King Saud University
College of Science
Department of Statistics & OR

STAT 109

BIOSTATISTICS

Outline of the course

Course Code & No:	STAT 109
Course Name	Biostatistics
Credits:	2 (2+0)
Pre-requisite:	None

Instructor: **Sana Abunasrah, Office :3rd floor ,3A03,Building 25**

Email: **Sabunasrah@ksu.edu.sa**

Week	Title
W1	Introduction to bio-statistics (1.1-1.4).
W2	Types of data and graphical representation (1.1-1.4).
W3	Descriptive statistics: measures of Central tendency- mean, median, mode (2.1 -2 .6 Excluding stem plot percentiles).
W4	Measures of dispersion-range, standard deviation, coefficient of variation (2.1 - 2.6 Excluding stem plot percentiles).
W5	Calculating measures from an ungrouped frequency table (2.1 - 2.6 Excluding stem plot percentiles).
W6	Basic probability, conditional probability, concept of independence, sensitivity, specificity (3.1 -3.6).
W7	Bayes theorem for predictive probabilities (3.1-3.6).
	First Midterm (30%- Sections 1.1 – 3.6)
W8	Some discrete probability distributions: cumulative probability (4.1-4.4).
W9	Binomial and Poisson -their mean and variance (4.1 - 4.4 Excluding the use of binomial and Poisson tables).
W10	Continuous probability distributions: Normal distribution (4.5 - 4.8).
W11	Standard normal distribution and t-distributions (4.5-4.8).
W12	Sampling with and without replacement, sampling distribution of one and two sample means and one and two proportions (5.1 - 5.7 Excluding sampling without replacement).
	Second Midterm (30% - Sections 4.5 – 5.7)
W13	Statistical inference: Point and interval estimation, Type of errors, Concept of P-value (6.2 - 6.6, 7.1 - 7.6 Excluding variances not equal page 181-182).
W14	Testing hypothesis about one and two samples means and proportions including paired data – differentcases under normality. (6.2 -6.6, 7.1 - 7.6, Excluding variances not equal page 181-182).
	Final Exam (40% - chapter 6 & 7)

Course Delivery	Lectures
Grading Policy	First Midterm:. 1: 30 Marks (1.5 hours)
	Second Midterm :30 Marks (1.5 hours)
	Final exam :40 Marks (2 hours)
Text Book	Foundations of Biostatistics by Islam, M. Ataharul, Al-Shiha, Abdullah

Marking scheme:

Mid1 exam : 30% (1.5 hours)

Mid2 exam : 30% (1.5 hours)

Final exam : 40% (2 hours)

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 bechanged 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 ofsitting the final exam.

Further References:

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

CHAPTER 1: Getting Acquainted with Biostatistics

1.1 Introduction:

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: nationalities, occupations, …).

Statistics: (1) Collection, organization, summarization, and analysis of data. (Descriptive Statistics)
(2) 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 = 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.

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.

Sample Size (*n*):

The number of elements in the sample is called the sample size and is denoted by ***n***.

Example: Suppose that we are interested in the weights of students enrolled in the college of engineering at KSU. If we are randomly select 50 students from engineering college at KSU and measure their weights. **Identify the population and the sample in the study?**

*The **population** consists of the weights of all of these students, and our variable of interest is the weight.*
*The weights of these 50 students forms a **sample**.*

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

(a) Discrete Variables:

There are jumps or gaps between the values.

Examples: - Family size ($x = 1, 2, 3, \dots$)

- Number of patients ($x = 0, 1, 2, 3, \dots$)

(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 ($10 < x < 15$)

- hemoglobin level (g\dl)

- Blood type
- Nationality
- Students Grades
- Educational level

(a) Nominal Qualitative Variables:

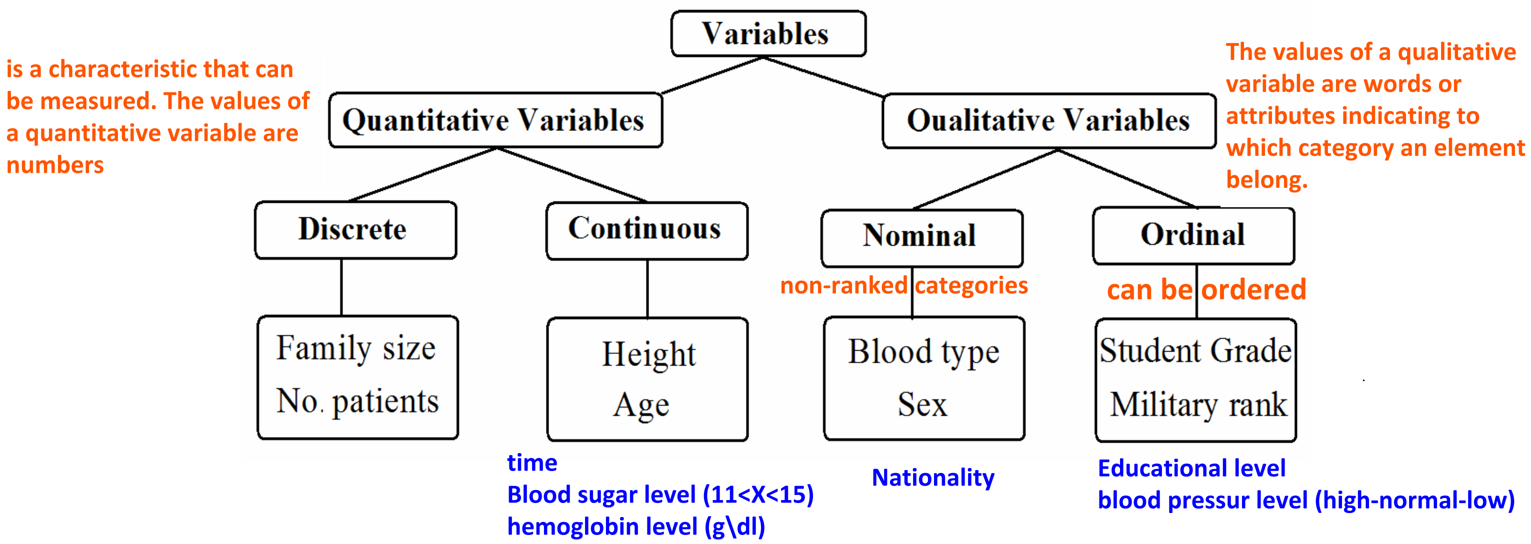
A nominal variable classifies the observations into various mutually exclusive and collectively non-ranked categories.

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

Examples:
- Blood pressure level (high- normal- low)
- Educational level (elementary, intermediate, …)
- Students grade (A, B, C, D, F)
- Military rank



1.4 Sampling and Statistical Inference:

(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.

Note: Explanation of (level) in the variables

- blood pressure level: *ordinal qualitative variable*

- blood sugar level(10< x <15): *continuous quantitative variable*

- hemoglobin level (g\dl): *continuous quantitative variable*

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:

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: Study of the hemoglobin level (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

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

Variable = X = hemoglobin level (continuous, quantitative)
Sample size = *n* = 50
Min= 13.5 Max= 18.3

Class Interval	Tally	Frequency
13.0 – 13.9	|||	3
14.0 – 14.9	𝍸	5
15.0 – 15.9	𝍸 𝍸 𝍸	15
16.0 – 16.9	𝍸 𝍸 𝍸 | –	16
17.0 – 17.9	𝍸 𝍸	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

what is the variable?
what is the type of variable ?

what is the sample size ?
you will probably face missing value in the Frq ? if i remove 10
how can i find missing value ?

