Elementary Biostatistics with Applications

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.

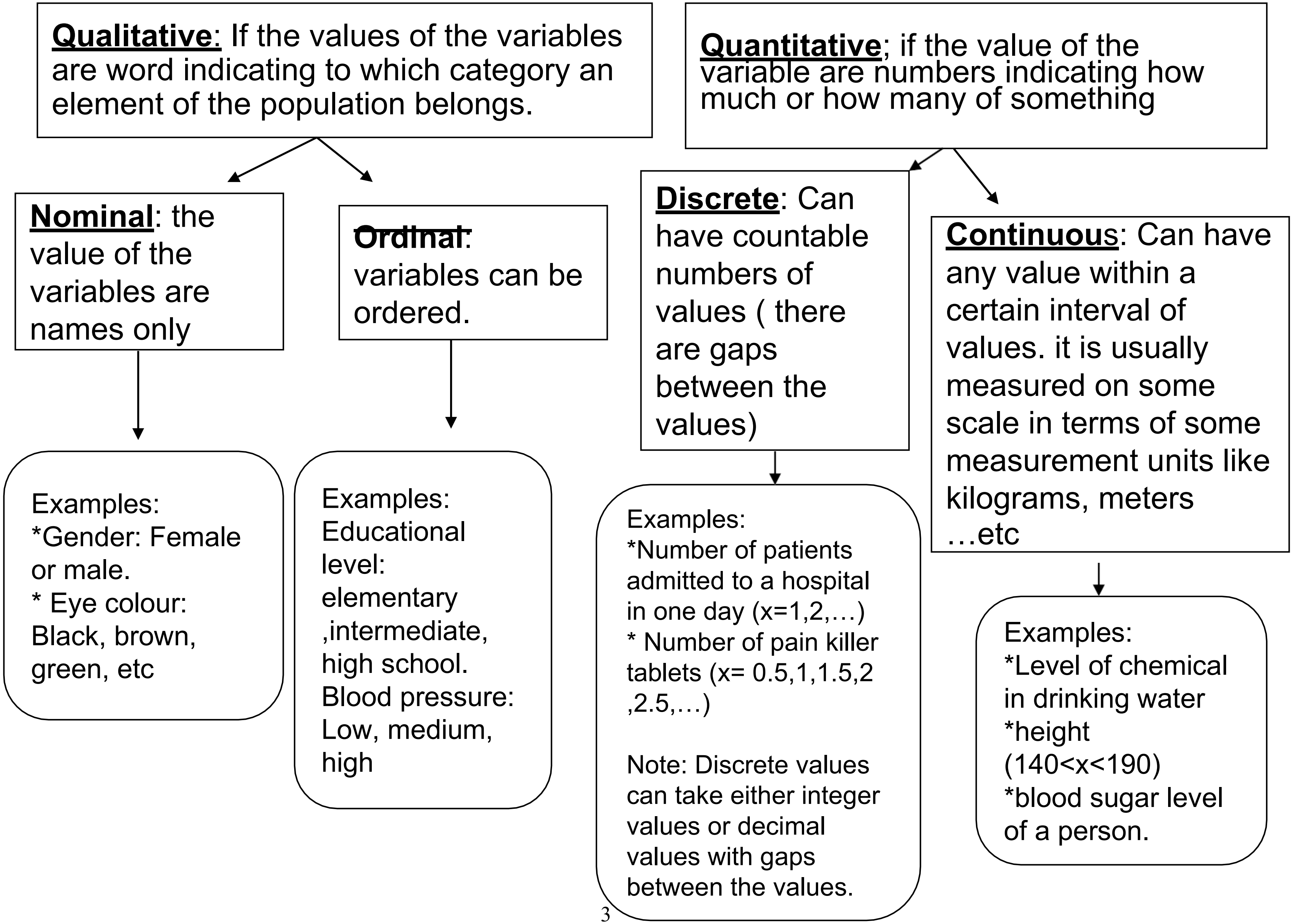
Biostatistics: Is a branch of statistics that interested in information obtained from biological and medical sciences.

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.

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

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

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”.

