

$= 1 - 0.8365$

$= 0.1635$

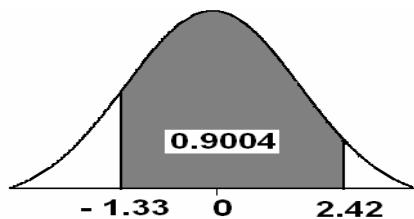| Z	0.00	...	0.08
:	:	:	⇓
:	...	...	⇓
0.90⇒	⇒	⇒	0.8365

(3)

$P(-1.33 \leq Z \leq 2.42) =$

$P(Z \leq 2.42) - P(Z \leq -1.33)$

$= 0.9922 - 0.0918$

$= 0.9004$

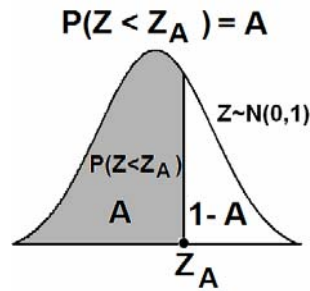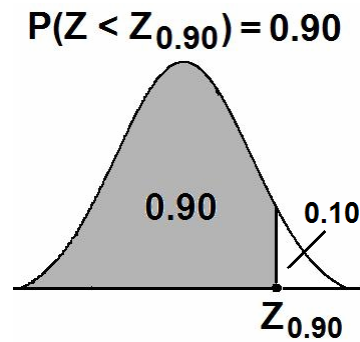| Z	...		−0.03
:	:		⇓
−1.30	⇒		0.0918
:			

(4) $P(Z \leq 0) = P(Z \geq 0) = 0.5$

Notation:

$P(Z \leq Z_A) = A$

$$P(Z < Z_A) = A$$

For example:

$$P(Z < Z_{0.025}) = 0.025$$

$$P(Z < Z_{0.90}) = 0.90$$
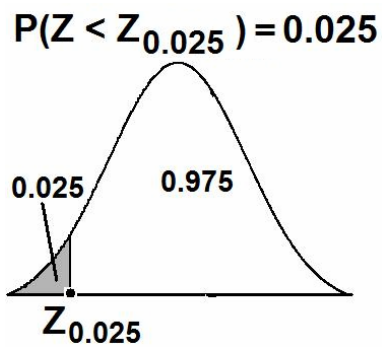
Result:

Since the pdf of Z~N(0,1) is symmetric about 0, we have:

$$Z_A = - Z_{1-A}$$

For example: $Z_{0.35} = - Z_{1-0.35} = - Z_{0.65}$

$Z_{0.86} = - Z_{1-0.86} = - Z_{0.14}$

Example:

Suppose that $Z \sim N(0,1)$.

If $P(Z \le a) = 0.9505$

Then $a = 1.65$

Z	…	0.05	…
:		⇑	
1.60	⇐	0.9505	
:			

$P(Z < a) = 0.9505$

$P(Z < Z_{0.9505}) = 0.9505$
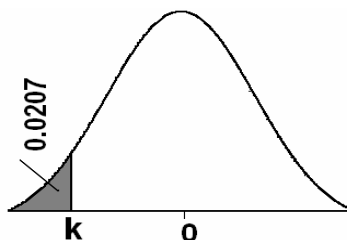
$a = Z_{0.9505}$

$a = Z_{0.9505} = 1.65$
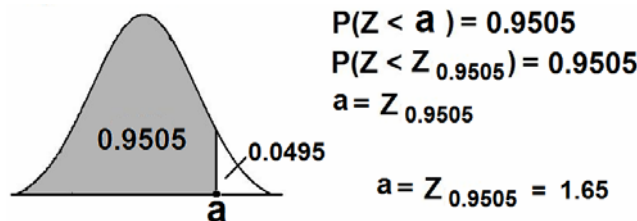
Example:

Suppose that Z~N(0,1). Find the value of k such that P(Z≤k)= 0.0207.

Solution:

$.k = -2.04$

Notice that $k = Z_{0.0207} = -2.04$

Z	…	−0.04	
⋮	⋮	⇑	
		⇑	
−2.0	⇐⇐	0.0207	
⋮			

Example:

If Z ~ N(0,1), then:

$Z_{0.90} = 1.285$

$Z_{0.95} = 1.645$

$Z_{0.975} = 1.96$

$Z_{0.99} = 2.325$
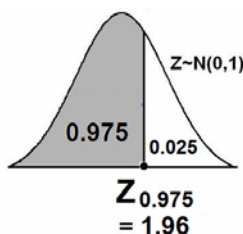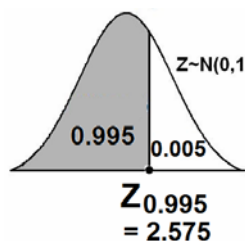
Using the result: $Z_A = - Z_{1-A}$

$Z_{0.10} = - Z_{0.90} = - 1.285$

$Z_{0.05} = - Z_{0.95} = - 1.645$

$Z_{0.025} = - Z_{0.975} = - 1.96$

$Z_{0.01} = - Z_{0.99} = - 2.325$

Calculating Probabilities of Normal (μ, σ^2):

- Recall the result:

$$X \sim \text{Normal}\,(\mu, \sigma^2) \quad \Leftrightarrow \quad Z = \frac{X - \mu}{\sigma} \sim \text{Normal}\,(0,1)$$

- $X \leq a \iff \dfrac{X - \mu}{\sigma} \leq \dfrac{a - \mu}{\sigma} \iff Z \leq \dfrac{a - \mu}{\sigma}$

1. $P(X \leq a) = P\left(Z \leq \dfrac{a - \mu}{\sigma}\right) = $ From the table.

2. $P(X \geq a) = 1 - P(X \leq a) = 1 - P\left(Z \leq \dfrac{a - \mu}{\sigma}\right)$

3. $P(a \leq X \leq b) = P(X \leq b) - P(X \leq a)$

$$= P\left(Z \leq \dfrac{b - \mu}{\sigma}\right) - P\left(Z \leq \dfrac{a - \mu}{\sigma}\right)$$

4. $P(X = a) = 0$, for every a.

4.7 Normal Distribution Application:

Example

Suppose that the hemoglobin levels of healthy adult males are approximately normally distributed with a mean of 16 and a variance of 0.81.

(a) Find that probability that a randomly chosen healthy adult male has a hemoglobin level less than 14.

(b) What is the percentage of healthy adult males who have hemoglobin level less than 14?

(c) In a population of 10,000 healthy adult males, how many would you expect to have hemoglobin level less than 14?

Solution:

  $X = $ hemoglobin level for healthy adults males

  Mean: $\mu = 16$
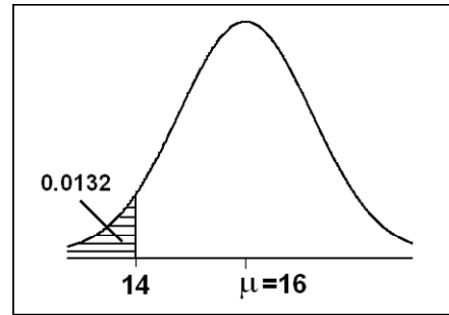
  Variance: $\sigma^2 = 0.81$

  Standard deviation: $\sigma = 0.9$

  $X \sim$ Normal (16, 0.81)

(a) The probability that a randomly chosen healthy adult male has hemoglobin level less than 14 is $P(X \leq 14)$.

$$P(X \leq 14) = P\left(Z \leq \frac{14 - \mu}{\sigma}\right)$$

$$= P\left(Z \leq \frac{14 - 16}{0.9}\right)$$

$$= P(Z \leq -2.22)$$

$$= 0.0132$$

(b) The percentage of healthy adult males who have hemoglobin level less than 14 is:

$$P(X \leq 14) \times 100\% = 0.0132 \times 100\% = 1.32\%$$

(c) In a population of 10000 healthy adult males, we would expect that the number of males with hemoglobin level less than 14 to be:

$$P(X \leq 14) \times 10000 = 0.0132 \times 10000 = 132 \ males$$

Example:
Suppose that the birth weight of Saudi babies has a normal distribution with mean μ=3.4 and standard deviation σ=0.35.
(a) Find the probability that a randomly chosen Saudi baby has a birth weight between 3.0 and 4.0 kg.
(b) What is the percentage of Saudi babies who have a birth weight between 3.0 and 4.0 kg?
(c) In a population of 100000 Saudi babies, how many would you expect to have birth weight between 3.0 and 4.0 kg?

Solution:
X = birth weight of Saudi babies
Mean: μ = 3.4
Standard deviation: σ = 0.35
Variance: $\sigma^2 = (0.35)^2 = 0.1225$
X ~ Normal (3.4, 0.1225)
(a) The probability that a randomly chosen Saudi baby has a birth weight between 3.0 and 4.0 kg is $P(3.0 < X < 4.0)$

$$P(3.0 < X < 4.0) = P(X \leq 4.0) - P(X \leq 3.0)$$

$$= P\left(Z \leq \frac{4.0 - \mu}{\sigma}\right) - P\left(Z \leq \frac{3.0 - \mu}{\sigma}\right)$$

$$= P\left(Z \leq \frac{4.0 - 3.4}{0.35}\right) - P\left(Z \leq \frac{3.0 - 3.4}{0.35}\right)$$

$$= P(Z \leq 1.71) - P(Z \leq -1.14)$$

$$= 0.9564 - 0.1271 = 0.8293$$
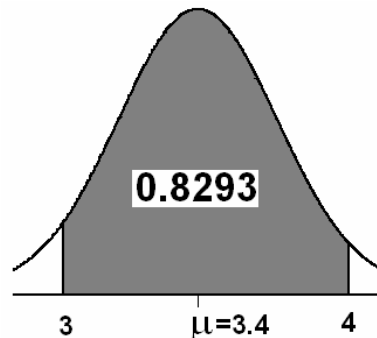
(b) The percentage of Saudi babies who have a birth weight between 3.0 and 4.0 kg is

P(3.0<X<4.0) × 100% = 0.8293× 100% = 82.93%

(c) In a population of 100,000 Saudi babies, we would expect that the number of babies with birth weight between 3.0 and 4.0 kg to be:

P(3.0<X<4.0) × 100000 = 0.8293× 100000 = 82930 babies

Standard Normal Table
Areas Under the Standard Normal Curve

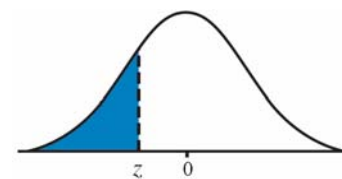| z	-0.09	-0.08	-0.07	-0.06	-0.05	-0.04	-0.03	-0.02	-0.01	-0.00	z
-3.50	0.00017	0.00017	0.00018	0.00019	0.00019	0.00020	0.00021	0.00022	0.00022	0.00023	**-3.50**
-3.40	0.00024	0.00025	0.00026	0.00027	0.00028	0.00029	0.00030	0.00031	0.00032	0.00034	**-3.40**
-3.30	0.00035	0.00036	0.00038	0.00039	0.00040	0.00042	0.00043	0.00045	0.00047	0.00048	**-3.30**
-3.20	0.00050	0.00052	0.00054	0.00056	0.00058	0.00060	0.00062	0.00064	0.00066	0.00069	**-3.20**
-3.10	0.00071	0.00074	0.00076	0.00079	0.00082	0.00084	0.00087	0.00090	0.00094	0.00097	**-3.10**
-3.00	0.00100	0.00104	0.00107	0.00111	0.00114	0.00118	0.00122	0.00126	0.00131	0.00135	**-3.00**
-2.90	0.00139	0.00144	0.00149	0.00154	0.00159	0.00164	0.00169	0.00175	0.00181	0.00187	**-2.90**
-2.80	0.00193	0.00199	0.00205	0.00212	0.00219	0.00226	0.00233	0.00240	0.00248	0.00256	**-2.80**
-2.70	0.00264	0.00272	0.00280	0.00289	0.00298	0.00307	0.00317	0.00326	0.00336	0.00347	**-2.70**
-2.60	0.00357	0.00368	0.00379	0.00391	0.00402	0.00415	0.00427	0.00440	0.00453	0.00466	**-2.60**
-2.50	0.00480	0.00494	0.00508	0.00523	0.00539	0.00554	0.00570	0.00587	0.00604	0.00621	**-2.50**
-2.40	0.00639	0.00657	0.00676	0.00695	0.00714	0.00734	0.00755	0.00776	0.00798	0.00820	**-2.40**
-2.30	0.00842	0.00866	0.00889	0.00914	0.00939	0.00964	0.00990	0.01017	0.01044	0.01072	**-2.30**
-2.20	0.01101	0.01130	0.01160	0.01191	0.01222	0.01255	0.01287	0.01321	0.01355	0.01390	**-2.20**
-2.10	0.01426	0.01463	0.01500	0.01539	0.01578	0.01618	0.01659	0.01700	0.01743	0.01786	**-2.10**
-2.00	0.01831	0.01876	0.01923	0.01970	0.02018	0.02068	0.02118	0.02169	0.02222	0.02275	**-2.00**
-1.90	0.02330	0.02385	0.02442	0.02500	0.02559	0.02619	0.02680	0.02743	0.02807	0.02872	**-1.90**
-1.80	0.02938	0.03005	0.03074	0.03144	0.03216	0.03288	0.03362	0.03438	0.03515	0.03593	**-1.80**
-1.70	0.03673	0.03754	0.03836	0.03920	0.04006	0.04093	0.04182	0.04272	0.04363	0.04457	**-1.70**
-1.60	0.04551	0.04648	0.04746	0.04846	0.04947	0.05050	0.05155	0.05262	0.05370	0.05480	**-1.60**
-1.50	0.05592	0.05705	0.05821	0.05938	0.06057	0.06178	0.06301	0.06426	0.06552	0.06681	**-1.50**
-1.40	0.06811	0.06944	0.07078	0.07215	0.07353	0.07493	0.07636	0.07780	0.07927	0.08076	**-1.40**
-1.30	0.08226	0.08379	0.08534	0.08691	0.08851	0.09012	0.09176	0.09342	0.09510	0.09680	**-1.30**
-1.20	0.09853	0.10027	0.10204	0.10383	0.10565	0.10749	0.10935	0.11123	0.11314	0.11507	**-1.20**
-1.10	0.11702	0.11900	0.12100	0.12302	0.12507	0.12714	0.12924	0.13136	0.13350	0.13567	**-1.10**
-1.00	0.13786	0.14007	0.14231	0.14457	0.14686	0.14917	0.15151	0.15386	0.15625	0.15866	**-1.00**
-0.90	0.16109	0.16354	0.16602	0.16853	0.17106	0.17361	0.17619	0.17879	0.18141	0.18406	**-0.90**
-0.80	0.18673	0.18943	0.19215	0.19489	0.19766	0.20045	0.20327	0.20611	0.20897	0.21186	**-0.80**
-0.70	0.21476	0.21770	0.22065	0.22363	0.22663	0.22965	0.23270	0.23576	0.23885	0.24196	**-0.70**
-0.60	0.24510	0.24825	0.25143	0.25463	0.25785	0.26109	0.26435	0.26763	0.27093	0.27425	**-0.60**
-0.50	0.27760	0.28096	0.28434	0.28774	0.29116	0.29460	0.29806	0.30153	0.30503	0.30854	**-0.50**
-0.40	0.31207	0.31561	0.31918	0.32276	0.32636	0.32997	0.33360	0.33724	0.3409	0.34458	**-0.40**
-0.30	0.34827	0.35197	0.35569	0.35942	0.36317	0.36693	0.37070	0.37448	0.37828	0.38209	**-0.30**
-0.20	0.38591	0.38974	0.39358	0.39743	0.40129	0.40517	0.40905	0.41294	0.41683	0.42074	**-0.20**
-0.10	0.42465	0.42858	0.43251	0.43644	0.44038	0.44433	0.44828	0.45224	0.45620	0.46017	**-0.10**
-0.00	0.46414	0.46812	0.47210	0.47608	0.48006	0.48405	0.48803	0.49202	0.49601	0.50000	**-0.00**

Standard Normal Table (continued)
Areas Under the Standard Normal Curve
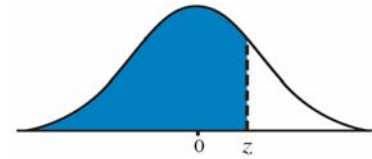
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	z
0.00	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586	**0.00**
0.10	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535	**0.10**
0.20	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409	**0.20**
0.30	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173	**0.30**
0.40	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793	**0.40**
0.50	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240	**0.50**
0.60	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490	**0.60**
0.70	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524	**0.70**
0.80	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327	**0.80**
0.90	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891	**0.90**
1.00	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214	**1.00**
1.10	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298	**1.10**
1.20	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147	**1.20**
1.30	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774	**1.30**
1.40	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189	**1.40**
1.50	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408	**1.50**
1.60	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449	**1.60**
1.70	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327	**1.70**
1.80	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062	**1.80**
1.90	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670	**1.90**
2.00	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169	**2.00**
2.10	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574	**2.10**
2.20	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899	**2.20**
2.30	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158	**2.30**
2.40	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361	**2.40**
2.50	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99492	0.99506	0.99520	**2.50**
2.60	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643	**2.60**
2.70	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.99720	0.99728	0.99736	**2.70**
2.80	0.99744	0.99752	0.99760	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807	**2.80**
2.90	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861	**2.90**
3.00	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.9990	**3.00**
3.10	0.99903	0.99906	0.99910	0.99913	0.99916	0.99918	0.99921	0.99924	0.99926	0.99929	**3.10**
3.20	0.99931	0.99934	0.99936	0.99938	0.99940	0.99942	0.99944	0.99946	0.99948	0.99950	**3.20**
3.30	0.99952	0.99953	0.99955	0.99957	0.99958	0.99960	0.99961	0.99962	0.99964	0.99965	**3.30**
3.40	0.99966	0.99968	0.99969	0.99970	0.99971	0.99972	0.99973	0.99974	0.99975	0.99976	**3.40**
3.50	0.99977	0.99978	0.99978	0.99979	0.99980	0.99981	0.99981	0.99982	0.99983	0.99983	**3.50**

CHAPTER 5: Probabilistic Features of the Distributions of Certain Sample Statistics

5.1 Introduction:

In this Chapter we will discuss the probability distributions of some statistics.

