106 Stat

Dr.Arwa M. Alshingiti
http://faculty.ksu.edu.sa/alshangiti

References

-*Biostatistics : A foundation in Analysis in the Health Science*
-*By :* Wayne W. Daniel
-*Elementary Biostatistics with Applications from Saudi Arabia*
By : Nancy Hasabelnaby

Chapter 1: Organizing and Displaying Data

1.1: Introduction

Here we will consider some basic definitions and terminologies

Statistics: Is the area of study that is interested in how to organize and summarize information and answer research questions.

Biostatistics: Is a branch of statistics that interested in information obtained from biological and medical sciences.

Population: Is the largest group of people or things in which we are interested in a particular time and about which we want to make some statement or conclusions.

Sample: A part of the population on which we collect data. The number of the element in the sample is called the sample size and denoted by *n*.

Variable: the characteristic to be measured on the elements of population or sample.

Types of variables

Qualitative: If the values of the variables are word indicating to which category an element of the population belongs.

Quantitative; if the value of the variable are numbers indicating how much or how many of something

Nominal: the value of the variables are names only

Examples:
*Gender: Female or male.
* Eye colour: Black, brown, green, etc

Ordinal: variables can be ordered.

Examples:
Educational level: elementary ,intermediate, high school.
Blood pressure: Low, medium, high

Discrete: Can have countable numbers of values (there are gaps between the values)

Examples:
*Number of patients admitted to a hospital in one day (x=1,2,…)
* Number of pain killer tablets (x= 0.5,1,1.5,2 ,2.5,…)

Note: Discrete values can take either integer values or decimal values with gaps between the values.

Continuous: Can have any value within a certain interval of values. it is usually measured on some scale in terms of some measurement units like kilograms, meters …etc

Examples:
*Level of chemical in drinking water
*height (140<x<190)
*blood sugar level of a person.

Example 1

Suppose we measure the amount of milk that a child drinks in a day (in ml) for a sample of 25 two-years children in Saudi Arabia.

The population: all two years children in Saudi Arabia

The variable: the amount of milk that a child drink in a day (in ml)

The variable is ***quantitative, continuous.***

The sample size is 25.

Example 2

Suppose we measure weather or not a child has a hearing loss for a sample of 20 young children with a history of repeated ear infections.

The population: all young children with a history of repeated ear infection.

The variable: whether or not a child has a hearing loss

The variable is ***qualitative, nominal.*** Since the values are either “yes’’ or ‘”no”.

The sample size is 20

Example 3

Suppose we measure the temperature for a sample of 25 animals having a certain disease.

The population

The variable

The type of the variable

The sample size

1.2 Organizing the Data

Suppose we collect a sample of size *n* from a population of interest. A first step in organizing is to order the data from smallest to largest (if it is not nominal). A further step is to count how many numbers are the same (if any). The last step is to organize it into a table called frequency table (or frequency distribution).

The frequency distribution has two kinds

1) Simple (ungrouped) frequency distribution: for
   - Qualitative variables
   - Discrete quantitative with small number of different variables

2) Grouped frequency distribution: for
   - Continuous quantitative variables
   - Discrete quantitative with large number of different variables.

Example 1.2.1: (simple frequency distribution)

Suppose we are interested in the number of children that a Saudi woman has and we take a sample of 16 women and obtain the following data on the number of children

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

Q1: What is the variable? The population? and the sample size?. What are the different values of the variable?

-the different values are: 0,1,2,3,4,5

Q2: Obtain a simple frequency distribution (table)?

If we order the data we obtained

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

To obtain a simple frequency distribution (table) we have to know the following concepts

The frequency: is obtained by counting how often each number in the data set .

The sample size (n): is the sum of the frequencies.

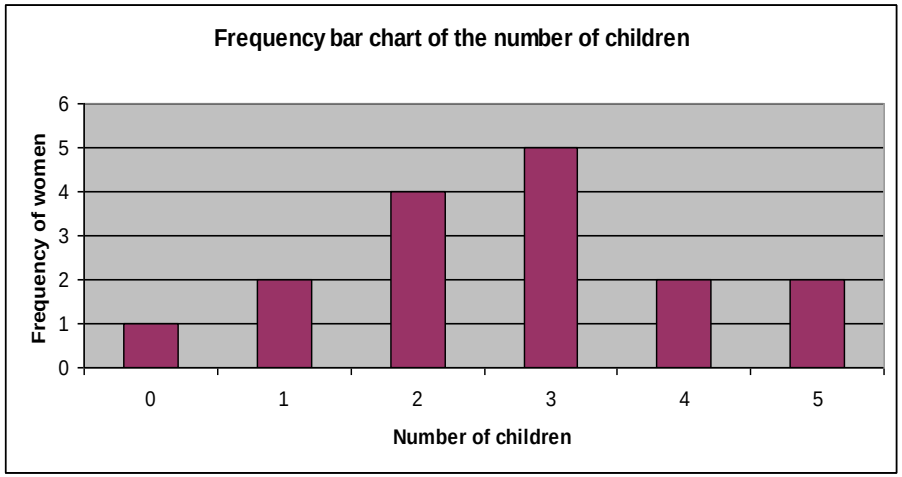
Relative frequency= frequency/n

Percentage frequency= Relative frequency*100= (frequency/n)*100.

Simple frequency table for the number of children.

Number of children (variable)	*frequency of women (frequency)*	*Relative frequency*	*Percentage frequency*
0			
1			
2			
3			
4			
5			
Total			

The simple frequency distribution has the *frequency bar chart* as graphical representation



Exercise: for more exercises and details about graphs
http://onlinestatbook.com/chapter2/graphing_qualitative.html

Example 1.2.2 :grouped frequency distribution

The following table gives the hemoglobin level (in g/dl) of a sample of 50 apparently healthy men aged 20-24. Find the grouped frequency distribution for the data.

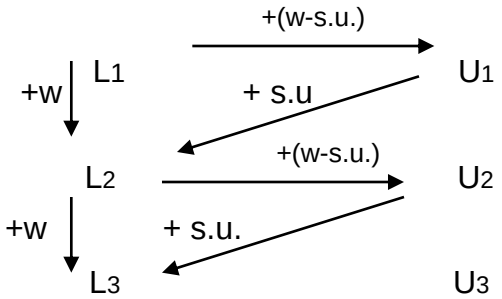
17	17.1	14.6	14	16.1	15.9	16.3	14.2	16.5
17.7	15.7	15.8	16.2	15.5	15.3	17.4	16.1	14.4
15.9	17.3	15.3	16.4	18.3	13.9	15	15.7	16.3
15.2	13.5	16.4	14.9	15.8	16.8	17.5	15.1	17.3
16.2	16.3	13.7	17.8	16.7	15.9	16.1	17.4	15.8

-What is the variable? The sample size?

- The max=18.8
-The min=13.5
-The range=max-min=18.8-13.5=4.8
-The smallest unit (s.u.)=0.1

Notes

1. The smallest unit is the smallest decimal used in the data. For example if we have data of the form (2.84, 1.34, 3.09,….) the smallest unit is 0.01. whereas the smallest unit of integer values like (5, 2, 4, , …) is 1.
2. In example 1.2.2 to group the data we use a set of intervals, called **class intervals**.
3. **The width (w)** is the distance from the lower or upper limit of one class interval to the same limit of the next class interval.
4. Let we denote the lower limit and upper limit of the class interval by L and U, that is the first class is L_1-U_1, the second class is L_2-U_2 …
5. To find ***the class intervals*** we use the following relationship



and so on

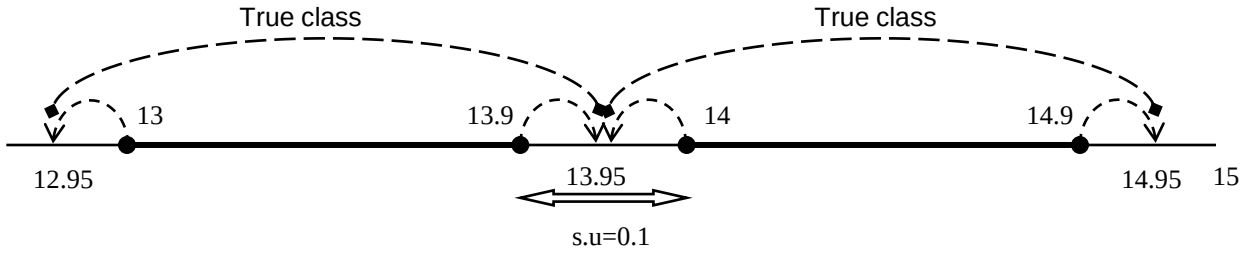
6. **Cumulative frequency**: is the number of values obtained in the class interval or before, which find by adding successfully the frequencies.
7. **Cumulative relative frequency**: is the proportion of values obtained in the class interval or before, which find by adding successfully the relative frequencies.
8. The Grouped frequency distribution for Example 1.2.2 is

Class Interval	Frequency	Relative frequency	Cumulative frequency	Cumulative relative frequency
13 - 13.9	3	0.06	3	0.06
14 - 14.9	5	0.1	8	0.16
15 - 15.9	15	0.3	23	0.46
16 - 16.9	16	0.32	39	0.78
17 - 17.9	10	0.2	49	0.98
18 - 18.9	1	0.02	50	1
Total	n=50	1		

1.3 True classes and displaying grouped frequency distributions

To Find the **true class intervals** we have two ways:

1) Subtract from the lower limit and add to the upper limit one- half of the smallest unit.
2) Decrease the last decimal place of the lower limit by 1 and put 5 after it, and for the upper limit we simply put 5 after the limit.