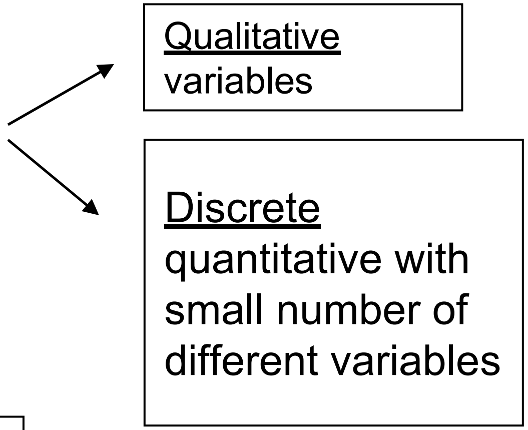
The sample size is 20

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

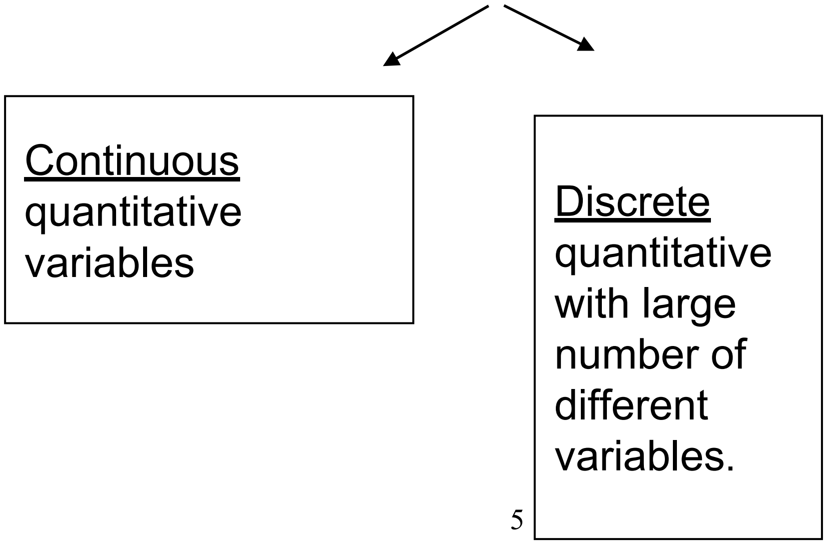
1) Simple (ungrouped) frequency distribution: for



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

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

Notes

1. In example 1.2.2 to group the data we use a set of intervals, called **class intervals**.
2. **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.
3. 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 …
4. To find ***the class intervals*** we use the following relationship

$$\begin{tikzcd}
L_1 \arrow[r] \arrow[d, "+w"'] & U_1 \arrow[d, "+w"] \arrow[dl] \\
L_2 \arrow[r] \arrow[d, "+w"'] & U_2 \arrow[d, "+w"] \\
L_3 & U_3
\end{tikzcd}$$

and so on

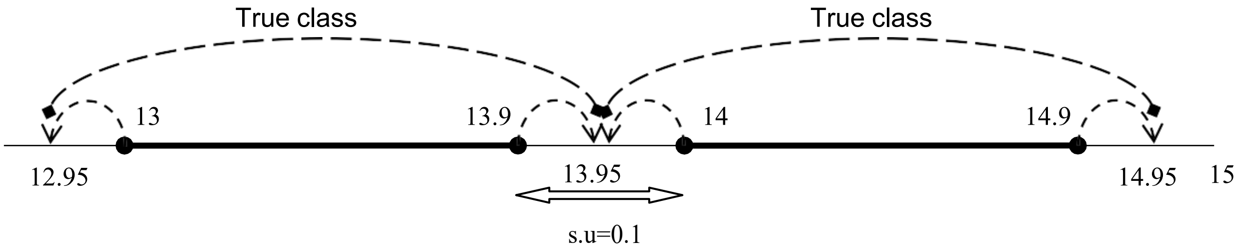
6. **Cumulative frequency**: is the number of values obtained in the class interval or before, which find by adding successfully the frequencies.
7. **Cumulative relative frequency**: is the proportion of values obtained in the class interval or before, which find by adding successfully the relative 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:

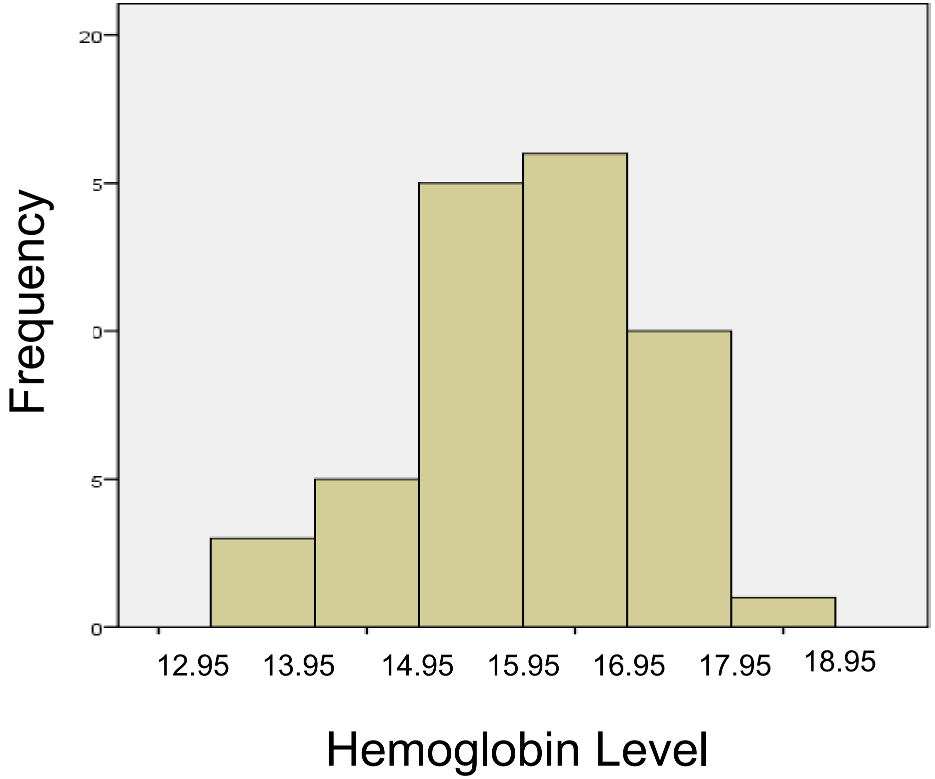
- Each upper limit of the true class interval ends with the same lower limit of the previous true class intervals
- The lower and upper limit of the true class interval must always end in 5, and they must always have one more decimal place than class limit.
- ***The mid point*** =(upper limit + lower 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

