Lectures' Notes

STAT – 106
BIOSTATISTICS

Summer Semester 1424/1425
STAT- 106

Teacher: Dr. Abdullah Al-Shiha
Office: 2B43 Building # 4
Department of Statistics and Operations Research
College of Science, King Saud University

Textbook:
Elementary Biostatistics with Applications from Saudi Arabia
By: Dr. Nancy Hasabelnaby

Chapter 1 : Organizing and Displaying Data

Introduction:

Statististcs:
Statistics is that area of study which is interested in learning how to collect, organize, and summarize information, and how to answer research questions and draw conclusions.

Bostatistics:
If the information is obtained from biological and medical sciences, then we use the term biostatistics.

Populations:
A population is the largest group of people or things in which we are interested at a particular time and about which we want to make some statements or conclusions.

Samples:
From the population, we select various elements (or individuals) on which we collect our information. This part of the population on which we collect data is called the sample.

Sample Size:
The number of elements in the sample is called the sample size and is denoted by n.

Variables:
The characteristics to be measured on the elements of the population or sample are called variables.

Example of Variables:

- Height No. of cars
- Sex Educational Level

Types of Variables:

(1) Quantitative Variables:

The values of a quantitative variable are numbers indicating how much or how many of something.

Examples:

- Weight
- Family Size
- Age

(2) Qualitative Variables:

The values of a qualitative variable are words or attributes indicating to which category an element of the population belong.

Examples:

- blood type
- educational level
- nationality

Types of Quantitative Variables:

Discrete Variables:

A discrete variable can have countable number of values (there are jumps between the values)

Examples: - family size ($x = 1, 2, 3, \ldots$)

- number of patients ($x = 0, 1, 2, 3, \ldots$)

Continuous Variables:

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

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

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

1.2. Organizing The Data

- Ungrouped (or Simple) frequency distributions for:
  - qualitative variables
  - discrete quantitative variables with a few different values
- grouped frequency distributions for:
  - continuous quantitative variables
  - discrete quantitative variables with large number of different values

Example: (Simple frequency distribution or ungrouped frequency distribution).
The following data represent the number of children of 16 Saudi women:

3, 5, 2, 4, 0, 1, 3, 5, 2, 3, 2, 3, 3, 2, 4, 1

- Variable = X = no. of children (discrete, quantitative)
- Sample size = n = 16
- The possible values of the variable are: 0, 1, 2, 3, 4, 5

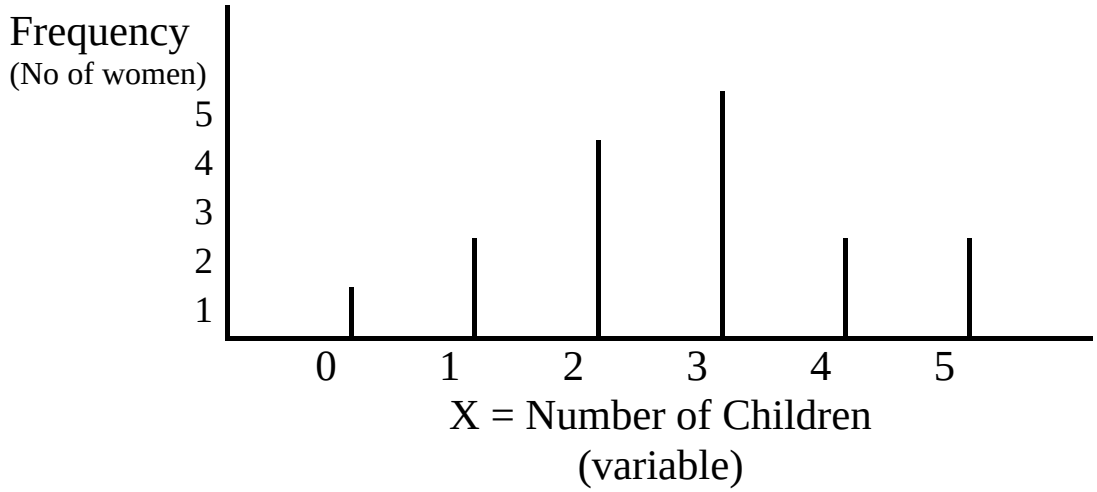
no. of children (variable)	**Frequency (no. of women)**	**Relative Freq. (R.F) (= Freq./n)**	**Percentage Freq. (= R.F. * 100%)**
0	1	0.0625	6.25%
1	2	0.125	12.5%
2	4	0.25	25%
3	5	0.3125	31.25%
4	2	0.125	12.5%
5	2	0.125	12.5%
Total	n=16	1.00	100%

Simple frequency distribution of the no. of children

Notes:

- The sample size = n = Total of the frequencies
- Relative frequency = frequency/n
- Percentage frequency = Relative frequency * 100%

- Frequency bar chart is a graphical representation for the simple frequency distribution.



Example 1.2: (grouped frequency distribution)
The following table gives 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

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

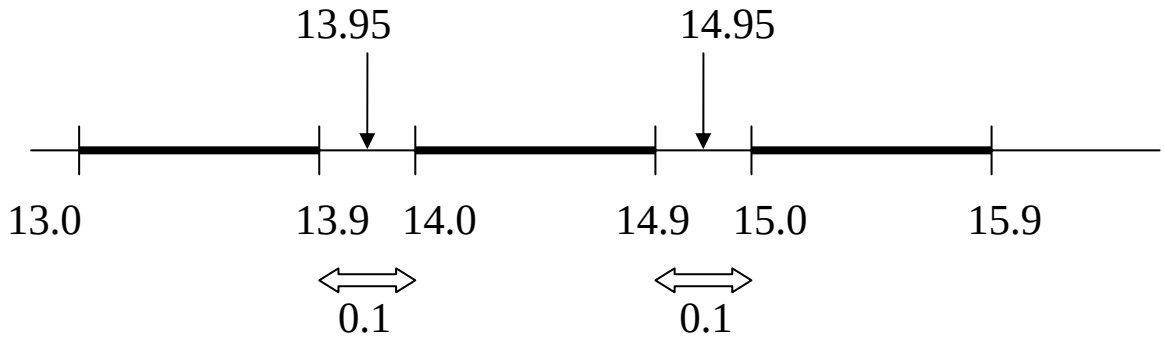
Class Interval (Hemoglobin level)	Frequency (no. of men)	Relative Frequency	Cumulative Frequency	Cumulative Relative Frequency
13.0 - 13.9	3	0.06	3	0.06
14.0 - 14.9	5	0.10	8	0.16
15.0 - 15.9	15	0.30	23	0.46
16.0 - 16.9	16	0.32	39	0.78
17.0 - 17.9	10	0.20	49	0.98
18.0 - 18.9	1	0.02	50 = n	1.00
Total	n=50	1.00		

Grouped frequency distribution
for the hemoglobin level of the 50 men

Notes:

- class interval = C. I.
- Cumulative frequency of a class interval = no. of values (frequency) obtained in that class interval or before.
- Mid-point (class mark) of a C.I. = $\frac{\text{upper limit} + \text{lower limit}}{2}$

- True Class Intervals:



Class Interval	True C. I.	Class mid-point	frequency
13.0 - 13.9	12.95 - 13.95	(13.0+13.9)/2 = 13.45	3
14.0 - 14.9	**13.95 - 14.95**	**(14.9+14.9)/2 = 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

↑ lower limits (L.L.)

↑ upper limits (U.L.)

↑ True lower limits

↑ True upper limits

a = Jump = 0.1

True U. L. = U. L. + $\frac{a}{2}$ = U. L. + 0.05

True L. L. = L. L. – $\frac{a}{2}$ = L. L. – 0.05

Width of a class interval (W) = True U. L. – True L. L.

In the previous example, W = 13.95 – 12.95 = 1.0

1.4. Displaying Grouped Frequency Distributions:-

For representing frequency or relative frequency distributions:

- Histograms
- Polygons
- Curves

For representing cumulative frequency or cumulative relative frequency distributions:

- Cumulative Curves
- Cumulative Polygon (ogives)

Example:

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

C.I. Age	True C. I.	Frequency No. of women	Cumulative Freq.	mid-points
15 - 19	14.5 - 19.5	8	8	17
20 - 24	19.5 - 24.5	16	24	22
25 - 29	24.5 - 29.5	32	56	27
30 - 34	29.5 - 34.5	28	84	32
35 - 39	34.5 - 39.5	12	96	37
40 - 44	39.5 - 44.5	4	100	42
Total		n=100		

$a = 20 - 19 = 1$

$\frac{a}{2} = 0.5$

True U. L. = U. L. + 0.5

True L. L. = L. L. – 0.5

W = 19.5 – 14.5 = 5

Organizing and Displaying Data

Organizing and Displaying Data

CHAPTER 2: Basic Summary Statistics

- Measures of Central Tendency (or location)
  - Mean – mode – median
- Measures of Dispersion (or Variation)
  - Range – Variance – Standard Deviation – Coefficient of Variation

2.1. Introduction:

For the population of interest, there is a population of values of the variable of interest.

- 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 $X_1, X_2, \ldots, X_N$ (parameters are unknown in general)

(ii) A **statistic** is a measure (or number) obtained from the sample values $x_1, x_2, \ldots, x_n$ (statistics are known in general)

- Since parameters are unknown, statistics are used to approximate (estimate) parameters.

2.2. Measures of Central Tendency: (Location)

- The values of a variable often tend to be concentrated around the center of the data.
- Some of these measures are: the mean, mode, and median.
- These measures are considered as representatives (or typical values) of the data.