To illustrate this let us find the true classes of example 1.2.2

Class Interval	True class interval	Mid points	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	16.95 - <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
Total			n=50

Notes:

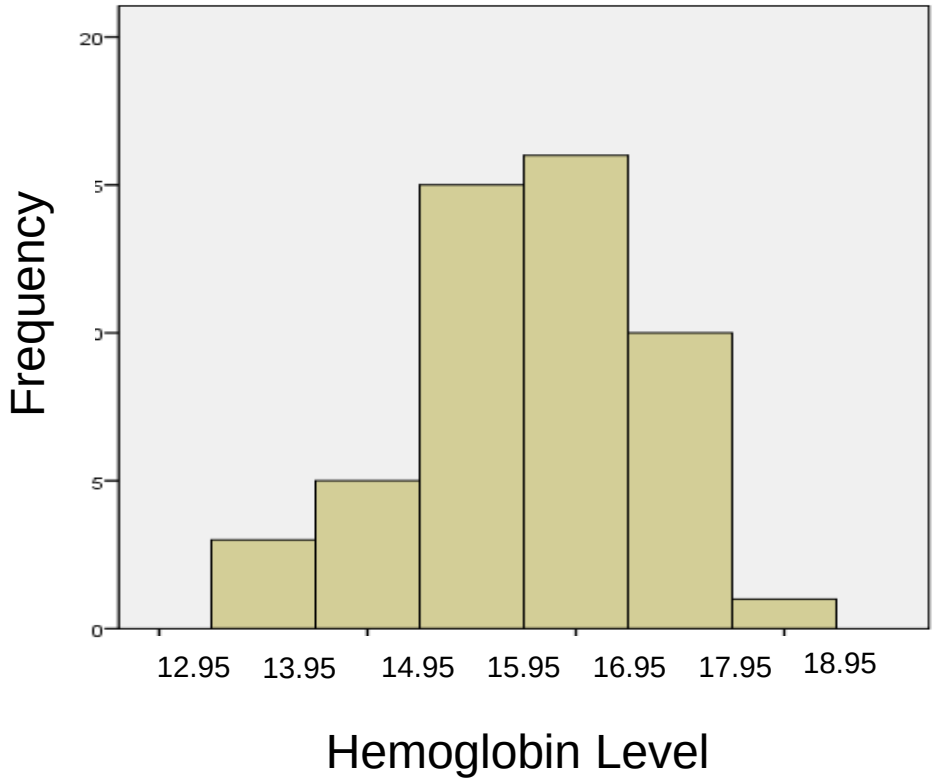
- Each upper limit of the true class interval ends with the same lower limit of the previous true class intervals
- The lower and upper limit of the true class interval must always end in 5, and they must always have one more decimal place than class limit.
- ***The mid point*** =(upper limit + lower limit)/2
- To find the midpoint of the interval we simply add the width to the previous midpoint.

1.4 Displaying grouped frequency distributions

Grouped frequency distributions can be displayed by

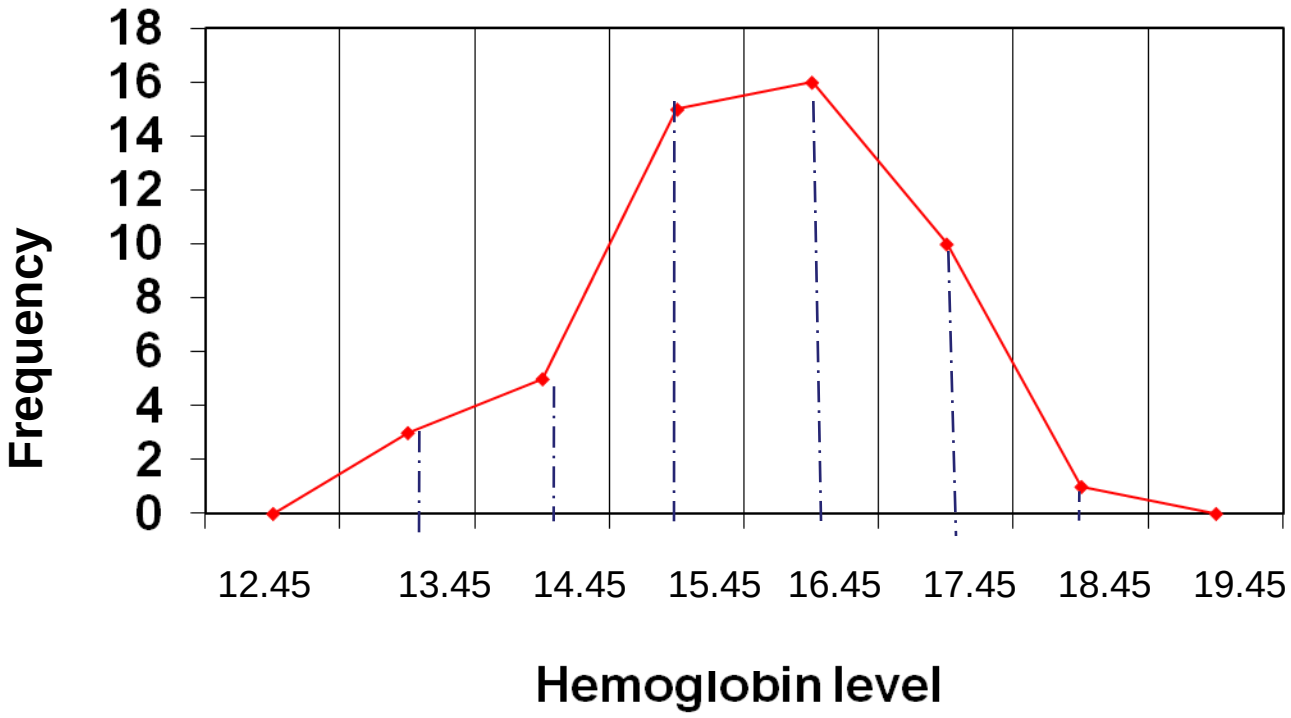
Histogram
Polygon
curves
} ⇨ For frequency or relative frequency distributions

Ogives ⇨ For cumulative or cumulative relative frequency distributions



⇦ Histogram

Polygon



Exercises: 1.R.1 (a-c-d-e) , 1.R.2 (a-c-d-e), 1.R.5 pg 25

	Horizontal access (x-access)	Vertical access (y-access)	Notes
Histogram	True classes	Frequency or relative frequency	
Polygon	Midpoints	Frequency or relative frequency	The ends are extended down to the x-access by midpoint of additional cells.
Curve	Midpoints	Frequency or relative frequency	

Revision: In the study, the blood glucose level (in mg/100 ml) was measured for a sample from all apparently healthy adult males.

a) Identify variable and the population in the study.

b) Complete the following table for this study, w= , s.u=

	Class interval (glucose level)	True class interval	Midpoints	Frequency	Relative frequency	Cumulative frequency
1.	70-79	69.5 –< 79.5	74.5	3		3
2.	80-89			12	0.16	
3.	90-99	89.5 –< 99.5	94.5	24	0.32	39
4.		99.5 –<109.5				
5.	110-119	109.5 –<119.5	114.5	6	0.08	75
	Total			75	1	

c) Using the above table, answer the questions:

i) What percentage of these males had blood glucose levels from 80 to 109 (mg/ 100 ml)?

ii) What number of these males has blood glucose levels of 99 (mg/ 100 ml) or less?

d) Make a relative frequency histogram and a relative frequency polygon

Chapter 2: Basic Summary Statistics

2.1: Introduction

This chapter concerns mainly about describing the "middle" of the observations and "how spread out" they are.

Measures of central tendency

Measures which are in some sense indicate where the "middle" or "centre" of the data is.
(e.g.Mean, median and mode)

Measures of dispersion

Measures which indicate how spread out the observation from each other.
(e.g. Range, variance, standard deviation and coefficient of variation)

Population

The population values of the variable of interest: $X_1, X_2, \ldots, X_N$ (usually they are unknown).

N=The population size

Sample

the sample values of the variable :

$x_1, x_2, \ldots, x_n$

***n*= the sample size.**

Any measure obtained from the population values of the variable of interest is called *a parameter*

Any measure obtained from the sample values of the variable of interest is called *a statistics*

2.2: Measures of central tendency

We use the term central tendency to refer to the natural fact that the values of the variable often tend to be more concentrated about the centre of the data. We will consider three such measures: the mean, the median and the mode.

Mean: (or average)

Population mean: let $X_1, X_2, \ldots, X_N$ be the population values of the variable (usually unknown), then the population mean is $\mu = \frac{X_1 + X_2 + \ldots + X_N}{N} = \frac{\sum X_i}{N}$

Sample mean :let $x_1, x_2, \ldots, x_n$ be the sample values of the variable, then the sample mean is $\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum x_i}{n}$

- The sample mean is an estimator of a population mean.
- Question: which one is a parameter and which one is a statistic?

Example 2.1: Consider a population consisting of the 5 nurses who work in a particular clinic, and we are interested in the age of these nurses in years

$X_1=30, X_2=22, X_3=35, X_4=27, X_5=41$

Then average nurse population is

$\mu = \frac{30+22+35+27+41}{5} = \frac{155}{5} = 31$ years.

Median (or med) The median is the middle value of the ordered observation

To find the median of a sample of n observation, we first order the data, then

1) If *n* is odd, the middle observation is the order (*n*+1)/2.

2) If *n* is even, the middle two observations are the *n*/2 and the next observation, the median is the average of them.

Example 2.2: Find the median of the following samples

a) 29, 30, 32, 31, 28, 29, 30, 42, 40, 40, 40.

First we order the data 28, 29, 29, 30, 30, 31, 32, 40, 40, 40, 42

n= 11, odd, the order of the median is (n+1)/2=(11+1)/2=6th

28, 29, 29, 30, 30, 31, 32, 40, 40, 40, 42

6th

med=31 (unit)

b) 1.5, 3.0, 18.5, 24.0, 12.0, 4.5, 6.0, 9.5, 10.5, 15.0, 11.0, 11.5

n=12, even, n/2=6th , hence we take the average of the 6th and the 7th value

The ordered sample is 1.5, 3.0, 4.5, 6.0, 9.5, 10.5, 11, 11.5, 12.0, 15.0, 18.5, 24.0

6th 7th

med=(10.5+11)/2=10.75 (unit)

Mode (or modal) The mode of set of values is that value which occurs with highest frequency .

Any data must has one of the three cases

- No mode: example: Data(1): 21, 15, 22 ,19, 14, 18

  Data(2): 3, 3, 5,5, 4, 4, 6, 6
- One mode, example :Data (1): 32, 15, 23, 17 , 22, 23, 19, 20, 22 .

  The mode=22 (unit)

  Data(2): 13.5, 12, 13.5, 15, 15, 14.6, 17, 12, 15

  The mode=15 (unit)
- More than one mode: example 18, 20, 19, 19, 21, 17, 20

  modes: 19 , 20 (unit)

Notes:

- Mean and median can only be found for quantitative variables, the mode can be found for quantitative and qualitative variables.
- There is only one mean and one median for any data set.
- The mean can be distorted by extreme values so much.
- A measure not affected so much by extreme values is the median.