As we mention earlier, a statistic is measure computed form the random sample. As the sample values vary from sample to sample, the value of the statistic varies accordingly.

A statistic is a random variable; it has a probability distribution, a mean and a variance.

5.2 Sampling Distribution:

The probability distribution of a statistic is called the sampling distribution of that statistic.

The sampling distribution of the statistic is used to make statistical inference about the unknown parameter.

5.3 Distribution of the Sample Mean:
(Sampling Distribution of the Sample Mean $\overline{X}$):

Suppose that we have a population with mean μ and variance σ^2. Suppose that $X_1, X_2, \ldots, X_n$ is a random sample of size (n) selected randomly from this population. We know that the sample mean is:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}.$$

Suppose that we select several random samples of size $n=5$.

	1st sample	2nd sample	3rd sample	…	Last sample
Sample values	28	31	14	.	17
	30	20	31	.	32
	34	31	25	.	29
	34	40	27	.	31
	17	28	32	.	30
Sample mean $\overline{X}$	28.4	29.9	25.8	…	27.8

- The value of the sample mean $\overline{X}$ varies from random sample to another.
- The value of $\overline{X}$ is random and it depends on the random sample.
- The sample mean $\overline{X}$ is a random variable.
- The probability distribution of $\overline{X}$ is called the sampling distribution of the sample mean $\overline{X}$.
- Questions:
  - o What is the sampling distribution of the sample mean $\overline{X}$?
  - o What is the mean of the sample mean $\overline{X}$?
  - o What is the variance of the sample mean $\overline{X}$?

Some Results about Sampling Distribution of $\overline{X}$:

Result (1): (mean & variance of $\overline{X}$)

If $X_1, X_2, \ldots, X_n$ is a random sample of size n from any distribution with mean μ and variance σ^2; then:

1. The mean of $\overline{X}$ is: $\mu_{\overline{X}} = \mu$.

2. The variance of $\overline{X}$ is: $\sigma_{\overline{X}}^2 = \dfrac{\sigma^2}{n}$.

3. The Standard deviation of $\overline{X}$ is call <u>the standard error</u> and is defined by: $\sigma_{\overline{X}} = \sqrt{\sigma_{\overline{X}}^2} = \dfrac{\sigma}{\sqrt{n}}$.

Result (2): (Sampling from normal population)

If $X_1, X_2, \ldots, X_n$ is a random sample of size n from a normal population with mean μ and variance σ^2; that is Normal(μ, σ^2), then the sample mean has a normal distribution with mean μ and variance σ^2/n, that is:

1. $\overline{X} \sim$ Normal $\left(\mu, \dfrac{\sigma^2}{n} \right)$.

2. $Z = \dfrac{\overline{X} - \mu}{\sigma / \sqrt{n}} \sim$ Normal $(0,1)$.

We use this result when sampling from normal distribution with known variance σ^2.

Result (3): (Central Limit Theorem: Sampling from Non-normal population)

Suppose that $X_1, X_2, \ldots, X_n$ is a random sample of size n from non-normal population with mean μ and variance σ^2. If the sample size n is large $(n \geq 30)$, then the sample mean has approximately a normal distribution with mean μ and variance σ^2/n, that is

1. $\overline{X} \approx$ Normal $\left(\mu, \dfrac{\sigma^2}{n}\right)$ (approximately)

2. $Z = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}} \approx$ Normal $(0,1)$ (approximately)

Note: "$\approx$" means "approximately distributed".
We use this result when sampling from non-normal distribution with known variance σ^2 and with large sample size.

Result (4): (used when σ^2 is unknown + normal distribution)

If $X_1, X_2, \ldots, X_n$ is a random sample of size n from a normal distribution with mean μ and unknown variance σ^2; that is Normal(μ, σ^2), then the statistic:

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

has a t- distribution with $(n-1)$ degrees of freedom, where S is the sample standard deviation given by:

$$S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}}$$

We write:

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

Notation: degrees of freedom = df = ν

The t-Distribution: (Section 6.3. pp 172-174)

- Student's t distribution.
- t-distribution is a distribution of a continuous random variable.
- Recall that, if X_1, X_2, …, X_n is a random sample of size n from a normal distribution with mean μ and variance σ^2, i.e. $N(\mu, \sigma^2)$, then

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

We can apply this result only when σ^2 is known!

- If σ^2 is unknown, we replace the population variance σ^2 with the sample variance $S^2 = \dfrac{\sum\limits_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$ to have the following statistic

$$T = \frac{\overline{X} - \mu}{S / \sqrt{n}}$$

Recall:

If X_1, X_2, …, X_n is a random sample of size n from a normal distribution with mean μ and variance σ^2, i.e. $N(\mu, \sigma^2)$, then the statistic:

$$T = \frac{\overline{X} - \mu}{S / \sqrt{n}}$$

has a t-distribution with $(n-1)$ degrees of freedom ($df = \nu = n-1$), and we write T~ t(ν) or T~ t(n–1).

Note:

- t-distribution is a continuous distribution.
- The value of t random variable range from $-\infty$ to $+\infty$ (that is, $-\infty < t < \infty$).
- The mean of t distribution is 0.
- It is symmetric about the mean 0.
- The shape of t-distribution is similar to the shape of the standard normal distribution.
- t-distribution $\rightarrow$ Standard normal distribution as n $\rightarrow \infty$.

Notation: (t $_\alpha$)

$P(T < t_\alpha) = \alpha$
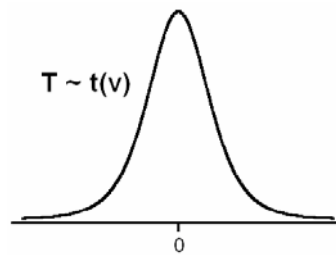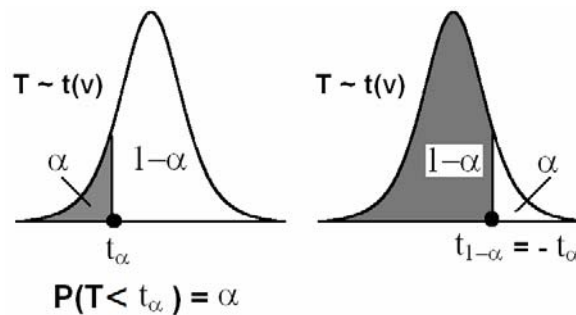
- t_α = The t-value under which we find an area equal to α
  = The t-value that leaves an area of α to the left.
- The value t_α satisfies: $P(T < t_\alpha) = \alpha$.
- Since the curve of the pdf of T~ t(v) is symmetric about 0, we have

$$t_{1-\alpha} = -t_\alpha$$

For example: $\quad t_{0.35} = -t_{1-0.35} = -t_{0.65}$

$$t_{0.82} = -t_{1-0.86} = -t_{0.14}$$

- Values of t_α are tabulated in a special table for several values of α and several values of degrees of freedom. (Table E, appendix p. A-40 in the textbook).

Example:
Find the t-value with v=14 (df) that leaves an area of:
- (a)   0.95 to the left.
- (b)   0.95 to the right.

Solution:
$v = 14$ (df); T~ t(14)
(a) The t-value that leaves an area of 0.95 to the left is
$t_{0.95} = 1.761$.

Table of t - Distribution
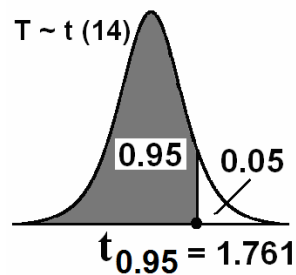
	0.95
14	1.761

$t_{0.95} = 1.761$

$t_{0.95} = 1.761$

(b) The t-value that leaves an area of 0.95 to the right is

$$t_{0.05} = -t_{1-0.05} = -t_{0.95} = -1.761$$

Table of t - Distribution

	0.05
14	−1.761
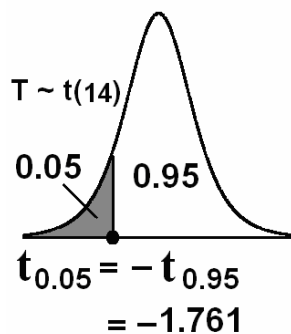
$t_{0.05} = -t_{0.95}$
$= -1.761$

$t_{0.05} = -1.761$

Note: Some t-tables contain values of α that are greater than or equal to 0.90. When we search for small values of α in these tables, we may use the fact that:

$$t_{1-\alpha} = -t_{\alpha}$$

Example:

For $v = 10$ degrees of freedom (df), find $t_{0.93}$ and $t_{0.07}$.

Solution:

$t_{0.93} = (1.372+1.812)/2 = 1.592$ (from the table)

$t_{0.07} = -t_{1-0.07} = -t_{0.93} = -1.592$ (using the rule: $t_{1-\alpha} = -t_{\alpha}$)

Table of t - Distribution
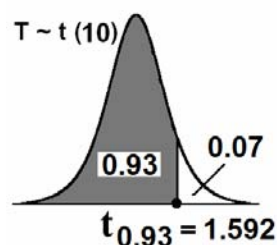
	0.90	0.95
10	1.372	1.812

$t_{0.93} = 1.592$

$t_{0.93} = \dfrac{1.372+1.812}{2}$
$= 1.592$

Critical Values of the t-distribution (t_α)
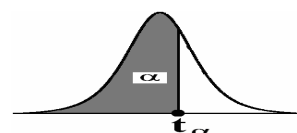
ν=df	$t_{0.90}$	$t_{0.95}$	$t_{0.975}$	$t_{0.99}$	$t_{0.995}$
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
35	1.3062	1.6896	2.0301	2.4377	2.7238
40	1.3030	1.6840	2.0210	2.4230	2.7040
45	1.3006	1.6794	2.0141	2.4121	2.6896
50	1.2987	1.6759	2.0086	2.4033	2.6778
60	1.2958	1.6706	2.0003	2.3901	2.6603
70	1.2938	1.6669	1.9944	2.3808	2.6479
80	1.2922	1.6641	1.9901	2.3739	2.6387
90	1.2910	1.6620	1.9867	2.3685	2.6316
100	1.2901	1.6602	1.9840	2.3642	2.6259
120	1.2886	1.6577	1.9799	2.3578	2.6174
140	1.2876	1.6558	1.9771	2.3533	2.6114
160	1.2869	1.6544	1.9749	2.3499	2.6069
180	1.2863	1.6534	1.9732	2.3472	2.6034
200	1.2858	1.6525	1.9719	2.3451	2.6006
∞	1.282	1.645	1.960	2.326	2.576

Application:

Example: (Sampling distribution of the sample mean)
Suppose that the time duration of a minor surgery is approximately normally distributed with mean equal to 800 seconds and a standard deviation of 40 seconds. Find the probability that a random sample of 16 surgeries will have average time duration of less than 775 seconds.

Solution:
X= the duration of the surgery
μ=800 , σ=40 , $\sigma^2 = 1600$
X~N(800, 1600)
Sample size: n=16
Calculating mean, variance, and standard error (standard deviation) of the sample mean $\overline{X}$:

Mean of $\overline{X}$: $\mu_{\overline{X}} = \mu$=800

Variance of $\overline{X}$: $\sigma_{\overline{X}}^2 = \dfrac{\sigma^2}{n} = \dfrac{1600}{16} = 100$

Standard error (standard deviation) of $\overline{X}$: $\sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{40}{\sqrt{16}} = 10$

Using the central limit theorem, $\overline{X}$ has a normal distribution with mean $\mu_{\overline{X}} = 800$ and variance $\sigma_{\overline{X}}^2 = 100$, that is:

$$\overline{X} \sim N(\mu, \frac{\sigma^2}{n}) = N(800, 100)$$

$$\Leftrightarrow Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{\overline{X} - 800}{10} \sim N(0,1)$$

The probability that a random sample of 16 surgeries will have an average time duration that is less than 775 seconds equals to:

$$P(\overline{X} < 775) = P\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < \frac{775 - \mu}{\sigma/\sqrt{n}}\right) = P\left(\frac{\overline{X} - 800}{10} < \frac{775 - 800}{10}\right)$$

$$= P\left(Z < \frac{775 - 800}{10}\right) = P(Z < -2.50) = 0.0062$$

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}) = N(800, 100)$$

$\mu_{\bar{X}} = 800$

$\sigma_{\bar{X}}^2 = \dfrac{\sigma^2}{n} = 100$

0.0062

$\sigma_{\bar{X}} = \dfrac{\sigma}{\sqrt{n}} = 10$

775   $\mu_{\bar{X}} = 800$
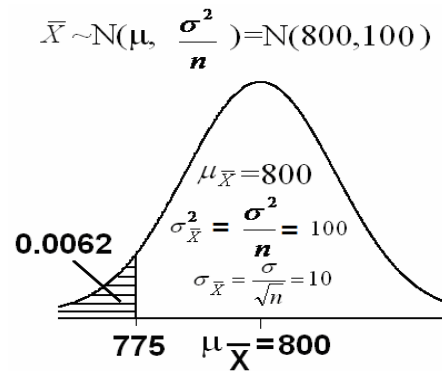
Example:

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

Solution:

X= the serum iron value

$\mu=120$, $\sigma=15$, $\sigma^2 = 225$

X~N(120, 225)

Sample size: $n=50$

Calculating mean, variance, and standard error (standard deviation) of the sample mean $\bar{X}$:

Mean of $\bar{X}$: $\mu_{\bar{X}} = \mu = 120$

Variance of $\bar{X}$: $\sigma_{\bar{X}}^2 = \dfrac{\sigma^2}{n} = \dfrac{225}{50} = 4.5$

Standard error (standard deviation) of $\bar{X}$: $\sigma_{\bar{X}} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{15}{\sqrt{50}} = 2.12$

Using the central limit theorem, $\bar{X}$ has a normal distribution with mean $\mu_{\bar{X}} = 120$ and variance $\sigma_{\bar{X}}^2 = 4.5$, that is:

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}) = N(120, 4.5)$$

$$\Leftrightarrow \ Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{X} - 120}{2.12} \sim N(0,1)$$

The probability that a random sample of 50 men will yield a mean between 115 and 125 microgram/100ml equals to:

$$P(115 < \bar{X} < 125) = P\left(\frac{115 - \mu}{\sigma/\sqrt{n}} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < \frac{125 - \mu}{\sigma/\sqrt{n}} \right)$$

$$= P\left(\frac{115-120}{2.12} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < \frac{125-120}{2.12}\right) = P(-2.36 < Z < 2.36)$$

$$= P(Z < 2.36) - P(Z < -2.36)$$

$$= 0.9909 - 0.0091$$

$$= 0.9818$$

5.4 Distribution of the Difference Between Two Sample Means ($\bar{X}_1 - \bar{X}_2$):

Suppose that we have two populations:

- 1-st population with mean μ_1 and variance σ_1^2
- 2-nd population with mean μ_2 and variance σ_2^2
- We are interested in comparing μ_1 and μ_2, or equivalently, making inferences about the difference between the means ($\mu_1-\mu_2$).
- We <u>independently</u> select a random sample of size n_1 from the 1-st population and another random sample of size n_2 from the 2-nd population:
- Let $\bar{X}_1$ and S_1^2 be the sample mean and the sample variance of the 1-st sample.
- Let $\bar{X}_2$ and S_2^2 be the sample mean and the sample variance of the 2-nd sample.
- The sampling distribution of $\bar{X}_1 - \bar{X}_2$ is used to make inferences about $\mu_1-\mu_2$.

The sampling distribution of $\overline{X}_1 - \overline{X}_2$:

Result:

The mean, the variance and the standard deviation of $\overline{X}_1 - \overline{X}_2$ are:

Mean of $\overline{X}_1 - \overline{X}_2$ is: $\qquad \mu_{\overline{X}_1 - \overline{X}_2} = \mu_1 - \mu_2$

Variance of $\overline{X}_1 - \overline{X}_2$ is: $\qquad \sigma^2_{\overline{X}_1 - \overline{X}_2} = \dfrac{\sigma^2_1}{n_1} + \dfrac{\sigma^2_2}{n_2}$

Standard error (standard) deviation of $\overline{X}_1 - \overline{X}_2$ is:

$$\sigma_{\overline{X}_1 - \overline{X}_2} = \sqrt{\sigma^2_{\overline{X}_1 - \overline{X}_2}} = \sqrt{\dfrac{\sigma^2_1}{n_1} + \dfrac{\sigma^2_2}{n_2}}$$

Result:

If the two random samples were selected from normal distributions (or non-normal distributions with large sample sizes) with known variances σ^2_1 and σ^2_2, then the difference between the sample means $(\overline{X}_1 - \overline{X}_2)$ has a normal distribution with mean $(\mu_1 - \mu_2)$ and variance $((\sigma^2_1 / n_1) + (\sigma^2_2 / n_2))$, that is:

- $\overline{X}_1 - \overline{X}_2 \sim N\left(\mu_1 - \mu_2 , \dfrac{\sigma^2_1}{n_1} + \dfrac{\sigma^2_2}{n_2}\right)$

- $Z = \dfrac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma^2_1}{n_1} + \dfrac{\sigma^2_2}{n_2}}} \sim N(0,1)$

Application:

Example:

Suppose it has been established that for a certain type of client (type A) the average length of a home visit by a public health nurse is 45 minutes with standard deviation of 15 minutes, and that for second type (type B) of client the average home visit is 30 minutes long with standard deviation of 20 minutes. If a nurse randomly visits 35 clients from the first type and 40

clients from the second type, what is the probability that the average length of home visit of first type will be greater than the average length of home visit of second type by 20 or more minutes?