Mean:

(1) Population mean μ:
If $X_1, X_2, \ldots, X_N$ are the population values, then the population mean is:

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

- The population mean μ is a parameter (it is usually unknown)

(2) Sample mean $\bar{x}$:
If $x_1, x_2, \ldots, x_n$ are the sample values, then the sample mean is

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

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

Example:

Consider the following population values:

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

Suppose that the sample values obtained are:

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

Then:

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

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

Notes:

- The mean is simple to calculate.
- There is only one mean for given sample data.
- The mean can be distorted by extreme values.
- 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** set into two equal parts.

Let $x_1, x_2, \ldots, x_n$ be the sample values.

(1) If the sample size, n, is odd:

- The median is the middle value of the **ordered** observations.
- The middle observation is the ordered $\frac{n+1}{2}$ observation.
- The median = The $\left(\frac{n+1}{2}\right)^{th}$ order observation.

Ordered set → (smallest to largest)	*	*	…	Middle value = MEDIAN	…	*
Rank (or order) →	1	2	…	$\frac{n+1}{2}$	…	n

Example:

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

Solution:

.$n = 5$ (odd number)

The rank of the middle value (median) = $\frac{n+1}{2} = \frac{5+1}{2} = 3$

Ordered set →	10	21	**38**	53	54
Rank (or order) →	1	2	$\frac{n+1}{2}$ **= 3**	4	5

The median =38 (unit)

(2) If the sample size, *n*, is even:

- The median is the mean (average) of the two middle values of the **ordered** observations.
- The middle two values are the ordered $\frac{n}{2}$ and $\frac{n}{2}+1$ observations.
- The median =

$$\frac{1}{2}\left\{\left(\frac{n}{2}\right)th \text{ ordered observation} + \left(\frac{n}{2}+1\right)th \text{ ordered observation}\right\}$$

Ordered set →	*	*	…	Middle value	Middle value	…	*
Rank (or order) →	1	2	…	$\frac{n}{2}$	$\frac{n}{2}+1$	…	n

Example:

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

Solution:

.$n = 6$ (even number)

The rank of the middle values are

$\frac{n}{2} = 6/2 = 3$ and $\frac{n}{2}+1 = (6/2) + 1 = 4$

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

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

Notes:

- The median is simple to calculate.
- There is only one median for given data.
- The median is not affected too much by extreme values.
- The median can only be found for quantitative variables.

Mode:

The mode of a set of values is that value which occurs with the highest frequency.

- If all values are different or have the same frequency, there is no mode.
- A set of data may have more than one mode.

Example:

Data set	Mode(s)
26, 25, 25, 34	25 (unit)
3, 7, 12, 6, 19	No mode
3, 3, 7, 7, 12, 12, 6, 6, 19, 19	No mode
3, 3, 12, 6, 8, 8	3 and 8 (unit)

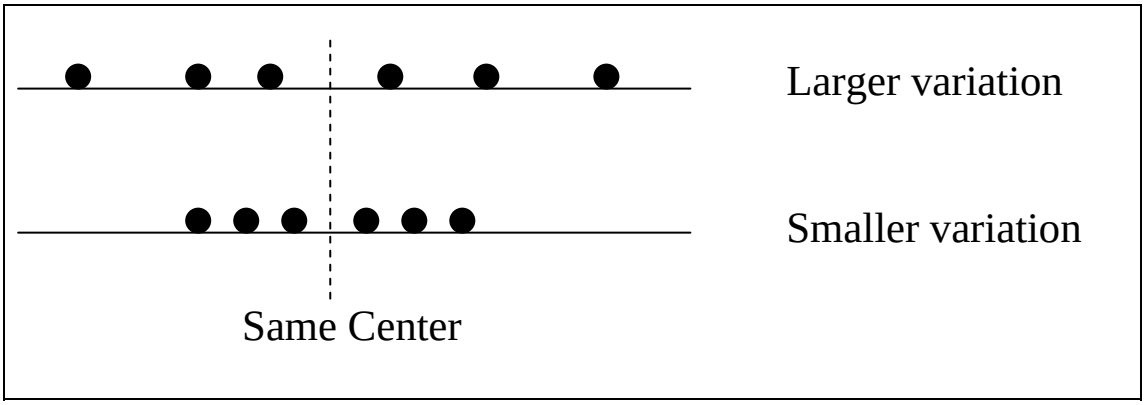
Notes:

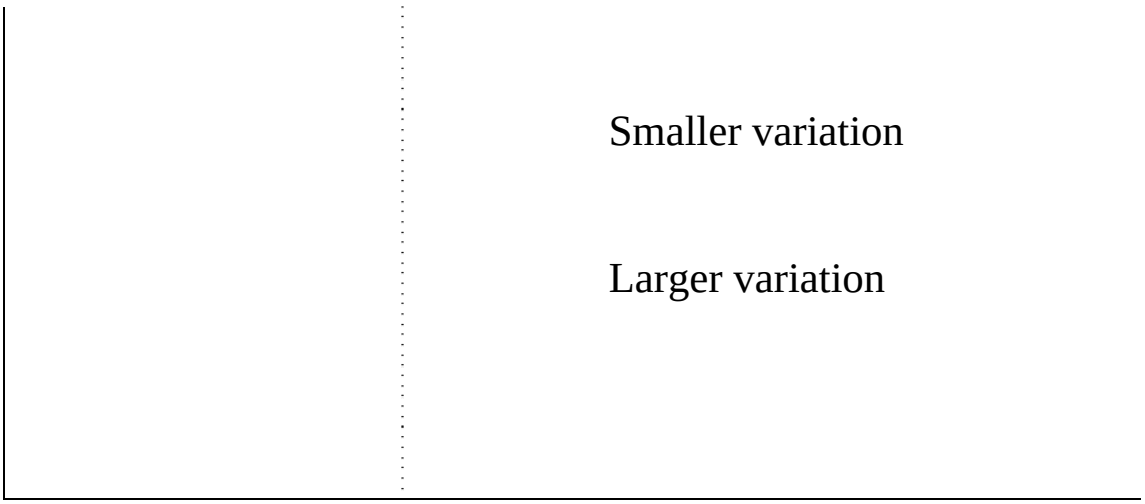
- The mode is simple to calculate but it is not "good".
- The mode is not affected too much by extreme values.
- The mode may be found for both quantitative and qualitative variables.

2.3. Measures of Dispersion (Variation):

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

- The variation is small when the values are close together.
- There is no variation if the values are the same.





Some measures of dispersion:

Range – Variance – Standard deviation – Coefficient of variation

Range:

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

Range (R) = Max – Min

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

Solution:
Range (R) = 35 – 25 = 10 (unit)

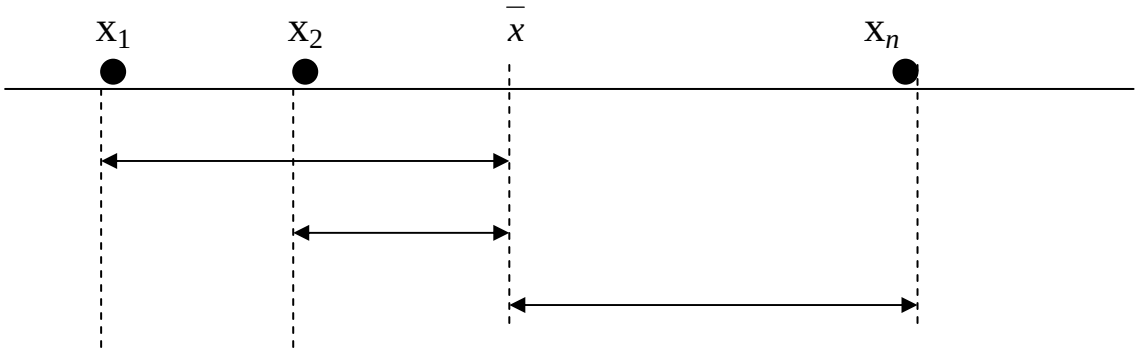
Note:
The range is not useful as a measure of the variation since it only takes into account two of the values. (it is not good)

Variance:

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

- The variance is small when all values are close to the mean.
- The variance is large when all values are spread out from the mean.

Squared deviations from the mean:



$(x_1 - \bar{x})^2$ $(x_2 - \bar{x})^2$ $(x_n - \bar{x})^2$

(1) Population variance σ^2 :

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

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

where $\mu = \frac{\sum_{i=1}^{N} X_i}{N}$ is the population mean.

Notes:

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

(2) Sample Variance S^2:

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

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

where $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ is the sample mean.

Notes:

- S^2 is a statistic because it is obtained from the sample values (it is known).
- S^2 is used to approximate (estimate) σ^2.
- $S^2 \geq 0$

Example:

We want to compute the sample variance of the following sample values: 10, 21, 33, 53, 54.

Solution:

n=5

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

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

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

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

Another Method:

x_i	$\left(x_i - \bar{x}\right) = \left(x_i - 34.2\right)$	$\left(x_i - \bar{x}\right)^2 = \left(x_i - 34.2\right)^2$
10	-24.2	585.64
21	-13.2	174.24
33	-1.2	1.44
53	18.8	353.44
54	19.8	392.04
$\sum_{i=1}^{5} x_i = 171$	$\sum_{i=1}^{5} \left(x_i - \bar{x}\right) = 0$	$\sum \left(x_i - \bar{x}\right)^2 = 1506.8$

$$\bar{x} = \frac{\sum_{i=1}^{5} x_i}{5} = \frac{171}{5} = 34.2$$

$$S^2 = \frac{1506.8}{4} = 376.7$$

Calculating Formula for S^2:

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

* Simple

* More accurate

Note:

To calculate S^2 we need:

- n = sample size
- $\sum x_i =$ The sum of the values
- $\sum x_i^2 =$ The sum of the squared values

For the above example:

x_i	10	21	33	53	54	$\sum x_i = 171$
x_i^2	100	441	1089	2809	2916	$\sum x_i^2 = 7355$

$$S^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1} = \frac{7355 - (5)(34.2)^2}{5-1} = \frac{1506.8}{4} = 376.7 \ \text{(unit)}^2$$

Standard Deviation *S*:

- The standard deviation is another measure of variation.
- It is the square root of the variance.