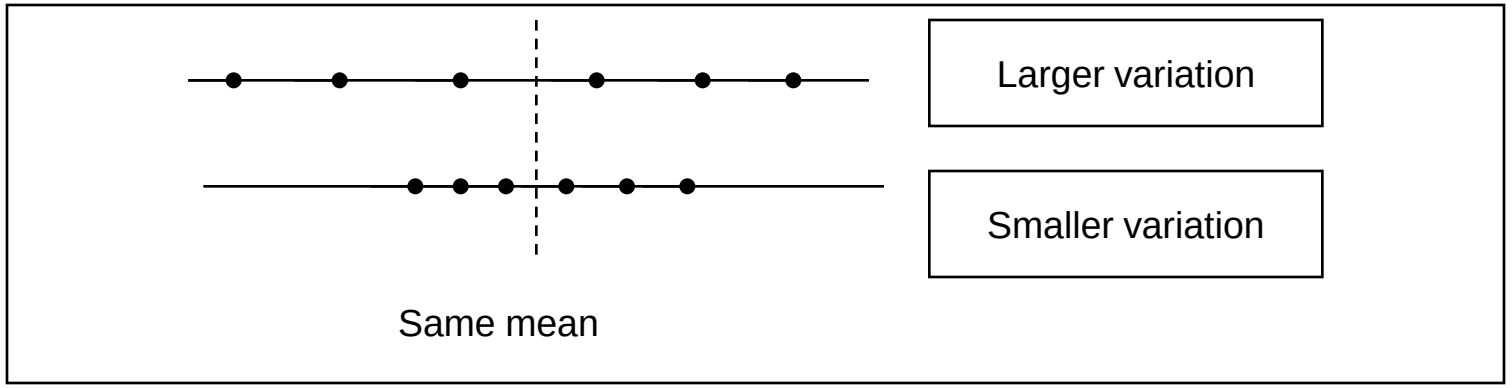
Animated example on the web:http://standards.nctm.org/document/eexamples/chap6/6.6/index.htm

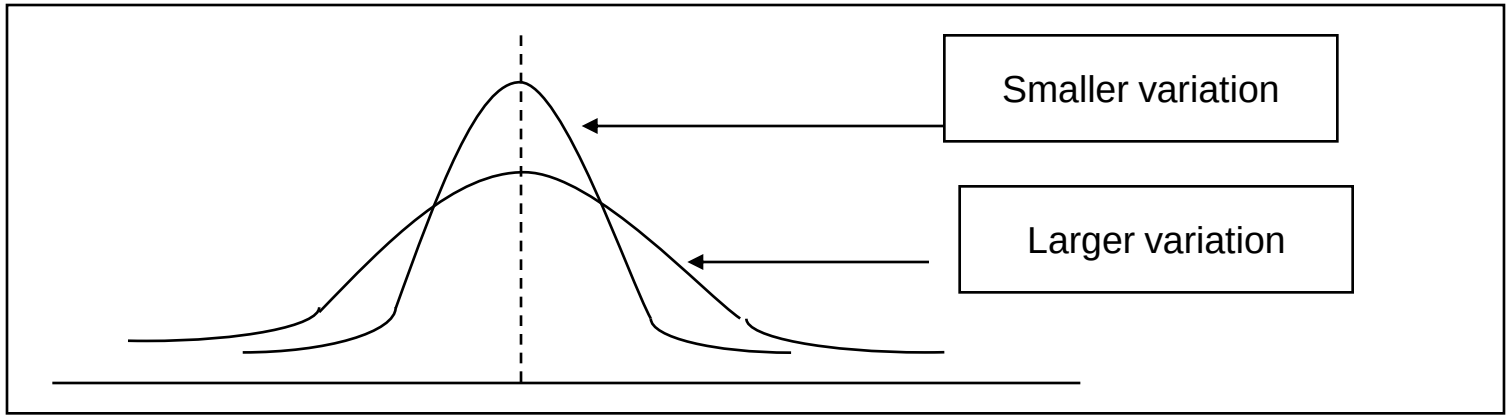
2.3: Measure of dispersion

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

-When the variation is small, this means that the observations are close to each other (but not the same).

- Can you mention a case when there is no variation?





We will consider four measures of dispersion: the range, the variance, the standard deviation and the coefficient of variation.

Range (R): Is the difference between the largest and smallest values in the set of values

Example 2.3 (q2.6- pg 35): Below are the birth weights (in kg) for a sample of babies born in Saudi Arabia:

1.69, 1.79, 3.32, 3.26, 2.71, 2.42, 2.59, 1.05, 3.19, 3.40, 3.23, 3.37, 3.6, 3.63

- Find the mean, mod and median.
- Find the range.

Note: The range is easy to calculate but it is not useful as a measure of variation since it only takes into account two of the values.

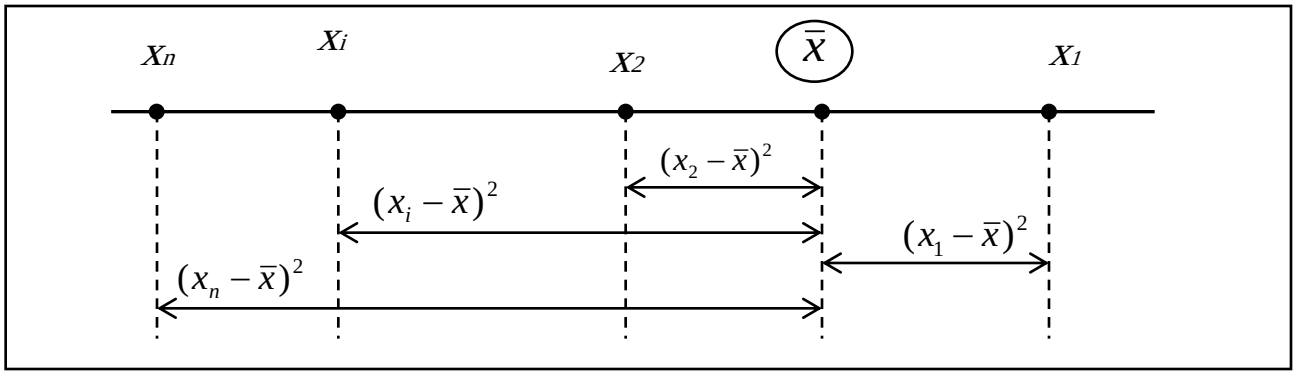
Variance: Is a measure which uses the mean as point of reference.

- Population variance: let $X_1, X_2, \ldots, X_N$ be the population values of the variable (usually unknown), then the population variance is $\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}$ where μ is the population mean.

- Sample Variance :let $x_1, x_2, \ldots, x_n$ be the sample values of the variable, then the sample variance is $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$ where $\bar{x}$ is the sample mean.

Notes:

- The variance is less when all the values are close to the mean, while it is more when all the values are spread out of the mean.



- The variance is always a nonnegative value ($\sigma^2 \geq 0, s^2 \geq 0$).
- Population variance σ^2 is usually unknown (parameter), hence it is estimated by the sample variance s^2 (statistic).
- A simpler formula to use for calculating sample variance is $\boxed{s^2 = \frac{\sum_{i=1}^{n} x_i^2 - n(\bar{x})^2}{n-1}}$
- The variance is expressed in squared unit.

Example 2.4: Consider a sample with observations of length in centimeters such as 10, 21, 33, 53, 54 . Find the mean and the variance?

Solution: To find the variance of the sample we need to calculate $\sum x_i$, and $\sum x_i^2$

n=5 (the sample size)

$\sum x_i$= 10+21+33+53+54=171

$\sum x_i^2$=(10) ^{2} + + (21) ^{2} + (33) ^{2} + (53) ^{2} + (54) ^{2} =7355

$\bar{x}$ = $\sum x_i / n$ =171/5= 34.2 cm

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

Standard deviation (std. dev.): The standard deviation is defined to be the root of the variance.

Population standard deviation

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

Sample standard deviation

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

Example: Find the standard deviation in example 2.4 above?

Solution: $s = \sqrt{s^2} = \sqrt{376.7} = 19.41$ cm

Coefficient of variation (CV):

- The variance and standard deviation are useful as measures of variation of the values of single variable for a single population.
- If we want to compare the variation in two data set the variance and standard deviations may give misleading results because:
  - The two variable may have different units as kilogram and centimeters which cannot be compared.
  - Although the same units are used, the mean of the two may be quit different in size.
- The coefficient of variation (CV) is used to compare the relative variation in two data set and it dose not depend on either the unit or how large the values are, the formula of CV is given by

$$\boxed{CV = \frac{s}{\bar{x}} \times 100(\%)}$$

- Suppose we have two data set as the following and we want to compare the variation

	mean	Std.dev.	CV
Set 1	$\bar{x}_1$	s_1	$CV_1 = \frac{s_1}{\bar{x}_1} \times 100\,(\%)$
Set2	$\bar{x}_2$	s_2	$CV_2 = \frac{s_2}{\bar{x}_2} \times 100\,(\%)$

Then we say that the variability in the first data set is larger than the variability in the second data set if $CV_1 > CV_2$ (and vice versa).

Example 2.5

Suppose two sets of samples of human males of different ages give the following results weight

set1: on males aged 29: $\bar{x}_1$=66kg s_1=4.5kg ⇒ CV_1=----------

set2: on males aged 10: $\bar{x}_2$=36kg s_1=4.5kg ⇒ CV_2=----------

Since $CV_2 > CV_1$, the variability in the weight of the 2nd set (10-years old) is greater than the variability in the 1st data set (29-years old).

Examples: 2.9 +2.11 pg 41

Some properties of mean, variance and std. dev.:

- If we multiply the data by constant then:

  the new mean would be multiplied by that constant, the new std. dev. would be multiplied by the absolute value of the constant, and the variance would be multiplied by the square of that constant.

- If we add or subtract a constant from the data then: the constant would be added to or subtracting from the new mean, whereas the new variance and std. dev. Would be the same.

2.4: Calculating measures from an ungrouped frequency tables:

Suppose we have the following frequency table, where m_i is the i^{th} value in the ungrouped frequency table or the midpoint in the grouped frequency table, and f_i is the i^{th} frequency. The formulas for sample mean and variance will be modified as follows:

Value (or midpoint)	frequency
m_1 m_2 ⋮ m_k	$f1$ f_2 ⋮ f_k
	$\sum f_i = n$

$n = \sum f_i$ (the sample size= the sum of frequencies)

k=number of distinct values (or number of intervals)

$$\sum_{i=1}^{n} x_i = \sum_{i=1}^{k} m_i\ f_i \quad , \quad \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{k} m_i^2\ f_i$$

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

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - n(\bar{x})^2}{n-1} \iff s^2 \approx \frac{\sum_{i=1}^{k} m_i^2\ f_i - n(\bar{x})^2}{n-1}$$

For using calculator to find the mean, variance and standard deviation, you can visit the site

http://faculty.ksu.edu.sa/alshangiti

Notes:

- When data are grouped we cannot determine from the frequency distribution what the actual data values are but only how many of them are in the class interval.