Solution:

For the first type:

$$\mu_1 = 45$$
$$\sigma_1 = 15$$
$$\sigma_1^2 = 225$$
$$n_1 = 35$$

For the second type:

$$\mu_2 = 30$$
$$\sigma_2 = 20$$
$$\sigma_2^2 = 400$$
$$n_2 = 40$$

The mean, the variance and the standard deviation of $\overline{X}_1 - \overline{X}_2$ are:

Mean of $\overline{X}_1 - \overline{X}_2$ is:

$$\mu_{\overline{X}_1 - \overline{X}_2} = \mu_1 - \mu_2 = 45 - 30 = 15$$

Variance of $\overline{X}_1 - \overline{X}_2$ is:

$$\sigma_{\overline{X}_1 - \overline{X}_2}^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \frac{225}{35} + \frac{400}{40} = 16.4286$$

Standard error (standard) deviation of $\overline{X}_1 - \overline{X}_2$ is:

$$\sigma_{\overline{X}_1 - \overline{X}_2} = \sqrt{\sigma_{\overline{X}_1 - \overline{X}_2}^2} = \sqrt{16.4286} = 4.0532$$

The sampling distribution of $\overline{X}_1 - \overline{X}_2$ is:

$$\overline{X}_1 - \overline{X}_2 \sim N(15, 16.4286)$$

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - 15}{\sqrt{16.4286}} \sim N(0,1)$$

The probability that the average length of home visit of first type will be greater than the average length of home visit of second type by 20 or more minutes is:

$$P(\overline{X}_1 - \overline{X}_2 > 20) = P\left(\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} > \frac{20 - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \right)$$

$$= P\left(Z > \frac{20 - 15}{4.0532} \right) = P(Z>1.23) = 1 - P(Z<1.23)$$

$$= 1 - 0.8907$$

$$= 0.1093$$

5.5 Distribution of the Sample Proportion ($\hat{p}$):
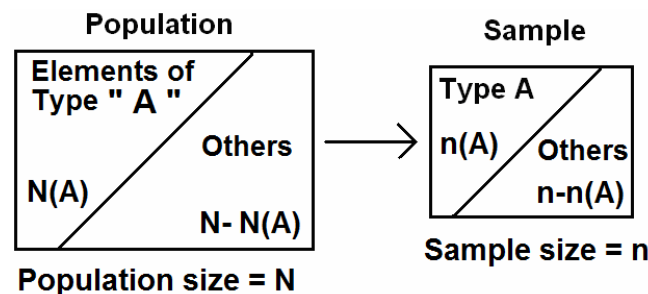
- For the population:
    $N(A)$ = number of elements in the population
        with a specified characteristic "A"
    N = total number of elements in the population
        (population size)
The population proportion is
$$p = \frac{N(A)}{N} \qquad \text{(p is a parameter)}$$

- For the sample:
    $n(A)$ = number of elements in the sample with the same
        characteristic "A"
    n = sample size
The sample proportion is
$$\hat{p} = \frac{n(A)}{n} \qquad (\hat{p} \text{ is a statistic})$$

- The sampling distribution of $\hat{p}$ is used to make inferences

about p.

Result:

The mean of the sample proportion ($\hat{p}$) is the population proportion (p); that is:

$$\mu_{\hat{p}} = p$$

The variance of the sample proportion ($\hat{p}$) is:

$$\sigma_{\hat{p}}^2 = \frac{p(1-p)}{n} = \frac{pq}{n}. \qquad \text{(where q=1 −p)}$$

The standard error (standard deviation) of the sample proportion ($\hat{p}$) is:

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{pq}{n}}$$

Result:

For large sample size ($n \geq 30, np > 5, nq > 5$), the sample proportion ($\hat{p}$) has approximately a normal distribution with mean $\mu_{\hat{p}} = p$ and a variance $\sigma_{\hat{p}}^2 = pq/n$, that is:

$$\hat{p} \sim N\left(p, \frac{pq}{n}\right) \qquad \text{(approximately)}$$

$$Z = \frac{\hat{p} - p}{\sqrt{\dfrac{pq}{n}}} \sim N(0,1) \qquad \text{(approximately)}$$

Example:

Suppose that 45% of the patients visiting a certain clinic are females. If a sample of 35 patients was selected at random, find the probability that:

1. the proportion of females in the sample will be greater than 0.4.
2. the proportion of females in the sample will be between 0.4 and 0.5.

Solution:

- .n = 35 (large)
- p = The population proportion of females = $\frac{45}{100} = 0.45$

- $\hat{p}$ = The sample proportion
  (proportion of females in the sample)
- The mean of the sample proportion ($\hat{p}$) is p = 0.45
- The variance of the sample proportion ($\hat{p}$) is:

$$\frac{p(1-p)}{n} = \frac{pq}{n} = \frac{0.45(1-0.45)}{35} = 0.0071.$$

- The standard error (standard deviation) of the sample proportion ($\hat{p}$) is:

$$\sqrt{\frac{p(1-p)}{n}} = \sqrt{0.0071} = 0.084$$

- $n \geq 30$, $np = 35 \times 0.45 = 15.75 > 5$, $nq = 35 \times 0.55 = 19.25 > 5$

1. The probability that the sample proportion of females ($\hat{p}$) will be greater than 0.4 is:

$$P(\hat{p} > 0.4) = 1 - P(\hat{p} < 0.4) = 1 - P\left(\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} < \frac{0.4-p}{\sqrt{\frac{p(1-p)}{n}}}\right)$$

$$= 1 - P\left(Z < \frac{0.4 - 0.45}{\sqrt{\frac{0.45(1-0.45)}{35}}}\right) = 1 - P(Z < -0.59)$$

$$= 1 - 0.2776 = 0.7224$$

2. The probability that the sample proportion of females ($\hat{p}$) will be between 0.4 and 0.5 is:

$$P(0.4 < \hat{p} < 0.5) = P(\hat{p} < 0.5) - P(\hat{p} < 0.4)$$

$$= P\left(\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} < \frac{0.5-p}{\sqrt{\frac{p(1-p)}{n}}}\right) - 0.2776$$

$$= P\left(Z < \frac{0.5 - 0.45}{\sqrt{\frac{0.45(1-0.45)}{35}}}\right) - 0.2776$$

$$= P(Z < 0.59) - 0.2776$$
$$= 0.7224 - 0.2776$$
$$= 0.4448$$

5.6 Distribution of the Difference Between Two Sample Proportions ($\hat{p}_1 - \hat{p}_2$):
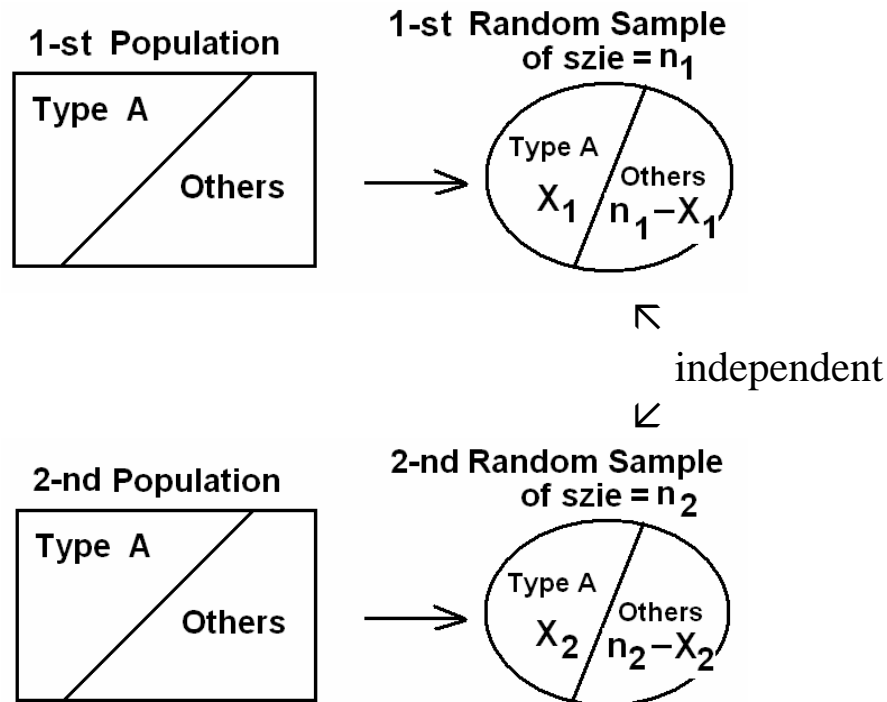
Suppose that we have two populations:

- p_1 = proportion of elements of type (A) in the 1-st population.
- p_2 = proportion of elements of type (A) in the 2-nd population.
- We are interested in comparing p_1 and p_2, or equivalently, making inferences about $p_1 - p_2$.
- We <u>independently</u> select a random sample of size n_1 from the 1-st population and another random sample of size n_2 from the 2-nd population:
- Let X_1 = no. of elements of type (A) in the 1-st sample.
- Let X_2 = no. of elements of type (A) in the 2-nd sample.
- $\hat{p}_1 = \dfrac{X_1}{n_1}$ = sample proportion of the 1-st sample

- $\hat{p}_2 = \dfrac{X_2}{n_2}$ = sample proportion of the 2-nd sample

- The sampling distribution of $\hat{p}_1 - \hat{p}_2$ is used to make inferences about $p_1 - p_2$.

The sampling distribution of $\hat{p}_1 - \hat{p}_2$:
Result:
The mean, the variance and the standard error (standard deviation) of $\hat{p}_1 - \hat{p}_2$ are:

- Mean of $\hat{p}_1 - \hat{p}_2$ is:

$$\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2$$

- Variance of $\hat{p}_1 - \hat{p}_2$ is:

$$\sigma^2_{\hat{p}_1 - \hat{p}_2} = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}$$

- Standard error (standard deviation) of $\hat{p}_1 - \hat{p}_2$ is:

$$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

- $q_1 = 1 - p_1$ and $q_2 = 1 - p_2$

Result:
For large samples sizes
$(n_1 \geq 30, n_2 \geq 30, n_1 p_1 > 5, n_1 q_1 > 5, n_2 p_2 > 5, n_2 q_2 > 5)$, we have that $\hat{p}_1 - \hat{p}_2$ has approximately normal distribution with mean $\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2$ and variance $\sigma^2_{\hat{p}_1 - \hat{p}_2} = \dfrac{p_1 q_1}{n_1} + \dfrac{p_2 q_2}{n_2}$, that is:

$$\hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2, \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2} \right) \quad \text{(Approximately)}$$

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\dfrac{p_1 q_1}{n_1} + \dfrac{p_2 q_2}{n_2}}} \sim N(0,1) \quad \text{(Approximately)}$$

Example:

Suppose that 40% of Non-Saudi residents have medical insurance and 30% of Saudi residents have medical insurance in a certain city. We have randomly and independently selected a sample of 130 Non-Saudi residents and another sample of 120 Saudi residents. What is the probability that the difference between the sample proportions, $\hat{p}_1 - \hat{p}_2$, will be between 0.05 and 0.2?

Solution:

p_1 = population proportion of non-Saudi with medical insurance.
p_2 = population proportion of Saudi with medical insurance.
$\hat{p}_1$ = sample proportion of non-Saudis with medical insurance.
$\hat{p}_2$ = sample proportion of Saudis with medical insurance.

$p_1 = 0.4$ $\qquad$ $n_1 = 130$
$p_2 = 0.3$ $\qquad$ $n_2 = 120$

$$\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2 = 0.4 - 0.3 = 0.1$$

$$\sigma^2_{\hat{p}_1 - \hat{p}_2} = \frac{p_1\, q_1}{n_1} + \frac{p_2\, q_2}{n_2} = \frac{(0.4)(0.6)}{130} + \frac{(0.3)(0.7)}{120} = 0.0036$$

$$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1\, q_1}{n_1} + \frac{p_2\, q_2}{n_2}} = \sqrt{0.0036} = 0.06$$

The probability that the difference between the sample proportions, $\hat{p}_1 - \hat{p}_2$, will be between 0.05 and 0.2 is:

$$P(0.05 < \hat{p}_1 - \hat{p}_2 < 0.2) = P(\hat{p}_1 - \hat{p}_2 < 0.2) - P(\hat{p}_1 - \hat{p}_2 < 0.05)$$

$$= P\left(\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\dfrac{p_1\, q_1}{n_1} + \dfrac{p_2\, q_2}{n_2}}} < \frac{0.2 - (p_1 - p_2)}{\sqrt{\dfrac{p_1\, q_1}{n_1} + \dfrac{p_2\, q_2}{n_2}}} \right)$$

$$- P \left(\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\dfrac{p_1 \, q_1}{n_1} + \dfrac{p_2 \, q_2}{n_2}}} < \frac{0.05 - (p_1 - p_2)}{\sqrt{\dfrac{p_1 \, q_1}{n_1} + \dfrac{p_2 \, q_2}{n_2}}} \right)$$

$$= P \left(Z < \frac{0.2 - 0.1}{0.06} \right) - P \left(Z < \frac{0.05 - 0.1}{0.06} \right)$$

$$= P(Z < 1.67) - P(Z < -0.83)$$

$$= 0.9515 - 0.2033$$

$$= 0.7482$$

CHAPTER 6: Using Sample Data to Make Estimations About Population Parameters

6.1 Introduction:

Statistical Inferences: (Estimation and Hypotheses Testing)

It is the procedure by which we reach a conclusion about a population on the basis of the information contained in a sample drawn from that population.

There are two main purposes of statistics;

- Descriptive Statistics: (Chapter 1 & 2): Organization & summarization of the data
- Statistical Inference: (Chapter 6 and 7): Answering research questions about some unknown population parameters.

(1) Estimation: (chapter 6)

Approximating (or estimating) the actual values of the unknown parameters:

– **Point Estimate:** A point estimate is single value used to estimate the corresponding population parameter.

– **Interval Estimate (or Confidence Interval):** An interval estimate consists of two numerical values defining a range of values that most likely includes the parameter being estimated with a specified degree of confidence.

(2) Hypothesis Testing: (chapter 7)

Answering research questions about the unknown parameters of the population (confirming or denying some conjectures or statements about the unknown parameters).

6.2 Confidence Interval for a Population Mean (μ) :

In this section we are interested in estimating the mean of a certain population (μ).
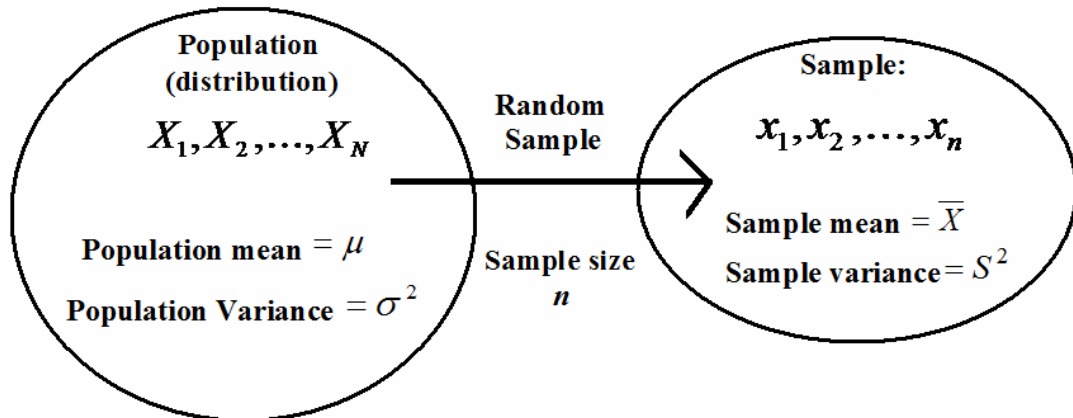
Population:

Population Size = N

Population Values: $X_1, X_2, \ldots, X_N$

Population Mean: $\mu = \dfrac{\sum_{i=1}^{N} X_i}{N}$

Population Variance: $\sigma^2 = \dfrac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}$

Sample:

Sample Size = n

Sample values: $x_1, x_2, \ldots, x_n$

Sample Mean: $\overline{X} = \dfrac{\sum_{i=1}^{n} x_i}{n}$

Sample Variance: $S^2 = \dfrac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$

(i) Point Estimation of μ:

A point estimate of the mean is a single number used to estimate (or approximate) the true value of μ .