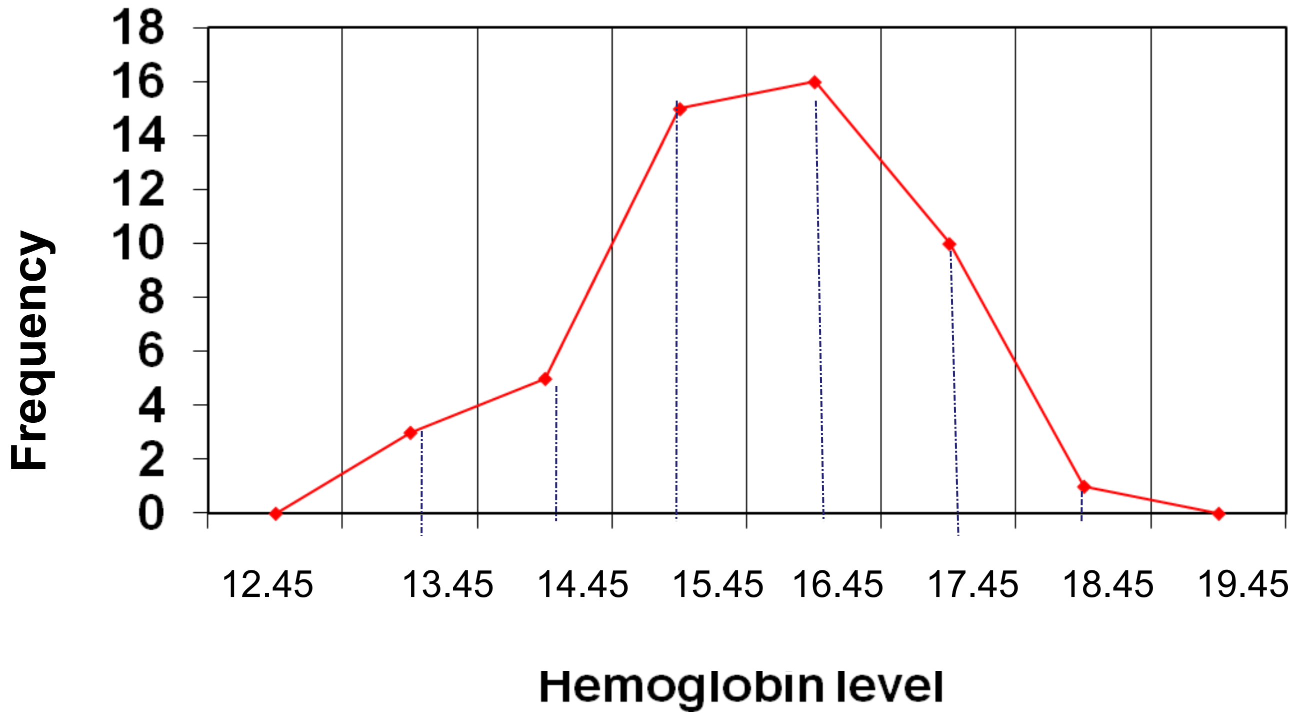
Histogram
Polygon
curves

For frequency or relative frequency distributions



Histogram

Polygon


Frequency
0
2
4
6
8
10
12
14
16
18
12.45
13.45
14.45
15.45
16.45
17.45
18.45
19.45
Hemoglobin level

Exammple 1.4: 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) From the table, find

	Class interval (glucose level)	Frequency	Relative frequency	Cumulative frequency
1.	70-79	3	0.04	3
2.	80-89	12	0.16	15
3.	90-99	24	0.32	39
4.	100-109	30	0.4	69
5.	110-119	6	0.08	75
	Total	75	1	

1) w= 2) n=

3) The number of healthy males with glucose level 80-89 mg/100 ml

4) The percentage of healthy males with glucose level less than 100-109 mg/100 ml

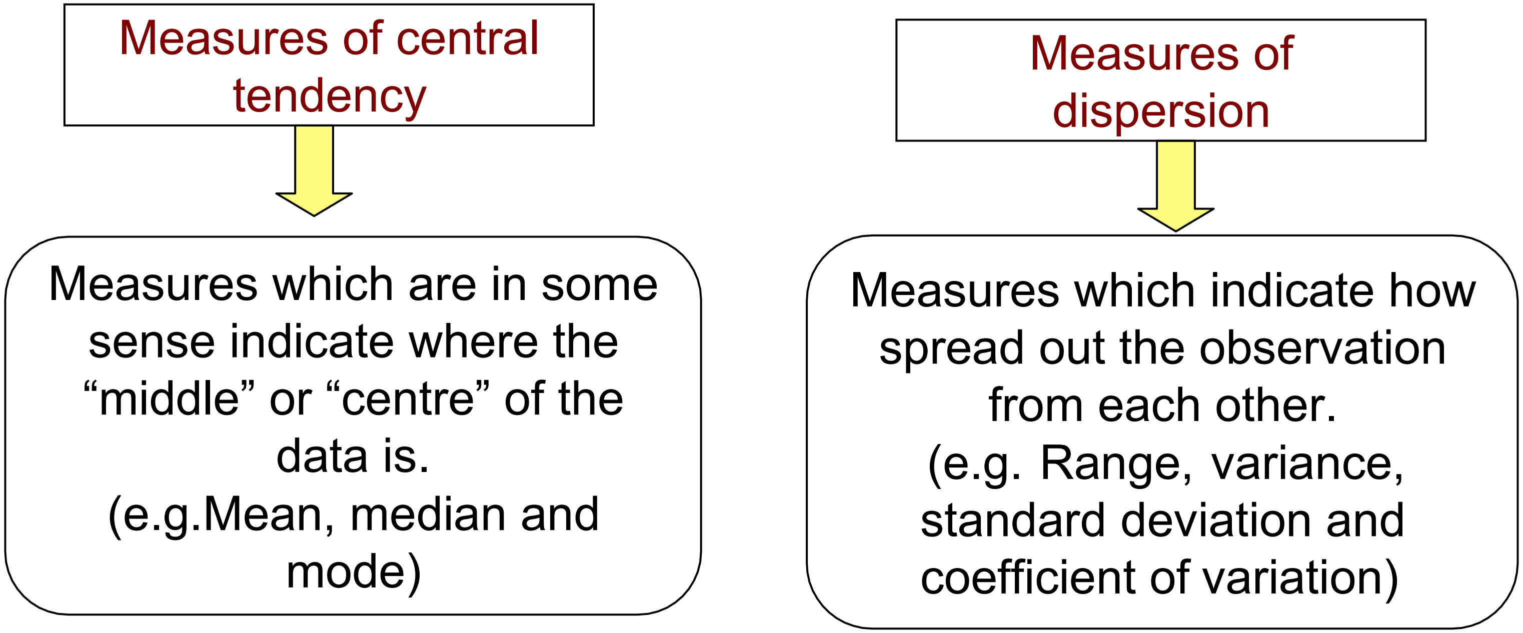
5) The number of healthy males with glucose level less than 99 mg/100 ml