- We can't find the actual values for the sample mean and sample variance but we can find approximation of them.

- For grouped data we assume that all values in particular class interval are located at the midpoint of the interval (m_i) because the mid point is best representative for whole interval

Example 2.6:

Suppose that in a study on drug consumption by pregnant Saudi women, the number of different drugs taking during pregnancy was determined for a sample of Saudi women who took at least one medication obtaining:

Value m_i	Frequency f_n	Cumulative frequency	$m_i f_n$	$m_i^2 f_n$
1	5	5	5	5
2	11	16	22	44
3	7	23	21	63
4	3	26	12	48
5	2	28	10	50
6	1	29	6	36
7	1	30	7	49
Total	n=30		83	295

Find the measure of central tendency and dispersion.

Solution: n=30

- $\bar{x}$=---------- drugs
- To find the median: since n=30 is even, the order of the two middle values is $n/2$=15th and 16th, from the cumulative frequency the 16th and 15th ordered observation is 2, and hence
- Med=---------- drugs

- The mode is 2 since it has the highest frequency.

The variance $s^2 = \frac{\sum m_i^2 f_i - n(\bar{x})^2}{n-1} = \ldots\ldots\ldots$

- The range : R=……
- The standard deviation s=……
- The coefficient of variation: CV=…….

Note: we didn't put any unit here since the variable is discrete, the word (drug) is just an indicator of what we are counting

==========

Example 2.7: The following are the ages of a sample of 100 women having children who were admitted to a particular hospital in Madinah in particular month.

Class Interval	Mid points	Frequency
15-19	17	8
20-24	22	16
25-29	27	32
30-34	32	28
35-39	37	12
40-44	42	4
Total		n=100

Find the mean, the variance, and the coefficient of variation.

Solution: n=--- (the sample size), k=--- (number of intervals)

$$\sum m_i f_i = 2860 \ , \quad \sum_{i=1}^{k} m_i^2 f_i = 85540$$

- The mean $\bar{x} \approx \frac{\sum_{i=1}^{k} m_i f_i}{n} = \frac{2860}{100} = 28.6$ years
- The variance $s^2 \approx \frac{\sum_{i=1}^{k} m_i^2 f_i - n(\bar{x})^2}{n-1} = \frac{85540 - (100)(28.6)^2}{99} = 37.81$ (years)^2
- The standard deviation $s = \sqrt{37.818182} = 6.14964893$ years
- The coefficient of variation $CV = \frac{6.14964893}{28.6} = 21.5\ \%$

A site that explains the concepts in Arabic
http://www.jmasi.com/ehsa/

A site that explains how to use SPSS for descriptive statistics
http://academic.udayton.edu/gregelvers/psy216/spss/descript1.htm

Chapter 3: Some Basic Probability Concepts

3.1 General view of probability

Probability: The probability of some event is the likelihood (chance) that this event will occur.

An experiment: Is a description of some procedure that we do.

The universal set (Ω): Is the set of all possible outcomes,

An event: Is a set of outcomes in Ω which all have some specified characteristic.

Notes:

1. Ω (the universal set) is called sure event

2. ϕ (the empty set) is called impossible event

Example (3.1)

Consider a set of 6 balls numbered 1, 2, 3, 4, 5, and 6. If we put the sex balls into a bag and without looking at the balls, we choose one ball from the bag, then this is an experiment which is has 6 outcomes.

- Ω ={1, 2, 3, 4, 5 ,6 }
- Consider the following events
  - E_1=the event that an even number occurs={2, 4, 6}.
  - E_2=the event of getting number greater than 2={3,4, 5, 6}.
  - E_3=the event that an odd number occurs={1, 3, 5}.
  - E_4=the event that a negative number occurs={}= ϕ .

Equally likely outcomes:

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

Probability of equally likely events

consider an experiment which has N equally likely outcomes, and let the numbers of outcomes in an event E given by ***n(E),*** then the probability of E is given by

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

Notes

1. For any event A , $0 \leq P(A) \leq 1$ (why?)

   That is, probability is always between 0 and 1.

2. $P(\Omega)=1$, and $P(\phi)=0$ (why?)

1 means the event is a certainty, 0 means the event is impossible

Example (3.2)

In the ball experiment we have
$n(\Omega)=6$, $n(E_1)=3$, $n(E_2)=4$, $n(E_2)=3$

$P(E_1)$=--------

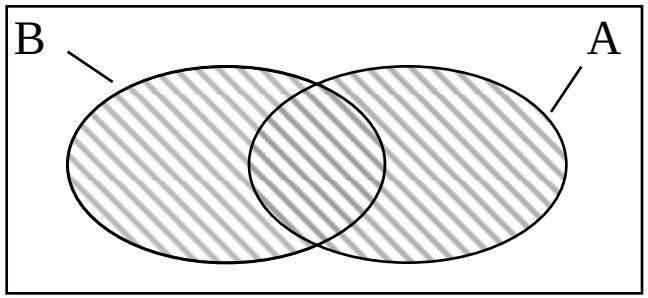
$P(E_2)$=----------

$P(E_3)$=----------

$P(E_4)$=----------

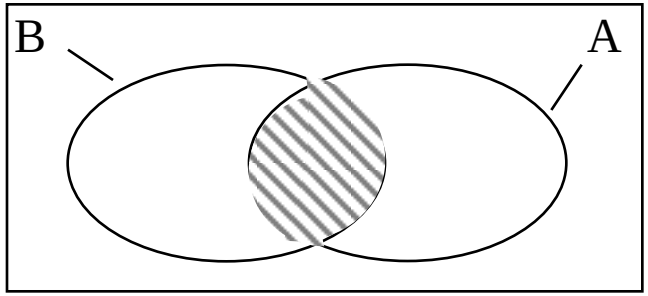
Relationships between events

❖ **Union** : A ∪ B, consists of all those outcomes in A or in B or in both A and B



A ∪ B

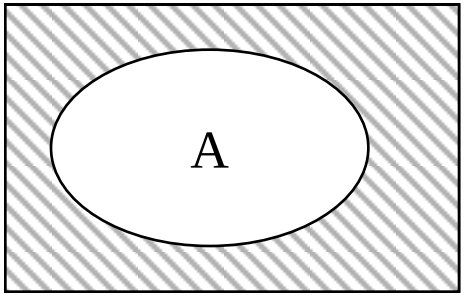
❖ **Intersection** : A ∩ B, consists of all those outcomes in both A and B



A ∩B

❖ **Complement** : A^c (or A`)

Consists of all outcomes that are in Ω but not in A



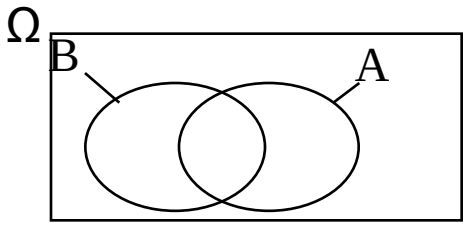
A^c

Notes:

1- $n(A \cup B) = n(A)+n(B)-n(A \cap B)$

and hence

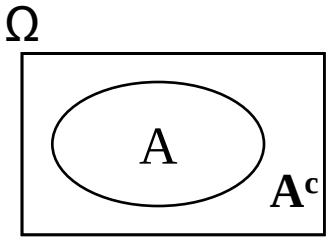
$P(A \cup B) = P(A)+P(B)-P(A \cap B)$



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

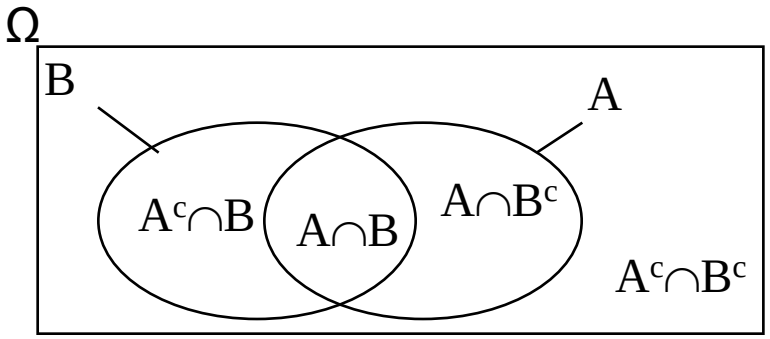
So that

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

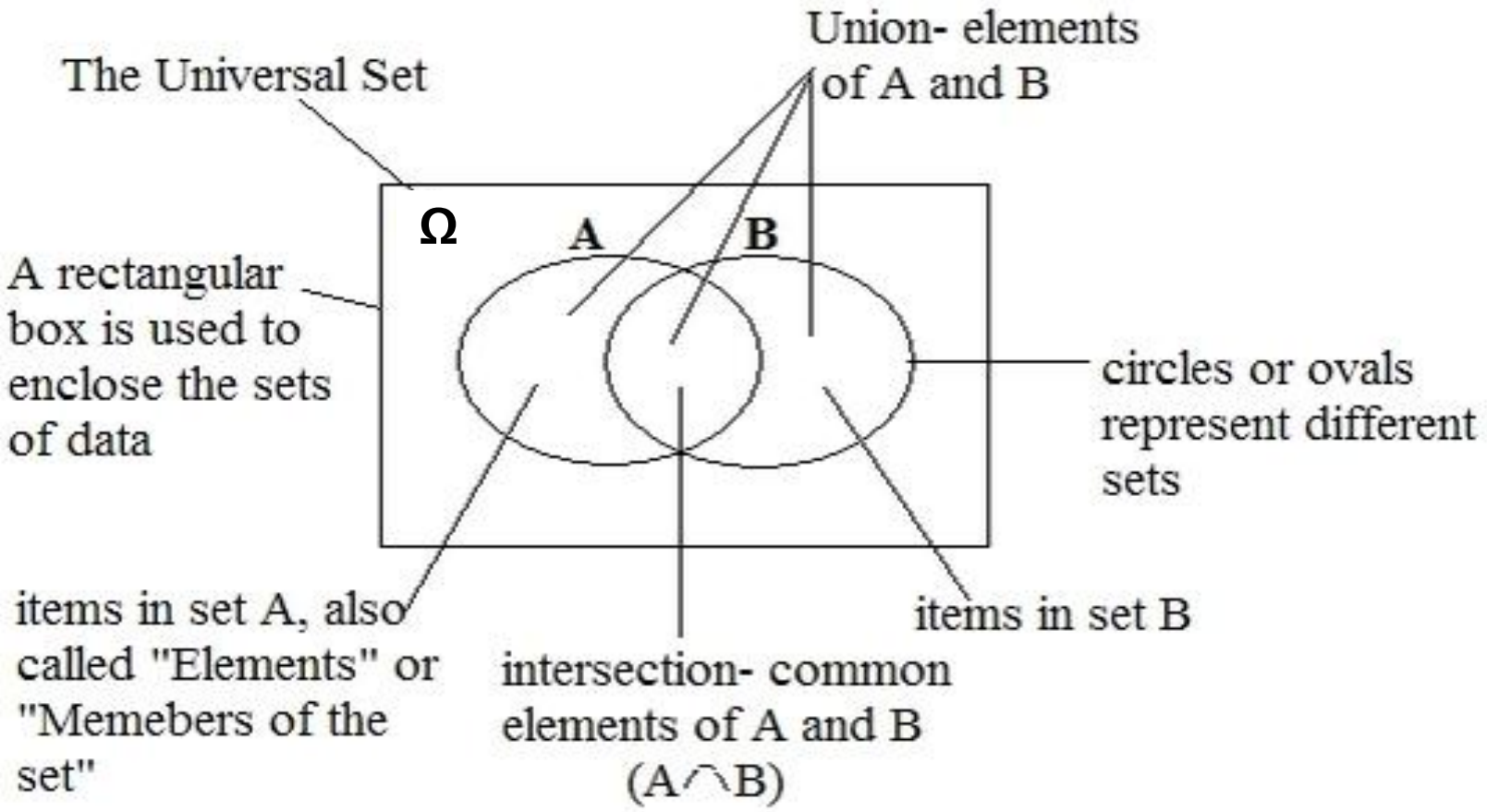


Sets (events) can be represented by

<u>Venn Diagram</u>



A Venn Diagram:



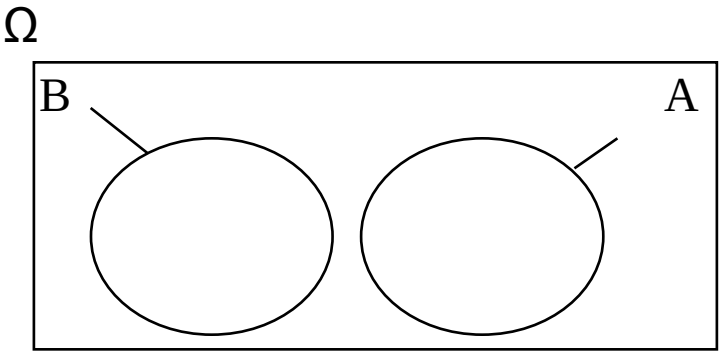
Disjoint events

Two events A and B are said to be disjoint (mutually exclusive) if $A \cap B = \phi$.

- In the case of disjoint events

$P(A \cap B)=0$

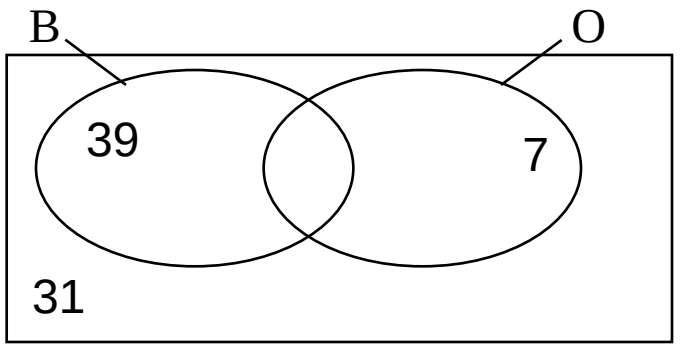
$P(A \cup B)= P(A)+P(B)$



Example 3.3

From a population of 80 babies in a certain hospital in the last month, let the even B="is a boy", and O="is over weight" we have the following incomplete Venn diagram.

- It is a boy
- It is not a boy and overweight
- It is a boy or it is overweight



Conditional probability:

the conditional probability of A given B is equal to the probability of $A \cap B$ divided by the probability of B, providing the probability of B is not zero.

That is

$P(A|B)=P(A \cap B)/ P(B)$, $P(B) \neq 0$

Notes:

1. $P(A|B)$ is the probability of the event A if we know that the event B has occurred
2. $P(B|A)=P(A \cap B)/ P(A)$, $P(A) \neq 0$

Example

Referring to example 3.3 what is the probability that

- He is a boy knowing that he is over weight?

- If we know that she is a girl, what is the probability that she is not overweight?

Independent events

-Two events A and B are said to be independent if the occurrence of one of them has no effect on the occurrence of the other.

Multiplication rule for independent events

If A and B are independent then

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

Notes

The multiplication rule is equivalent to

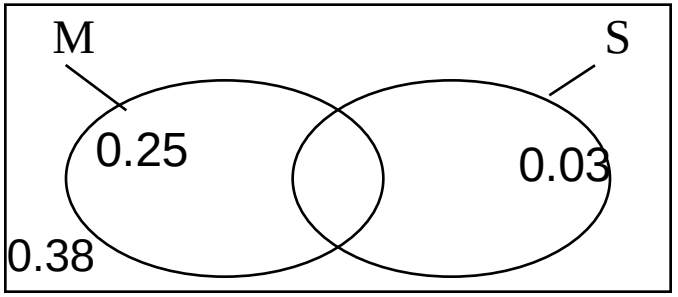
$P(A \mid B)= P(A)$ and $P(B \mid A)= P(B)$

Example 3.4

In a population of people with a certain disease, let M="Men" and
S="suffer from swollen leg "

We have the following incomplete Venn diagram

If we randomly choose one person



- Complete the Venn diagram

- Find the probability that this person

1- Is a man and suffer from swollen leg ?

2- Is a women?

3- Is a women and does not suffer from swollen leg ?

4- Does not suffering from swollen leg?

Marginal prbability:

Definition: Given some variable that can be broken down into m categories designated by $A_1, A_2,\ldots,A_m$ and another jointly accurance variable that is broken down into n categories designated by $B_1, B_2,\ldots,B_n$, the *marginal probability* of A_i , called $P(A_i)$, is equal to the sum of the joint probabilities of A_i with all categories of B. That is

$P(A_i)=\sum P(A_i \cap B_j)$, for all values of j.

This will be clear in the following example

Example 3.5:

The following table shows 1000 nursing school applicants classified according to scores made on a college entrance examination and the quality of the high school form which they graduated, as rated by the group of educators.

Score		Quality of high school			
		Poor (P)	Average (A)	Superior (S)	total
	Low(L)	105	60	55	
	Medium(M)	70	175	145	
	High(H)	25	65	300	
	total				

- How many marginal probabilities can be calculated from these data? State each probability notation and do calculations.
- Calculate the probability that an applicant picked at random from this group:

1-Made a low score on the examination

2- Graduated from superior high school.

3- Made a low score on the examination given that he or she graduated from Superior high school

5- Made a high score or graduated from a superior high school.

- Calculate the following probabilities
  1. P(A)
  2 . P(S)
  3. P(M)
  4. P(M∩P)
  5. P(A∪ L
  6. P(P ∩ S)
  7.P(L∪ H)
  8. P(H/S)

Chapter 4: Probability Distribution

4.1 Probability Distribution of Discrete Random Variables

- Random variable: is a variable that measured on population where each element must have an equal chance of being selected.

- let X be a discrete random variable, and suppose we are able to count the number of population where X=x, then the value of x together with the probability P(X=x) are called probability distribution of the discrete random variable X.

Example 4.1

Suppose we measure the number of complete days that a patient spends in the hospital after a particular type of operation in Dammam hospital in one year, obtaining the following results.

Number of days, x	Frequency
1	5
2	22
3	15
4	8
N	50

The probability of the event { X=x } is the relative frequency

$$P(X=x)=\frac{n(X=x)}{n(S)}=\frac{n(X=x)}{N}$$

That is: P(X=1)=5/50=0.1

P(X=2)=22/50=0.44

P(X=3)=15/50=0.3

P(X=4)=8/50=0.16

- What is the value of $\sum P(X=x)$?

Number of days, x	P(X=x)
1	0.1
2	0.44
3	0.3
4	0.16
Sum	1

The probability distribution must satisfy the conditions

`

1- $0 \le P(X = x) \le 1$

2- $\sum P(X = x) = 1$

The first condition must be satisfied since P(X=x) is a probability, and the second condition must be satisfied since the events {X=x} are mutually exclusive and there union is the sample space.

-**Population mean for a discrete random variable**: If we know the distribution function P(X=x) for each possible value x of a discrete random variable, then we the population mean (or the expected value of the random variable X) is

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

Example: The expected number of complete days that a patient spends in the hospital after a particular type of operation in Dammam hospital in one year (example 3.1) is

$\mu = \sum x\, P(X = x)$ =1(0.1)+2(0.44)+3(0.3)+4(0.16)=--------

-Cumulative distributions : the cumulative distribution or the cumulative probability distribution of a random variable is $P(X \le x)$

It is obtained in a way similar to finding the cumulative relative frequency distribution for samples.

-referring to example 3.1

P(X≤1)=0.1

P(X≤2)=P(X=1)+P(X=2)=0.1+0.44=0.54

$P(X\leq 3)=P(X=1)+P(X=2)+P(X=3)=0.1+0.44+0.3=0.84$

$P(X\leq 4)=P(X=1)+P(X=2)+P(X=3)+ P(X=4)=0.1+0.44+0.3+0.16=1$

The cumulative probability distribution can be displayed in the following table

Number of days x	$P(X=x)$	$P(X\leq x)$
1	0.1	0.1
2	0.44	0.54
3	0.3	0.84
4	0.16	1
Sum	1	

-From the table find:

1-$P(X<3)=P(X\leq 2)=0.54$

2-$P(2\leq X\leq 4)=P(X=4)+P(X=3)+P(X=2)=0.9$

Or $P(2 \leq X\leq 4)=P(X\leq 4)-P(X<2)=1-0.1=0.9$

3-$P(X>2)= P(X=3)+P(X=4)=0.46$

Or $P(X>2)= 1-P(X\leq 2)=1-0.54=0.46$

In general we can use the following rules for integer number a and b

1- $P(X \leq a)$ is a cumulative distribution probability

2-$P(X < a)=P(X \leq a-1)$

3-$P(X \geq b)=1-P(X< b)=1-P(X \leq b-1)$

4-$P(X>b)= 1-P(X \leq b)$

5- $P(a \leq X \leq b)=P(X \leq b) – P(X<a)= P(X \leq b) – P(X \leq a-1)$

6- $P(a<X \leq b) = P(X \leq b) - P(X \leq a)$

7-$P(a \leq X<b) = P(X \leq b-1) - P(X \leq a-1)$

8- $P(a<X<b)=P(X \leq b-1)-P(X \leq a)$

4.2 Binomial Distribution

The binomial distribution is a discrete distribution that is used to model the following experiment

1-The experiment has a finite number of trials *n*.