– Draw a random sample of size n from the population:

$$x_1, x_2, \ldots, x_n$$

– Compute the sample mean: $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$

Result:

The sample mean $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$ is a "good" point estimator of the population mean (μ).

(ii) Confidence Interval (Interval Estimate) of μ:

An interval estimate of μ is an interval (L, U) containing the true value of μ "with a probability of $1 - \alpha$".

* $1 - \alpha$ = is called the confidence coefficient (level)
*   L = lower limit of the confidence interval
*   U = upper limit of the confidence interval

Result: (For the case when σ is known)
(a) If $X_1, X_2 \ldots, X_n$ is a random sample of size n from a normal distribution with mean μ and known variance σ^2, then:
A $(1 - \alpha)100\%$ confidence interval for μ is:

$$\overline{X} \pm Z_{1-\frac{\alpha}{2}} \; \sigma_{\overline{X}}$$

$$\overline{X} \pm Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

$$\left(\overline{X} - Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \;, \; \overline{X} + Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right)$$

$$\overline{X} - Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \; < \mu < \; \overline{X} + Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

(b) If $X_1, X_2 \ldots, X_n$ is a random sample of size n from a non-normal distribution with mean μ and known variance σ^2, and if the sample size n is large $(n \geq 30)$, then:
An approximate $(1 - \alpha)100\%$ confidence interval for μ is:

$$\overline{X} \pm Z_{1-\frac{\alpha}{2}} \; \sigma_{\overline{X}}$$

$$\overline{X} \pm Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

$$\left(\overline{X} - Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \;, \; \overline{X} + Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right)$$

$$\overline{X} - Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \; < \mu < \; \overline{X} + Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

Note that:
1. We are $(1-\alpha)100\%$ confident that the true value of μ belongs to the interval $(\overline{X} - Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \; , \; \overline{X} + Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}})$.

2. Upper limit of the confidence interval $= \overline{X} + Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$

3. Lower limit of the confidence interval $= \overline{X} - Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$

4. $Z_{1-\frac{\alpha}{2}} =$ Reliability Coefficient

5. $Z_{1-\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}} =$ margin of error $=$ precision of the estimate

6. In general the interval estimate (confidence interval) may be expressed as follows:

$$\overline{X} \pm Z_{1-\frac{\alpha}{2}} \; \sigma_{\overline{X}}$$

estimator $\pm$ (reliability coefficient) $\times$ (standard Error)

estimator $\pm$ margin of error

6.3 The t Distribution:
(Confidence Interval Using t)

We have already introduced and discussed the t distribution.
Result: (For the case when σ is unknown + normal population)
If $X_1, X_2 ..., X_n$ is a random sample of size n from a normal distribution with mean μ and unknown variance σ^2, then:
A $(1-\alpha)100\%$ confidence interval for μ is:

$$\overline{X} \pm t_{1-\frac{\alpha}{2}} \; \hat{\sigma}_{\overline{X}}$$

$$\overline{X} \pm t_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}$$

$$\left(\overline{X} - t_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \; , \; \overline{X} + t_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \right)$$

where the degrees of freedom is:

$$df = \nu = n-1.$$

Note that:

1. We are $(1-\alpha)100\%$ confident that the true value of μ belongs to the interval $\left(\overline{X} - t_{1-\frac{\alpha}{2}} \dfrac{S}{\sqrt{n}} , \ \overline{X} + t_{1-\frac{\alpha}{2}} \dfrac{S}{\sqrt{n}} \right)$.

2. $\hat{\sigma}_{\overline{X}} = \dfrac{S}{\sqrt{n}}$ (estimate of the standard error of $\overline{X}$)

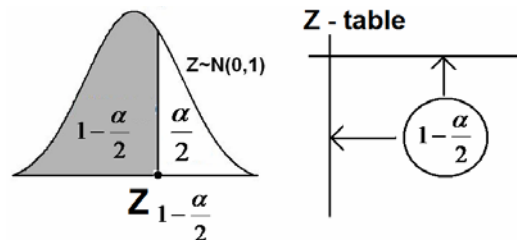3. $t_{1-\frac{\alpha}{2}} = $ Reliability Coefficient

4. In this case, we replace σ by S and Z by t.

5. In general the interval estimate (confidence interval) may be expressed as follows:

  Estimator $\pm$ (Reliability Coefficient) $\times$ (Estimate of the Standard Error)

$$\overline{X} \pm t_{1-\frac{\alpha}{2}} \ \hat{\sigma}_{\overline{X}}$$

Notes: (Finding Reliability Coefficient)

(1) We find the reliability coefficient $Z_{1-\frac{\alpha}{2}}$ from the Z-table as follows:

(2) We find the reliability coefficient $t_{1-\frac{\alpha}{2}}$ from the t-table as follows: $(df = \nu = n-1)$
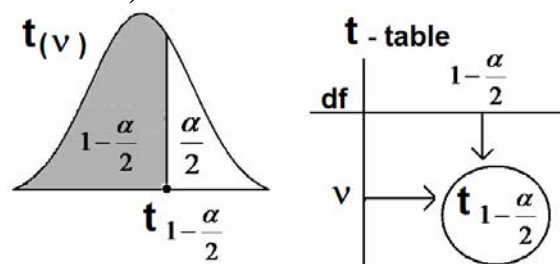
Example:

Suppose that $Z \sim N(0,1)$. Find $Z_{1-\frac{\alpha}{2}}$ for the following cases:

(1) $\alpha = 0.1$ (2) $\alpha = 0.05$ (3) $\alpha = 0.01$

Solution:

(1) For $\alpha = 0.1$:

$$1 - \frac{\alpha}{2} = 1 - \frac{0.1}{2} = 0.95 \qquad \Rightarrow \qquad Z_{1-\frac{\alpha}{2}} = Z_{0.95} = 1.645$$
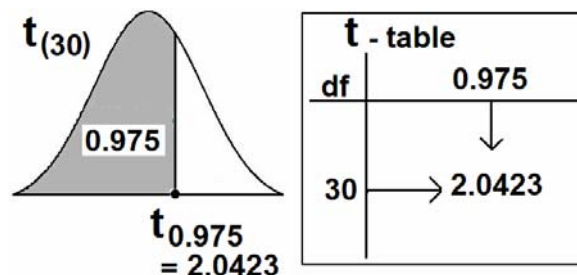
(2) For $\alpha = 0.05$:

$$1 - \frac{\alpha}{2} = 1 - \frac{0.05}{2} = 0.975 \qquad \Rightarrow \qquad Z_{1-\frac{\alpha}{2}} = Z_{0.975} = 1.96.$$

(3) For $\alpha = 0.01$:

$$1 - \frac{\alpha}{2} = 1 - \frac{0.01}{2} = 0.995 \qquad \Rightarrow \qquad Z_{1-\frac{\alpha}{2}} = Z_{0.995} = 2.575.$$
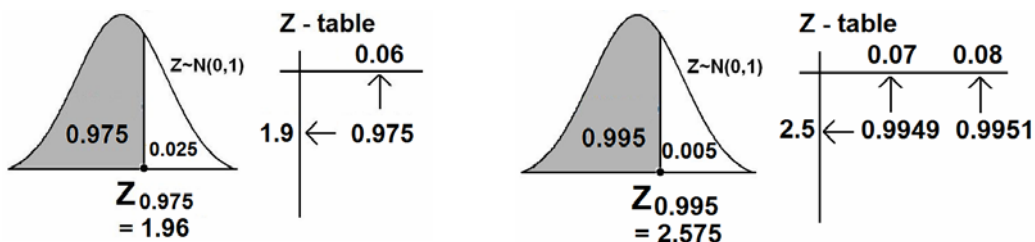
Example:

Suppose that $t \sim t(30)$. Find $t_{1-\frac{\alpha}{2}}$ for $\alpha = 0.05$.

Solution:

$$df = \nu = 30$$

$$1 - \frac{\alpha}{2} = 1 - \frac{0.05}{2} = 0.975 \qquad \Rightarrow \qquad t_{1-\frac{\alpha}{2}} = t_{0.975} = 2.0423$$

Example: (The case where σ^2 is known)

Diabetic ketoacidosis is a potential fatal complication of diabetes mellitus throughout the world and is characterized in part by very high blood glucose levels. In a study on 123 patients living in Saudi Arabia of age 15 or more who were admitted for diabetic ketoacidosis, the mean blood glucose level was 26.2 mmol/l. Suppose that the blood glucose levels for such patients have a normal distribution with a standard deviation of 3.3 mmol/l.

(1) Find a point estimate for the mean blood glucose level of such diabetic ketoacidosis patients.

(2) Find a 90% confidence interval for the mean blood glucose level of such diabetic ketoacidosis patients.

Solution:

Variable = X = blood glucose level (quantitative variable).

Population = diabetic ketoacidosis patients in Saudi Arabia of age 15 or more.

Parameter of interest is: μ = the mean blood glucose level.

Distribution is normal with standard deviation $\sigma = 3.3$.

σ^2 is known ($\sigma^2 = 10.89$)

X ~ Normal(μ , 10.89)

μ = ?? (unknown- we need to estimate μ)

Sample size: $n = 123$ (large)

Sample mean: $\overline{X} = 26.2$

(1) Point Estimation:

We need to find a point estimate for μ.

$\overline{X} = 26.2$ is a point estimate for μ.

$\mu \approx 26.2$

(2) Interval Estimation (Confidence Interval = C. I.):

We need to find 90% C. I. for μ.

$90\% = (1-\alpha)100\%$

$1-\alpha = 0.9 \Leftrightarrow \alpha = 0.1 \Leftrightarrow \dfrac{\alpha}{2} = 0.05 \quad \Leftrightarrow \quad 1-\dfrac{\alpha}{2} = 0.95$

The reliability coefficient is: $Z_{1-\frac{\alpha}{2}} = Z_{0.95} = 1.645$

90% confidence interval for μ is:

$$\left(\overline{X} - Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \ , \ \overline{X} + Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right)$$

$$\left(26.2 - (1.645)\frac{3.3}{\sqrt{123}} \ , \ 26.2 + (1.645)\frac{3.3}{\sqrt{123}} \right)$$

$$(26.2 - 0.4894714 \ , \ 26.2 + 0.4894714)$$

$$(25.710529 \ , \ 26.689471)$$

We are 90% confident that the true value of the mean μ lies in the interval $(25.71 \ , \ 26.69)$, that is:

$$25.71 < \mu < 26.69$$

Note: for this example even if the distribution is not normal, we may use the same solution because the sample size n=123 is large.

Example: (The case where σ^2 is unknown)

A study was conducted to study the age characteristics of Saudi women having breast lump. A sample of 121 Saudi women gave a mean of 37 years with a standard deviation of 10 years. Assume that the ages of Saudi women having breast lumps are normally distributed.

(a) Find a point estimate for the mean age of Saudi women having breast lumps.

(b) Construct a 99% confidence interval for the mean age of Saudi women having breast lumps

Solution:

X = Variable = age of Saudi women having breast lumps (quantitative variable).

Population = All Saudi women having breast lumps.

Parameter of interest is: $\mu =$ the age mean of Saudi women having breast lumps.

X ~ Normal(μ , σ^2)

$\mu = ??$ (unknown- we need to estimate μ)

$\sigma^2 = ??$ (unknown)

Sample size: $n = 121$

Sample mean: $\overline{X} = 37$

Sample standard deviation: $S = 10$

Degrees of freedom: df $= v = 121 - 1 = 120$

(a) Point Estimation: We need to find a point estimate for μ.

$\overline{X} = 37$ is a "good" point estimate for μ.
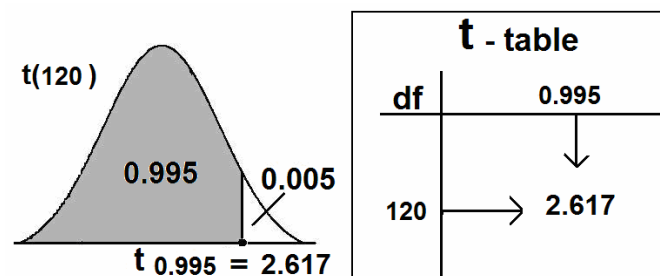
$\mu \approx 37 \ years$

(b) Interval Estimation (Confidence Interval = C. I.): We need to find 99% C. I. for μ.

$$99\% = (1 - \alpha)100\%$$

$$1 - \alpha = 0.99 \Leftrightarrow \alpha = 0.01 \quad \Leftrightarrow \quad \frac{\alpha}{2} = 0.005 \quad \Leftrightarrow \quad 1 - \frac{\alpha}{2} = 0.995$$

$v = df = 120$

The reliability coefficient is: $t_{1-\frac{\alpha}{2}} = t_{0.995} = 2.617$

99% confidence interval for μ is:

$$\overline{X} \pm t_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}$$

$$37 \pm (2.617) \frac{10}{\sqrt{121}}$$

$$37 \pm 2.38$$

$$(37 - 2.38, \ 37 + 2.38)$$

$$(34.62 \ , \ 39.38)$$

Another Way:

$$\left(\overline{X} - t_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \ , \ \overline{X} + t_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \right)$$

$$\left(37 - (2.617) \frac{10}{\sqrt{121}} \ , \ 37 + (2.617) \frac{10}{\sqrt{121}} \right)$$

$$(37 - 2.38, \ 37 + 2.38)$$

$$(34.62 , 39.38)$$

We are 99% confident that the true value of the mean μ lies in the interval $(34.61 , 39.39)$, that is:

$$34.62 < \mu < 39.38$$

6.4 Confidence Interval for the Difference between Two Population Means ($\mu_1 - \mu_2$):

Suppose that we have two populations:
- 1-st population with mean μ_1 and variance σ_1^2
- 2-nd population with mean μ_2 and variance σ_2^2
- We are interested in comparing μ_1 and μ_2, or equivalently, making inferences about the difference between the means ($\mu_1 - \mu_2$).
- We <u>independently</u> select a random sample of size n_1 from the 1-st population and another random sample of size n_2 from the 2-nd population:
- Let $\bar{X}_1$ and S_1^2 be the sample mean and the sample variance of the 1-st sample.
- Let $\bar{X}_2$ and S_2^2 be the sample mean and the sample variance of the 2-nd sample.
- The sampling distribution of $\bar{X}_1 - \bar{X}_2$ is used to make inferences about $\mu_1 - \mu_2$.

Recall:

1. Mean of $\bar{X}_1 - \bar{X}_2$ is: $\qquad \mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2$

2. Variance of $\bar{X}_1 - \bar{X}_2$ is: $\qquad \sigma^2_{\bar{X}_1 - \bar{X}_2} = \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}$

3. Standard error of $\bar{X}_1 - \bar{X}_2$ is: $\quad \sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$

4. If the two random samples were selected from normal distributions (or non-normal distributions with large sample sizes) with known variances σ_1^2 and σ_2^2, then the difference between the sample means $(\bar{X}_1 - \bar{X}_2)$ has a normal distribution with mean $(\mu_1 - \mu_2)$ and variance $((\sigma_1^2 / n_1) + (\sigma_2^2 / n_2))$, that is:

- $\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2 , \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}\right)$

- $Z = \dfrac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \sim N(0,1)$

Point Estimation of $\mu_1 - \mu_2$:
Result:
$\quad \bar{X}_1 - \bar{X}_2$ is a "good" point estimate for $\mu_1 - \mu_2$.

Interval Estimation (Confidence Interval) of $\mu_1 - \mu_2$:
We will consider two cases.
(i) First Case: σ_1^2 and σ_2^2 are known:
$\quad$ If σ_1^2 and σ_2^2 are known, we use the following result to find an interval estimate for $\mu_1 - \mu_2$.
Result:
A $(1-\alpha)100\%$ confidence interval for $\mu_1 - \mu_2$ is:
$$(\bar{X}_1 - \bar{X}_2) \pm Z_{1-\frac{\alpha}{2}} \, \sigma_{\bar{X}_1 - \bar{X}_2}$$

$$(\overline{X}_1 - \overline{X}_2) \pm Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$\left((\overline{X}_1 - \overline{X}_2) - Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \;,\; (\overline{X}_1 - \overline{X}_2) + Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right)$$

$$(\overline{X}_1 - \overline{X}_2) - Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\overline{X}_1 - \overline{X}_2) + Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Estimator $\pm$ (Reliability Coefficient) $\times$ (Standard Error)

(ii) Second Case:

Unknown equal Variances: $(\sigma_1^2 = \sigma_2^2 = \sigma^2$ **is unknown):**

If σ_1^2 and σ_2^2 are equal but unknown $(\sigma_1^2 = \sigma_2^2 = \sigma^2)$, then the pooled estimate of the common variance σ^2 is

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

where S_1^2 is the variance of the 1-st sample and S_2^2 is the variance of the 2-nd sample. The degrees of freedom of S_p^2 is

$$df = \nu = n_1 + n_2 - 2.$$

We use the following result to find an interval estimate for $\mu_1 - \mu_2$ when we have normal populations with unknown and equal variances.

Result:

A $(1-\alpha)100\%$ confidence interval for $\mu_1 - \mu_2$ is:

$$(\overline{X}_1 - \overline{X}_2) \pm t_{1-\frac{\alpha}{2}} \sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}$$

$$\left((\overline{X}_1 - \overline{X}_2) - t_{1-\frac{\alpha}{2}} \sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}} \;,\; (\overline{X}_1 - \overline{X}_2) + t_{1-\frac{\alpha}{2}} \sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}} \right)$$

where reliability coefficient $t_{1-\frac{\alpha}{2}}$ is the t-value with

df=ν=n_1+n_2−2 degrees of freedom.