6) The number of healthy males with glucose level greater than 100

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)

2.2: Measures of central tendency

Firstly For the discrete variables

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 X1,X2, …, XN be the population values of the variable (usually unknown), then the population mean is

$$\boxed{\mu = \frac{X_1 + X_2 + \ldots + X_N}{N} = \frac{\sum X_i}{N}}$$

Sample mean :let *x1, x2, …, xn* be the sample values of the variable, then the sample mean is

$$\boxed{\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum x_i}{n}}$$

- 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 2.2.1: 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

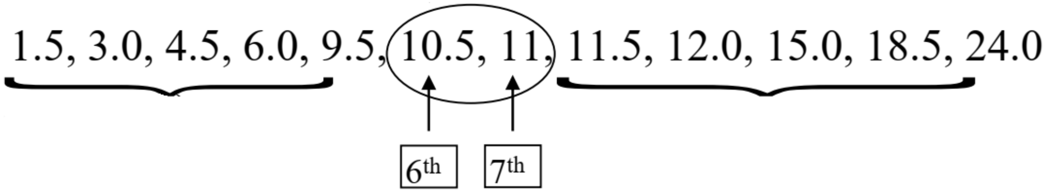
n= 11, odd, the order of the median is (n+1)/2=(11+1)/2=6th

$$\underbrace{28,\ 29,\ 29,\ 30,\ 30,}\ \underset{\boxed{\text{6th}}}{\underset{\uparrow}{\textcircled{31}}},\ \underbrace{32,\ 40,\ 40,\ 40,}\ 42$$