2- Each single trial has only two possible (mutually exclusive)outcomes of interest such as recovers or doesn't recover; lives or dies; needs an operation or doesn't need an operation. We will call having certain characteristic *success* and not having this characteristic *failure.*

3- The probability of a success is a constant π for each trial. The probability of a failure is 1- π.

4- The trials are independent; that is the outcome of one trial has no effect on the outcome of any other trial.

Then the discrete random variable X=the number of successes in *n* trials has a Binomial(n,π) distribution for which the probability distribution function is given by

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

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

Note

If the discrete random variable X has a binomial distribution , we write

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

The mean and variance for the binomial distribution:

- The mean for a Binomial(n , π) random variable is $\mu = \Sigma x\, P(X=x) = n\pi$

The variance $\sigma^2 = n\pi(1-\pi)$

Example 4.2

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

a- Find the probability distribution function for the number of men out of 6 with high blood pressure.

b- Find the probability that there are 4 men with high blood pressure?

c-Find the probability that all the 6 men have high blood pressure?

d-Find the probability that none of the 6 men have high blood pressure?

e- what is the probability that more than two men will have high blood pressure?

f-Find the expected number of high blood pressure.

Solution:

Let X=the number of men out of 6 with high blood pressure.

Then X has a binomial distribution (why ?).

Success= The man has a high blood pressure

Failure= The man doesn't have a high blood pressure

Probability of success= π=0.15 (and hence *Probability of failure=1-π=0.85)*

Number of trials=n=6

n=6 , *π=0.15* , *1- π=0.85*

- Then X has a Binomial distribution , *X~ Bin (6,0.15)*

a - the probability distribution function is $P(X=x)=\binom{6}{x}0.15^{x}(0.85)^{6-x}$ $x=0,1,\ldots,6$

---

b- the probability that 4 men will have high blood pressure

P(X=3)= $\binom{6}{4}0.15^{4}(0.85)^{2}$ =(15)(0.15)^4(0.85)^2=--------

---

C- the probability that all the 6 men have high blood pressure

P(X=6)= $\binom{6}{6}0.15^{6}(0.85)^{0}$ =-------

d-the probability that none of 6 men have high blood pressure is

P(X=0)= $\binom{6}{0}0.15^0(0.85)^6 = 0.85^6 = ------$

e- the probability that more than two men will have high blood pressure is

P(X>2)=1-P(X≤2)=1-[P(X=0)+P(X=1)+P(x=2)]

$=1-[\ 0.37715 + \binom{6}{1}0.15^1(0.85)^5 + \binom{6}{2}0.15^2(0.85)^4]$

$=1-[\ ------- + ------- + -------] \quad =1- \quad ----- = -----$

F- the <u>expected number of high blood pressure</u> is $\mu = n\pi = -------$

and <u>the variance</u> is $\sigma^2 = n\pi(1-\pi) = --------$

4.3 The Poisson Distribution

The Poisson distribution is a discrete distribution that is used to model the random variable X that represents the number of occurrences of some random event in the interval of time or space.

The probability that X will occur (the probability distribution function) is given by:

$$P(X=x)=\begin{cases}\dfrac{e^{-\lambda}\lambda^{x}}{x!}, & x=0,1,2,.....\\ 0 & \text{otherwise}\end{cases}$$

λ is the average number of occurrences of the random variable in the interval.

The mean $\mu=\lambda$

The variance $\sigma^2=\lambda$

If X has a Poisson distribution we write X~ Poisson (λ)

Examples of Poisson distribution:

- The number of patients in a waiting room in an hour.
- The number of serious injuries in a particular factory in a year.
- The number of times a three year old child has an ear infection in a year.

- **Example 4.3:**

Suppose we are interested in the number of snake bite cases seen in a particular Riyadh hospital *in a year*. Assume that the average number of snake bite cases at the hospital in a year is 6 .

1- What is the probability that in a randomly chosen year, the number of snake bites cases will be 7?

2- What is the probability that the number of cases will be less than 2 in 6 months?

3-What is the probability that the number of cases will be 13 in 2 year ?

4- What is Expected number of snake bites in a year? What is the variance of snake bites in a year?

Solution:

X= number of snake bite cases seen at this hospital *in a year.*

Then X~ Poisson (6)

First note the following

- The average number of snake bite cases at the hospital in a *year* $= \lambda = 6$

  X~ Poisson (6)

- The average number of snake bite cases at the hospital in 6 *months*

  = the average number of snake bite cases at the hospital in (1/2) *year* $= (1/2)\lambda = 3$

  y~ Poisson (3)

- The average number of snake bite cases at the hospital in 2 *years* $= 2\lambda = 12$

  V~ Poisson (12)

1- The probability that the number of snake bites will be 7 in *a year*

$\lambda=6$

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

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

2- The probability that the number of cases will be less than 2 in 6 months

$\lambda^{*}=3$

$$P(Y=y)=\frac{e^{-3}3^{x}}{3!}$$

$$P(Y<2)=P(Y=0)+P(Y=1)$$

$$=\frac{e^{-3}3^{0}}{0!}+\frac{e^{-3}3^{1}}{1!}\qquad =---+---\qquad =----$$

Remember

If X~ Poisson (λ)

$$P(X=x)=\frac{e^{-\lambda}\lambda^{x}}{x!}$$

$x=0,1,2,....$

3- The probability that the number of cases will be 13 in 2 years

$\lambda^{**}=12$

$$P(V=v)=\frac{e^{-12}12^{v}}{v!}$$

$$P(V=13)=\frac{e^{-12}12^{13}}{13!}\qquad =---$$

4- the expected number of snake bites in a year: $\mu=---$

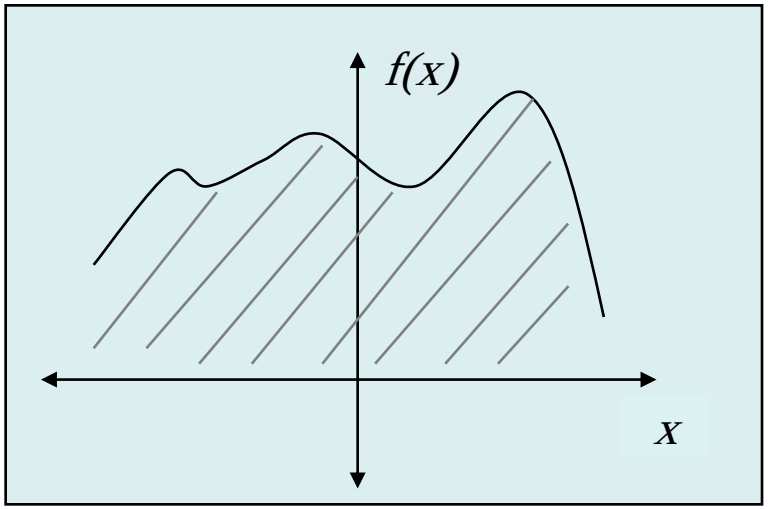
$\lambda=6$

the variance of snake bites in a year: $\sigma^{2}=---$

4.4 Probability Distribution of Continuous Random Variable

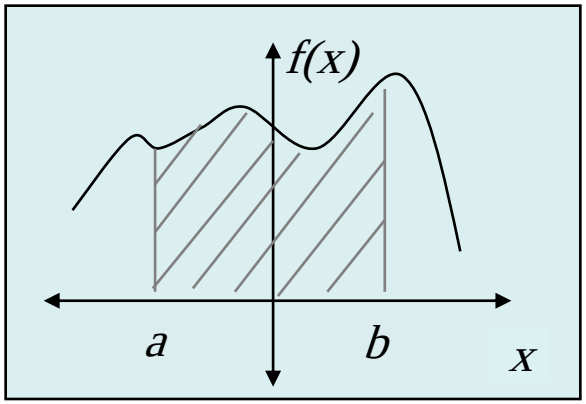
If X is a continuous random variable, then there exist a function *f(X)* called *probability density function* that has the following properties:

1- The area under the probability curve $f(x)=1$

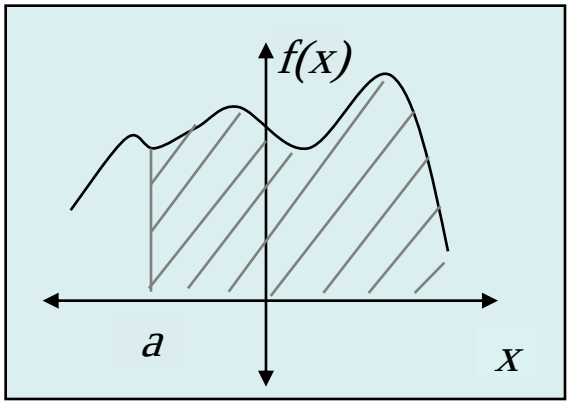


$$\text{area}=\int_{-\infty}^{\infty} f(x)dx=1$$

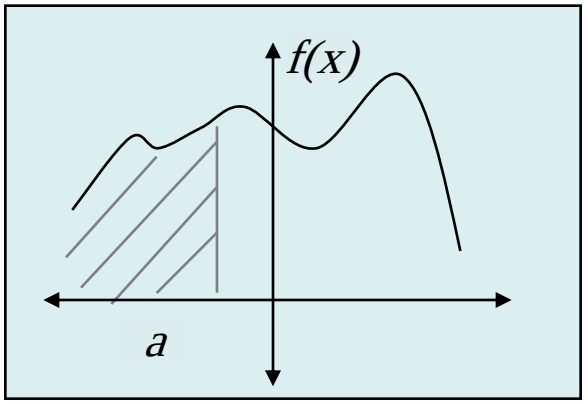
2- Probability of interval events are given by areas under the probability curve



$$P(a \le X \le b) = \int_a^b f(x)dx$$



$$P(X \ge a) = \int_a^{\infty} f(x)dx$$



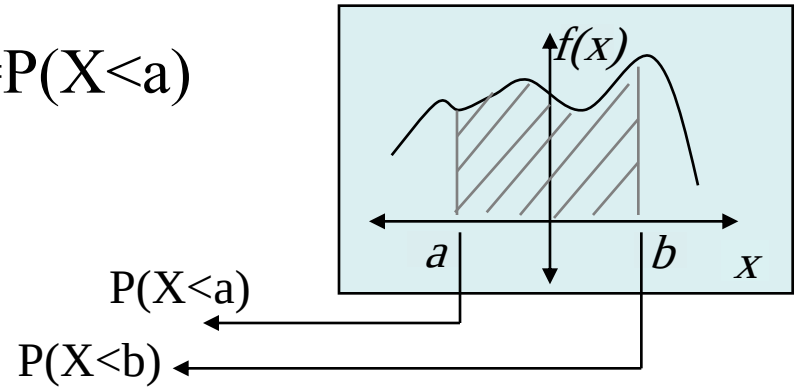
$$P(X \le a) = \int_{-\infty}^a f(x)dx$$

3- P(X=a)=0 (why?)