Notes:
1. Minimum value ∈ first interval.
2. Maximum value ∈ 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:

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

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

(12.95+13.95)/2=13.45

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%
	n=50		total =1		total = 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%
The percentage of people whose hemoglobin levels are less than 14.0 = 6%

The percentage of people whose hemoglobin levels are more than 16.9 = 22%
The percentage of people whose hemoglobin levels are more than or equal 16= 54%

Displaying Grouped Frequency Distributions:

For representing frequencies, we may use one of the following graphs:

- The Histogram
- The Frequency Polygon

Example: 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		

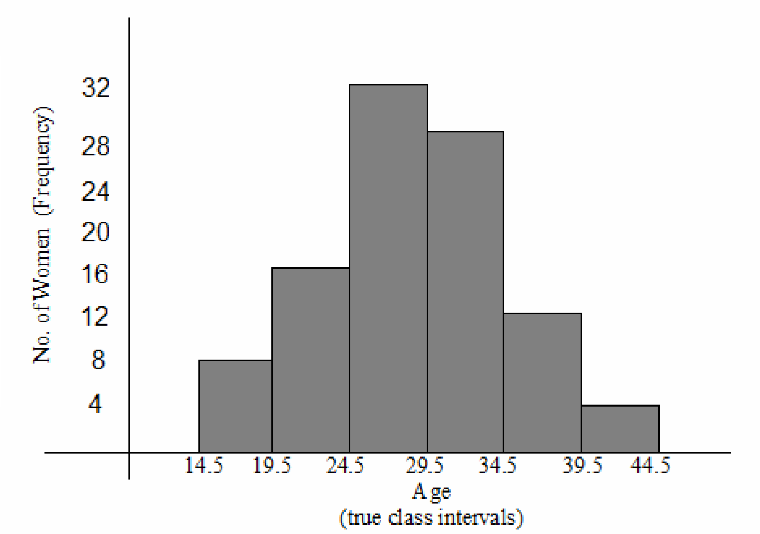
Width of the interval:

True Class Interval >> W =true upper limit – true lower limit = 19.5 – 14.5 = 5

Mid-points & Class interval & True Class W= lower limit - lower limit of the preceding interval

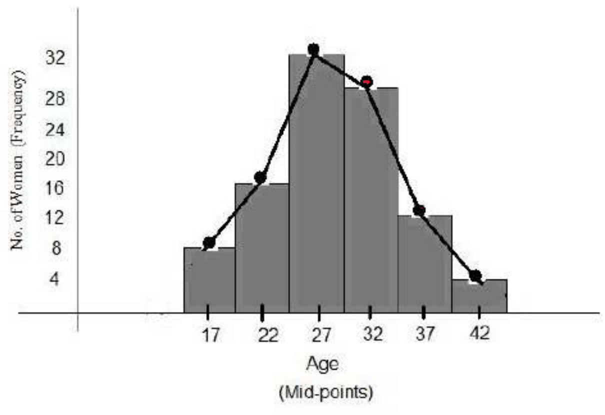
(1) Histogram: Organizing and Displaying Data using Histogram:

frequency or percentage frequency or relative frequency

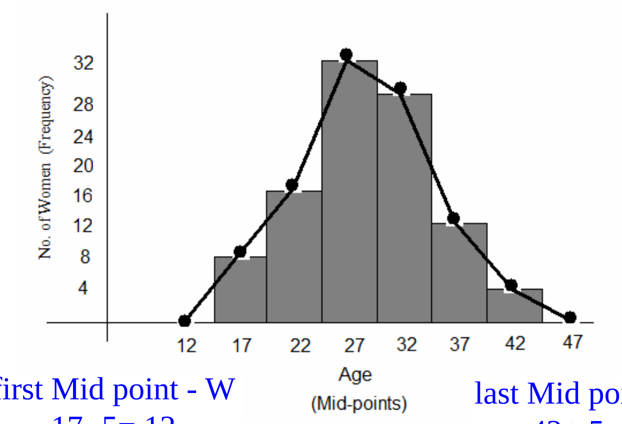


(2) Frequency Polygon: Organizing and Displaying Data using Polygon:

Polygon (Open)

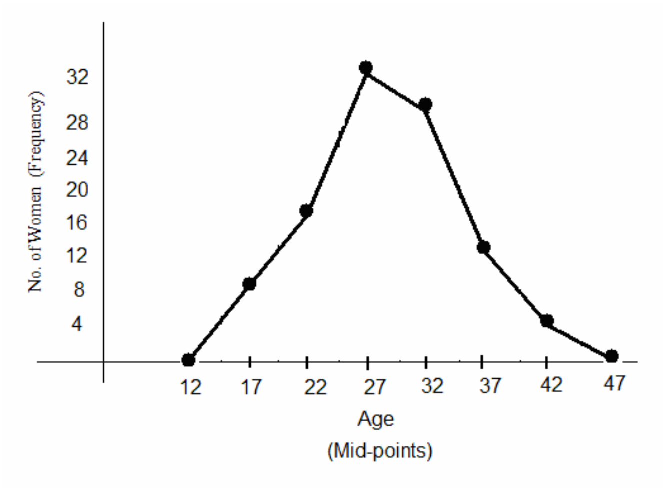


Polygon (Closed)



first Mid point - W
17- 5= 12

last Mid point + W
42+ 5= 47



To calculate width :

First method:

width = W= lower limit- lower limit of the previous interval

width =W= Mid point - Mid point of the previous interval.

Second method:

W=True upper limit - True lower limit

2.4 Descriptive Statistics: Measures of Central Tendency:

- Measures of Central Tendency (or location)

  Mean ; Mode ; Median

- Measures of Dispersion (or Variation)

  Range ; Variance ; Standard Deviation ; Coefficient of Variation

We 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:

$$\boxed{\mu = \frac{X_1 + X_2 + \cdots + X_N}{N} = \frac{\sum_{i=1}^{N} X_i}{N}} \quad \text{(unit)}$$

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:

$$\boxed{\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}} \quad \text{(unit)}$$

- 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 :

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

The population mean is: $\mu = \frac{41+30+35+22+27}{5} = \frac{155}{5} = 31$ (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:

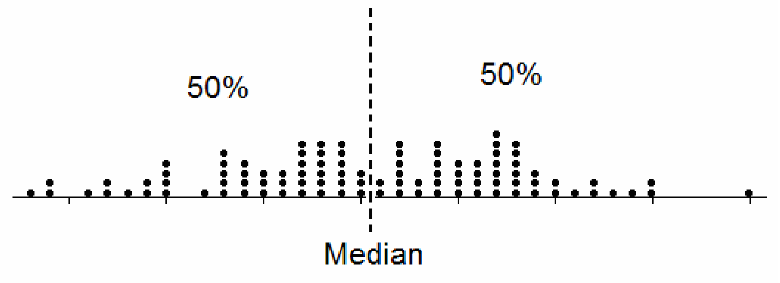
- Extreme values have an influence on the mean. Therefore, the mean may be distorted by extreme values.

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:

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

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 <u>even</u> 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 **<u>ordered</u>** observations:

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

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.

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

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.

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

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 = $= \dfrac{20+32}{2} = \dfrac{52}{2} = 26$ (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. with odd sample size

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 no 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:

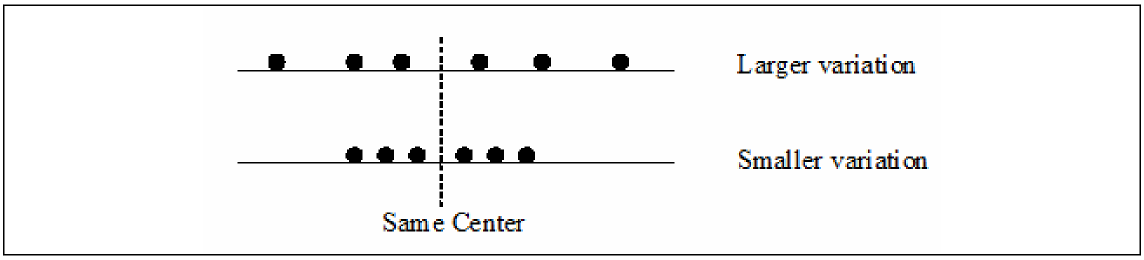
- The mode is not a "good" measure of location, because it depends on a few values of the data.
- The mode does not take into account all values of the sample.
- There might be no mode for a data set.
- There might be more than one mode for a data set.