Example: (1st Case: σ_1^2 and σ_2^2 are known)

An experiment was conducted to compare time length

(duration time) of two types of surgeries (A) and (B). 75 surgeries of type (A) and 50 surgeries of type (B) were performed. The average time length for (A) was 42 minutes and the average for (B) was 36 minutes.

(1) Find a point estimate for $\mu_A - \mu_B$, where μ_A and μ_B are population means of the time length of surgeries of type (A) and (B), respectively.

(2) Find a 96% confidence interval for $\mu_A - \mu_B$. Assume that the population standard deviations are 8 and 6 for type (A) and (B), respectively.

Solution:

Surgery	Type (A)	Type (B)
Sample Size	$n_A = 75$	$n_B = 50$
Sample Mean	$\overline{X}_A = 42$	$\overline{X}_B = 36$
Population Standard Deviation	$\sigma_A = 8$	$\sigma_B = 6$

(1) A point estimate for $\mu_A - \mu_B$ is:

$$\overline{X}_A - \overline{X}_B = 42 - 36 = 6.$$

(2) Finding a 96% confidence interval for $\mu_A - \mu_B$:

$\alpha = ??$

$96\% = (1-\alpha)100\% \Leftrightarrow 0.96 = (1-\alpha) \Leftrightarrow \alpha = 0.04 \Leftrightarrow \alpha/2 = 0.02$

Reliability Coefficient: $Z_{1-\frac{\alpha}{2}} = Z_{0.98} = 2.055$

A 96% C.I. for $\mu_A - \mu_B$ is:

$$(\overline{X}_A - \overline{X}_B) \pm Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}$$

$$6 \pm Z_{0.98} \sqrt{\frac{8^2}{75} + \frac{6^2}{50}}$$

$$6 \pm (2.055) \sqrt{\frac{64}{75} + \frac{36}{50}}$$

$$6 \pm 2.578$$

$$3.422 < \mu_A - \mu_B < 8.58$$

We are 96% confident that $\mu_A - \mu_B \in (3.42, 8.58)$.

Note: Since the confidence interval does not include zero, we conclude that the two population means are not equal ($\mu_A - \mu_B \neq 0$ $\Leftrightarrow \mu_A \neq \mu_B$). Therefore, we may conclude that the mean time length is not the same for the two types of surgeries.

Example: (2nd Case: $\sigma_1^2 = \sigma_2^2$ unknown)

To compare the time length (duration time) of two types of surgeries (A) and (B), an experiment shows the following results based on two independent samples:

Type *A*:   140, 138, 143, 142, 144, 137
Type *B*:   135, 140, 136, 142, 138, 140

(1) Find a point estimate for $\mu_A - \mu_B$, where μ_A (μ_B) is the mean time length of type *A* (*B*).

(2) Assuming normal populations with equal variances, find a 95% confidence interval for $\mu_A - \mu_B$.

Solution:

First we calculate the mean and the variances of the two samples, and we get:

Surgery	Type (A)	Type (B)
Sample Size	$n_A = 6$	$n_B = 6$
Sample Mean	$\overline{X}_A = 140.67$	$\overline{X}_B = 138.50$
Sample Variance	$S_A^2 = 7.87$	$S_B^2 = 7.10$

(1) A point estimate for $\mu_A - \mu_B$ is:
$$\overline{X}_A - \overline{X}_B = 140.67 - 138.50 = 2.17.$$

(2) Finding 95% Confidence interval for $\mu_A - \mu_B$:

95% = $(1-\alpha)100\%$ $\Leftrightarrow 0.95 = (1-\alpha) \Leftrightarrow \alpha = 0.05 \Leftrightarrow \alpha/2 = 0.025$

.df $= \nu = n_A + n_B - 2 = 10$

Reliability Coefficient: $t_{1-\frac{\alpha}{2}} = t_{0.975} = 2.228$

The pooled estimate of the common variance is:

$$S_p^2 = \frac{(n_A - 1)S_A^2 + (n_B - 1)S_B^2}{n_A + n_B - 2}$$

$$= \frac{(6-1)(7.87) + (6-1)(7.1)}{6+6-2} = 7.485$$

A 95% C.I. for $\mu_A - \mu_B$ is:

$$(\overline{X}_A - \overline{X}_B) \pm t_{1-\frac{\alpha}{2}} \sqrt{\frac{S_p^2}{n_A} + \frac{S_p^2}{n_B}}$$

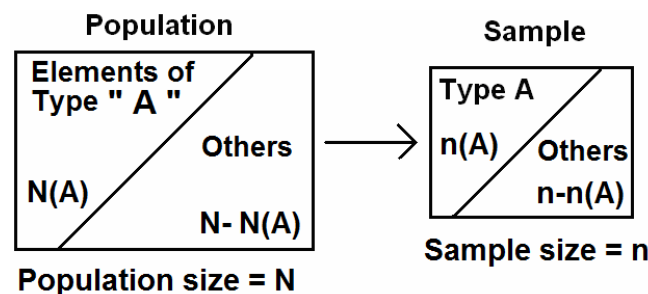$$2.17 \pm (2.228) \sqrt{\frac{7.485}{6} + \frac{7.485}{6}}$$

$$2.17 \pm 3.519$$

$$-1.35 < \mu_A - \mu_B < 5.69$$

We are 95% confident that $\mu_A - \mu_B \in (-1.35, 5.69)$.

Note: Since the confidence interval includes zero, we conclude that the two population means may be equal ($\mu_A - \mu_B = 0 \Leftrightarrow \mu_A = \mu_B$). Therefore, we may conclude that the mean time length is the same for both types of surgeries.

6.5 Confidence Interval for a Population Proportion (p):

Recall:

1. For the population:

   $N(A) =$ number of elements in the population with a specified characteristic "A"

   N = total number of elements in the population (population size)

   The population proportion is:

   $$p = \frac{N(A)}{N} \qquad \text{(p is a parameter)}$$

2. For the sample:

$n(A) =$ number of elements in the sample with the same characteristic "A"

n = sample size

The sample proportion is:

$$\hat{p} = \frac{n(A)}{n} \qquad (\hat{p} \text{ is a statistic})$$

3. The sampling distribution of the sample proportion ($\hat{p}$) is used to make inferences about the population proportion (p).

4. The mean of ($\hat{p}$) is: $\qquad \mu_{\hat{p}} = p$

5. The variance of ($\hat{p}$) is: $\quad \sigma_{\hat{p}}^2 = \dfrac{p(1-p)}{n}$

6. The standard error (standard deviation) of ($\hat{p}$) is:

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \, .$$

7. For large sample size ($n \geq 30, np > 5, n(1-p) > 5$), the sample proportion ($\hat{p}$) has approximately a normal distribution with mean $\mu_{\hat{p}} = p$ and a variance $\sigma_{\hat{p}}^2 = p(1-p)/n$, that is:

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right) \qquad \text{(approximately)}$$

$$Z = \frac{\hat{p} - p}{\sqrt{\dfrac{p(1-p)}{n}}} \sim N(0,1) \qquad \text{(approximately)}$$

(i) Point Estimate for (p):
Result:

A good point estimate for the population proportion (p) is the sample proportion ($\hat{p}$).

(ii) Interval Estimation (Confidence Interval) for (p):
Result:

For large sample size ($n \geq 30, np > 5, n(1-p) > 5$), an approximate $(1-\alpha)100\%$ confidence interval for (p) is:

$$\hat{p} \pm Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$\left(\hat{p} - Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \ , \quad \hat{p} + Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$$

Estimator $\pm$ (Reliability Coefficient) $\times$ (Standard Error)

Example:

In a study on the obesity of Saudi women, a random sample of 950 Saudi women was taken. It was found that 611 of these women were obese (overweight by a certain percentage).
(1) Find a point estimate for the true proportion of Saudi women who are obese.
(2) Find a 95% confidence interval for the true proportion of Saudi women who are obese.

Solution:

Variable: whether or not a women is obese (qualitative variable)
Population: all Saudi women
Parameter: p =the proportion of women who are obese.

Sample:
$n = 950$ (950 women in the sample)
$n(A) = 611$ (611 women in the sample who are obese)
The sample proportion (the proportion of women who are obese in the sample.) is:

$$\hat{p} = \frac{n(A)}{n} = \frac{611}{950} = 0.643$$

(1) A point estimate for p is: $\hat{p} = 0.643$.
(2) We need to construct 95% C.I. for the proportion (p).

$$95\% = (1-\alpha)100\% \iff 0.95 = 1-\alpha \iff \alpha = 0.05 \iff \frac{\alpha}{2} = 0.025 \iff 1-\frac{\alpha}{2} = 0.975$$

The reliability coefficient: $Z_{1-\frac{\alpha}{2}} = z_{0.975} = 1.96$.

A 95% C.I. for the proportion (p) is:

$$\hat{p} \pm Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$0.643 \pm (1.96)\sqrt{\frac{(0.643)(1-0.643)}{950}}$$

$$0.643 \pm (1.96)(0.01554)$$

$$0.643 \pm 0.0305$$

$$(0.6127 , 0.6735)$$

We are 95% confident that the true value of the population proportion of obese women, p, lies in the interval $(0.61, 0.67)$, that is:

$$0.61 < p < 0.67$$

6.6 Confidence Interval for the Difference Between Two Population Proportions ($p_1 - p_2$):
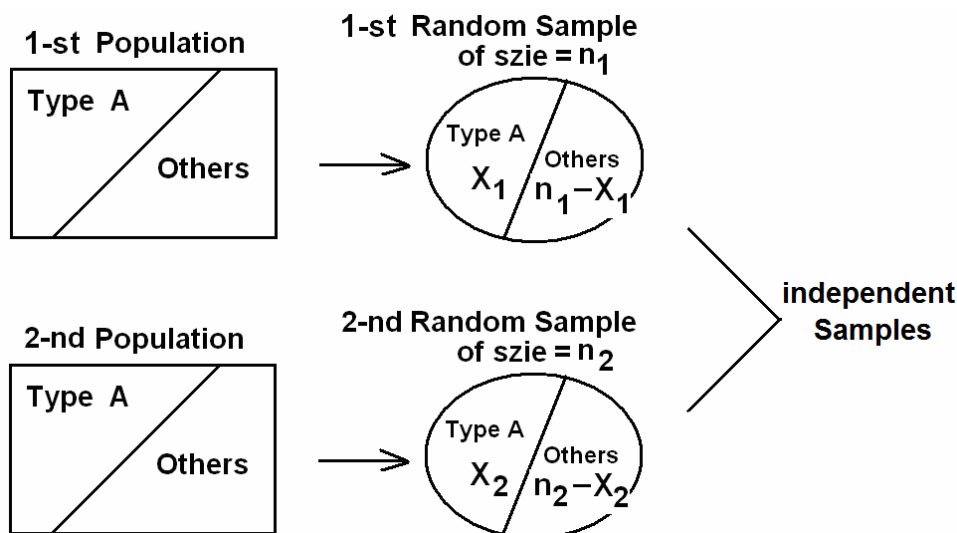
Suppose that we have two populations with:

- p_1 = population proportion of elements of type (A) in the 1-st population.
- p_2 = population proportion of elements of type (A) in the 2-nd population.
- We are interested in comparing p_1 and p_2, or equivalently, making inferences about $p_1 - p_2$.
- We independently select a random sample of size n_1 from the 1-st population and another random sample of size n_2 from the 2-nd population:

- Let X_1 = no. of elements of type (A) in the 1-st sample.
- Let X_2 = no. of elements of type (A) in the 2-nd sample.
- $\hat{p}_1 = \dfrac{X_1}{n_1}$ = the sample proportion of the 1-st sample
- $\hat{p}_2 = \dfrac{X_2}{n_2}$ = the sample proportion of the 2-nd sample
- The sampling distribution of $\hat{p}_1 - \hat{p}_2$ is used to make inferences about $p_1 - p_2$.

Recall:

1. Mean of $\hat{p}_1 - \hat{p}_2$ is: $\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2$

2. Variance of $\hat{p}_1 - \hat{p}_2$ is: $\sigma^2_{\hat{p}_1 - \hat{p}_2} = \dfrac{p_1 q_1}{n_1} + \dfrac{p_2 q_2}{n_2}$

3. Standard error (standard deviation) of $\hat{p}_1 - \hat{p}_2$ is:

$$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\dfrac{p_1 q_1}{n_1} + \dfrac{p_2 q_2}{n_2}}$$

4. For large samples sizes
$(n_1 \geq 30, n_2 \geq 30, n_1 p_1 > 5, n_1 q_1 > 5, n_2 p_2 > 5, n_2 q_2 > 5)$, we have that $\hat{p}_1 - \hat{p}_2$ has approximately normal distribution with mean $\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2$ and variance $\sigma^2_{\hat{p}_1 - \hat{p}_2} = \dfrac{p_1 q_1}{n_1} + \dfrac{p_2 q_2}{n_2}$, that is:

$$\hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2, \dfrac{p_1 q_1}{n_1} + \dfrac{p_2 q_2}{n_2} \right) \quad \text{(Approximately)}$$

$$Z = \dfrac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\dfrac{p_1 q_1}{n_1} + \dfrac{p_2 q_2}{n_2}}} \sim N(0,1) \quad \text{(Approximately)}$$

Note: $q_1 = 1 - p_1$ and $q_2 = 1 - p_2$.

Point Estimation for $p_1 - p_2$:
Result:

A good point estimator for the difference between the two proportions, $p_1 - p_2$, is:

$$\hat{p}_1 - \hat{p}_2 = \frac{X_1}{n_1} - \frac{X_2}{n_2}$$

Interval Estimation (Confidence Interval) for $p_1 - p_2$:
Result:

For large n_1 and n_2, an approximate $(1-\alpha)100\%$ confidence interval for $p_1 - p_2$ is:

$$(\hat{p}_1 - \hat{p}_2) \pm Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1\,\hat{q}_1}{n_1} + \frac{\hat{p}_2\,\hat{q}_2}{n_2}}$$

$$\left((\hat{p}_1 - \hat{p}_2) - Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1\,\hat{q}_1}{n_1} + \frac{\hat{p}_2\,\hat{q}_2}{n_2}}\,,\;\; (\hat{p}_1 - \hat{p}_2) + Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1\,\hat{q}_1}{n_1} + \frac{\hat{p}_2\,\hat{q}_2}{n_2}} \right)$$

Estimator $\pm$ (Reliability Coefficient) $\times$ (Standard Error)

Example:

A researcher was interested in comparing the proportion of people having cancer disease in two cities (A) and (B). A random sample of 1500 people was taken from the first city (A), and another independent random sample of 2000 people was taken from the second city (B). It was found that 75 people in the first sample and 80 people in the second sample have cancer disease.

(1) Find a point estimate for the difference between the proportions of people having cancer disease in the two cities.

(2) Find a 90% confidence interval for the difference between the two proportions.

Solution:

p_1 = population proportion of people having cancer disease in the first city (A)

p_2 = population proportion of people having cancer disease in the second city (B)

$\hat{p}_1$ = sample proportion of the first sample

$\hat{p}_2$ = sample proportion of the second sample

X_1 = number of people with cancer in the first sample

X_2 = number of people with cancer in the second sample

For the first sample we have:

$$n_1 = 1500\,, \qquad X_1 = 75$$

$$\hat{p}_1 = \frac{X_1}{n_1} = \frac{75}{1500} = 0.05 \quad , \qquad \hat{q}_1 = 1 - 0.05 = 0.95$$

For the second sample we have:

$$n_2 = 2000 \quad , \qquad X_2 = 80$$

$$\hat{p}_2 = \frac{X_2}{n_2} = \frac{80}{2000} = 0.04 \qquad , \qquad \hat{q}_2 = 1 - 0.04 = 0.96$$

(1) Point Estimation for $p_1 - p_2$:

A good point estimate for the difference between the two proportions, $p_1 - p_2$, is:

$$\hat{p}_1 - \hat{p}_2 = 0.05 - 0.04$$
$$= 0.01$$

(2) Finding 90% Confidence Interval for $p_1 - p_2$:

90% = $(1-\alpha)100\% \Leftrightarrow 0.90 = (1-\alpha) \Leftrightarrow \alpha = 0.1 \Leftrightarrow \alpha/2 = 0.05$

The reliability coefficient: $Z_{1-\frac{\alpha}{2}} = z_{0.95} = 1.645$

A 90% confidence interval for $p_1 - p_2$ is:

$$(\hat{p}_1 - \hat{p}_2) \pm Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$$

$$(\hat{p}_1 - \hat{p}_2) \pm Z_{0.95} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$$

$$0.01 \pm 1.645 \sqrt{\frac{(0.05)(0.95)}{1500} + \frac{(0.04)(0.96)}{2000}}$$

$$0.01 \pm 0.01173$$

$$-0.0017 < p_1 - p_2 < 0.0217$$

We are 90% confident that $p_1 - p_2 \in (-0.0017, 0.0217)$.

Note: Since the confidence interval includes zero, we may conclude that the two population proportions are equal ($p_1 - p_2 = 0 \Leftrightarrow p_1 = p_2$). Therefore, we may conclude that the proportion of people having cancer is the same in both cities.

CHAPTER 7: Using Sample Statistics To Test Hypotheses About Population Parameters:

In this chapter, we are interested in testing some hypotheses about the unknown population parameters.