4-P(X≥a)=P(X>a) and P(X≤ a)=P(X<a)

5- P(X≥a)= 1- P(X≤a)

6-P(a<X<b)=P(X<b)-P(X<a)



4.5 The Normal Distribution:

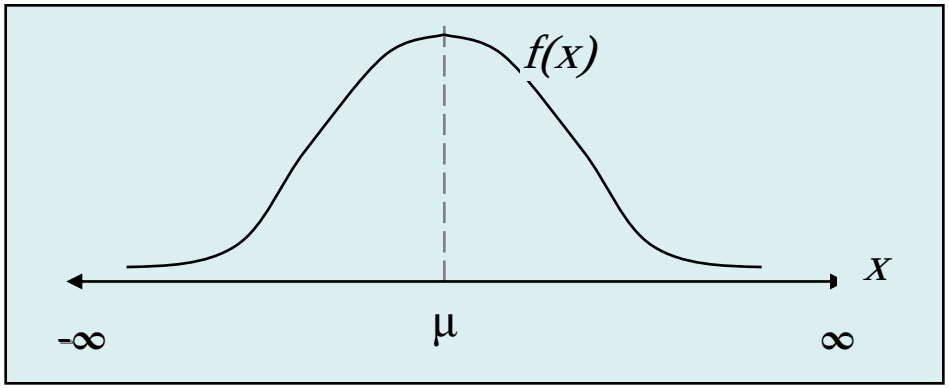
The normal distribution is one of the most important **continuous distribution** in statistics.

It has the following characteristics

1- X takes values from $-\infty$ to ∞.

2- The population mean is μ and the population variance is σ^2, and we write $X \sim N(\mu, \sigma^2)$.

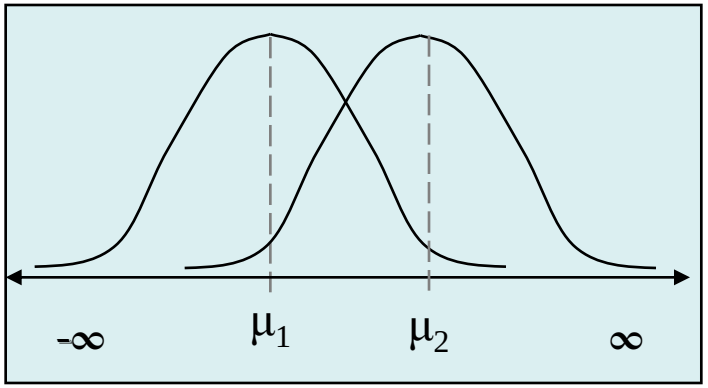
3- The graph of the density of a normal distribution has a bell shaped curve, that is symmetric about μ



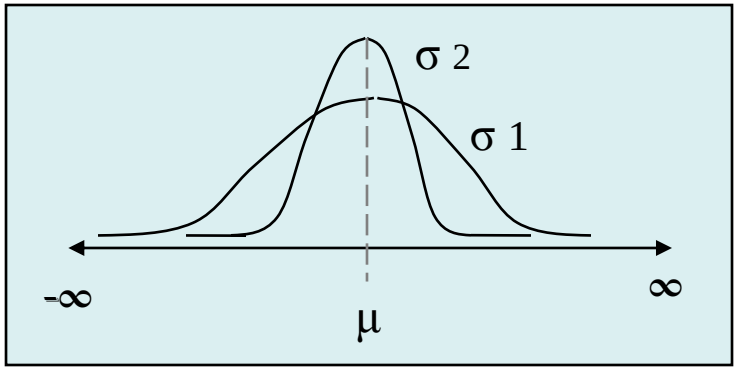
4- μ= mean=mode=median of the normal distribution.

5-The location of the distribution depends on μ (location parameter).

The shape of the distribution depends on σ (shape parameter).



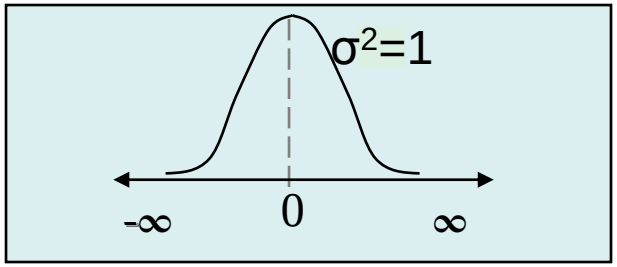
$\mu_1 < \mu_2$



$\sigma_1 > \sigma_2$

Standard normal distribution:

– The *standard normal distribution* is a normal distribution with mean $\mu=0$ and variance $\sigma^2=1$.



Result

– If $X \sim N(\mu, \sigma^2)$ then $Z=\frac{X-\mu}{\sigma} \sim N(\mu, \sigma^2)$.

Notes

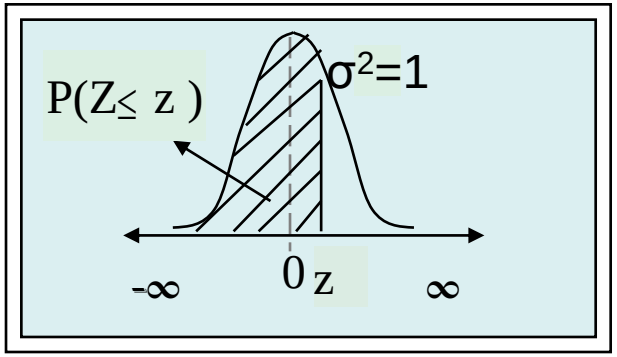
- The probability $A= P(Z \leq z)$ is the area to the left of z under the standard normal curve.

-There is a Table gives values of $P(Z \leq z)$ for different values of z.

Calculating probabilities from Normal (0,1)

- $P(Z \leq z)$ From the table

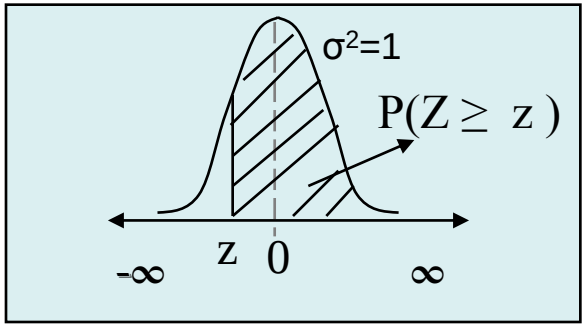
(the area under the curve to the left of z)



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

↑ From the table

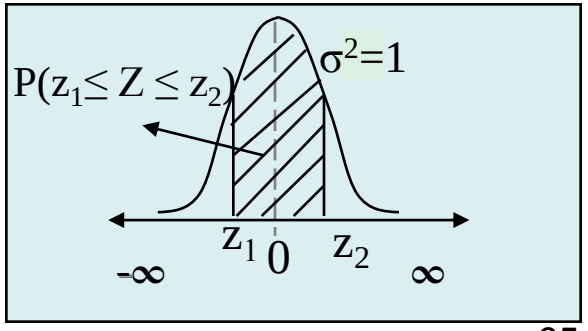
(the area under the curve to the right of z)



- $P(z_1 \leq Z \leq z_2) = P(Z \leq z_2) - P(Z \leq z_1)$

↑ ↑ From the table

(the area under the curve between z_1 and z_2)



Notes:

- $P(Z \leq 0) = P(Z \geq 0) = 0.5$ (why?)
- $P(Z = z) = 0$ for any z.
- $P(Z \leq z) = P(Z < z)$ and $P(Z \geq z) = P(Z > z)$
- If $z \leq -3.49$ then $P(Z \leq z) = 0$, and if $z \geq 3.49$ then $P(Z \leq z) = 1$.

Example 4.1 :

- $P(Z \leq 1.5) = 0.9332$
- $P(-1.33 \leq Z \leq 2.42) = P(Z \leq 2.42) - P(Z < 1.33) =$

  $= 0.9922 - 0.0918 = 0.9004$
- $P(Z \geq 0.98) = 1 - P(Z \leq 0.98) = 1 - 0.8365 = 0.1635$

Z	0.00	0.01	…
:	⇓		
1.5 ⇒	0.933		
:			

Example 4.2 :

Suppose that the hemoglobin level for healthy adult males are approximately normally distributed with mean 16 and variance of 0.81. Find the probability that a randomly chosen healthy adult male has hemoglobin level

a) Less than 14. b) Greater than 15. C) Between 13 and 15

Solution

Let X= the hemoglobin level for healthy adult male, then

$X \sim N(\mu=16, \sigma^2=0.81)$.

a) Since $\mu=16$, $\sigma^2=0.81$, we have $\sigma=\sqrt{0.81}=0.9$

$P(X<14)= P(Z<\frac{14-\mu}{\sigma})= P(Z<\frac{14-16}{0.9})= P(Z< -2.22)=$

b) $P(X>15)= P(Z>\frac{15-\mu}{\sigma})= P(Z>\frac{15-16}{0.9})= P(Z> -1.11)= 1- P(Z\leq -1.11)=$

c) $P(13<X<15)= P(\frac{13-\mu}{\sigma}<Z<\frac{15-\mu}{\sigma})= P(Z<\frac{15-16}{0.9})- P(Z<\frac{13-16}{0.9})$

$= P(Z\leq -1.11) - P(Z \leq -3.33)$

Result(1)

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from N (μ, σ^2), then

1) $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \sim$ N ($\mu, \sigma^2/n$)

2) $$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \sim N(0, 1).$$

Central Limit Theorem

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from any distribution with mean μ and variance σ^2, and if n is large ($n \geq 30$), then

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \approx N(0, 1).$$

(that is, Z has approximately standard normal distribution)

Result (2)

If σ^2 is unknown in the central limit theorem, then s (the sample standard deviation) can be used instead of σ, that is

$$Z = \frac{\bar{x} - \mu}{s / \sqrt{n}} \approx N(0, 1).$$

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

Chapter 5: Statistical Inference

5.1 Introduction: There are two main purposes in statistics

-Organizing and summarizing data (descriptive statistics).