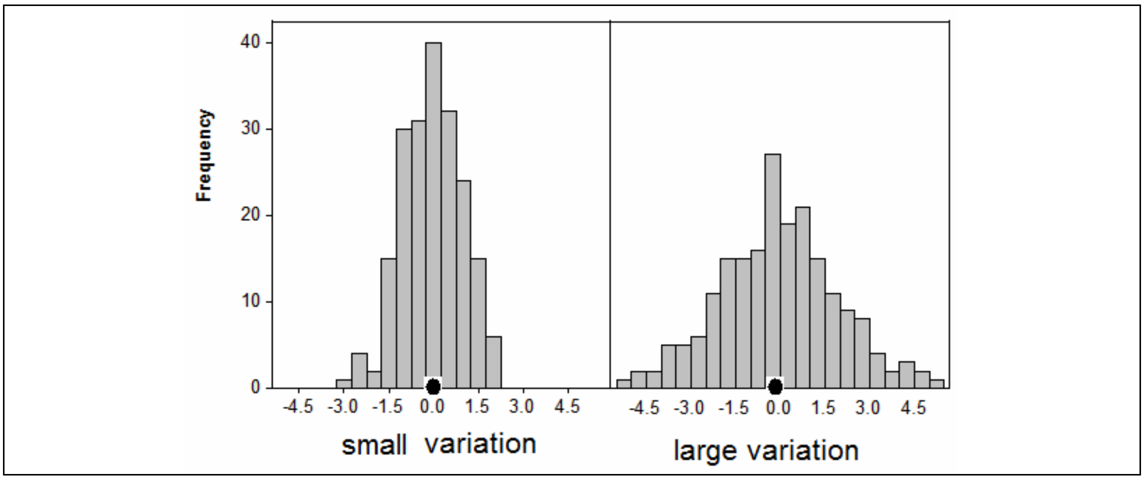
2.6 Descriptive Statistics: Measures of Dispersion (Measures of Variation):

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.





<u>The Range:</u>

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:

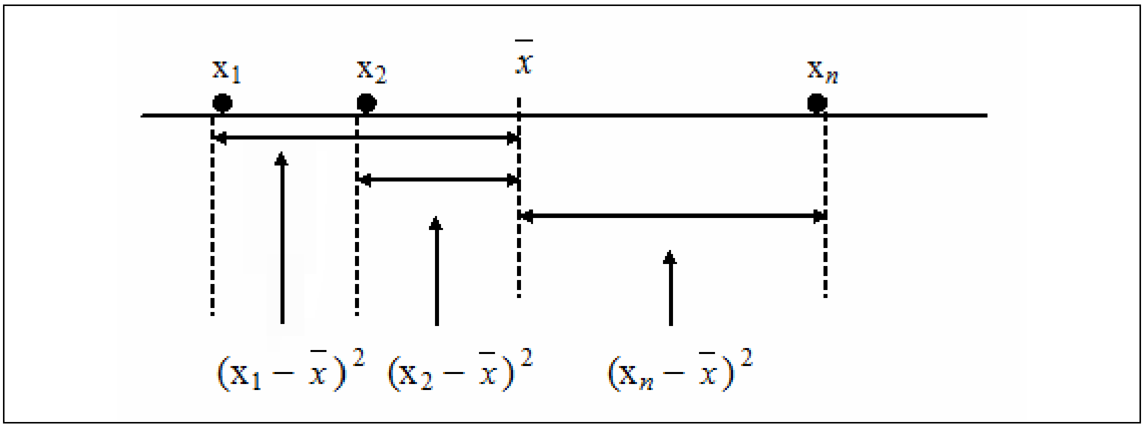
$$x_i - \bar{x}$$

The squared deviation is:

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

The sum of squared deviations is: $\boxed{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

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



(1) The Population Variance σ^2 : "Sigma Square"

(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:

$$\boxed{\sigma^2 = \frac{\sum_{i=1}^{N}(X_i-\mu)^2}{N} = \frac{(X_1-\mu)^2+(X_2-\mu)^2+\cdots+(X_N-\mu)^2}{N} \quad (unit)^2}$$

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:

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$
$$= \frac{(x_1-\bar{x})^2 + (x_2-\bar{x})^2 + \cdots + (x_n-\bar{x})^2}{n-1} \qquad (unit)^2$$
$$= \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1} \qquad \text{(practical formula)}$$

where $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ 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.
- $S^2 \ge 0$
- $S^2 = 0 \quad \Leftrightarrow$ all observation have the same value
  $\Leftrightarrow$ 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: $n = 5$

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

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

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

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

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

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

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)

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

Example:

For the previous example, the sample standard deviation is

$$S = \sqrt{S^2} = \sqrt{376.7} = 19.41 \qquad \text{(unit)}$$

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:

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

- 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.

Descriptive statistics

Measures of

Central tendency or (location)

- Mean (unit)
- Median (unit)
- Mode (unit)

Dispersion or (Variation)

- Range (unit)
- Variance = $(standard\ deviation)^2$ $(unit)^2$
- Standard deviation =$\sqrt{variance}$ (unit)
- Coefficient of variation C.V. $= \frac{S}{\bar{x}} \times 100$ (unit-less)

Population:

$X_1, X_2, \ldots, X_N$

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

Sample:

$x_1, x_2, \ldots, x_n$

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

	Population	Sample
Size	N	n
Mean	$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
Variance	$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}, \quad \sigma^2 \geq 0$	$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}, \quad S^2 \geq 0$
Standard deviation	$\sigma = \sqrt{\sigma^2}, \quad \sigma \geq 0$	$S = \sqrt{S^2}, \quad S \geq 0$

Chapter 3: Probability The Basis of 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. called "Omega"

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: 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:

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

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.

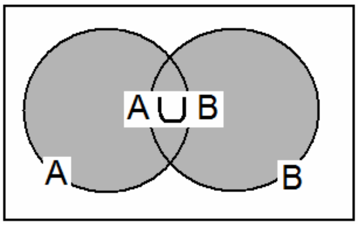
$$\therefore P(E_1) = \frac{3}{6}, \qquad P(E_2) = \frac{3}{6}, \qquad P(E_3) = \frac{2}{6},$$

Some Operations on Events:

Let A and B be two events defined on the sample space Ω.

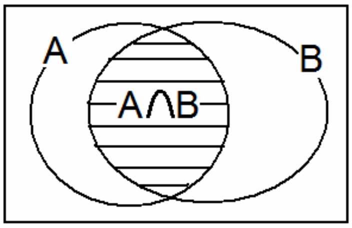
Union of Two events: $(A \cup B)$ or $(A+B)$

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.



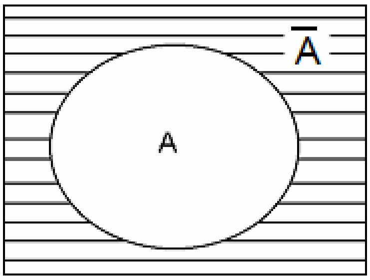
Intersection of Two Events: $(A \cap B)$

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:

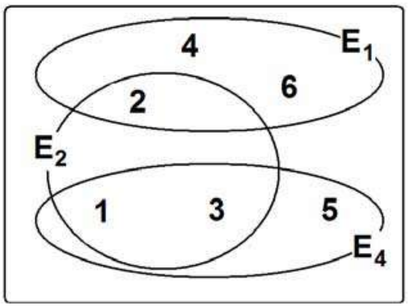
Experiment: Selecting a ball from a box containing 6 balls numbered 1, 2, 3, 4, 5, and 6 randomly.

Define the following events:

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

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

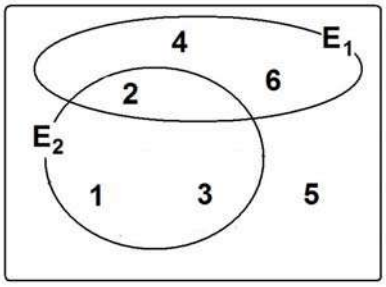
$E_4 = \{1, 3, 5\}$ = getting an odd number.



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

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

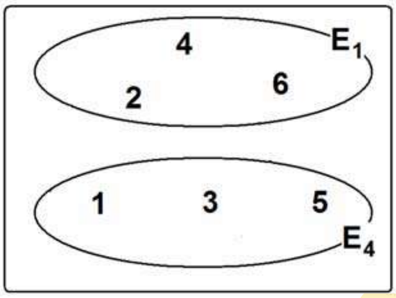
$$P(E_1 \cup E_2) = \frac{n(E_1 \cup E_2)}{n(\Omega)} = \frac{5}{6}$$



(2) $E_1 \cup E_4 = \{1, 2, 3, 4, 5, 6\} = \Omega$

= getting an even number **or** an odd number.