7.1 Introduction:

Consider a population with some unknown parameter θ. We are interested in testing (confirming or denying) some conjectures about θ. For example, we might be interested in testing the conjecture that $\theta > \theta_o$, where θ_o is a given value.

- A hypothesis is a statement about one or more populations.
- A research hypothesis is the conjecture or supposition that motivates the research.
- A statistical hypothesis is a conjecture (or a statement) concerning the population which can be evaluated by appropriate statistical technique.
- For example, if θ is an unknown parameter of the population, we might be interested in testing the conjecture sating that $\theta \geq \theta_o$ against $\theta < \theta_o$ (for some specific value θ_o).
- We usually test the null hypothesis (H_o) against the alternative (or the research) hypothesis (H_1 or H_A) by choosing one of the following situations:

  (i) $H_o: \theta = \theta_o$ against $H_A: \theta \neq \theta_o$

  (ii) $H_o: \theta \geq \theta_o$ against $H_A: \theta < \theta_o$

  (iii) $H_o: \theta \leq \theta_o$ against $H_A: \theta > \theta_o$

- Equality sign must appear in the null hypothesis.
- H_o is the null hypothesis and H_A is the alternative hypothesis. (H_o and H_A are complement of each other)
- The null hypothesis (H_o) is also called "the hypothesis of no difference".
- The alternative hypothesis (H_A) is also called the research hypothesis.

- There are 4 possible situations in testing a statistical hypothesis:

<div align="center">

Condition of Null Hypothesis H_o
(Nature/reality)

</div>

Possible Action (Decision)		H_o is true	H_o is false
	Accepting H_o	Correct Decision	Type II error (β)
	Rejecting H_o	Type I error (α)	Correct Decision

- There are two types of Errors:
  - Type I error = Rejecting H_o when H_o is true
    P(Type I error) = P(Rejecting Ho | Ho is true) = α
  - Type II error = Accepting Ho when Ho is false
    P(Type II error) = P(Accepting Ho | Ho is false) = β

- The level of significance of the test is the probability of rejecting true H_o:
  α = P(Rejecting H_o | H_o is true) = P(Type I error)

- There are 2 types of alternative hypothesis:
  - One-sided alternative hypothesis:
    - $H_O: \theta \geq \theta_o$ against $H_A: \theta < \theta_o$
    - $H_O: \theta \leq \theta_o$ against $H_A: \theta > \theta_o$
  - Two-sided alternative hypothesis:
    - $H_O: \theta = \theta_o$ against $H_A: \theta \neq \theta_o$

- We will use the terms "accepting" and "not rejecting" interchangeably. Also, we will use the terms "acceptance" and "nonrejection" interchangeably.
- We will use the terms "accept" and "fail to reject" interchangeably

The Procedure of Testing H_o (against H_A):
   The test procedure for rejecting H_o (accepting H_A) or accepting H_o (rejecting H_A) involves the following steps:

1. Determining a test statistic (T.S.)

We choose the appropriate test statistic based on the point estimator of the parameter.

The test statistic has the following form:

$$\text{Test statistic} = \frac{Estimate\ -\ hypothesized\ parameter}{Sta.ndard\ Error\ of\ the\ Estimate}$$

2. Determining the level of significance (α):

$$\alpha = 0.01,\ 0.025,\ 0.05,\ 0.10$$

3. Determining the rejection region of H_o (R.R.) and the acceptance region of H_o (A.R.).

The R.R. of H_o depends on H_A and α:

- H_A determines the direction of the R.R. of H_o
- α determines the size of the R.R. of H_o

(α = the size of the R.R. of H_o = shaded area)

$H_A: \theta \neq \theta_o$	$H_A: \theta > \theta_o$	$H_A: \theta < \theta_o$
Two-sided alternative	One-sided alternative	One-sided alternative
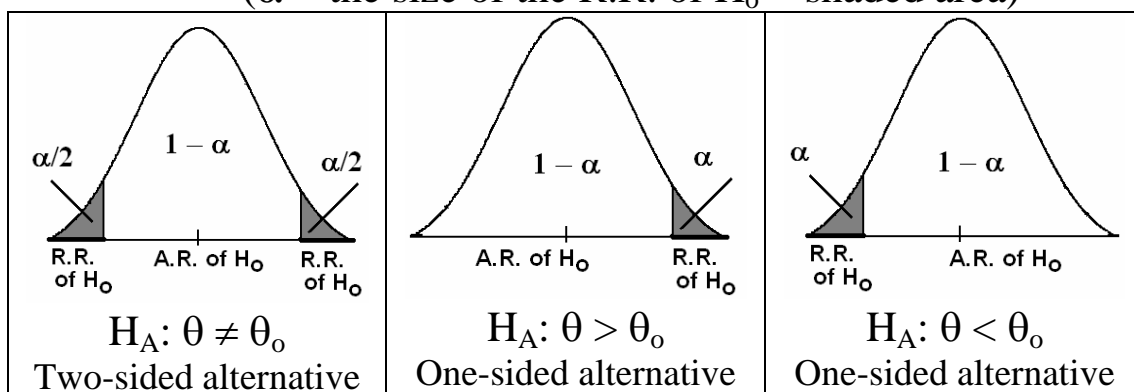
4. Decision:

We reject H_o (and accept H_A) if the value of the test statistic (T.S.) belongs to the R.R. of H_o , and vice versa.

Notes:

1. The rejection region of H_o (R.R.) is sometimes called "the critical region".

2. The values which separate the rejection region (R.R.) and the acceptance region (A.R.) are called "the critical values".

7.2 Hypothesis Testing: A Single Population Mean (μ):

Suppose that $X_1, X_2, \ldots, X_n$ is a random sample of size n from a distribution (or population) with mean μ and variance σ^2.

We need to test some hypotheses (make some statistical inference) about the mean (μ).

Recall:

1. $\overline{X}$ is a "good" point estimate for μ.

2. Mean of $\overline{X}$ is: $\qquad \mu_{\overline{X}} = \mu$.

3. Variance of $\overline{X}$ is: $\qquad \sigma^2_{\overline{X}} = \dfrac{\sigma^2}{n}$.

4. Standard error (standard deviation) of $\overline{X}$ is:

$$\sigma_{\overline{X}} = \sqrt{\sigma^2_{\overline{X}}} = \frac{\sigma}{\sqrt{n}}.$$

5. For the case of normal distribution with any sample size or the case of non-normal distribution with large sample size, and for known variance σ^2, we have:

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

6. For the case of normal distribution with unknown variance σ^2 and with any sample size, we have:

$$t = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(n-1).$$

where $S = \sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2/(n-1)}$ and $\text{df} = \nu = n - 1$.

The Procedure for hypotheses testing about the mean (μ):
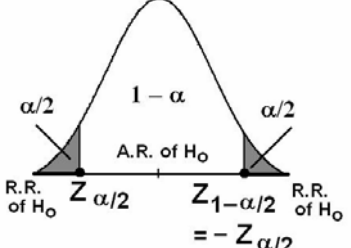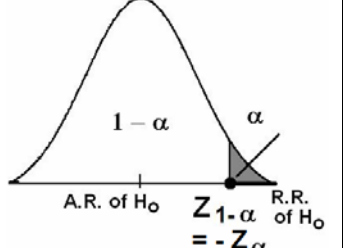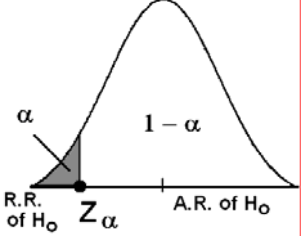
Let μ_o be a given known value.

(1) First case:
Assumptions:
- The variance σ^2 is known.
- Normal distribution with any sample size, or
- Non-normal distribution with large sample size.

Test Procedures:

Hypotheses	$H_o: \mu = \mu_o$ $H_A: \mu \neq \mu_o$	$H_o: \mu \leq \mu_o$ $H_A: \mu > \mu_o$	$H_o: \mu \geq \mu_o$ $H_A: \mu < \mu_o$
Test Statistic (T.S.)	Calculate the value of: $Z = \dfrac{\bar{X} - \mu_o}{\sigma/\sqrt{n}} \sim N(0,1)$		
R.R. & A.R. of H_o			
Critical value (s)	$Z_{\alpha/2}$ and $-Z_{\alpha/2}$	$Z_{1-\alpha} = -Z_\alpha$	Z_α
Decision:	We reject H_o (and accept H_A) at the significance level α if:		
	$Z < Z_{\alpha/2}$ or $Z > Z_{1-\alpha/2} = -Z_{\alpha/2}$ Two-Sided Test	$Z > Z_{1-\alpha} = -Z_\alpha$ One-Sided Test	$Z < Z_\alpha$ One-Sided Test

(2) Second case:

Assumptions:
- The variance σ^2 is unknown.
- Normal distribution.
- Any sample size.

Test Procedures:

Hypotheses	$H_o: \mu = \mu_o$ $H_A\ \mu \neq \mu_o$	$H_o: \mu \leq \mu_o$ $H_A: \mu > \mu_o$	$H_o: \mu \geq \mu_o$ $H_A: \mu < \mu_o$
Test Statistic (T.S.)	Calculate the value of: $\ t = \dfrac{\overline{X} - \mu_o}{S/\sqrt{n}} \ \sim t(n{-}1)$ $(df = \nu = n{-}1)$		
R.R. & A.R. of H_o			
Critical value (s)	$t_{\alpha/2}\ $ and $-t_{\alpha/2}$	$t_{1-\alpha} = -t_\alpha$	t_α
Decision:	We reject H_o (and accept H_A) at the significance level α if:		
	$t < t_{\alpha/2}$ or $t > t_{1-\alpha/2} = -t_{\alpha/2}$ Two-Sided Test	$t > t_{1-\alpha} = -t_\alpha$ One-Sided Test	$t < t_\alpha$ One-Sided Test

Example: (first case: variance σ^2 is known)

A random sample of 100 recorded deaths in the United States during the past year showed an average of 71.8 years. Assuming a population standard deviation of 8.9 year, does this seem to indicate that the mean life span today is greater than 70 years? Use a 0.05 level of significance.

Solution:

.n=100 (large), $\overline{X}$ =71.8, σ=8.9 (σ is known)

μ=average (mean) life span

μ_o=70

α=0.05

Hypotheses:

H_o: $\mu \leq 70$ ($\mu_o = 70$)

H_A: $\mu > 70$ (research hypothesis)

Test statistics (T.S.) :

$$Z = \frac{\overline{X} - \mu_o}{\sigma / \sqrt{n}} = \frac{71.8 - 70}{8.9 / \sqrt{100}} = 2.02$$

Level of significance:

α=0.05

Rejection Region of H_o (R.R.): (critical region)

$- Z_{\alpha} = -Z_{0.05} = 1.645$ (critical value)

We should reject H_o if:

$Z > -Z_{\alpha} = -Z_{0.05} = 1.645$
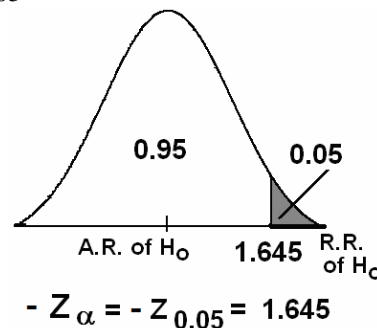
Decision:

Since Z=2.02 ∈R.R., i.e., Z=2.02>−$Z_{0.05}$, we reject H_o:μ≤70 at α=0.05 and accept H_A:μ>70. Therefore, we conclude that the mean life span today is greater than 70 years.

Note: Using P- Value as a decision tool:

P-value is the smallest value of α for which we can reject the null hypothesis H_o.

Calculating P-value:

* Calculating P-value depends on the alternative hypothesis H_A.

* Suppose that $Z_c = \frac{\overline{X} - \mu_o}{\sigma / \sqrt{n}}$ is the computed value of the test Statistic.

* The following table illustrates how to compute P-value, and how to use P-value for testing the null hypothesis:

Alternative Hypothesis:	$H_A: \mu \neq \mu_o$	$H_A: \mu > \mu_o$	$H_A: \mu < \mu_o$
P-Value =	$2 \times P(Z > |z_c|)$	$P(Z > z_c)$	$P(Z < z_c)$
Significance Level =	α		
Decision:	Reject H_o if P-value $< \alpha$.		

Example:

For the previous example, we have found that:

$$Z_C = \frac{\overline{X} - \mu_o}{\sigma / \sqrt{n}} = 2.02$$

The alternative hypothesis was $H_A: \mu > 70$.

$P - Value = P(Z > Z_C)$

$\qquad = P(Z > 2.02) = 1 - P(Z < 2.02) = 1 - 0.9783 = 0.0217$

The level of significance was $\alpha = 0.05$.

Since P-value $< \alpha$, we reject H_o.

Example: (second case: variance σ^2 is unknown)

The manager of a private clinic claims that the mean time of the patient-doctor visit in his clinic is 8 minutes. Test the hypothesis that $\mu=8$ minutes against the alternative that $\mu \neq 8$ minutes if a random sample of 50 patient-doctor visits yielded a mean time of 7.8 minutes with a standard deviation of 0.5 minutes. It is assumed that the distribution of the time of this type of visits is normal. Use a 0.01 level of significance.

Solution:

The distribution is normal.

.$n=50$ (large)

$\overline{X} = 7.8$

S=0.5 (sample standard deviation)

σ is unknown

$\mu=$ mean time of the visit

$\mu_o=8$

$\alpha=0.01$ ($\alpha/2 = 0.005$)

Hypotheses:

$H_o: \mu = 8$ ($\mu_o=8$)

$H_A: \mu \neq 8$ (research hypothesis)

Test statistics (T.S.):

$$t = \frac{\overline{X} - \mu_o}{S / \sqrt{n}} = \frac{7.8 - 8}{0.5 / \sqrt{50}} = -2.83$$

df $= \nu = $ n-1 $= $ 50-1$=$49

Level of significance:

$\alpha = 0.01$

Rejection Region of H$_o$ (R.R.):   (critical region)

$t_{\alpha/2} = t_{0.005} (= -t_{0.995}) = -2.678$ (1st critical value)

$-t_{\alpha/2} = -t_{0.005} = 2.678$ (2nd critical value)

We should reject H$_o$ if:

$t < t_{\alpha/2} = t_{0.005} = -2.678$

or

$t > -t_{\alpha/2} = -t_{0.005} = 2.678$

0.005   **0.99**   **0.005**

A.R. of H$_o$

R.R. $t_{\alpha/2}$ $-t_{\alpha/2}$ R.R.
of H$_o$ of H$_o$

-2.678   2.678

$t_{\alpha/2} = -t_{1-\alpha/2}$

$t_{\alpha/2} = t_{0.005} = -2.678$
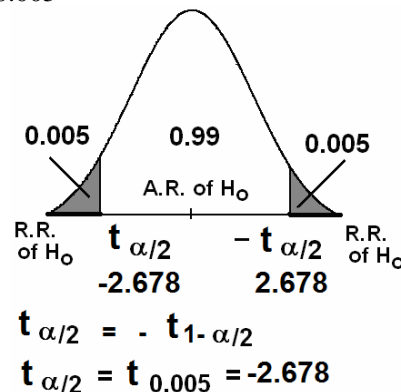
Decision:

Since t$= -2.83 \in$R.R., i.e., t$= -2.83 < $ t$_{0.005}$, we reject H$_o$:μ=8 at α=0.01 and accept H$_A$:μ≠8. Therefore, we conclude that the claim is not correct.

Note:

For the case of non-normal population with unknown variance, and when the sample size is large (n≥30), we may use the following test statistic:

$$Z = \frac{\overline{X} - \mu_o}{S / \sqrt{n}} \sim N(0,1)$$

That is, we replace the population standard deviation (σ) by the sample standard deviation (S), and we conduct the test as described for the first case.

7.3 Hypothesis Testing: The Difference Between Two Population Means: (Independent Populations)

Suppose that we have two (independent) populations:

- 1-st population with mean μ_1 and variance σ_1^2
- 2-nd population with mean μ_2 and variance σ_2^2
- We are interested in comparing μ_1 and μ_2, or equivalently, making inferences about the difference between the means $(\mu_1-\mu_2)$.
- We <u>independently</u> select a random sample of size n_1 from the 1-st population and another random sample of size n_2 from the 2-nd population:
- Let $\overline{X}_1$ and S_1^2 be the sample mean and the sample variance of the 1-st sample.
- Let $\overline{X}_2$ and S_2^2 be the sample mean and the sample variance of the 2-nd sample.
- The sampling distribution of $\overline{X}_1 - \overline{X}_2$ is used to make inferences about $\mu_1-\mu_2$.

We wish to test some hypotheses comparing the population means.

Hypotheses:

We choose one of the following situations:

    (i) H_o: $\mu_1 = \mu_2$ against H_A: $\mu_1 \neq \mu_2$

    (ii) H_o: $\mu_1 \geq \mu_2$ against H_A: $\mu_1 < \mu_2$

    (iii) H_o: $\mu_1 \leq \mu_2$ against H_A: $\mu_1 > \mu_2$

or equivalently,

    (i) H_o: $\mu_1-\mu_2 = 0$ against H_A: $\mu_1 - \mu_2 \neq 0$

    (ii) H_o: $\mu_1-\mu_2 \geq 0$ against H_A: $\mu_1 - \mu_2 < 0$

    (iii) H_o: $\mu_1-\mu_2 \leq 0$ against H_A: $\mu_1 - \mu_2 > 0$

Test Statistic:

(1) First Case:

    For normal populations (or non-normal populations with large sample sizes), and if σ_1^2 and σ_2^2 are known, then the test statistic is:

$$Z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

(2) Second Case:

For normal populations, and if σ_1^2 and σ_2^2 are unknown but equal ($\sigma_1^2 = \sigma_2^2 = \sigma^2$), then the test statistic is:

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\dfrac{S_p^2}{n_1} + \dfrac{S_p^2}{n_2}}} \sim t(n_1+n_2-2)$$

where the pooled estimate of σ^2 is

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

and the degrees of freedom of S_p^2 is df= $v=n_1+n_2-2$.