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

Example:
For the previous example, the sample standard deviation is

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

Coefficient of Variation (C.V.):

- The variance and the standard deviation are useful as measures of variation of the values of a single variable for a single population (or sample).
- 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 the **relative variation** that will not depend on either the units or on how large the values are. This measure is the coefficient of variation (C.V.) which is defined by:

$$\text{C.V.} = \frac{S}{\bar{x}} * 100\% \qquad \text{(free of unit or unit less)}$$

	Mean	St.dev.	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 relative variability in the 1st data set is larger than the relative variability in the 2nd data set if $C.V_1 > C.V_2$ (and vice versa).

Example:

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 $C.V_1 < C.V_2$, the relative variability in the 2nd data set is larger than the relative variability in the 1st data set.

Notes: (Some Properties of $\bar{x}$, S, and S^2):

Sample values are : $x_1, x_2, \ldots, x_n$
a and *b* are constants.

Sample Data	Sample mean	Sample st.dev.	Sample Variance
$x_1, x_2, \ldots, x_n$	$\bar{x}$	S	S^2
$ax_1, ax_2, \ldots, ax_n$	$a\bar{x}$	$\lvert a\rvert S$	a^2S^2
$x_1 + b, , \ldots, x_n + b$	$\bar{x} + b$	S	S^2
$ax_1 + b, \ldots, ax_n + b$	$a\bar{x} + b$	$\lvert a\rvert S$	a^2S^2

Absolute value:

$$|a| = \begin{cases} a & if \ \ a \geq 0 \\ -a & if \ \ a < 0 \end{cases}$$

Example:

	Sample	Sample mean	Sample St..dev.	Sample Variance
	1,3,5	3	2	4
(1)	-2, -6, -10	-6	4	16
(2)	11, 13, 15	13	2	4
(3)	8, 4, 0	4	4	16

Data (1) $-2x_1, -2x_2, -2x_3$ ($a = -2$)

(2) $x_1 + 10, x_2 + 10, x_3 + 10$ ($b = 10$)

(3) $-2x_1 + 10, -2x_2 + 10, -2x_3 + 10$ ($a = -2$, b = 10)

2.4. Calculating Measures from An Ungrouped (Simple) Frequency Table:

For the general ungrouped (simple) frequency table of the data: $x_1, x_2, \ldots, x_n$:

Value m_i	Freq. f_i	$m_i f_i$	m_i^2	$m_i^2 f_i$
m_1	f_1	$m_1 f_1$	m_1^2	$m_1^2 f_1$
m_2	f_2	$m_2 f_2$	m_2^2	$m_2^2 f_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
m_k	f_k	$m_k f_k$	m_k^2	$m_k^2 f_k$
	$n = \sum f$	$\sum mf$		$\sum m^2 f$

Note: $n = \sum_{i=1}^{k} f_i$ = no. of observations

$\sum x = \sum_{i=1}^{k} m_i f_i$ = sum of the observations

$\sum x^2 = \sum_{i=1}^{k} m_i^2 f_i$ = sum of the squared observations

For calculating $\bar{x}$ and S^2, we need:

$n =$ the sample size $= \sum_{i=1}^{k} f_i$

$\sum x =$ the sum of the values $= \sum_{i=1}^{k} m_i f_i$

$\sum x^2 =$ the sum of the squared values $= \sum_{i=1}^{k} m_i^2 f_i$

Sample Mean:

$$\bar{x} = \frac{\sum_{i=1}^{k} m_i f_i}{\sum_{i=1}^{k} f_i} \quad \Leftrightarrow \quad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Sample Variance:

$$S^2 = \frac{\sum_{i=1}^{k} m_i^2 f_i - \left(\sum_{i=1}^{k} f_i\right)\bar{x}^2}{\left(\sum_{i=1}^{k} f_i\right) - 1} \quad \Leftrightarrow \quad S^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}$$

Example: (p.41)

Data (x_i) : 1, 2, 1, 2, 2, 2, 3, 3, 4, 5, 1, 2, 2, 2, 2, 3, 3, 4, 6, 5, 1,1, 2, 2, 2, 3, 3, 3, 4, 7

Simple Frequency Distribution:

Value m_i	Freq. f_i	$m_i f_i$	m_i^2	$m_i^2 f_i$
1	5	5	1	5
2	11	22	4	44
3	7	21	9	63
4	3	12	16	48
5	2	10	25	50
6	1	6	36	36
7	1	7	49	49
	$n = \sum f$ =30	$\sum mf$ =83		$\sum m^2 f$ =295

- Mean: $\bar{x} = \frac{\sum mf}{\sum f} = \frac{83}{30} = 2.8$ (unit)
- Variance:

$$S^2 = \frac{\sum m^2 f - \left(\sum f\right)\bar{x}^2}{\left(\sum f\right) - 1} = \frac{295 - (30)(2.8)^2}{30 - 1} = 2.3 \text{ (unit)}^2$$

- Standard Deviation: $S = \sqrt{2.3} = 1.517$ (unit)
- Coefficient of variation:

$$\text{C.V.} = \frac{S}{\bar{x}} * 100\% = \left(\frac{1.517}{2.8}\right) * 100\% = 54.16\%$$

- Mode: mode = 2 (unit)
- Median: (n = 30 even)

$\frac{n}{2} = 15$ and $\frac{n}{2}+1 = 16$

$$\text{median} = \frac{\left(\frac{n}{2}\right)^{th} \text{order obs.} + \left(\frac{n}{2}+1\right)^{th} \text{ordered obs.}}{2}$$

$$= \frac{15^{th} \text{ ordered obs.} + 16^{th} \text{ ordered obs.}}{2}$$

$$= \frac{2+2}{2} = 2 \text{ (unit)}$$

2.5. Approximating Measures From Grouped Data:

For grouped data:

- We do not know the actual values.
- We know how many of the values in each class interval
- Thus, we cannot find the actual values for $\bar{x}$ and S^2.
- We assume that all values in a particular class interval are located at the mid point of that interval.

Recall that, for calculating $\bar{x}$ and S^2, we need:

$$n, \qquad \sum x, \qquad \text{and} \quad \sum x^2$$

Let k = the number of class intervals
m_i = the mid point of the i-th C.I.
f_i = the frequency of the i-th C.I.

C.I.	mid-point m_i	Freq. f_i	$m_i f_i$	m_i^2	$m_i^2 f_i$
1st C.I.	m_1	f_1	$m_1 f_1$	m_1^2	$m_1^2 f_1$
2nd C.I.	m_2	f_2	$m_2 f_2$	m_2^2	$m_2^2 f_2$
⋮	⋮	⋮	⋮	⋮	⋮
kth C.I	m_k	f_k	$m_k f_k$	m_k^2	$m_k^2 f_k$
		$n = \sum f_i$	$\sum x = \sum m_i f_i$		$\sum x^2 = \sum m_i^2 f_i$

Therefore, the approximation of $\bar{x}$ and S^2 are :

$$\bar{x} = \frac{\sum m_i f_i}{n} \qquad ; \; n = \sum_{i=1}^{k} f_i$$

$$S^2 = \frac{\sum_{i=1}^{k} m_i^2 f_i - n\bar{x}^2}{n-1}$$

Example:

Class Interval (ages in year)	Freq. f_i	mid point m_i	$m_i f_i$	m_i^2	$m_i^2 f_i$
15 - 19	8	17	136	289	2312
20 - 24	16	22	352	484	7744
25 - 29	32	27	864	729	23328
30 - 34	28	32	896	1024	28672
35 - 39	12	37	444	1369	16428
40 - 44	4	42	168	1764	7056
	$n = \sum f_i$ =100		$\sum mf_i$ = 2860		$\sum m_i^2 f_i$ = 85540

$$\bar{x} = \frac{\sum mf}{n} = \frac{2860}{100} = 28.6 \qquad \text{(year)}$$

$$S^2 = \frac{\sum_{i=1}^{k} m_i^2 f_i - n\bar{x}^2}{n-1} = \frac{85540 - (100)(28.6)^2}{100-1} = 37.8 \qquad (\text{year})^2$$

$$S = \sqrt{S^2} = \sqrt{37.8} = 6.1 \quad \text{(year)}$$

$$\text{C.V.} = \frac{S}{\bar{x}} * 100\% = \frac{6.1}{28.6} * 100\% = 21.5\%$$

Chapter 3: Basic Probability Concepts

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:
is some procedure (or process) that we do.

Sample Space:
The set of all possible outcomes of an experiment is called the sample space (or Universal set) and is denoted by Ω.

An Event:
is a subset of the sample space Ω.

- $\phi \subseteq \Omega$ is an event (impossible event)
- $\Omega \subseteq \Omega$ is an event (sure event)

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

- This experiment has 6 possible outcomes $\Omega = \{1, 2, 3, 4, 5, 6\}$.
- Consider the following events:

E_1 = getting an event 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 occurrences of the outcomes have the same chance.

Probability of An Event:

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

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

Example: In the ball experiment in the previous example, suppose the ball is selected randomly.

$$\Omega=\{1, 2, 3, 4, 5, 6\}\ ;\ n(\Omega)=6$$
$$E_1=\{2, 4, 6\}\ ;\ n(E_1)=3$$
$$E_2=\{1, 2, 3\}\ ;\ n(E_2)=3$$
$$E_3=\{1, 3\}\ ;\ n(E_3)=2$$

The outcomes are equally likely.