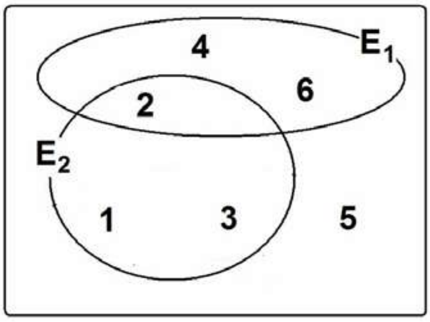
$$P(E_1 \cup E_4) = \frac{n(E_1 \cup E_4)}{n(\Omega)} = \frac{6}{6} = 1$$



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.

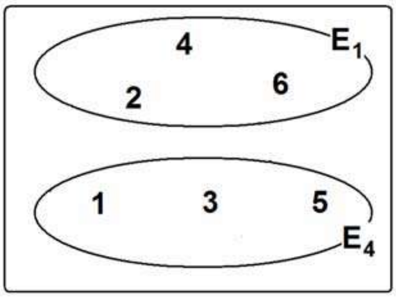
(3) $E_1 \cap E_2 = \{2\}$ = getting an even number **and** a number less than 4.

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



(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

$\overline{E}_1$ = not getting an even number = $\overline{\{2,4,6\}}$ = $\{1, 3, 5\}$

= getting an odd number.

= E_4

Mutually exclusive (disjoint) Events:

The events A and B are disjoint (or mutually exclusive) if:

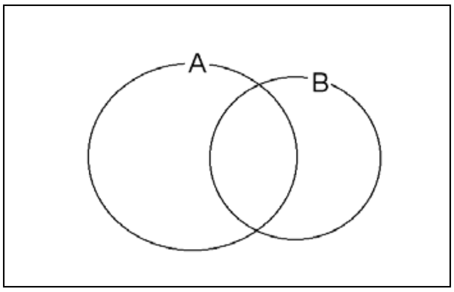
$$A \cap B = \phi.$$

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) $P(A \cup B) = P(A) + P(B)$

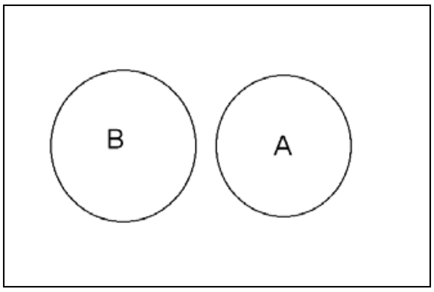
If $A \cap B \neq \phi$, then A and B are not mutually exclusive (not disjoint).



$A \cap B \neq \phi$

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:

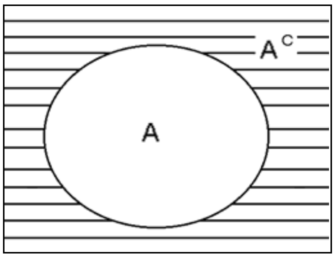
The events $A_1, A_2, \ldots, A_n$ are exhaustive events if:

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

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

The Addition Rule:

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

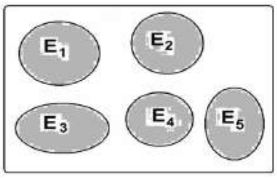
Special Cases:

1. For mutually exclusive (disjoint) events A and B

$$P(A \cup B) = P(A) + P(B)$$

2. For mutually exclusive (disjoint) events $E_1, E_2, \ldots, E_n$:

$$\boxed{P(E_1 \cup E_2 \cup \ldots \cup E_n) = P(E_1) + P(E_2) + \cdots + P(E_n)}$$



Note: If the events $A_1, A_2, \ldots, A_n$ are exhaustive and mutually exclusive (disjoint) events, then:

$$P(A_1 \cup A_2 \cup \ldots \cup A_n) = P(A_1) + P(A_2) + \cdots + P(A_n) = P(\Omega) = 1$$

Marginal Probability:

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	$\cdots$	B_n	Total
A_1	$n(A_1 \cap B_1)$	$n(A_1 \cap B_2)$	$\cdots$	$n(A_1 \cap B_n)$	$n(A_1)$
A_2	$n(A_2 \cap B_1)$	$n(A_2 \cap B_2)$	$\cdots$	$n(A_2 \cap B_n)$	$n(A_2)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_m	$n(A_m \cap B_1)$	$n(A_m \cap B_2)$	$\cdots$	$n(A_m \cap B_n)$	$n(A_m)$
Total	$n(B_1)$	$n(B_2)$	$\cdots$	$n(B_n)$	$n(\Omega)$

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

	B_1	B_2	$\cdots$	B_n	Marginal Probability
A_1	$P(A_1 \cap B_1)$	$P(A_1 \cap B_2)$	$\cdots$	$P(A_1 \cap B_n)$	P(A_1)
A_2	$P(A_2 \cap B_1)$	$P(A_2 \cap B_2)$	$\cdots$	$P(A_2 \cap B_n)$	P(A_2)
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_m	$P(A_m \cap B_1)$	$P(A_m \cap B_2)$	$\cdots$	$P(A_m \cap B_n)$	P(A_m)
Marginal Probability	P(B_1)	P(B_2)	$\cdots$	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:

$$\boxed{P(A_i) = P(A_i \cap B_1) + P(A_i \cap B_2) + \ldots + P(A_i \cap B_n) = \sum_{j=1}^{n} P(A_i \cap B_j)}$$

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 $P(B_j)$ 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	n(A₁)
A_2	20	70	10	100	n(A₂)
A_3	30	100	120	250	n(A₃)
Total	100	200	200	500	
	n(B₁)	n(B₂)	n(B₃)	n(Ω)	

Table of probabilities of each event:

	B_1	B_2	B_3	Marginal Probability	
A_1	0.1	0.06	0.14	0.3	P(A₁)
A_2	0.04	0.14	0.02	0.2	P(A₂)
A_3	0.06	0.2	0.24	0.5	P(A₃)
Marginal Probability	0.2	0.4	0.4	1	P(Ω)

P(B₁) P(B₃)

P(B₂)

For example: $P(A_2) = P(A_2 \cap B_1) + P(A_2 \cap B_2) + P(A_2 \cap B_n)$

$$= 0.04 + 0.14 + 0.02$$

$$= 0.2$$

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"*

since $E_2 \cap E_4 = \phi$

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

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, $E_i \cap E_j = \phi \quad (i \neq j)$.
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

Age	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

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

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

Example : (Page 39)

(سؤال اضافي)

What is the probability that the selected physician is not 40-49 years old and smokers occasionally ?

$P(A_3^C \cap B_2) = P(A_1 \cap B_2) + P(A_2 \cap B_2) + P(A_4 \cap B_2)$

$= 9/339 + 30/339 + 0/339$

$= 39/339$

$= 0.11504$

What is the probability that the selected physician is 30-39 years old and not a daily smoker ?

$P(A_2 \cap B_1^C) = P(A_2 \cap B_2) + P(A_2 \cap B_3)$

$= 30/339 + 49/339$

$= 79/339$

$= 0.2330$

- $B_2 =$ 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

Since $A_2 \cap A_3 = \phi$

$$P(A_2 \cup A_3)=\frac{n(A_2 \cup A_3)}{n(\Omega)}=\frac{189+79}{339}=\frac{268}{339}=0.7906$$

or

$$P(A_2 \cup A_3)=P(A_2)+P(A_3)=\frac{189}{339}+\frac{79}{339}=0.7906$$

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 and 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 and used some sort of medication.

The complement events:

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

~~Percentages:~~ $\cancel{\%(D) = 10\%}$ $\cancel{\%(M) = 25\%}$ $\cancel{\%(D \cap M) = 5\%}$

The probabilities of the given events are:

$P(D) = 0.1$ $P(M) = 0.25$ $P(D \cap M) = 0.05$

Q:Complete the table, then answer:

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

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

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

Calculating probabilities of some events:

$D \cup M$ = the selected woman delivered prematurely or used medication.

$$P(D \cup M) = P(D) + (M) - P(D \cap M) = 0.1 + 0.25 - 0.05 = 0.3$$

$\overline{M}$ = The selected woman did not use medication

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

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

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

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

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

$\overline{D} \cap \overline{M}$ = 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.