Summary of Testing Procedure:

Hypotheses	$H_o: \mu_1 - \mu_2 = 0$ $H_A: \mu_1 - \mu_2 \neq 0$	$H_o: \mu_1 - \mu_2 \leq 0$ $H_A: \mu_1 - \mu_2 > 0$	$H_o: \mu_1 - \mu_2 \geq 0$ $H_A: \mu_1 - \mu_2 < 0$
Test Statistic For the First Case:	$Z = \dfrac{\overline{X}_1 - \overline{X}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \sim N(0,1)$ {if σ_1^2 and σ_2^2 are known}		
R.R. and A.R. of H_o (For the First Case)			
Test Statistic For the Second Case:	$T = \dfrac{\overline{X}_1 - \overline{X}_2}{\sqrt{\dfrac{S_p^2}{n_1} + \dfrac{S_p^2}{n_2}}} \sim t(n_1+n_2-2)$ {if $\sigma_1^2 = \sigma_2^2 = \sigma^2$ is unknown}		
R.R. and A.R. of H_o (For the Second Case)			
Decision:	Reject H_o (and accept H_A) at the significance level α if:		
	T.S. $\in$ R.R. Two-Sided Test	T.S. $\in$ R.R. One-Sided Test	T.S. $\in$ R.R. One-Sided Test

Example: (σ_1^2 and σ_2^2 are known)

Researchers wish to know if the data they have collected provide sufficient evidence to indicate the difference in mean serum uric acid levels between individuals with Down's syndrome and normal individuals. The data consist of serum uric acid on 12 individuals with Down's syndrome and 15 normal individuals. The sample means are $\overline{X}_1 = 4.5$ mg/100ml and $\overline{X}_2 = 3.4$ mg/100ml. Assume the populations are normal with variances $\sigma_1^2 = 1$ and $\sigma_2^2 = 1.5$. Use significance level $\alpha = 0.05$.

Solution:

μ_1 = mean serum uric acid levels for the individuals with Down's syndrome.

μ_2 = mean serum uric acid levels for the normal individuals.

$n_1 = 12$ $\overline{X}_1 = 4.5$ $\sigma_1^2 = 1$

$n_2 = 15$ $\overline{X}_2 = 3.4$ $\sigma_2^2 = 1.5$.

Hypotheses:

H_o: $\mu_1 = \mu_2$ against H_A: $\mu_1 \neq \mu_2$

or

H_o: $\mu_1 - \mu_2 = 0$ against H_A: $\mu_1 - \mu_2 \neq 0$

Calculation:

$\alpha = 0.05$

$Z_{0.75} = 1.96$ (1st critical value)

$-Z_{0.75} = -1.96$ (2nd critical value)

Test Statistic (T.S.):

$$Z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} = \frac{4.5 - 3.4}{\sqrt{\dfrac{1}{12} + \dfrac{1.5}{15}}} = 2.569$$
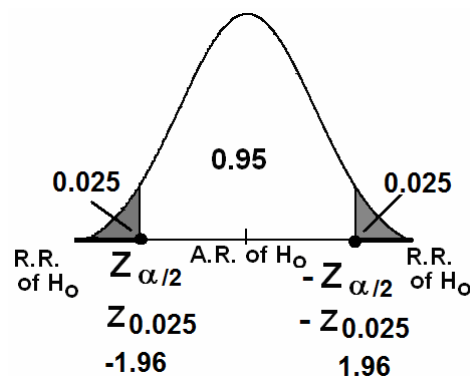
Decision:

Since Z=2.569 $\in$ R.R. we reject H_o: $\mu_1 = \mu_2$ and we accept (do not reject) H_A: $\mu_1 \neq \mu_2$ at $\alpha = 0.05$. Therefore, we conclude that the two population means are not equal.

Notes:

1. We can easily show that a 95% confidence interval for ($\mu_1 - \mu_2$) is (0.26, 1.94), that is:

$$0.26 < \mu_1 - \mu_2 < 1.94$$

Since this interval does not include 0, we say that 0 is not a candidate for the difference between the population means ($\mu_1 - \mu_2$), and we conclude that $\mu_1 - \mu_2 \neq 0$, i.e., $\mu_1 \neq \mu_2$. Thus we arrive at the same conclusion by means of a confidence interval.

2. $P - Value = 2 \times P(Z > | Z_C |)$

$$= 2P(Z > 2.57) = 2[1 - P(Z < 2.57)] = 2(1 - 0.9949) = 0.0102$$

The level of significance was $\alpha = 0.05$.
Since P-value $< \alpha$, we reject H_o.

Example: ($\sigma_1^2 = \sigma_2^2 = \sigma^2$ is unknown)

An experiment was performed to compare the abrasive wear of two different materials used in making artificial teeth. 12 pieces of material 1 were tested by exposing each piece to a machine measuring wear. 10 pieces of material 2 were similarly tested. In each case, the depth of wear was observed. The samples of material 1 gave an average wear of 85 units with a sample standard deviation of 4, while the samples of materials 2 gave an average wear of 81 and a sample standard deviation of 5. Can we conclude at the 0.05 level of significance that the mean abrasive wear of material 1 is greater than that of material 2? Assume normal populations with equal variances.

Solution:

Material 1	material 2
$n_1 = 12$	$n_2 = 10$
$\overline{X}_1 = 85$	$\overline{X}_2 = 81$
$S_1 = 4$	$S_2 = 5$

Hypotheses:
  H_o: $\mu_1 \leq \mu_2$
  H_A: $\mu_1 > \mu_2$
Or equivalently,
  H_o: $\mu_1 - \mu_2 \leq 0$
  H_A: $\mu_1 - \mu_2 > 2$
Calculation:
$\alpha = 0.05$

0.95   0.05

A.R. of H_o $t_{0.05}$ R.R. of H_o
= 1.725
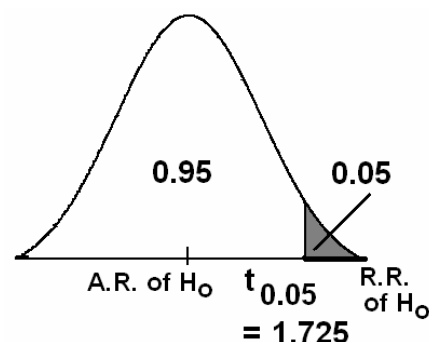
$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} = \frac{(12 - 1)(4)^2 + (10 - 1)(5)^2}{12 + 10 - 2} = 20.05$$

.df $= \nu = n_1 + n_2 - 2 = 12 + 10 - 2 = 20$

$t_{0.05} = 1.725$ (critical value)

Test Statistic (T.S.):

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\dfrac{S_p^2}{n_1} + \dfrac{S_p^2}{n_2}}} = \frac{85 - 81}{\sqrt{\dfrac{20.05}{12} + \dfrac{20.05}{10}}} = 1.04$$

Decision:

Since T=1.04 $\in$ A.R. (T=1.04< $t_{0.05}$ = 1.725), we accept (do not reject) H_o and we reject H_A:$\mu_1 - \mu_2 > 0$ (H_A:$\mu_1 > \mu_2$) at α=0.05. Therefore, we conclude that the mean abrasive wear of material 1 is <u>not</u> greater than that of material 2.

7.4 Paired Comparisons:

- In this section, we are interested in comparing the means of two related (non-independent/dependent) normal populations.
- In other words, we wish to make statistical inference for the difference between the means of two related normal populations.
- Paired t-Test concerns about testing the equality of the means of two related normal populations.

Examples of related populations are:

1. Height of the father and height of his son.
2. Mark of the student in MATH and his mark in STAT.
3. Pulse rate of the patient before and after the medical treatment.
4. Hemoglobin level of the patient before and after the medical treatment.

Example: (effectiveness of a diet program)

Suppose that we are interested in studying the effectiveness of a certain diet program. Let the random variables X and Y are as follows:

X = the weight of the individual before the diet program

Y = the weight of the same individual after the diet program

We assume that the distributions of these random variables are normal with means μ_1 and μ_2, respectively.

These two variables are related (dependent/non-independent) because they are measured on the same individual.

Populations:

1-st population (X): weights before a diet program

mean = μ_1

2-nd population (Y): weights after the diet program

mean = μ_2

Question:

Does the diet program have an effect on the weight?

Answer is:

No if $\mu_1 = \mu_2$ ($\mu_1 - \mu_2 = 0$)

Yes if $\mu_1 \neq \mu_2$ ($\mu_1 - \mu_2 \neq 0$)

Therefore, we need to test the following hypotheses:

Hypotheses:

H_o: $\mu_1 = \mu_2$ (H_o: the diet program has no effect on weight)

H_A: $\mu_1 \neq \mu_2$ (H_A: the diet program has an effect on weight)

Equivalently we may test:

H_o: $\mu_1 - \mu_2 = 0$

H_A: $\mu_1 - \mu_2 \neq 0$

Testing procedures:

- We select a random sample of n individuals. At the beginning of the study, we record the individuals' weights before the diet program (X). At the end of the diet program, we record the individuals' weights after the program (Y). We end up with the following information and calculations:

Individual i	Weight before X_i	Weight after Y_i	Difference $D_i = X_i - Y_i$
1	X_1	Y_1	$D_1 = X_1 - Y_1$
2	X_2	Y_2	$D_2 = X_2 - Y_2$
.	.	.	
.	.	.	

Individual i	Weight before X_i	Weight after Y_i	Difference $D_i = X_i - Y_i$
.	.	.	
n	X_n	Y_n	$D_n = X_n - Y_n$

- Hypotheses:

  H_o: the diet program has no effect on weight

  H_A: the diet program has an effect on weight

  Equivalently,

  H_o: $\mu_1 = \mu_2$

  H_A: $\mu_1 \neq \mu_2$

  Equivalently,

  H_o: $\mu_1 - \mu_2 = 0$

  H_A: $\mu_1 - \mu_2 \neq 0$

  Equivalently,

  H_o: $\mu_D = 0$

  H_A: $\mu_D \neq 0$

  where:

$$\mu_D = \mu_1 - \mu_2$$

- We calculate the following quantities:
  - The differences (D-observations):

    $$D_i = X_i - Y_i \quad (i=1, 2, \ldots, n)$$
  - Sample mean of the D-observations (differences):

$$\overline{D} = \frac{\sum_{i=1}^{n} D_i}{n} = \frac{D_1 + D_2 + \ldots + D_n}{n}$$

  - Sample variance of the D-observations (differences):

$$S_D^2 = \frac{\sum_{i=1}^{n}\left(D_i - \overline{D}\right)^2}{n-1} = \frac{\left(D_1 - \overline{D}\right)^2 + \left(D_2 - \overline{D}\right)^2 + \cdots + \left(D_n - \overline{D}\right)^2}{n-1}$$

  - Sample standard deviation of the D-observations:

$$S_D = \sqrt{S_D^2}$$

- Test Statistic:

  We calculate the value of the following test statistic:

$$t = \frac{\overline{D}}{S_D / \sqrt{n}} \sim t(n-1)$$

This statistic has a t-distribution with df $= \nu = n-1$.

- Rejection Region of H_o:

Critical values are: $t_{\alpha/2}$ and $t_{1-\alpha/2} = -t_{\alpha/2}$.
The rejection region (critical region) at the significance level α is:
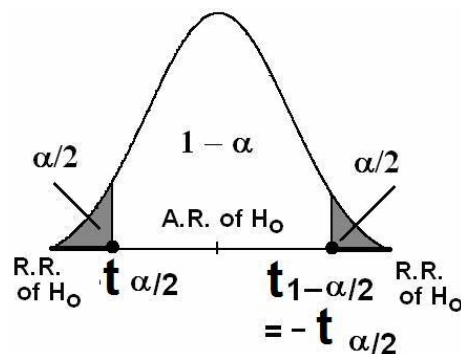
$$t < t_{\alpha/2} \quad \text{or} \quad t > t_{1-\alpha/2} = -t_{\alpha/2}$$

- Decision:

We reject H_o and accept H_A at the significance level α if $T \in$ R.R. ,i.e., if:

$$t < t_{\alpha/2} \quad \text{or} \quad t > t_{1-\alpha/2} = -t_{\alpha/2}$$

Numerical Example:

In the previous example, suppose that the sample size was 10 and the data were as follows:

Individual (i)	1	2	3	4	5	6	7	8	9	10
Weight before (X_i)	86.6	80.2	91.5	80.6	82.3	81.9	88.4	85.3	83.1	82.1
Weight after (Y_i)	79.7	85.9	81.7	82.5	77.9	85.8	81.3	74.7	68.3	69.7

Does these data provide sufficient evidence to allow us to conclude that the diet program is effective? Use $\alpha=0.05$ and assume that the populations are normal.

Solution:

μ_1 = the mean of weights before the diet program
μ_2 = the mean of weights after the diet program
Hypotheses:

H_o: $\mu_1 = \mu_2$ (H_o: the diet program is not effective)

$H_A: \mu_1 \neq \mu_2$ (H_A: the diet program is effective)

Equivalently,

$H_o: \mu_D = 0$

$H_A: \mu_D \neq 0$ (where: $\mu_D = \mu_1 - \mu_2$)

Calculations:

i	X_i	Y_i	$D_i = X_i - Y_i$
1	86.6	79.7	6.9
2	80.2	85.9	-5.7
3	91.5	81.7	9.8
4	80.6	82.5	-1.9
5	82.3	77.9	4.4
6	81.9	85.8	-3.9
7	88.4	81.3	7.1
8	85.3	74.7	10.6
9	83.1	68.3	14.8
10	82.1	69.7	12.4
sum	$\sum X = 842$	$\sum Y = 787.5$	$\sum D = 54.5$

$$\overline{D} = \frac{\sum_{i=1}^{n} D_i}{n} = \frac{54.5}{10} = 5.45$$

$$S_D^2 = \frac{\sum_{i=1}^{n}(D_i - \overline{D})^2}{n-1} = \frac{(6.9 - 5.45)^2 + ... + (12.4 - 5.45)^2}{10-1} = 50.3283$$

$$S_D = \sqrt{S_D^2} = \sqrt{50.3283} = 7.09$$

Test Statistic:

$$t = \frac{\overline{D}}{S_D / \sqrt{n}} = \frac{5.45}{7.09 / \sqrt{10}} = 2.431$$

Degrees of freedom:

$$df = \nu = n-1 = 10-1 = 9$$

Significance level: $\alpha = 0.05$

Rejection Region of H_o:

Critical values: $t_{0.025} = -2.262$ and $t_{0.975} = -t_{0.025} = 2.262$

Critical Region: $t < -2.262$ or $t > 2.262$
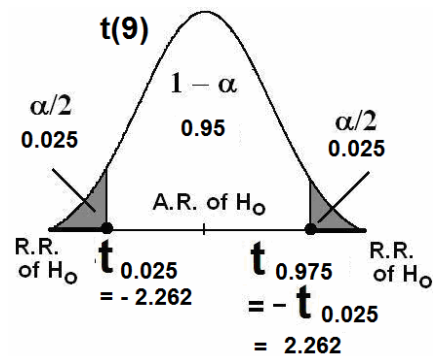
Decision:

Since t= 2.43 ∈R.R., i.e., t=2.43 > $t_{0.975}$=−$t_{0.025}$=2.262, we reject:

   H_o: $\mu_1 = \mu_2$ (the diet program is not effective)

and we accept:

   H_1: $\mu_1 \neq \mu_2$ (the diet program is effective)

Consequently, we conclude that the diet program is effective at α=0.05.

Note:

- The sample mean of the weights before the program is $\overline{X} = 84.2$

- The sample mean of the weights after the program is $\overline{Y} = 78.75$

- Since the diet program is effective and since $\overline{X} = 84.2 > \overline{Y} = 78.75$, we can conclude that the program is effective in reducing the weight.

Confidence Interval for the Difference between the Means of Two Related Normal Populations ($\mu_D = \mu_1 - \mu_2$):

   In this section, we consider constructing a confidence interval for the difference between the means of two related (non-independent) normal populations. As before, let us define the difference between the two means as follows:

$$\mu_D = \mu_1 - \mu_2$$

where μ_1 is the mean of the first population and μ_2 is the mean of the second population. We assume that the two normal populations are not independent.

Result:

A $(1-\alpha)$100% confidence interval for $\mu_D = \mu_1 - \mu_2$ is:

$$\overline{D} \pm t_{1-\frac{\alpha}{2}} \frac{S_D}{\sqrt{n}}$$

$$\overline{D} - t_{1-\frac{\alpha}{2}} \frac{S_D}{\sqrt{n}} < \mu_D < \overline{D} + t_{1-\frac{\alpha}{2}} \frac{S_D}{\sqrt{n}}$$

where:

$$\overline{D} = \frac{\sum_{i=1}^{n} D_i}{n} \quad, \quad S_D = \sqrt{S_D^2}, \quad S_D^2 = \frac{\sum_{i=1}^{n}(D_i - \overline{D})^2}{n-1}, \quad df = \nu = n-1.$$

Example:

Consider the data given in the previous numerical example:

Individual (i)	1	2	3	4	5	6	7	8	9	10
Weight before (X$_i$)	86.6	80.2	91.5	80.6	82.3	81.9	88.4	85.3	83.1	82.1
Weight after (Y$_i$)	79.7	85.9	81.7	82.5	77.9	85.8	81.3	74.7	68.3	69.7

Find a 95% confidence interval for the difference between the mean of weights before the diet program (μ_1) and the mean of weights after the diet program (μ_2).