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

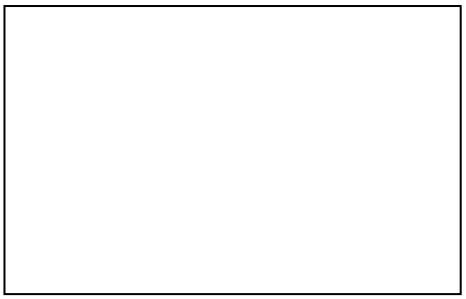
Some Operations on Events:

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

Union: $A\cup B$

- $A\cup B$ Consists of all outcomes in A **or** in B **or** in both A and B.

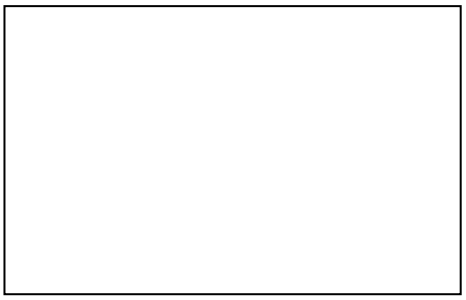
- $A\cup B$ Occurs if A occurs, **or** B occurs, **or** both A and B occur.



Intersection: $A\cap B$

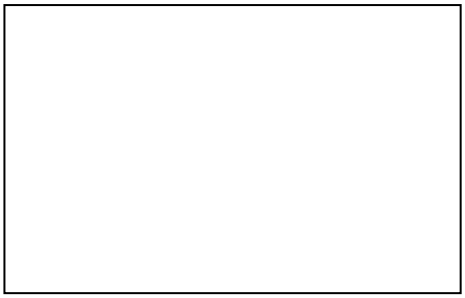
- $A\cap B$ Consists of all outcomes in both A **and** B.

- $A\cap B$ Occurs if both A **and** B occur .



Complement: A^c

- A^c is the complement of A.
- A^c consists of all outcomes of Ω but are not in A.
- A^c 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.

$E_3 = \{1, 3\}$ = getting 1 or 3.

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

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

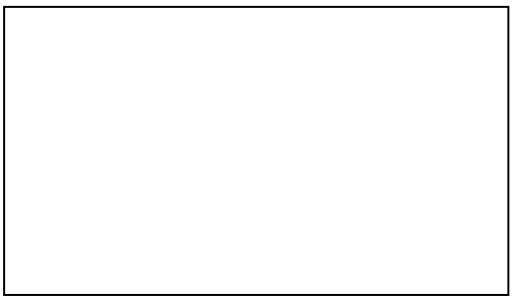
= getting an even no. **or** a no. less than 4.

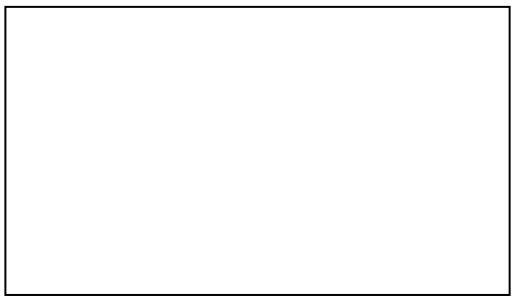
$$P(E_1 \cup E_2) = \frac{n(E_1 \cup E_2)}{n(\Omega)} = \frac{5}{6}$$

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

= getting an even no. **or** an odd no.

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





Note: $E_1 \cup E_4 = \Omega$

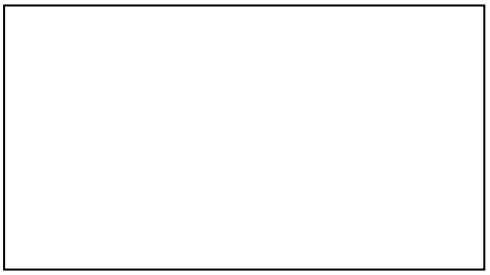
E_1 and E_4 are called exhaustive events.

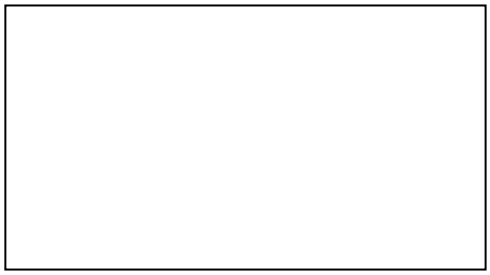
(3) $E_1 \cap E_2 = \{2\}$ = getting an even no. **and** a no. less than 4.

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

(4) $E_1 \cap E_4 = \phi$ = getting an even no. **and** an odd no.

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





Note: $E_1 \cap E_4 = \phi$

E_1 and E_4 are called disjoint (or mutually exclusive) events.

(5) E_1^c = <u>not</u> getting an even no. = $\{2, 4, 6\}^c = \{1, 3, 5\}$

= getting an odd no.

= E_4

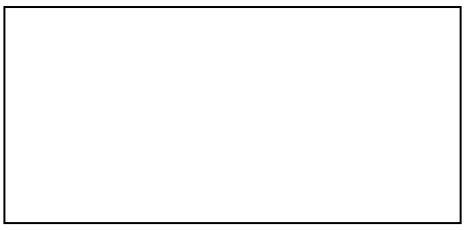
<u>Notes:</u>

1. The event $A_1, A_2, \ldots, A_n$ are <u>exhaustive</u> events if $A_1 \cup A_2 \cup \ldots \cup A_n = \Omega$.
2. The events A and B are <u>disjoint</u> (or <u>mutually exclusive</u>) if $A \cap B = \phi$.

   In this case :

   (i) $P(A \cap B) = 0$

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

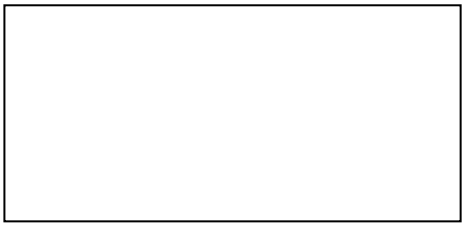


3. $A \cup A^c = \Omega$, A and A^c are exhaustive events.

   $A \cap A^c = \phi$, A and A^c are disjoint events.

4. $n(A^c) = n(\Omega) - n(A)$

   $P(A^c) = 1 - P(A)$



General Probability Rules:

1. $0 \le P(A) \le 1$

2. $P(\Omega) = 1$

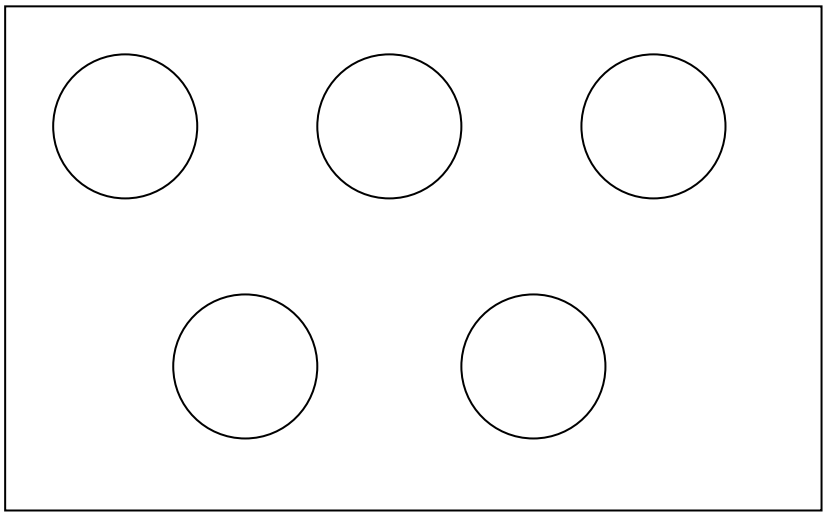
3. $P(\phi) = 0$

4. $P(A^c) = 1 - P(A)$

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

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

7. For disjoint events $E_1, E_2, \ldots, E_n$
$P(E_1 \cup E_2 \cup \ldots \cup E_n) = P(E_1) + P(E_2) + \cdots + P(E_n)$



2.3. Probability Applied to Health Data:-

Example 3.1:

630 patients are classified as follows : (Simple frequency table)

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 $\therefore n(\Omega) = 630$

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

$$n(E_1) = 284, \quad n(E_2) = 258, \quad n(E_3) = 63, \quad n(E_4) = 25.$$

$$P(E_1) = \frac{284}{630}, \quad P(E_2) = \frac{258}{630}, \quad P(E_3) = \frac{63}{630}, \quad P(E_4) = \frac{25}{630},$$

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

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

(since $E_2 \cap E_4 = \phi$)

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 3.2:

339 physicians are classified as follows.

		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

Experiment: Selecting a physician at random

$n(\Omega) = 339$ equally likely outcomes

Events:

- $A_3 =$ the selected physician is aged 40 - 49

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

- $B_2 =$ the selected physician smokes occasionally

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

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

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

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

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

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

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

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

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

(Since $A_2 \cap A_3 = \phi$)

3.3. (Percentage/100) As Probabilities And the Use of Venn Diagrams:

$$P(E)=\frac{n(E)}{n(\Omega)}=\ \textbf{??}$$

$n(\Omega)=\textbf{??}$ unknown $\qquad n(E)=$?? unknown

Suppose that $\%(E)=$ Percentage of elements of E relative to the elements of Ω, $n(\Omega)$, is known.

$$\%(E)=\frac{n(E)}{n(\Omega)}\times 100\%$$

$$n(E)=\frac{\%(E)*n(\Omega)}{100\%}$$

$$\therefore P(E)=\frac{n(E)}{n(\Omega)}=\frac{\%(E)*n(\Omega)}{100\%*n(\Omega)}=\frac{\%(E)}{100\%}$$

Example 3.3: (p.72)

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 some sort of medication.
- $D\cap M=$ The selected woman delivered prematurely and used some sort of medication.