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

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

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

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

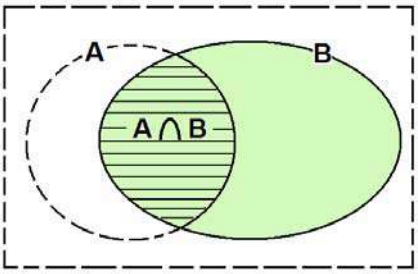
$$\begin{aligned} P(\overline{D} \cup \overline{M}) &= P(\overline{D}) + (\overline{M}) - P(\overline{D} \cap \overline{M}) \\ &= 0.9 + 0.75 - 0.70 = 0.95 \end{aligned} \qquad \text{(by the rule)}$$

<u>Conditional Probability:</u>

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

$$\boxed{P(A \mid B) = \frac{P(A \cap B)}{\underbrace{P(B)}_{\text{Known given}}} \quad ; \; P(B) \neq 0}$$

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



Notes: For calculating $P(A|B)$, we may use any one of the following:

(i) $P(A|B)=\dfrac{P(A\cap B)}{P(B)} = \dfrac{P(A \cap B)}{P(B)} = \dfrac{n(A \cap B)/n(\Omega)}{n(B)/n(\Omega)}$

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

(iii) Using the restricted table directly.

Multiplication Rules of Probability:

For any two events A and B, we have:

$$\boxed{\begin{array}{l} P(A\cap B) = P(B)P(A|B) \\ P(A\cap B) = P(A)P(B|A) \end{array}}$$

Example:

Smoking Habit

		Daily (B_1)	Occasionally (B_2)	Not at all (B_3)	Total
Age	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

- Consider the event (B_1 | A_2) knowing
  (B_1 | A_2) = the selected physician smokes daily given that his age is between 30 and 39

- $P(B_1) = \dfrac{n(B_1)}{n(\Omega)} = \dfrac{176}{339} = 0.519$
- $P(B_1 \mid A_2) = \dfrac{P(B_1 \cap A_2)}{P(A_2)} = \dfrac{0.324484}{0.557522} = 0.5820$

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

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:

$$P(B_1) = 0.519$$
$$P(B_1 \mid A_2) = 0.5820$$
$$P(B_1 \mid A_2) > P(B_1) \ , \quad P(B_1) \neq P(B_1 \mid A_2) \quad !$$

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

Example: (Multiplication Rule of Probability)

Example : (Page 44)

If we have P(A) = 0.9 , P (B/A) =0.8 find P(A ∩ B) = ?

Solution :

$$P(B/A) = \frac{P(A \cap B)}{P(A)}$$

$$0.8 = \frac{P(A \cap B)}{0.9}$$

$$P(A \cap B) = 0.8 \times 0.9$$
$$= 0.72$$

~~B = the event of passing the second part~~

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

~~From the given information:~~

~~The probability of passing the first part is:~~

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

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

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

~~Now, we use the multiplication rule to find $P(A \cap B)$~~ as follows:

~~$P(A \cap B) = P(A)\,P(B|A) = (0.9)(0.8) = 0.72$~~

~~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.~~

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

$$(i)\ \ P(A\,|\,B) = P(A) \iff (ii)\ \ P(B\,|\,A) = P(B) \iff (iii)\ \ P(B \cap A) = P(A)P(B)$$

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

Example: Suppose that A and B are two events such that:

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

Theses two events are not independent (they are dependent) because: $P(A)\,P(B) = 0.5\times 0.6 = 0.3 \qquad P(A\cap B)=0.2.$

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

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:

Complete the table, then answer:

	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*

Add some questions

Q1: Are A and B independent events?

Q2: Are A and B disjoint events?

Q3: Are A and B Exhaustive events?

Note: The Addition Rule for Independent Events: If the events A and B are independent, then by the addition rule,

$$\begin{aligned} P(A\cup B) &= P(A) + P(B) - P(A\cap B) \\ &= P(A) + P(B) - P(A)\,P(B) \end{aligned}$$

Example : (Page 46)

Q1: Are A and B independent events?

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

$$P(A \cap B) = P(A) \cdot P(B)$$

$$0.2 \neq (0.5)(0.6)$$

$$0.2 \neq 0.3$$

So A,B are not independent.

Q2: Are A and B disjoint events?

$$P(A \cup B) = P(A) + P(B) \quad \textbf{Or} \quad P(A \cap B) = 0$$
$$P(A \cap B) = 0.2 \neq 0$$

So A,B are not disjoint.

Q3: Are A and B Exhaustive events?

$$P(A \cup B) = P(\Omega) = 1$$

$$P(A \cup B) = 0.9 \neq 1$$

So A,B are not Exhaustive.

Solution in exercises

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) = $\frac{n(C)}{n(\Omega)}=\frac{3}{12}=0.25$

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

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

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

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

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

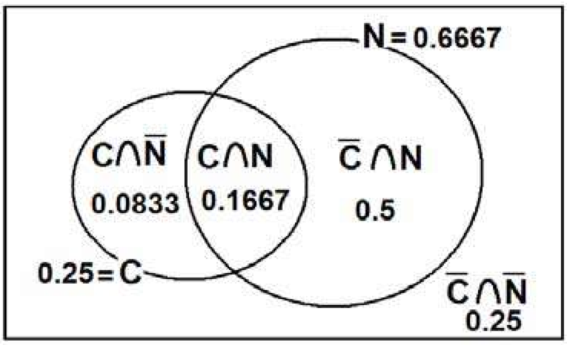
b) The probability of choosing a nurse who woks at night given that she has children:

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

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

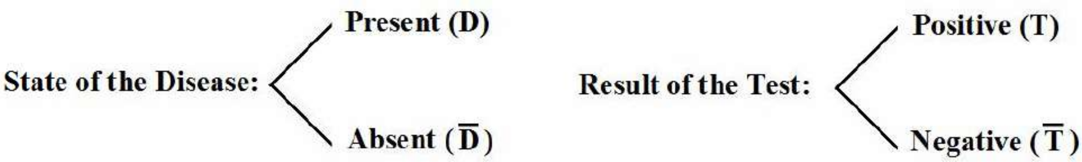
d) The events C and N not are disjoint because $C \cap N \neq \phi$. (Note: $n(C \cap N) = 2$)

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 the test	+ve (T)	Correct diagnosing	false positive result
	-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:

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

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

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

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.

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

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

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

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:

	Disease		
Test Result	**Present (D)**	**Absent ($\bar{D}$)**	**Total**
Positive (T)	a	b	a + b = n(T)
Negative ($\bar{T}$)	c	d	c + d = n($\bar{T}$)
Total	a + c = n(D)	b + d = n($\bar{D}$)	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: $P(T \mid \bar{D}) = \dfrac{n(T \cap \bar{D})}{n(\bar{D})} = \dfrac{b}{b+d}$
2. The probability of false negative result: $P(\bar{T} \mid D) = \dfrac{n(\bar{T} \cap D)}{n(D)} = \dfrac{c}{a+c}$
3. The sensitivity of the screening test: $P(T \mid D) = \dfrac{n(T \cap D)}{n(D)} = \dfrac{a}{a+c}$
4. The specificity of the screening test: $P(\bar{T} \mid \bar{D}) = \dfrac{n(\bar{T} \cap \bar{D})}{n(\bar{D})} = \dfrac{d}{b+d}$

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, $P(D\,|\,T)$:

$$\mathrm{P(D\,|\,T)} = \frac{\mathrm{P(T \cap D)}}{\mathrm{P(T\,)}}$$

Proof:

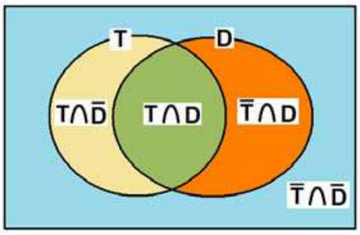
But we know that:

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

$$P(T \cap D) = P(T \mid D)\, P(D) \quad \text{(multiplication rule)}$$

$$P(T \cap \overline{D}) = P(T \mid \overline{D})\, P(\overline{D}) \quad \text{(multiplication rule)}$$

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



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

$$P(D \mid T) = \frac{P(T \mid D)\, P(D)}{P(T \mid D)\, P(D) \; + \; P(T \mid \overline{D})\, P(\overline{D})} \quad \ldots\ldots\ldots\ldots (1)$$

Note:

$P(T \mid D)$ = sensitivity.