Solution:

We need to find a 95% confidence interval for $\mu_D = \mu_1 - \mu_2$:

$$\overline{D} \pm t_{1-\frac{\alpha}{2}} \frac{S_D}{\sqrt{n}}$$

We have found:

$$\overline{D} = 5.45 \quad, \quad S_D^2 = 50.3283 \quad, \quad S_D = \sqrt{S_D^2} = 7.09$$

The value of the reliability coefficient $t_{1-\frac{\alpha}{2}}$ ($df = \nu = n-1 = 9$) is

$$t_{1-\frac{\alpha}{2}} = t_{0.975} = 2.262.$$

Therefore, a 95% confidence interval for $\mu_D = \mu_1 - \mu_2$ is

$$5.45 \pm (2.262) \frac{7.09}{\sqrt{10}}$$

$$5.45 \pm 5.0715$$

$$0.38 < \mu_D < 10.52$$

$$0.38 < \mu_1 - \mu_2 < 10.52$$

We are 95% confident that $\mu_D = \mu_1 - \mu_2 \in (0.38, 10.52)$.

Note: Since this interval does not include 0, we say that 0 is not a candidate for the difference between the population means $(\mu_1 - \mu_2)$, and we conclude that $\mu_1 - \mu_2 \neq 0$, i.e., $\mu_1 \neq \mu_2$. Thus we arrive at the same conclusion by means of a confidence interval.

7.5 Hypothesis Testing: A Single Population Proportion (p):

In this section, we are interested in testing some hypotheses about the population proportion (p).
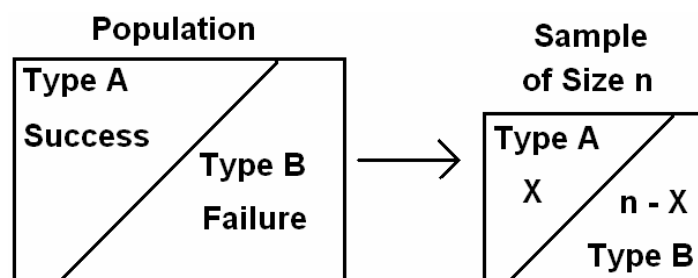
Recall:
- $p =$ Population proportion of elements of Type A in the population

$$p = \frac{no.\ of\ elements\ of\ type\ A\ in\ the\ population}{Total\ no.\ of\ elements\ in\ the\ population}$$

$$p = \frac{A}{N} \quad (N = population\ size)$$

- $.n =$ sample size
- $X =$ no. of elements of type A in the sample of size n.
- $\hat{p} =$ Sample proportion elements of Type A in the sample
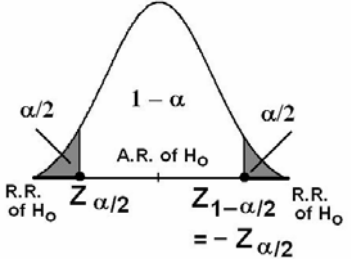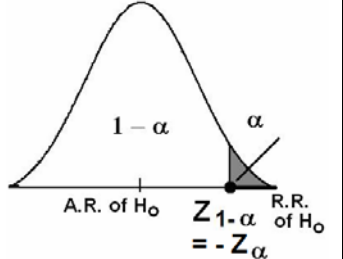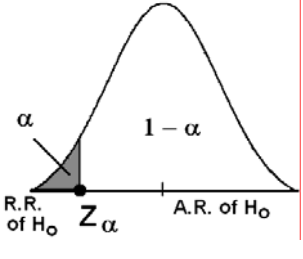
$$\hat{p} = \frac{no.\ of\ elements\ of\ type\ A\ in\ the\ sample}{no.\ of\ elements\ in\ the\ sample}$$

$$\hat{p} = \frac{X}{n} \quad (n = \text{sample size} = \text{no. of elements in the sample})$$

- $\hat{p}$ is a "good" point estimate for p.
- For large n, ($n \geq 30$, $np > 5$), we have

$$Z = \frac{\hat{p} - p}{\sqrt{\dfrac{p(1-p)}{n}}} \sim N(0,1)$$

- Let p_o be a given known value.
- Test Procedure:

Hypotheses	$H_o: p = p_o$ $H_A: p \neq p_o$	$H_o: p \leq p_o$ $H_A: p > p_o$	$H_o: p \geq p_o$ $H_A: p < p_o$
Test Statistic (T.S.)	$Z = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}} \sim N(0,1)$		
R.R. & A.R. of H_o			
Decision:	Reject H_o (and accept H_A) at the significance level α if:		
	$Z < Z_{\alpha/2}$ or $Z > Z_{1-\alpha/2} = -Z_{\alpha/2}$ Two-Sided Test	$Z > Z_{1-\alpha} = -Z_{\alpha}$ One-Sided Test	$Z < Z_{\alpha}$ One-Sided Test

Example:

A researcher was interested in the proportion of females in the population of all patients visiting a certain clinic. The researcher claims that 70% of all patients in this population are females. Would you agree with this claim if a random survey shows that 24 out of 45 patients are females? Use a 0.10 level of significance.

Solution:

p = Proportion of female in the population.

.n=45 (large)

X= no. of female in the sample = 24

$\hat{p}$ = proportion of females in the sample

$$\hat{p} \; = \; \frac{X}{n} = \frac{24}{45} = 0.5333$$

$$p_o = \frac{70}{100} = 0.7$$

$$\alpha = 0.10$$

Hypotheses:

H_o: $p = 0.7$ ($p_o = 0.7$)

H_A: $p \neq 0.7$

Level of significance:

$\alpha = 0.10$

Test Statistic (T.S.):

$$Z = \frac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1 - p_0)}{n}}}$$

$$= \frac{0.5333 - 0.70}{\sqrt{\dfrac{(0.7)(0.3)}{45}}} = -2.44$$

Rejection Region of H_o (R.R.):

Critical values:

$Z_{\alpha/2} = Z_{0.05} = -1.645$

$-Z_{\alpha/2} = -Z_{0.05} = 1.645$

We reject H_o if:

$Z < Z_{\alpha/2} = Z_{0.05} = -1.645$

or

$Z > -Z_{\alpha/2} = -Z_{0.05} = 1.645$
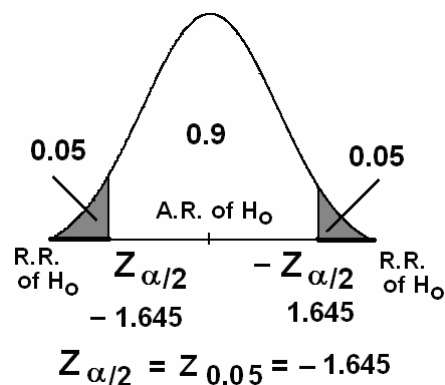
Decision:

Since Z= −2.44 ∈Rejection Region of H_o (R.R), we reject

H_o:p=0.7 and accept H_A:$p \neq$0.7 at α=0.1. Therefore, we do not agree with the claim stating that 70% of the patients in this population are females.

Example:

In a study on the fear of dental care in a certain city, a survey showed that 60 out of 200 adults said that they would hesitate to take a dental appointment due to fear. Test whether the proportion of adults in this city who hesitate to take dental appointment is less than 0.25. Use a level of significance of 0.025.

Solution:

p = Proportion of adults in the city who hesitate to take a dental appointment.

.n= 200 (large)

X= no. of adults who hesitate in the sample = 60

$\hat{p}$ = proportion of adults who hesitate in the sample

$$\hat{p} = \frac{X}{n} = \frac{60}{200} = 0.3$$

p_o=0.25

α=0.025

Hypotheses:

H$_o$: $p \geq 0.25$ (p_o=0.25)

H$_A$: $p < 0.25$ (research hypothesis)

Level of significance:

α=0.025

Test Statistic (T.S.):

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}} = \frac{0.3 - 0.25}{\sqrt{\frac{(0.25)(0.75)}{200}}} = 1.633$$

Rejection Region of H$_o$ (R.R.):

Critical value: $Z_\alpha = Z_{0.025} = -1.96$

Critical Region:

We reject H$_o$ if: $Z < Z_\alpha = Z_{0.025} = -1.96$
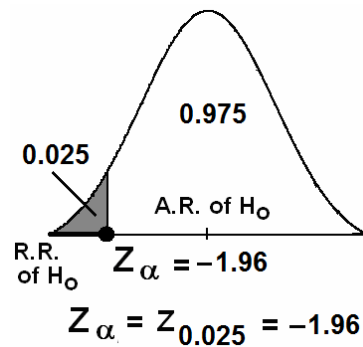
Decision:

Since $Z = 1.633 \in$ Acceptance Region of H_o (A.R.), we accept (do not reject) H_o: $p \geq 0.25$ and we reject H_A: $p < 0.25$ at $\alpha = 0.025$. Therefore, we do not agree with claim stating that the proportion of adults in this city who hesitate to take dental appointment is less than 0.25.

7.6 Hypothesis Testing: The Difference Between Two Population Proportions ($p_1 - p_2$):

In this section, we are interested in testing some hypotheses about the difference between two population proportions ($p_1 - p_2$).
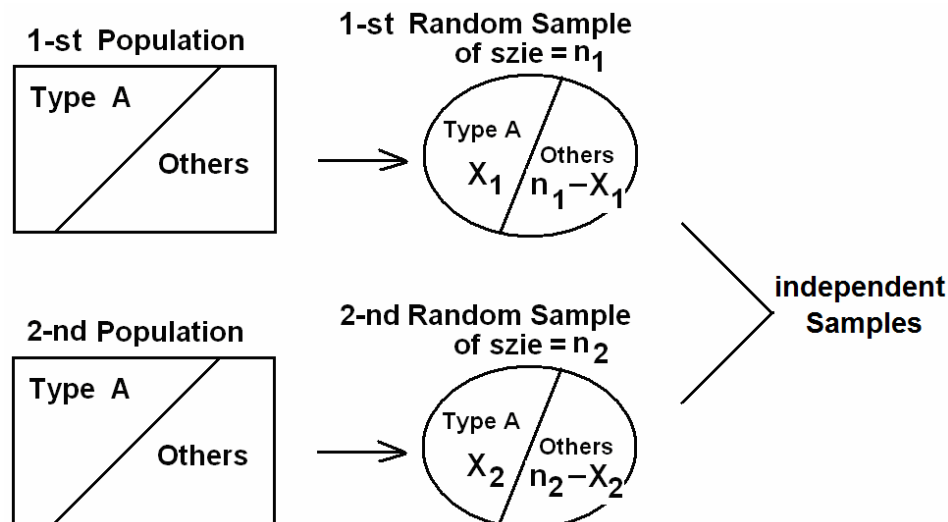
Suppose that we have two populations:

- $.p_1$ = population proportion of the 1-st population.
- $.p_2$ = population proportion of the 2-nd population.
- We are interested in comparing p_1 and p_2, or equivalently, making inferences about $p_1 - p_2$.
- We independently select a random sample of size n_1 from

the 1-st population and another random sample of size n_2 from the 2-nd population:

- Let X_1 = no. of elements of type A in the 1-st sample.
- Let X_2 = no. of elements of type A in the 2-nd sample.
- $\hat{p}_1 = \dfrac{X_1}{n_1}$ = the sample proportion of the 1-st sample
- $\hat{p}_2 = \dfrac{X_2}{n_2}$ = the sample proportion of the 2-nd sample
- The sampling distribution of $\hat{p}_1 - \hat{p}_2$ is used to make inferences about $p_1 - p_2$.
- For large n_1 and n_2, we have

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\dfrac{p_1 \, q_1}{n_1} + \dfrac{p_2 \, q_2}{n_2}}} \sim N(0,1) \quad \text{(Approximately)}$$

- $q = 1 - p$

Hypotheses:
We choose one of the following situations:

    (i) H_o: $p_1 = p_2$ against H_A: $p_1 \neq p_2$
    (ii) H_o: $p_1 \geq p_2$ against H_A: $p_1 < p_2$
    (iii) H_o: $p_1 \leq p_2$ against H_A: $p_1 > p_2$

or equivalently,

    (i) H_o: $p_1 - p_2 = 0$ against H_A: $p_1 - p_2 \neq 0$
    (ii) H_o: $p_1 - p_2 \geq 0$ against H_A: $p_1 - p_2 < 0$
    (iii) H_o: $p_1 - p_2 \leq 0$ against H_A: $p_1 - p_2 > 0$

Note, under the assumption of the equality of the two population proportions (H_o: $p_1 = p_2 = p$), the pooled estimate of the common proportion p is:

$$\overline{p} = \frac{X_1 + X_2}{n_1 + n_2} \qquad\qquad (\overline{q} = 1 - \overline{p})$$

The test statistic (T.S.) is

$$Z = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\dfrac{\overline{p}(1-\overline{p})}{n_1} + \dfrac{\overline{p}(1-\overline{p})}{n_2}}} \sim N(0,1)$$

Testing Procedure:

Hypotheses	$H_o: p_1 - p_2 = 0$ $H_A: p_1 - p_2 \neq 0$	$H_o: p_1 - p_2 \leq 0$ $H_A: p_1 - p_2 > 0$	$H_o: p_1 - p_2 \geq 0$ $H_A: p_1 - p_2 < 0$
Test Statistic (T.S.)	$Z = \dfrac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\dfrac{\overline{p}(1-\overline{p})}{n_1} + \dfrac{\overline{p}(1-\overline{p})}{n_2}}} \sim N(0,1)$		
R.R. and A.R. of H_o			
Decision:	Reject H_o (and accept H_1) at the significance level α if $Z \in$ R.R.:		
Critical Values	$Z > Z_{\alpha/2}$ or $Z < -Z_{\alpha/2}$ Two-Sided Test	$Z > Z_{\alpha}$ One-Sided Test	$Z < -Z_{\alpha}$ One-Sided Test

Example:

In a study about the obesity (overweight), a researcher was interested in comparing the proportion of obesity between males and females. The researcher has obtained a random sample of 150 males and another independent random sample of 200 females. The following results were obtained from this study.

	n	Number of obese people
Males	150	21
Females	200	48

Can we conclude from these data that there is a difference between the proportion of obese males and proportion of obese females? Use $\alpha = 0.05$.

Solution:

$.p_1$ = population proportion of obese males
$.p_2$ = population proportion of obese females
$\hat{p}_1$ = sample proportion of obese males
$\hat{p}_2$ = sample proportion of obese females

<div style="text-align:center">

Males   Females
$n_1 = 150$ $n_2 = 200$
$X_1 = 21$ $X_2 = 48$

</div>

$$\hat{p}_1 = \frac{X_1}{n_1} = \frac{21}{150} = 0.14 \qquad \hat{p}_2 = \frac{X_2}{n_2} = \frac{48}{200} = 0.24$$

The pooled estimate of the common proportion p is:

$$\bar{p} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{21 + 48}{150 + 200} = 0.197$$

Hypotheses:

H_o: $p_1 = p_2$
H_A: $p_1 \neq p_2$

or

H_o: $p_1 - p_2 = 0$
H_A: $p_1 - p_2 \neq 0$

Level of significance: $\alpha = 0.05$
Test Statistic (T.S.):

$$Z = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\dfrac{\bar{p}(1-\bar{p})}{n_1} + \dfrac{\bar{p}(1-\bar{p})}{n_2}}} = \frac{(0.14 - 0.24)}{\sqrt{\dfrac{0.197 \times 0.803}{150} + \dfrac{0.197 \times 0.803}{200}}} = -2.328$$

Rejection Region (R.R.) of H_o:

Critical values:

$Z_{\alpha/2} = Z_{0.025} = -1.96$
$Z_{1-\alpha/2} = Z_{0975} = 1.96$

Critical region:

Reject H_o if: $Z < -1.96$ or $Z > 1.96$
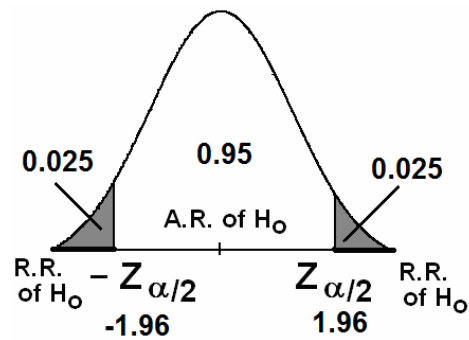
Decision:

Since Z= −2.328 ∈R.R., we reject H_o: $p_1 = p_2$ and accept H_A: $p_1 \neq p_2$ at α=0.05. Therefore, we conclude that there is a difference between the proportion of obese males and the proportion of obese females. Additionally, since, $\hat{p}_1 = 0.14$ < $\hat{p}_2 = 0.24$, we may conclude that the proportion of obesity for females is larger than that for males.

TABLE OF CONTENTS

TABLE OF CONTENTS

TABLE OF CONTENTS