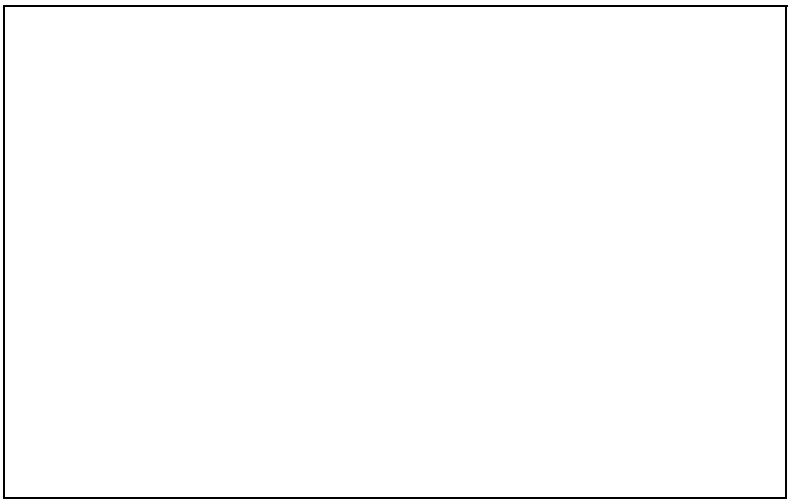
$\%(D)=10\%$ $\qquad \%(M)=25\%$ $\qquad \%(D\cap M)=5\%$

$$\therefore P(D)=\frac{\%(D)}{100\%}=\frac{10\%}{100\%}=0.1$$

$$P(M)=\frac{\%(M)}{100\%}=\frac{25\%}{100\%}=0.25$$

$$P(D\cap M)=\frac{\%(D\cap M)}{100\%}=\frac{5\%}{100\%}=0.05$$

A Venn diagram:



$P(D) = 0.1$
$P(M) = 0.25$
$P(D \cap M) = 0.05$
$P(D^c \cap M) = 0.2$
$P(D \cap M^c) = 0.05$
$P(D^c \cap M^c) = 0.70$
$P(D \cup M) = 0.30$

Probability given by a Venn diagram

A Two-way table:

	M	M^c	Total
D	**0.05**	0.05	**0.10**
D^c	0.20	0.70	0.90
Total	**0.25**	0.75	1.00

Probabilities given by a two-way table.

Calculating probabilities of some events:

M^c = The selected woman did not use medication

$$P(M^c) = 1 - P(M) = 1 - 0.25 = 0.75$$

$D^c \cap M^c$ = the selected woman did not deliver prematurely and did not use medication.

$$P(D^c \cap M^c) = 1 - P(D \cup M) = \text{ ??}$$

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

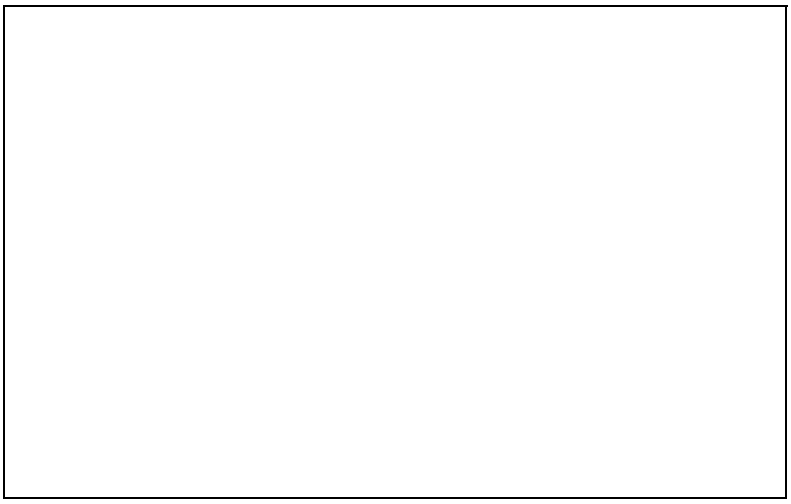
$$P(D \cup M) = P(D) + (M) - P(D \cap M)$$
$$= 0.1 + 0.25 - 0.05 = 0.3$$

$$P(D^c \cap M^c) = 1 - P(D \cup M) = 1 - 0.3 = 0.7$$

Note:

From the Venn diagram, it is clear that:

$$P(D) = P(D \cap M) + P(D \cap M^c)$$
$$P(M) = P(D \cap M) + P(D^c \cap M)$$
$$P(D \cap M^c) = P(D) - P(D \cap M)$$
$$P(D^c \cap M) = P(M) - P(D \cap M)$$
$$P(D^c \cap M^c) = 1 - P(D \cup M)$$



3.4. Conditional Probability:

- The conditional probability of the event A given the event B is defined by:

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

- P(A | B) = the probability of the event A if we know that the event B has occurred.
- Notes:

(1) $$P(A|B)=\frac{P(A\cap B)}{P(B)}$$
$$=\frac{n(A\cap B)/n(\Omega)}{n(B)/n(\Omega)}$$

$$\therefore P(A|B)=\frac{n(A\cap B)}{n(B)}$$

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

(3) $$\left.\begin{array}{l} P(A\cap B)=P(B)P(A|B) \\ P(A\cap B)=P(A)P(B|A) \end{array}\right\}$$ *multiplication rules*

Example:

		Smoking Habbit			
		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

- For calculating $P(A \mid B)$, we can use

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

(ii) Using the restricted table directly.

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

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

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

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

Or

$$P(B_1 \mid A_2) = \frac{n(B_1 \cap A_2)}{n(A_2)}$$

$$= \frac{110}{189} = 0.5820$$

Notice that $P(B_1) < P(B_1 \mid A_2)$!! … $P(B_1) \neq P(B_1 \mid A_2)$
What does this mean?

Independent Events

There are 3 cases:

- $P(A|B) > P(A)$
  which means that knowing B increases the probability of occurrence of A.
- $P(A|B) < P(A)$
  which means that knowing B decreases the probability of occurrence of A.

- $P(A|B) = P(A)$
  which means that knowing B has no effect on the probability of occurrence of A.
  In this case A is independent of B.

Independent Events:

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

$$(i)\ \ P(A|B) = P(A)$$
$$\Leftrightarrow (ii)\ \ P(B|A) = P(B)$$
$$\Leftrightarrow (iii)\ \ P(B \cap A) = P(A)P(B) \leftarrow \text{(multiplication rule)}$$

Example:

In the previous example, B_1 and A_2 are not independent because:

(1) $P(B_1) = 0.5192 \neq P(B_1 | A_2) = 0.5820$

also

(2) $P(B_1 \cap A_2) = 0.32448 \neq P(B_1)P(A_2) = 0.28945$

Combinations:

- **Notation:** n factorial is denoted by $n!$ and is defined by:

$$n! = n(n-1)(n-2)\cdots(2)(1) \qquad for \quad n \geq 1$$
$$0! = 1$$

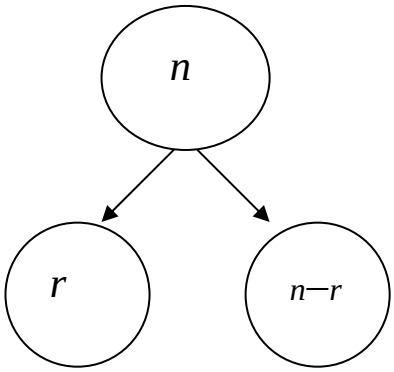
Example: $5! = (5)(4)(3)(2)(1) = 120$

- **Combinations:**

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

$$\binom{n}{r} = \frac{n!}{r!\ (n-r)!}; \qquad r = 0, 1, 2, \ldots, n$$

$\binom{n}{r}$ is read as " n " choose " r ".



$$\binom{n}{n} = 1 \qquad \binom{n}{0} = 1 \qquad \binom{n}{r} = \binom{n}{n-r}$$

Example 3.9:

If we have 10 equal–priority operations and only 4 operating rooms, in how many ways can we choose the 4 patients to be operated on first?

Answer:

$n = 10 \qquad r = 4$

The number of different ways for selecting 4 patients from 10 patients is

$$\binom{10}{4} = \frac{10!}{4!\,(10-4)!} = \frac{10!}{4!\,6!} = \frac{(10)(9)(8)\cdots(2)(1)}{(4)(3)(2)(1)\quad(6)(5)(4)(3)(2)(1)}$$
$$= 210 \qquad (different\ ways)$$

Chapter 4: Probability Distributions

Some events can be defined using random variables.

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

4.2. Probability Distributions of Discrete R.V.'s:-

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 person had a cold in January 1998 in Saudi Arabia.

Suppose we are able to count the no. of people in Saudi Arabia for which $X = x$

x (no. of times a person had a cold in January 1998 in S. A.)	Frequency of x (no. of people who had a cold in January 1998 in S.A.)
0	10,000,000
1	3,000,000
2	2,000,000
3	1,000,000
Total	$N =$ 16,000,000

Simple frequency table of no. of times a person had a cold in January 1998 in Saudi Arabia.

Experiment: Selecting a person at random
Define the event:

$(X = x)$ = The event that the selected person had a cold x times.
In particular,

$(X = 0)$=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.

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

$$\therefore P(X = x) = \frac{n(X = x)}{n(\Omega)} = \frac{\begin{pmatrix} no.\ of\ \ people\ \ who\ \ had\ \ x \\ colds\ \ in\ \ Januay\ \ 1998 \end{pmatrix}}{16{,}000{,}000}$$

x	freq. of x $n(X = x)$	$P(X = x) = n(X = x)/1600000$
0	10000000	0.6250
1	3000000	0.1875
2	2000000	0.1250
3	1000000	0.0625
Total	16000000	1.0000

Note:

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

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

This table is called the probability distribution of the discrete random variable X .