-Answer research questions about population parameter (statistical inference).

There are two general areas of statistical inference:

- Hypothesis testing: answering questions about population parameters.
- Estimation: approximating the actual values of population parameters.

there are two kinds of estimation:

- Point estimation.
- Interval estimation (confidence interval).

Here we will consider two types of population parameters

Population mean: μ (for quantitative variable)

μ=The average (mean) value for some qualitative variable.

Examples:
-The mean life span for some bacteria
- The income mean of government employee in Saudi Arabia.
- The mean of weight of all new born babies in some country.

Population proportion π

$$\pi = \frac{\text{no. of element in the population with some charachtaristic}}{\text{Total no. of element in the population}}$$

Examples:
-The proportion of Saudi people who have some disease
- The proportion of smokers in Riyadh.
- The proportion of Children in Saudi Arabia.

5.2: Estimation of Population Mean: μ

1) Point Estimation:

- A point estimate is a single number used to estimate the corresponding population parameter.
- $\bar{x}$ is a point estimate of μ

That is , the sample mean is a point estimate of the population mean.

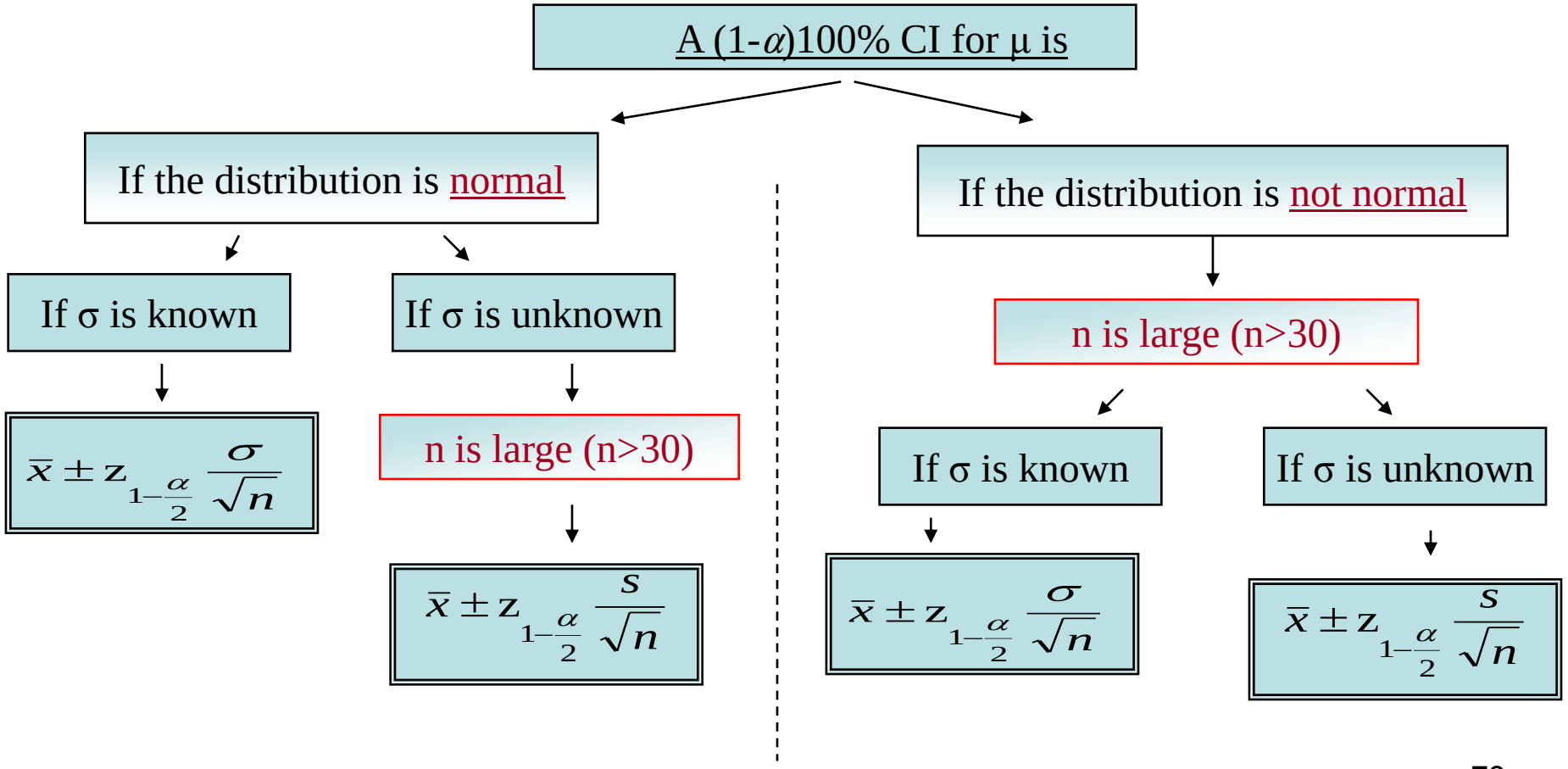
Interval Estimation (Confidence Interval:CI) of μ

Definition: (1-α)100% Confidence Interval: Is an interval of numbers (L,U), defined by lower L and upper U limits that contains the population parameter with probability (1-α).

1-α: the confidence coefficient.

L: Lower limit of the confidence interval.

U : upper limit of the confidence interval.



Note: The CI $\bar{x} \pm z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ means

(L , U)= $(\bar{x} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \ \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}})$

- Similarly for $\bar{x} \pm z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$, (L , U)= $(\bar{x} - z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}, \ \bar{x} + z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}.)$

- Interpretation of the CI: We are (1-α)100% confident that the (mean μ) of the (population) is between L and U.

Example:

Let Z~N(0, 1)

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

Here we have the probability (the area) and we want to find the exact value of z. hence we can use the table of standard normal but in the opposite direction.

a) α=0.05

$\alpha/2=0.025$

$1-\alpha/2=0.975$

From the standard normal table $Z_{0.975} = 1.96$

b) α=0.1

$\alpha/2=0.05$

$1-\alpha/2=0.95$

$Z_{0.95} = 1.645$

Z	…	0.06	...
:	:	⇑ ⇑	
1.9	⇐⇐	0.975	
:			

Example 5.2: On 123 patient of diabetic ketoacidosis patient in Saudi Arabia , the mean blood glucose level was 26.2 with a standard deviation of 3.3 mm0l/l. Find the 90% confidence interval for the mean blood glucose level of such diabetic ketoacidosis patient.

Solution:

Variable: blood glucose level (in mmol/l)

Population: Diabetic ketoacodosis patient in Saudi Arabia.

Parameter: μ (the average blood glucose level)

n=123, $\bar{x} = 26.2$ s=3.3

- σ^2 unknown , n=123>30 (large) ⇒ the 90% CI for μ is given by

$$\bar{x} \pm z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$

90% = (1- α)100% $\Rightarrow$ *1- α=0.9*

α=0.1 $\Rightarrow$ *α/2=0.05* $\Rightarrow$ *1-α/2=0.95*

$Z_{0.95} = 1.645$

The 90% CI for μ is $\bar{x} \pm z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$

Which is can be written as $(\bar{x} - z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}},\ \bar{x} + z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}})$

$$= (\ldots\ldots - (\ldots\ldots) \frac{\ldots\ldots}{\sqrt{\ldots\ldots}},\ \ldots\ldots + (\ldots\ldots) \frac{\ldots\ldots}{\sqrt{\ldots\ldots}})$$

$$= (\ldots\ldots,\ \ldots\ldots\ldots)$$

Interpretation: We are 90 % confident that the mean blood glucose level of diabetic ketoacodosis patient in Saudi Arabia is between ……. and …….

5.3: Estimation of Population Proportion π

- **Recall that, the population proportion**

$$\pi = \frac{\text{no. of element in the population with some charachtaristic}}{\text{Total no. of element in the population}} \leftarrow N$$

- **To estimate the population proportion we take a sample of size n from the population and find the sample proportion p**

$$p = \frac{\text{no. of element in the sample with some charachtaristic}}{\text{Total no. of element in the sample}} \leftarrow n$$

Result: when both $n\pi > 5$ and $n(1-\pi) > 5$ then

$$p \approx N(\pi, \pi(1-\pi)/n).$$

and hence

$$Z = \frac{p-\pi}{\sqrt{\pi(1-\pi)/n}} \approx N(0,1).$$

Estimation for π

1) Point Estimation:

A point estimator of π (population proportion) is p (sample proportion)

1) Interval Estimation: If np>5 and n(1-p)>5,

The (1-α)100% Confidence Interval for π is given by

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

Note:1) $p \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}$ can be written as

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

2) np= the number in the sample with the characteristic

n(1-p)= the number in the sample wich did not have the characteristic.

Example 5.2

In the study on the fear of dental care in Riyadh, 22% of 347 adults said they would hesitate to take a dental appointment due to fear. Find the point estimate and the 95% confidence interval for proportion of adults in Riyadh who hesitate to take dental appointments.

Solution:

Variable: whether or not the person would hesitate to take a dental appointment out of fear.

Population: adults in Riyadh.

Parameter: π, the proportion who would hesitate to take an appointment.

n= ….. , p= …..=……

np=(……)(…..)=…… >5 and n(1-p)=(…..)(……)=2…..>5

1- point estimation of π is p=….

2- 95% CI for π is $p \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}$

1-α=0.95 ⇒ α=0.05 ⇒ α/2=0.025 ⇒1- α/2=0.975

$$Z_{1-\alpha/2} = Z_{0.975} = 1.96$$

The 95 % CI for π is

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

$$= \left(\ldots\ldots - (\ldots\ldots)\sqrt{\frac{\ldots.}{\ldots\ldots}}, \ \ \ldots\ldots + (\ldots.)\sqrt{\frac{\ldots.}{\ldots.}} \right)$$

$$= (\ldots \ \ - (\ldots.)(\ldots \ldots\ldots), \ \ldots. + (\ldots..)(\ldots \ldots..))$$

$$= (\ldots\ldots\ldots \ \ , \ \ \ \ldots\ldots\ldots)$$

Interpretation: we are 95% confident that the true proportion of adult in Riyadh who hesitate to take a dental appointment is between ….. and ……