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



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

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

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

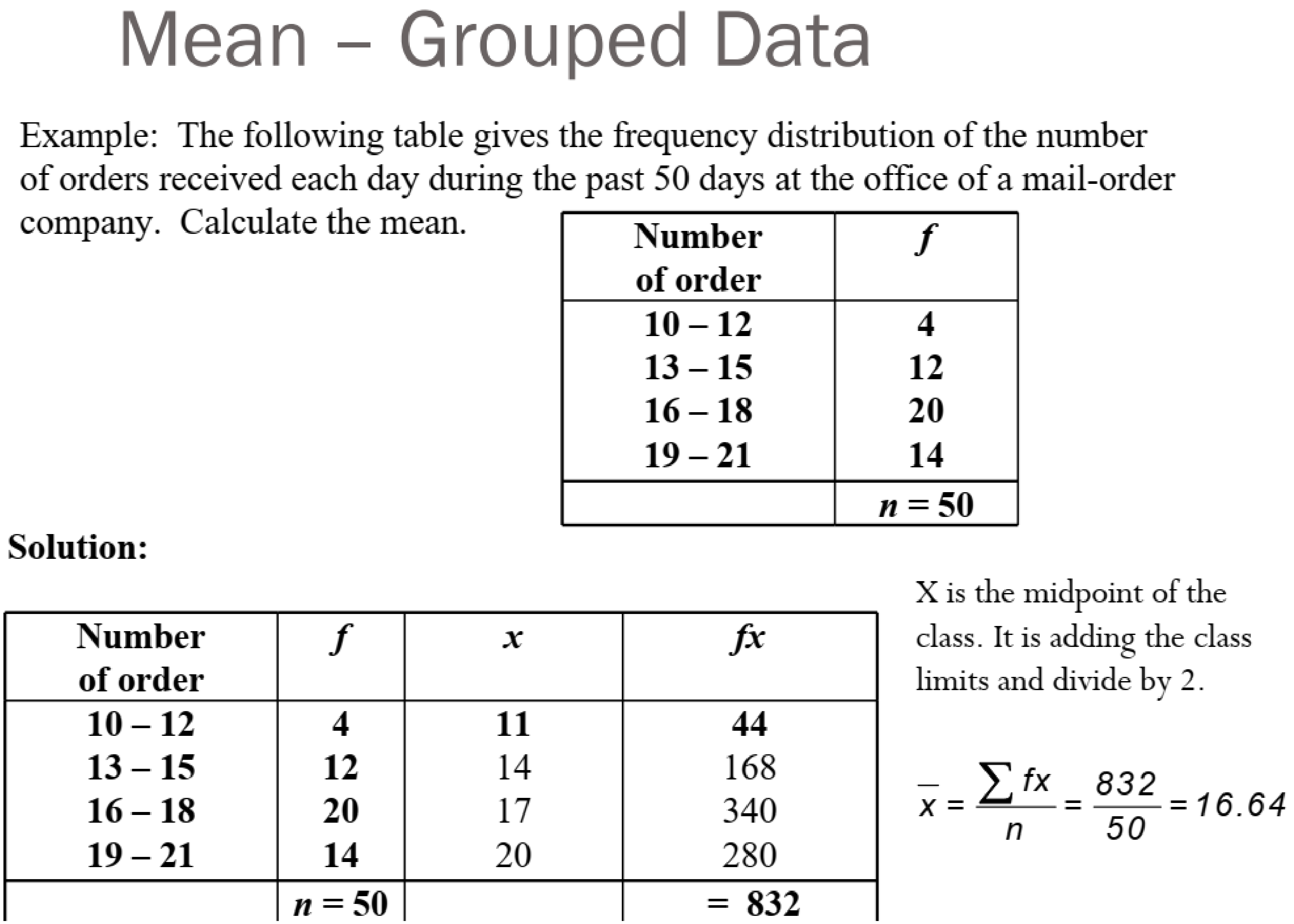
- No mode: example: Data(1): 21, 15, 22 ,19, 14, 18
  Data(2): 3, 3, 5,5, 4, 4, 6, 6
- One mode, example :Data (1): 32, 15, 23, 17 , 22, 23, 19, 20, 22, 22 .
  The mode=22 (unit)
  Data(2): 13.5, 12, 13.5, 15, 15, 14.6, 17, 12, 15
  The mode=15 (unit)
- More than one mode: example 18, 20, 19, 19, 21, 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.
- measures that not affected so much by extreme values are the median and the mode.

Secondly For the continuous variables



Mean – Grouped Data

Example: The following table gives the frequency distribution of the number of orders received each day during the past 50 days at the office of a mail-order company. Calculate the mean.

Number of order	***f***
10 – 12	**4**
13 – 15	**12**
16 – 18	**20**
19 – 21	**14**
	***n* = 50**

Solution:

Number of order	***f***	***x***	***fx***
10 – 12	**4**	**11**	**44**
13 – 15	**12**	14	168
16 – 18	**20**	17	340
19 – 21	**14**	20	280
	***n* = 50**		**= 832**

X is the midpoint of the class. It is adding the class limits and divide by 2.

$$\overline{x} = \frac{\sum fx}{n} = \frac{832}{50} = 16.64$$

Median and Interquartile Range – Grouped Data

Step 1: Construct the cumulative frequency distribution.