Notes:

- $(X=0), (X=1), (X=2), (X=3)$ are mutually exclusive (disjoint) events.
- $(X=0), (X=1), (X=2), (X=3)$ are exhaustive events.
- 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) $P(X \geq 2) = P(X = 2) + P(X = 3) = 0.1250 + 0.0625 = 0.1875$

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

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

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

or

$P(X \leq 2) = 1 - P\left((X \leq 2)^c\right)$
$= 1 - P(X > 2) = 1 - P(X = 3) = 1 - 0.625 = 0.9375$

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

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

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

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

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

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

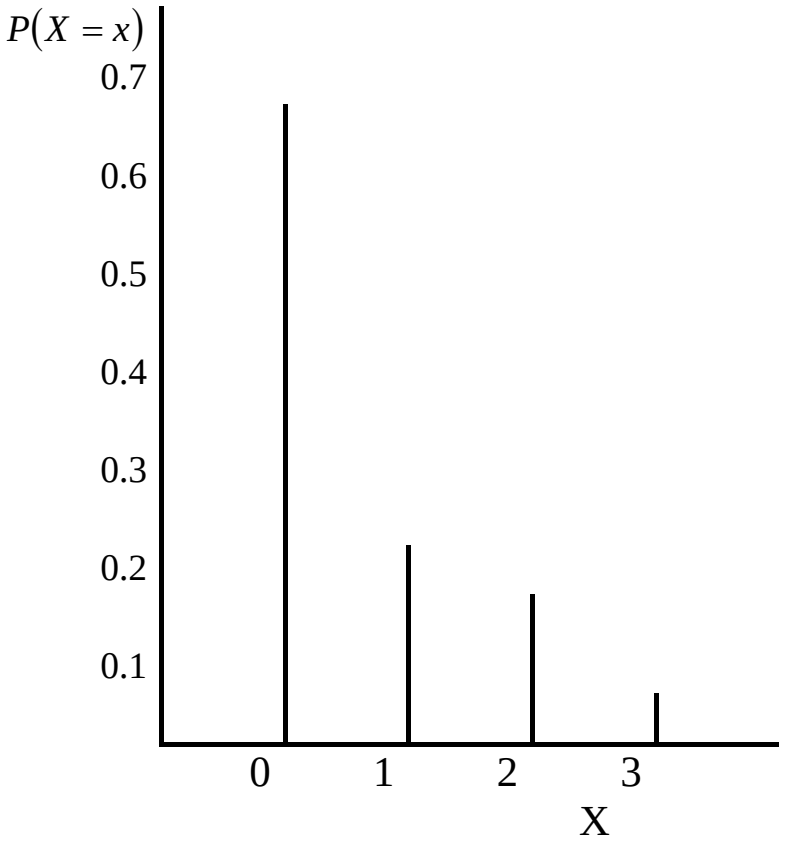
(11) The probability that the selected person had <u>more than</u> 2 colds $= P(X > 2) = P(X = 3) = 0.0625$

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

Graphical Presentation:

The probability distribution of a discrete r. v. X can be graphically presented as follows

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



Population Mean of a Discrete Random Variable

The mean of a discrete random variable X is denoted by μ and defined by:

$$\mu = \sum_x x\, P(X = x) \qquad \text{[mean = expected value]}$$

Example: We wish to calculate the mean μ of the discrete r. v. X in the previous example.

x	$P(X = x)$	$x\,P(X = x)$
0	0.6250	0.0
1	0.1875	0.1875
2	0.1250	0.2500
3	0.0625	0.1875
Total	$\sum P(X = x) = 1.00$	$\mu = \sum x\, P(X = x)$ =0.625

$$\mu = \sum_x x\, P(X = x) = (0)(0.625) + (1)(0.1875) + (2)(0.125) + (3)(0.0625) = 0.625$$

Cumulative Distributions:

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

$$P(X \le x) = \sum_{a \le x} P(X = a)$$

Example: The cumulative distribution of X in the previous example is:

x	$P(X \le x)$
0	0.6250
1	0.8125
2	0.9375
3	1.0000

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

Binomial Distribution:

- It is discrete distribution.
- It is used to model an experiment for which:

1. The experiment has *n* trials.
2. Two possible outcomes for each trial :
   S = success and *F*= failure
   (boy or girl, Saudi or non-Saudi,…)
3. The probability of success: $P(S)=\pi$ 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

The discrete r. v.:

X = The number of successes in the *n* trials

has a binomial distribution with parameter *n* and π , and we write:

$$X \sim Binomial(n,\pi)$$

The probability distribution of *X* is given by:

$$P(X=x)=\begin{cases}\binom{n}{x}\pi^{x}\,(1-\pi)^{n-x} & \text{for } x=0,\,1,\,2,\,\ldots,\,n\\ 0 & \text{otherwise}\end{cases}$$

where $$\binom{n}{x}=\frac{n!}{x!\,(n-x)!}$$

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

x	$P(X=x)$
0	$\binom{n}{0}\pi^0(1-\pi)^{n-0}=(1-\pi)^n$
1	$\binom{n}{1}\pi^1(1-\pi)^{n-1}$
2	$\binom{n}{2}\pi^2(1-\pi)^{n-2}$
⋮	⋮
$n-1$	$\binom{n}{n-1}\pi^{n-1}(1-\pi)^1$
n	$\binom{n}{n}\pi^n(1-\pi)^0=\pi^n$
Total	1.00

Result:

If $X\sim$ Binomial(n, π) , then

- The mean: $\mu = n\pi$ (expected value)
- The variance: $\sigma^2 = n\pi(1-\pi)$

Example: 4.2 (p.106)

Suppose that the probability that a Saudi man has high blood pressure is 0.15. If we randomly select 6 Saudi men, find the probability distribution of the number of men out of 6 with high blood pressure. Also, find the expected number of men with high blood pressure.

Solution:

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

S = Success: The man has high blood pressure
F = failure: The man does not have high blood pressure.

- Probability of success $P(S) = \pi = 0.15$
- no. of trials $n = 6$

$X \sim$ Binomial (6, 0.15) $\quad \begin{bmatrix} \pi = 0.15 \\ 1-\pi = 0.85 \\ n = 6 \end{bmatrix}$

The probability distribution of X is:

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

$$P(X=0) = \binom{6}{0}(0.15)^0(0.85)^6 = (1)(0.15)^0(0.85)^6 = 0.37715$$
$$P(X=1) = \binom{6}{1}(0.15)^1(0.85)^5 = (6)(0.151)(0.85)^5 = 0.39933$$
$$P(X=2) = \binom{6}{2}(0.15)^2(0.85)^4 = (15)(0.15)^2(0.85)^4 = 0.17618$$
$$P(X=3) = \binom{6}{3}(0.15)^3(0.85)^3 = (20)(0.15)^3(0.85)^3 = 0.04145$$
$$P(X=4) = \binom{6}{4}(0.15)^4(0.85)^2 = (15)(0.15)^4(0.85)^2 = 0.00549$$
$$P(X=5) = \binom{6}{5}(0.15)^5(0.85)^1 = (6)(0.15)^5(0.85)^1 = 0.00039$$
$$P(X=6) = \binom{6}{6}(0.15)^6(0.85)^0 = (1)(0.15)^6(1)^0 = 0.00001$$

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

The expected number (mean) of men out of 6 with high blood pressure is: $\mu = n\pi = (6)(0.15) = 0.9$

The variance is: $\sigma^2 = n\pi(1-\pi) = (6)(0.15)(0.85) = 0.765$

Poisson Distribution:

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

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

where e = 2.71828 (the natural number).

We write:

$X \sim$ Poisson (λ)

- The mean (average) of Poisson (λ) is : $\mu = \lambda$
- The variance is: $\sigma^2 = \lambda$

- The Poisson distribution is used to model a discrete r. v. which is a count of how many times a specified random event occurred in an interval of time or space.

Example:

- No. of patients in a waiting room in an hours.
- No. of serious injuries in a particular factory in a month.
- No. of calls received by a telephone operator in a day.
- No. of rats in each house in a particular city.

Note:

λ is the average (mean) of the distribution.

If X = The number of calls received in a month and

$X \sim$ Poisson (λ)

then:

(i) Y = The no. calls received in a year.

$Y \sim$ Poisson (λ^*), where $\lambda^* = 12\lambda$

$Y \sim$ Poisson (12λ)

(ii) W = The no. calls received in a day.

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

$W \sim \text{Poisson}\left(\frac{\lambda}{30}\right)$

Example:

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

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

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

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

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

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

Solution:

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

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

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

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

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

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

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

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

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

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

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

4.3. Probability Distributions of Continuous Random Variables:

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

(i) The total area under the curve of $f(x)=1$.

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

(ii) Probability of an interval event is given by the area under the curve of $f(x)$ and above that interval.