$P(T \mid \overline{D}) = 1 - P(\overline{T} \mid \overline{D})$ = 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, $P(\overline{D} \mid \overline{T})$:

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

$$P(\overline{D} \mid \overline{T}) = \frac{P(\overline{T} \mid \overline{D})\, P(\overline{D})}{P(\overline{T} \mid \overline{D})\, P(\overline{D}) \; + \; P(\overline{T} \mid D)\, P(D)} \quad \ldots\ldots\ldots\ldots (2)$$

Note:

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

$P(\overline{T} \mid D) = 1 - P(T \mid D)$ = 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:

	Alzheimer Disease		
Test Result	**Present (D)**	**Absent ($\bar{D}$)**	**Total**
Positive (T)	436	5	441
Negative ($\bar{T}$)	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:

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

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):

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

From the tabulated data we compute:

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

$$P(T\,|\,\overline{D}) = \frac{n(T \cap \overline{D})}{n(\overline{D})} = \frac{5}{500} = 0.01 \quad = \text{1-Specivicity} = 1\text{- }0.99$$

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

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

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

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):

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

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

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

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

$$P(\overline{T}\,|\,D) = \frac{n(\overline{T} \cap D)}{n(D)} = \frac{14}{450} = 0.0311 \quad = \text{1-Sensitivity} = 1\text{- }0.9689$$

Substitution in Equation (2) gives:

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

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

$$= 0.996$$

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

Bayes Theorm pages 48-52

		Has the disease (D)	Dose not have the disease ($\bar{D}$)	Total
The result of the test	Positive (T)	Correct decision $n(T \cap D)$ Sensitivity $P(T \mid D) = \dfrac{n(T \cap D)}{n(D)}$	False decision $n(T \cap \bar{D})$ false positive result $P(T \mid \bar{D}) = \dfrac{n(T \cap \bar{D})}{n(\bar{D})}$	n(T)
	Negative ($\bar{T}$)	False decision $n(\bar{T} \cap D)$ false negative result $P(\bar{T} \mid D) = \dfrac{n(\bar{T} \cap D)}{n(D)}$	Correct decision $n(\bar{T} \cap \bar{D})$ Specificity $P(\bar{T} \mid \bar{D}) = \dfrac{n(\bar{T} \cap \bar{D})}{n(\bar{D})}$	n($\bar{T}$)
	Total	n(D)	n($\bar{D}$)	n(Ω)

Note that from the table:

$$P(\bar{T} \mid D) + P(T \mid D) = 1 \quad \text{and} \quad P(\bar{T} \mid \bar{D}) + P(T \mid \bar{D}) = 1$$

i.e. false negative + Sensitivity = 1 and Specificity + false positive = 1

The probability of the relevant disease in the general population, **P(D)** [or **P(D')** = 1 - P(D)] which is obtained from another independent study.

Predictive value Positive:

$$P(D|\,T) = \frac{P(\mathrm{T}|\ \mathrm{D}\ )*P(D)}{\text{نفس البسط} + \text{نفس البسط}\ (\ D \to \bar{D})}$$

$$= \frac{P(\ T|D\ )*P(D)}{P(\ T|D\ )*P(D)\ +P(\ T|\bar{D}\ )*P(\bar{D})} = \frac{Sensitivity*P(D)}{Sensitivity*P(D)+(1-Specificity)*P(\bar{D})}$$

Predictive value Negative:

$$P(\bar{D}|\,\bar{T}) = \frac{P(\ \bar{T}|\bar{D}\ )*P(\bar{D})}{\text{نفس البسط} + \text{نفس البسط}\ (\ \bar{D} \to D)}$$

$$= \frac{P(\ \bar{T}|\bar{D}\ )*P(\bar{D})}{P(\ \bar{T}|\bar{D}\ )*P(\bar{D})\ +P(\ \bar{T}|D\ )*P(D)} = \frac{Specificity*P(\bar{D})}{Specificity*P(\bar{D})+(1-Sensitivity)*P(D)}$$

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:

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

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:

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

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

Note: The table may contain a missing value.

Notes:

- The probability distribution of any discrete random variable X must satisfy the following two properties:
  (1) $0 \le P(X=x) \le 1$
  (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) $P(X \ge 2) = P(X=2) + P(X=3) = 0.1250 + 0.0625 = 0.1875$

(2) $P(X > 2) = P(X=3) = 0.0625$ [note: $P(X>2) \ne P(X \ge 2)$]

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

(4) $P(X \le 2) = P(X=0) + P(X=1) + P(X=2)$

$= 0.6250 + 0.1875 + 0.1250 = 0.9375$

another solution:

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

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

(5) $P(-1 \le X < 2) = P(X=0) + P(X=1)$

$= 0.6250 + 0.1875 = 0.8125$

(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 at least 2 cold:

$$P(X \ge 2) = P(X = 2) + P(X = 3) = 0.1875$$

(10) The probability that the selected person had at most 2 colds:

$$P(X \le 2) = 0.9375$$

(11) The probability that the selected person had more than 2 colds:

$$P(X > 2) = P(X = 3) = 0.0625$$

(12) The probability that the selected person had less than 2 colds:

$$P(X < 2) = P(X = 0) + P(X = 1) = 0.8125$$

Graphical Presentation:

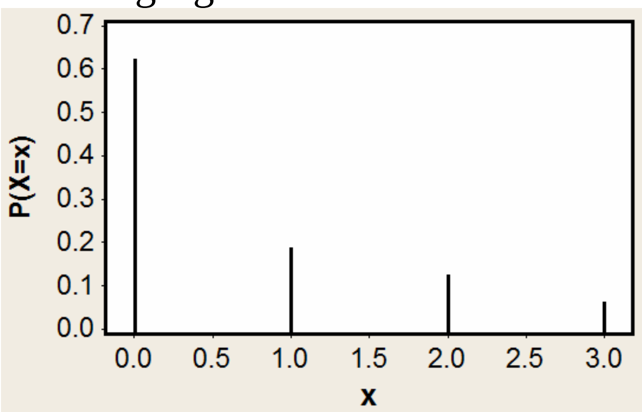
The probability distribution of a discrete r. v. X can be graphically represented.

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:



Find the Probability that the selected person had no cold in January 2010 ?

possible values of the random variable X are ....

Mean and Variance of a Discrete Random Variable

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	1	$\mu = \sum x\, P(X = x) = 1.9$			$\sigma^2 = \sum (x-\mu)^2 P(X = x) = 0.69$

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

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

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

$$= 0.69$$

<u>Cumulative Distributions:</u>

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

$$P(X \le x) = \sum_{a \le x} P(X = a) \qquad \text{(Sum over all values } \le x\text{)}$$

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

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

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($\overline{(X \ge 1)}$) = 1-P(X<1) = 1- P(X≤0)

= 1- 0.05 = 0.95

probability distribution $\longrightarrow$ cumulative distribution

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

x	$P(X \le x)$
0	0.05
1	0.05+0.25= 0.30
2	0.05+0.25+0.45= 0.75
3	0.05+0.25+0.45+0.25=1

cumulative distribution $\longrightarrow$ probability distribution

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

x	$P(X = x)$
0	0.05
1	0.30-0.05= 0.25
2	0.75 -0.30 = 0.45
3	1 – 0.75 =0.25

Complement of probability:

- $P(X \le a) = 1 - P(X > a)$
- $P(X < a) = 1 - P(X \ge a)$

- $P(X \ge a) = 1 - P(X < a)$
- $P(X > a) = 1 - P(X \le a)$

In exercise

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. $P(X \leq 3)$
   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) $1 = \sum P(X = x) = 0.25 + 0.35 + 0.20 + 0.15 + K$

$1 = 0.95 + K$

$$K = 0.05$$

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. $P(X \leq 3) = P(X=1)+P(X=2)+P(X=3) = 0.8$
   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

$$P(X\geq 4) = P(X=4)+P(X=5) =0.15+0.05=0.2$$

d) The probability that the patient has at most 2 defective teeth

$$P(X\leq 2) = P(X=1)+P(X=2) = 0.25+0.35=0.6$$

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	$\mu=2.4$			$\sigma^2=\sum(x-\mu)^2P(X=x)$ $=1.34$

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

Combinations: Notation (n!):

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

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

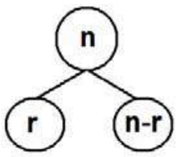
$$0! = 1$$

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. ${}_nC_n = 1, \qquad {}_nC_0 = 1,$
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, 2, … , 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:

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

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:

If X~ Binomial(n, p) , then

The mean: $\mu = np$ (expected value)

The variance: $\sigma^2 = npq$

Example: Suppose that the robability 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 highblood 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 highblood 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, 2, 3, 4, 5, 6

(1) The probability distribution of X is:

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

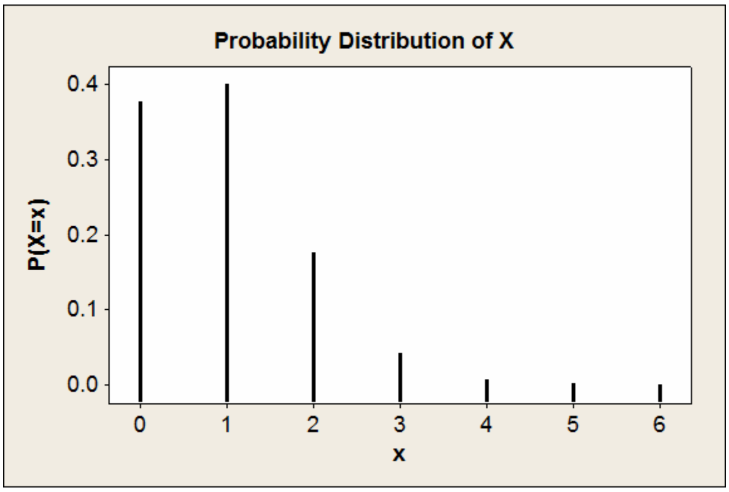
The probabilities of all values of X are:

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

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
Total	1

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:

$$\begin{aligned} P(X \le 2) &= P(X=0) + P(X=1) + P(X=2) \\ &= 0.37715 + 0.39933 + 0.17618 \\ &= 0.95266 \end{aligned}$$

(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:

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

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

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

$P(X\geq2) = P(X=2)+P(X=3)+P(X=4)+P(X=5)$
$= 0.26367+ 0.08789+ 0.01465+ 0.00098$
$= 0.\ 0.36719$

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

$P(X\leq3) = P(X=0)+P(X=1)+P(X=2)+P(X=3)$
$= 0.23730+ 0.39551+ 0.26367+ 0.08789$
$= 0.98437$

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

$$\mu = n\,p = 5\times0.25 = 1.25$$

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

$$\sigma^2 = n\,pq = 5\times0.25\times0.75 = 0.9375$$

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 . We write :

$X \sim$ Poisson (λ)

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: $\sigma^2=\lambda$

Stander deviation = sqrt(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 ~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 = λ/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 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?

There are additional questions !!

Solution:

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

$X \sim$ Poisson (6) $\qquad (\lambda=6)$

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

(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

$Y \sim$ 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.

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

$$P(W=w)=\frac{e^{-0.5}(0.5)^{w}}{w!}: \quad w=01,2,\ldots.$$

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

Extra questions (**Page 71)**:

(4) Find the probability that there will be more than or equal one snake bite cases in a month $\lambda^{*}=\frac{\lambda}{12}=\frac{6}{12}=0.5$

$$P(X\geq 1)=1-P(x<1)$$
$$=1-P(X=0)$$
$$=1-\frac{e^{-0.5}(0.5)^{0}}{0!}=1-0.6065=0.3935$$

(5) The mean of snake bite cases in a year

$$\mu=\lambda=6$$

(6) The variance of snake bite cases in a month

$$\sigma^{2}=\lambda^{*}=\frac{\lambda}{12}=\frac{6}{12}=0.5$$

(7) The standard deviation of snake bite cases in 2 years

$$\sigma=\sqrt{\lambda^{*}}=\sqrt{2\lambda}=\sqrt{2(6)}=\sqrt{12}=3.4641$$

(8) Find the probability that there will be more than

3 snake bite cases in 2 years_$\lambda^* = 2\lambda = 2(6) = 12$

$$P(x > 3) = 1 - P(X \leq 3)$$

$$= 1 - [P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)]$$

$$= 1 - 0.0023$$

$$= 0.9977$$

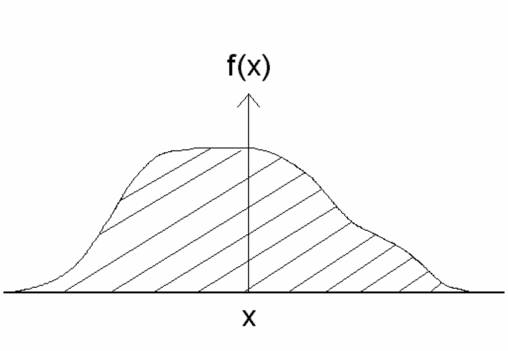
By calculator:

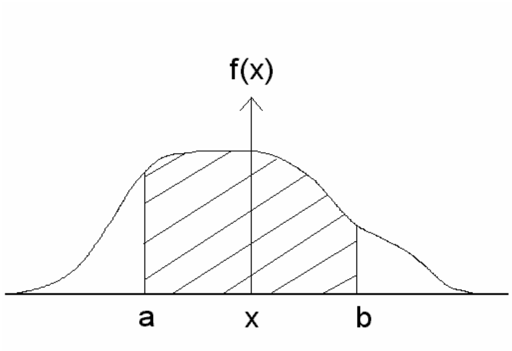
$\sum_{x=0}^{3}\left(\frac{e^{-12} \times 12^{x}}{x!}\right) = 2.29 \times 10^{-3} = 0.00229 \approx 0.0023$

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.



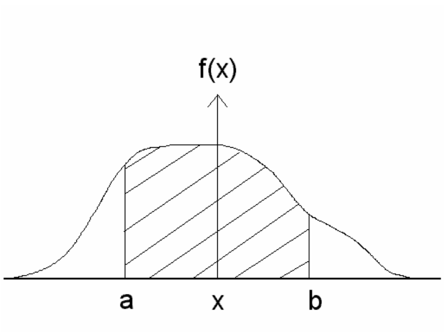


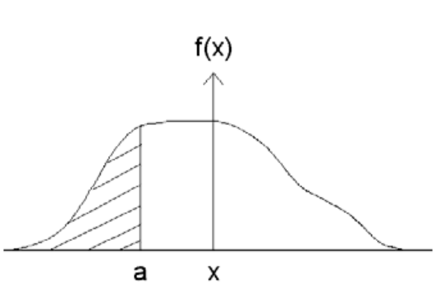
$$\text{Total area} = \int_{-\infty}^{\infty} f(x)\,dx = 1$$

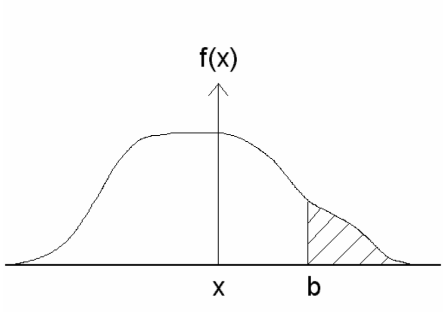
$$P(a \le X \le b) = \int_a^b f(x)dx = \text{area}$$

(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.







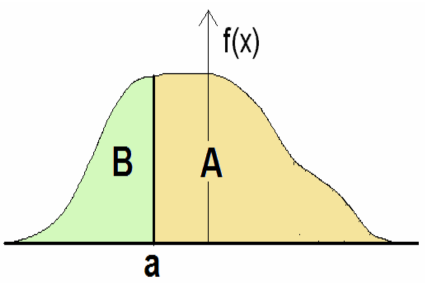
$$P(a \le X \le b) = \int^b f(x)dx = \text{area}$$

$$P(X \le a) = \int_{-\infty}^a f(x)dx = \text{area}$$

$$P(X \ge b) = \int_b^{\infty} f(x)\,dx = \text{area}$$

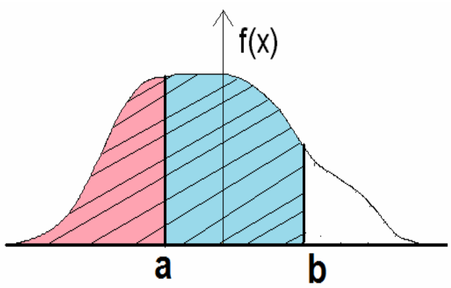
Note: If X is continuous r.v. then:

1. $P(X=a)=0$ for any a.
2. $P(X \le a)=P(X<a)$
3. $P(X \ge b)=P(X>b)$
4. $P(a \le X \le b)=P(a \le X < b)=P(a < X \le b)=P(a<X<b)$
5. $P(X \le x)=$ cumulative probability
6. $P(X \ge a)=1-P(X<a)=1-P(X \le a)$
7. $P(a \le X \le b)=P(X \le b)-P(X \le a)$



$P(X \ge a)=1-P(X \le a)$

$A = 1 - B$, Total area = 1



$$P(a \le X \le b)=P(X \le b)-P(X \le a)$$
$$\int_a^b f(x)dx = \int_{-\infty}^{b} f(x)dx - \int_{-\infty}^{a} f(x)dx$$

4.6 The Normal Distribution:

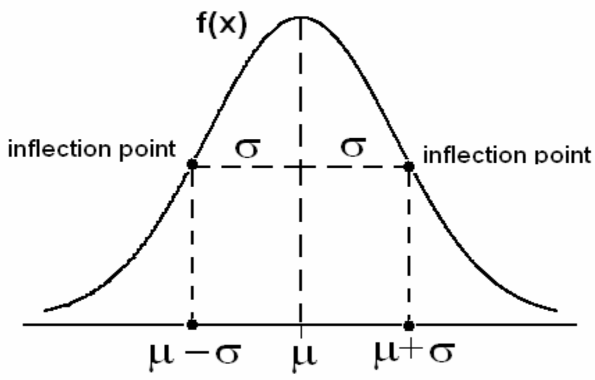
- One of the most important continuous distributions.
- Many measurable characteristics are normally or approximately normally distributed.
 (Examples: height, weight, …)
- The probability density function of the normal distribution is given by:

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

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 = μ
standard deviation = σ
variance = σ^2

3. The highest point of the curve of f(x) at the mean μ.
   (Mode = μ)
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 = μ (determines the location)

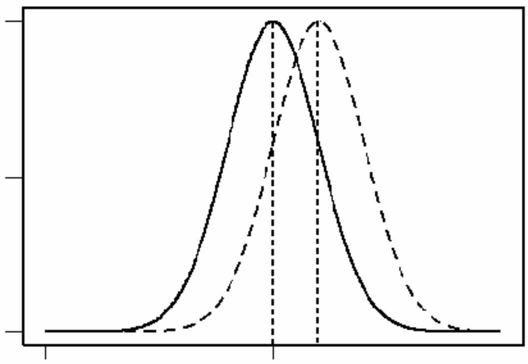
   standard deviation = σ (determines the shape)
6. If the r.v. X is normally distributed with mean μ and standard deviation σ (variance σ^2), we write:

   $X \sim \text{Normal}\left(\mu, \sigma^2\right)$ or $X \sim \text{N}\left(\mu, \sigma^2\right)$
7. The location of the normal distribution depends on μ. The shape of the normal distribution depends on σ.

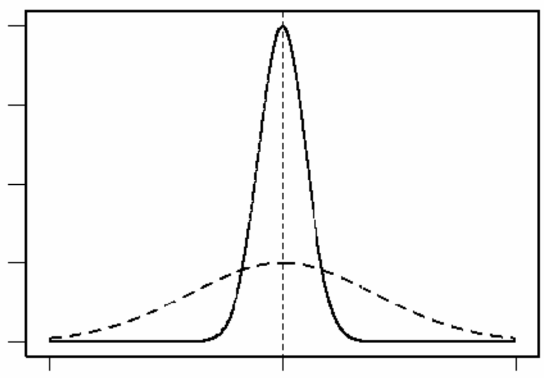
Note: The location of the normal distribution depends on μ and its shape depends on σ.
Suppose we have two normal distributions:

________ $N(\mu_1, \sigma_1)$

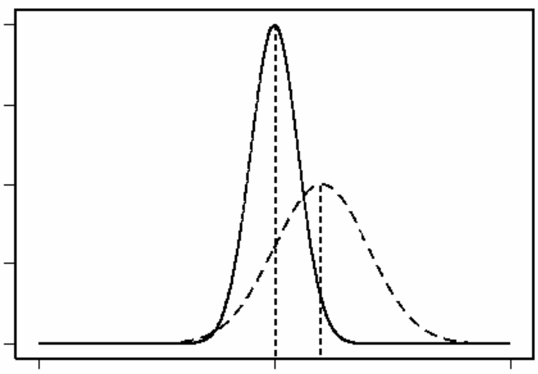
----------- $N(\mu_2, \sigma_2)$



$\mu_1 < \mu_2, \sigma_1=\sigma_2$



$\mu_1 = \mu_2, \sigma_1<\sigma_2$



$\mu_1 < \mu_2, \sigma_1<\sigma_2$

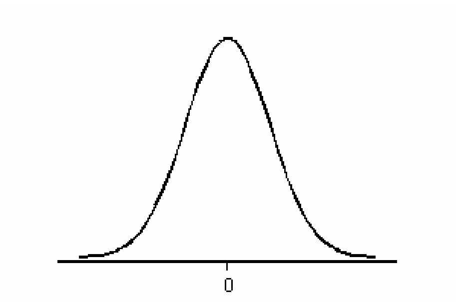
<u>The Standard Normal Distribution:</u>

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:

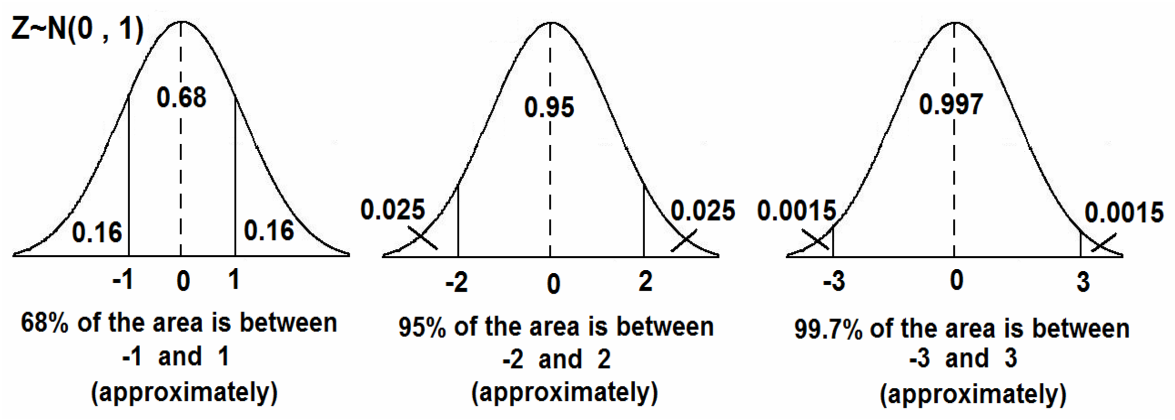
$$Z \sim N(0, 1)$$

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

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



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.



Result:

If $X \sim$ Normal μ, σ^2 , then $Z = \frac{X -}{\quad} \sim$ Normal (0,1).

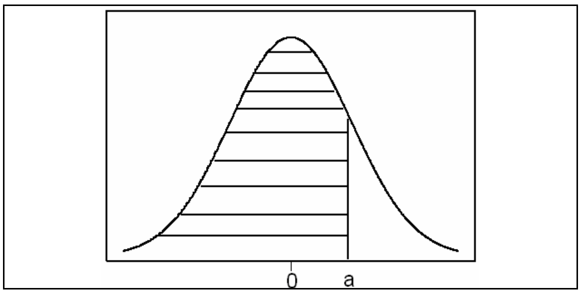
Calculating Probabilities of Normal (0,1):

Suppose Z ~ Normal (0,1).

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

$$P(Z \leq a)$$

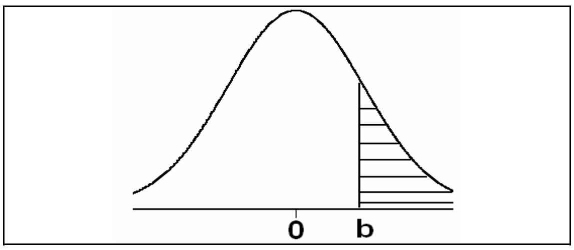
(i) $P(Z \leq a) =$ From the table



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

Where:

$P(Z \leq b \quad) =$ 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