Step 2: Decide the class that contain the median.

Class Median is the first class with the value of cumulative frequency equal at least n/2.

Step 3: Find the median by using the following formula:

$$\text{Median} = L_m + \left(\frac{\frac{n}{2} - F}{f_m} \right) i$$

Where:

n = the **total frequency**

F = the **cumulative frequency *before*** class median

f_m = the **frequency** of the class median

i = the class width

L_m = the **lower boundary** of the class median

Example: Based on the grouped data below, find the median:

Time to travel to work	Frequency
1 – 10	8
11 – 20	14
21 – 30	12
31 – 40	9
41 – 50	7

Solution:

1st Step: Construct the cumulative frequency distribution

Time to travel to work	Frequency	Cumulative Frequency
1 – 10	8	8
11 – 20	14	22
21 – 30	12	34
31 – 40	9	43
41 – 50	7	50

$\frac{n}{2} = \frac{50}{2} = 25$ ⟶ class median is the 3rd class

So, $F = 22$, $f_m = 12$, $L_m = 20.5$ and $i = 10$

Therefore,

$$\text{Median} = L_m + \left(\frac{\frac{n}{2} - F}{f_m} \right) i$$

$$= 21.5 + \left(\frac{25 - 22}{12} \right) 10$$

$$= 24$$

Thus, 25 persons take less than 24 minutes to travel to work and another 25 persons take more than 24 minutes to travel to work.

Mode – Grouped Data

Mode

- Mode is the value that has the highest frequency in a data set.
- For grouped data, class mode (or, modal class) is the class with the highest frequency.
- To find mode for grouped data, use the following formula:

$$\text{Mode} = L_{mo} + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right) i$$

Where:

i is the class width

Δ_1 is the difference between the frequency of class mode and the frequency of the class **after** the class mode

Δ_2 is the difference between the frequency of class mode and the frequency of the class **before** the class mode

L_{mo} is the **lower boundary** of class mode

Example: Based on the grouped data below, find the mode

Time to travel to work	Frequency
1 – 10	8
11 – 20	14
21 – 30	12
31 – 40	9
41 – 50	7

Solution:

Based on the table,

$L_{mo} = 10.5$, $\Delta_1 = (14 - 8) = 6$, $\Delta_2 = (14 - 12) = 2$ and $i = 10$

$$\text{Mode} = 10.5 + \left(\frac{6}{6+2}\right)10 = 17.5$$

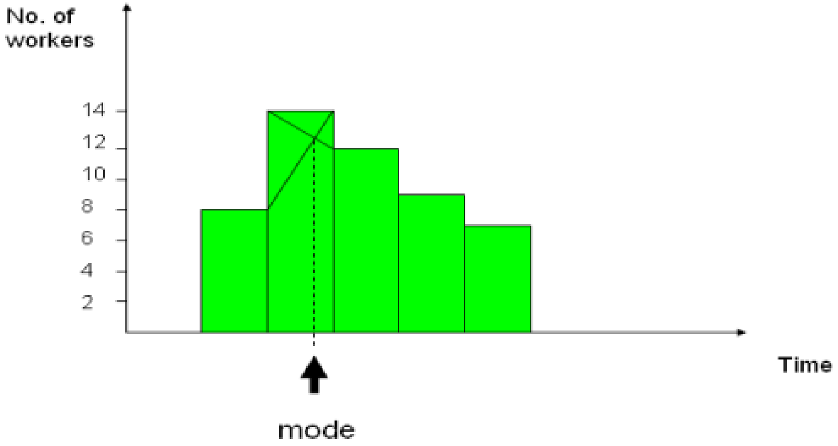
Mode can also be obtained from a histogram.

Step 1: Identify the modal class and the bar representing it

Step 2: Draw two cross lines as shown in the diagram.

Step 3: Drop a perpendicular from the intersection of the two lines until it touch the horizontal axis.