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

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

(i) $P(X = x) = 0$ for any x
(ii) $P(X \le a) = P(X < a)$
(iii) $P(X \ge b) = P(X > b)$
(iv) $P(a \le X \le b) = P(a \le X < b) = P(a < X \le b) = P(a < X < b)$
(v) $P(X \le x) =$ cumulative probability

(vi) $P(X \geq a) = 1 - P(X < a) = 1 - P(X \leq a)$
(vii) $P(a \leq X \leq b) = P(X \leq b) - P(X \leq a)$

$P(X \geq a) = 1 - P(X \leq a)$ $A = 1 - B$ Total area = 1	$P(a \leq X \leq b) = P(X \leq b) - P(X \leq a)$ $\int_a^b f(x)dx = \int_{-\infty}^{b} f(x)dx - \int_{-\infty}^{a} f(x)dx$

4.4. The Normal Distribution:

- One of the most important continuous distributions.
- Many measurable characteristics are normally or approximately normally distributed.
(examples: height, weight, …)

- 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
variance = σ^2

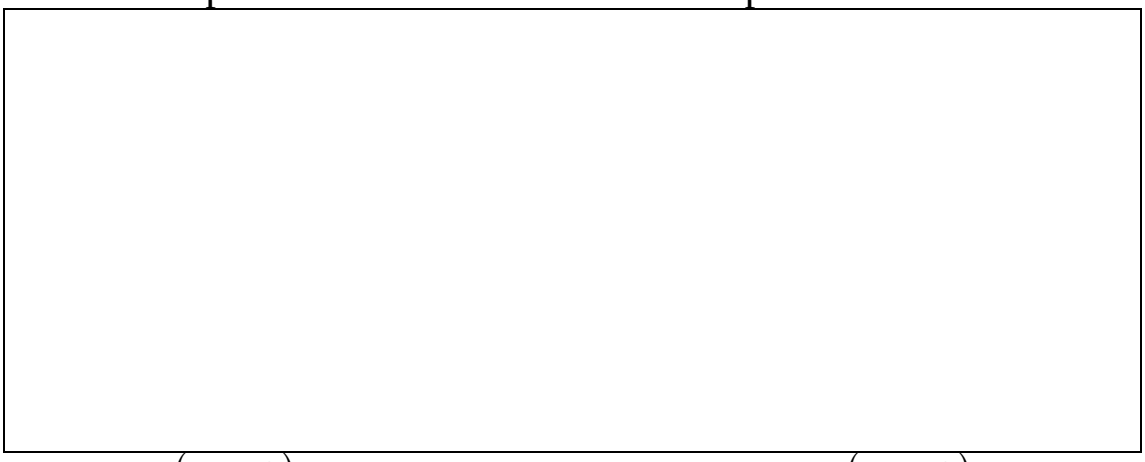
(3) The highest point of the curve of $f(x)$ at the mean μ.
The curve of $f(x)$ is symmetric about the mean μ.

$\therefore \mu =$ mean = mode = median

(4) The normal distribution depends on two parameters:
mean = μ and variance = σ^2

(5) If the r.v. X is normally distributed with mean μ and variance σ^2, we write:
$X \sim$ Normal (μ, σ^2) or $X \sim \mathrm{N}(\mu, \sigma^2)$

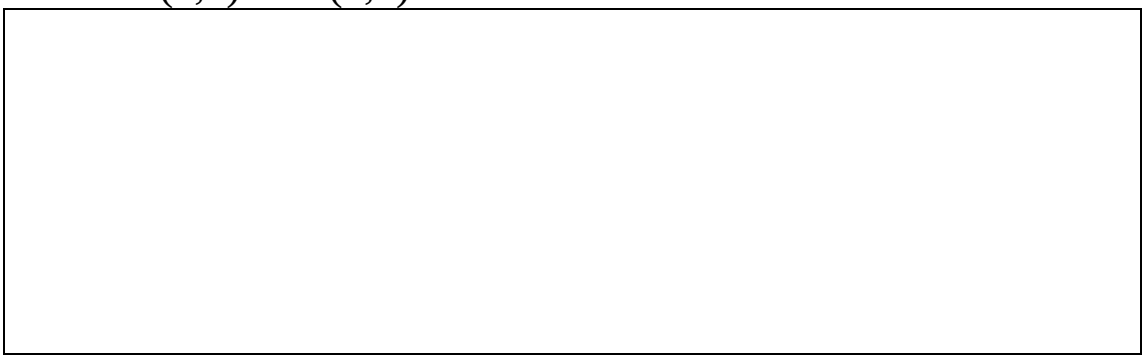
(6) The location of the normal distribution depends on μ
The shape of the normal distribution depends on σ^2



Normal (μ_1, σ_1^2) Normal (μ_2, σ_2^2)

The Standard Normal Distribution:

The normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$ is called the standard normal distribution and is denoted by Normal (0,1) or N(0,1)



- The standard normal 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 $Z = \dfrac{X-\mu}{\sigma} \sim$ Normal (0,1)

Calculating Probabilities of Normal (0,1):

Suppose Z ~ Normal (0,1).

(i) $P(Z \le a) =$ From Table (A) page 223, 224

(ii) $P(Z \ge a) = 1 - P(Z \le a)$ (Table A)

(iii) $P(a \le Z \le b) = P(Z \le b) - P(z \le a)$ (Table A)

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

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

(1)

$P(Z \le 1.50) = 0.9332$

Z	0.00	0.01	…
:	⇓		
1.5 ⇒	0.9332		
:			

(2)

$P(Z \ge 0.98) = 1 - P(Z \le 0.98)$

$= 1 - 0.8365$

$= 0.1635$

Z	0.00	…	0.08
:	:	:	⇓
:	…	…	⇓
0.9⇒	⇒	⇒	0.8365

(3)

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

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

$= 0.9922 - 0.0918$

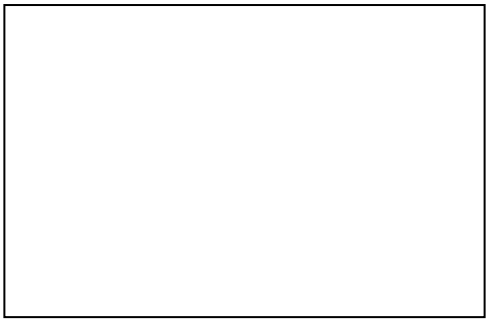
$= 0.9004$

Z	…	0.02	0.03
:	:	⇓	⇓
–1.3	⇒		0.0918
:		⇓	
2.4	⇒	0.9922	

(4) $P(Z \le 0) = P(Z \ge 0) = 0.5$

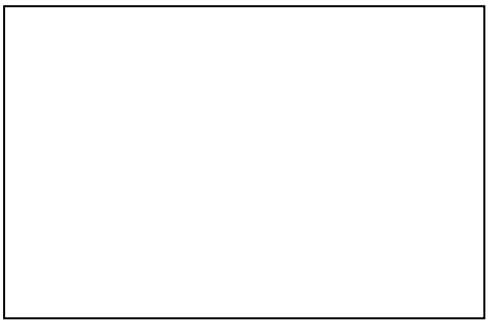
Notation:

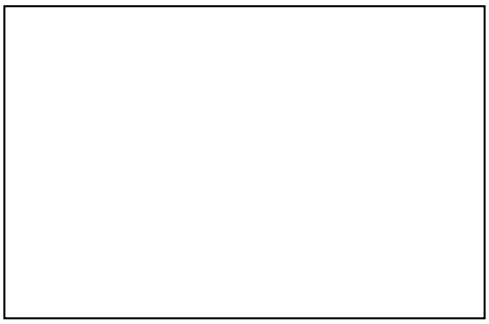
$P(Z \le Z_A) = A$



For example:

$P(Z \le Z_{0.025}) = 0.025$ $P(Z \le Z_{0.90}) = 0.90$





Example: Z ~ N(0,1)

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

Then $a = 1.65$

Z	…	0.05	…
:		⇑	
1.6	⇐	0.9505	
:			

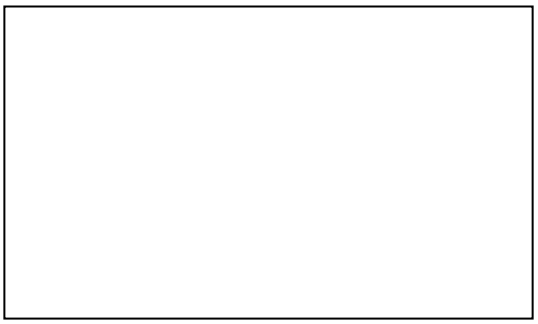
Example: Z ~ N(0,1)

$Z_{0.90} = 1.285$

$Z_{0.95} = 1.645$

$Z_{0.975} = 1.96$

$Z_{0.99} = 2.325$



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

- $X \sim$ Normal $(\mu, \sigma^2) \Leftrightarrow Z = \dfrac{X-\mu}{\sigma} \sim$ Normal (0,1)
- $X \le a \Leftrightarrow \dfrac{X-\mu}{\sigma} \le \dfrac{a-\mu}{\sigma} \Leftrightarrow Z \le \dfrac{a-\mu}{\sigma}$

(i) $P(X \le a) = P\left(Z \le \dfrac{a-\mu}{\sigma}\right)$

(ii) $P(X \ge a) = 1 - P(X \le a) = 1 - P\left(Z \le \dfrac{a-\mu}{\sigma}\right)$

(iii)

$$P(a \le X \le b) = P(X \le b) - P(X \le a) = P\left(Z \le \frac{b-\mu}{\sigma}\right) - P\left(Z \le \frac{a-\mu}{\sigma}\right)$$

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

Example 4.8 (p.124)

X = hemoglobin level for healthy adults males
$\mu = 16$ $\sigma^2 = 0.81$ $\sigma = 0.9$
$X \sim$ Normal (16, 0.81)

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

$$P(X \le 14) = P\left(Z \le \frac{14-\mu}{\sigma}\right)$$
$$= P\left(Z \le \frac{14-16}{0.9}\right)$$
$$= P(Z \le -2.22)$$
$$= 0.0132$$

$\therefore$ 1.32% of healthy adult males have hemoglobin level less than 14.

Example 4.9 :

X = birth weight of Saudi babies
$\mu = 3.4$ $\sigma = 0.35$ $\sigma^2 = (0.35)^2 = 0.1225$
X ~ Normal (3.4, 0.1225)

The probability that a randomly chosen Saudi baby has a birth weight between 3.0 and 4.0 kg is $P(3.0 < X < 4.0)$

$$
\begin{aligned}
P(3.0 < X < 4.0) &= P(X \le 4.0) - P(X \le 3.0) \\
&= P\left(Z \le \frac{4.0-\mu}{\sigma}\right) - P\left(Z \le \frac{3.0-\mu}{\sigma}\right) \\
&= P\left(Z \le \frac{4.0-3.4}{0.35}\right) - P\left(Z \le \frac{3.0-3.4}{0.35}\right) \\
&= P(Z \le 1.71) - P(Z \le -1.14) \\
&= 0.9564 - 0.1271 = 0.8293
\end{aligned}
$$

$\therefore$ 82.93% of Saudi babies have birth weight between 3.0 and 4.0 kg.

Standard Normal Curve
Z ~ Normal(0,1)

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

SOME RESULTS:

Result (1):

If $X_1, X_2, \ldots, X_n$ is random sample of size n from Normal(μ, σ^2), then:

(i) $\overline{X} \sim$ Normal $\left(\mu, \frac{\sigma^2}{n}\right)$

(ii) $Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \sim$ Normal (0,1) where $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$.

Result (2): (σ^2 is known): (Central Limit Theorem)

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

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \approx \text{Normal (0,1)} \qquad \text{(approximately)}$$

when the sample size n is large $(n \geq 30)$.

Note: "$\approx$" means "approximately distributed"

Result (3): (σ^2 is unknown)

If $X_1, X_2, \ldots, X_n$ is a random sample of size n from any distribution with mean μ, then:

$$Z = \frac{\overline{X} - \mu}{S / \sqrt{n}} \approx \text{Normal (0,1)}$$

when n is large $(n \geq 30)$.

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

Chapter 5: Hypothesis Testing and Estimation:

There are two main purposes of statistics;

- Descriptive Statistics: (Chapter 1 & 2)
  Organization & summarization of the data

- Statistical Inference: (Chapter 5)
  Answering research questions about some population parameters
  - $\Rightarrow$ (1) Hypothesis Testing:
    Answering questions about the population parameters
  - $\Rightarrow$ (2) Estimation:
    Approximating the actual values of parameters;
    - Point Estimation
    - Interval Estimation or
      Confidence Interval (C. I.)

We will consider two types of population parameters:

(1) Population means (for quantitative variables):

$\mu =$ The average (mean) value of some quantitative variable.

Example:
- The mean life span of some bacteria.
- The income mean of government employees in Saudi Arabia.

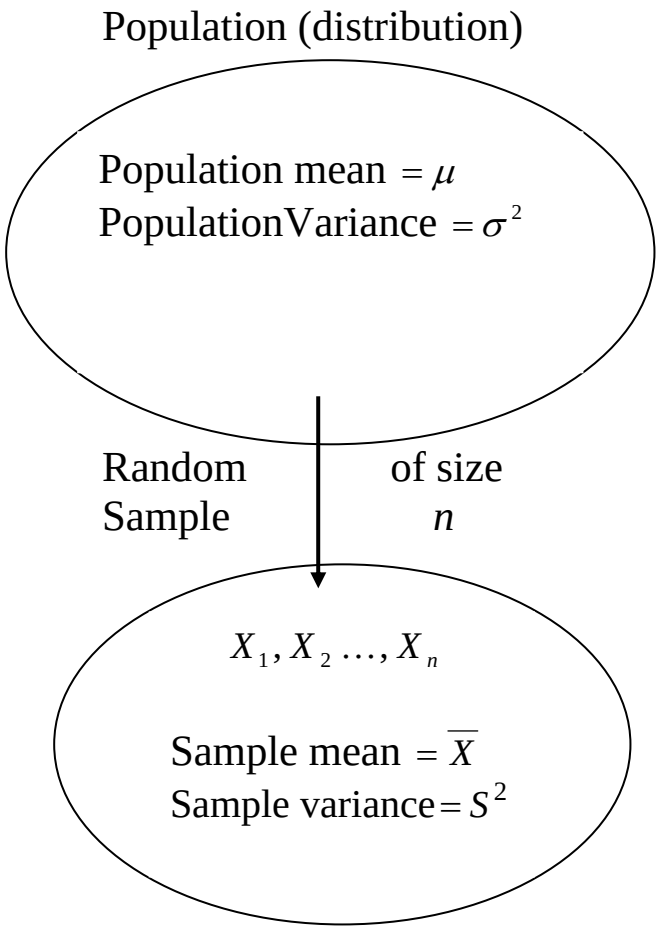
(2) Population proportions (for qualitative variables):

$$\pi = \frac{\text{no. of elements in the population with some specified characteristic}}{\text{Total no. of elements in the population}}$$

Example:

- The proportion of Saudi people who have some disease.
- The proportion of smokers in Riyadh
- The proportion of females in Saudi Arabia

(1) Estimation of Population Mean μ :



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

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

We are interested in estimating the mean of a population (μ)

(i) Point Estimation:

A point estimate is a single number used to estimate (approximate) the true value of μ .

- Draw a random sample of size *n* from the population:

$X_1, X_2 \ldots, X_n$

$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is used as a point estimator of μ .

(ii) Interval Estimation (Confidence Interval = C. I.):

An interval estimate of μ is an interval (*L*,*U*) containing the true value of μ "with probability $1-\alpha$ "

$1-\alpha$ is called the confidence coefficient
L = lower limit of the confidence interval
U= upper limit of the confidence interval

- Draw a random sample of size *n* from the population $X_1, X_2 \ldots, X_n$.

Result:

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

A $(1-\alpha)100\%$ confidence interval for μ is :

(i) if σ is known:

$$\left(\overline{X} - Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \ , \ \overline{X} + Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right) \quad \Leftrightarrow \quad \overline{X} \pm Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

(ii) if σ is unknown.

$$\left(\overline{X} - Z_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \ , \ \overline{X} + Z_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \right) \quad \Leftrightarrow \quad \overline{X} \pm Z_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}$$

Example:

Z ~ N(0,1)

$Z_{1-\frac{\alpha}{2}}$ = ??

$Z_{0.95} = 1.645$ ($\alpha = 0.1$)

$Z_{0.975} = 1.96$ ($\alpha = 0.05$)

Example:

Variable = blood glucose level (quantitative variable)

Population = diabetic ketoacidosis patients in Saudi Arabia of age 15 or more

parameter = μ = the average blood glucose level

$n = 123$ (large)

$\overline{X} = 26.2$

$S = 3.3$ (σ^2 is unknown)

(i) Point Estimation:

We need to find a point estimate for μ.

$\overline{X} = 26.2$ is a point estimate for μ

$\mu \approx 26.2$

(ii) Interval Estimation (Confidince Interval = C. I.):

We need to find 90% C. I. for μ.

90% = $(1-\alpha)100\%$

$1-\alpha = 0.9 \Leftrightarrow \alpha = 0.1$

$\frac{\alpha}{2} = 0.05 \quad \Leftrightarrow \quad 1-\frac{\alpha}{2} = 0.95$

$Z_{1-\frac{\alpha}{2}} = Z_{0.95} = 1.645$

$\therefore$ 90% confidence interval for μ is:

$$\left(\overline{X} - Z_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \, , \, \overline{X} + Z_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \right)$$

or

$$\left(26.2 - (1.645)\frac{3.3}{\sqrt{123}} \, , \, 26.2 + (1.645)\frac{3.3}{\sqrt{123}} \right)$$

or

$$(26.2 - 0.4894714 \, , \, 26.2 + 0.4894714)$$

or $(25.710529 \, , \, 26.689471)$

we are 90% confident that the mean μ lies in $(25.71 \, , \, 26.69)$ or

$$25.71 < \mu < 26.69$$

5.4. Estimation for a Population Proportion:-

- The population proportion is

$$\pi = \frac{N(A)}{N} \qquad (\pi \text{ is a parameter})$$

where

$N(A)$ = number of elements in the population with a specified characteristic "A"

N = total number of element in the population (population size)

- The sample proportion is

$$p = \frac{n(A)}{n} \qquad (p \text{ is a statistic})$$

where

$n(A)$ = number of elements in the sample with the same characteristic "A"

n = sample size

Result: For large sample sizes $(n \geq 30)$, we have

$$Z = \frac{p - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \approx \text{ Normal } (0,1)$$

Estimation for π :

(1) Point Estimation:

A good point estimate for π is p.

(2) Interval Estimation (confidence interval):

A $(1-\alpha)100\%$ confidence interval for π is

$$p \pm Z_{1-\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}} \qquad \text{(Approximation for large } n\text{)}$$

or

$$\left(p - Z_{1-\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}} \;,\quad p + Z_{1-\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}} \right),$$

Example 5.6 (p.156)

Variable=whether or not a women is obese (qualitative variable)
Population=all adult Saudi women in the western region seeking care at primary health centers
Parameter = π = The proportion of women who are obese

Sample:
n = 950 women in the sample
$n(A)$ = 611 women in the sample who are obese

$$\therefore\ p = \frac{n(A)}{n} = \frac{611}{950} = 0.643$$

is the proportion of women who are obese in the sample.

(1) A point estimate for π is p = 0.643

(2) We need to construct 95% C.I. about π.

$$95\% = (1-\alpha)100\% \quad \Leftrightarrow 0.95 = 1-\alpha$$
$$\Leftrightarrow \alpha = 0.05$$
$$\Leftrightarrow \frac{\alpha}{2} = 0.025$$
$$\Leftrightarrow 1-\frac{\alpha}{2} = 0.975$$

$$\therefore Z_{1-\frac{\alpha}{2}} = z_{0.975} = 1.96$$

$\therefore$ 95% C.I. about π is

$$p \pm Z_{1-\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}$$

or $$0.643 \pm (1.96)\sqrt{\frac{(0.643)(1-0.643)}{950}}$$
$$0.643 \pm (1.96)(0.01554)$$

or $$0.643 \pm 0.0305$$

or $$(0.6127\ ,\ 0.6735)$$

We are 95% confident that the proportion of obese women, π, lies in the interval $(0.61,\ 0.67)$ or:

$$0.61 < \pi < 0.67$$