Step 4: Read the mode from the horizontal axis

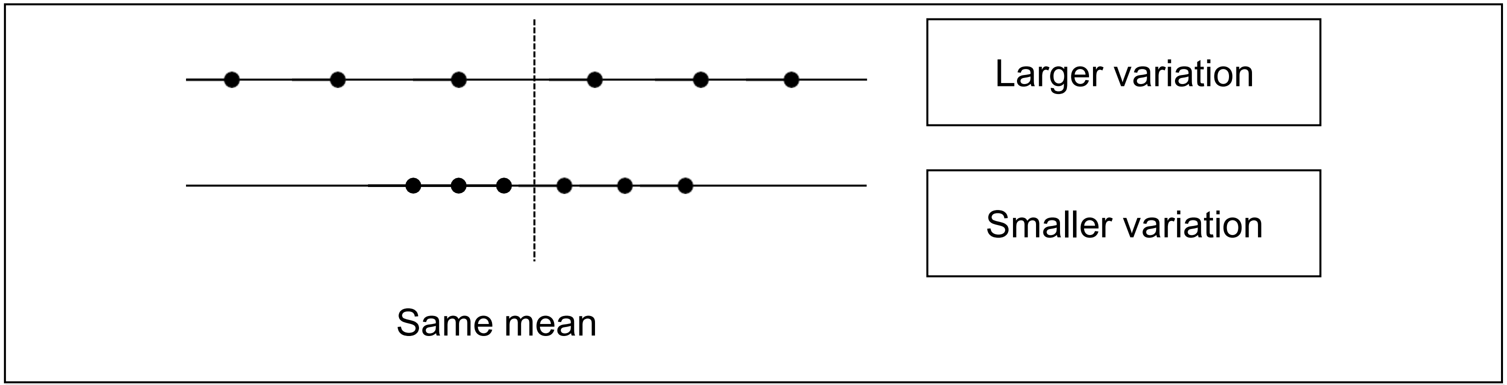


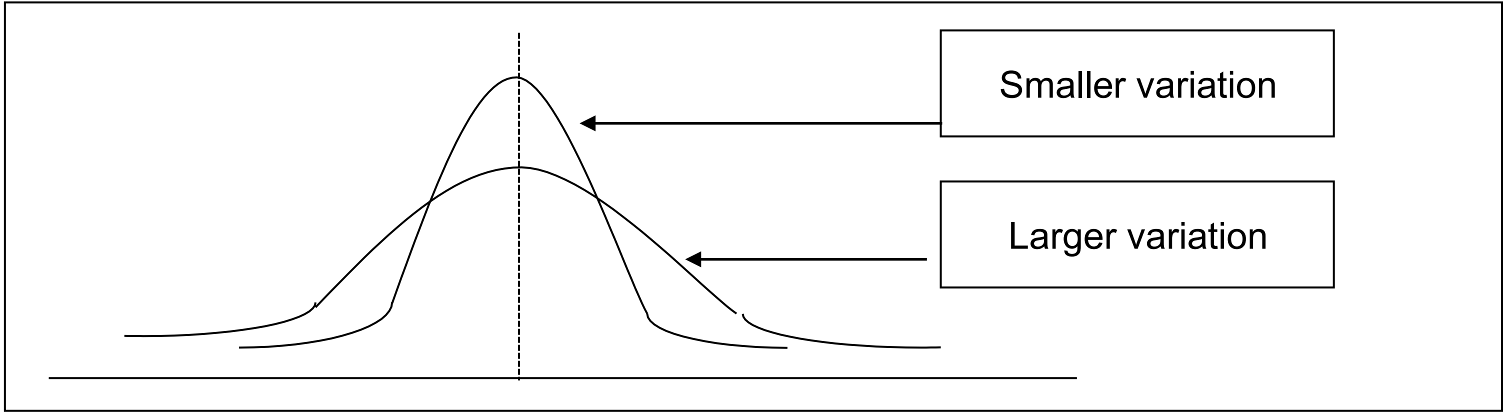
2.3: Measure of dispersion

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.
- R=3.63-1.05=2.58.

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.

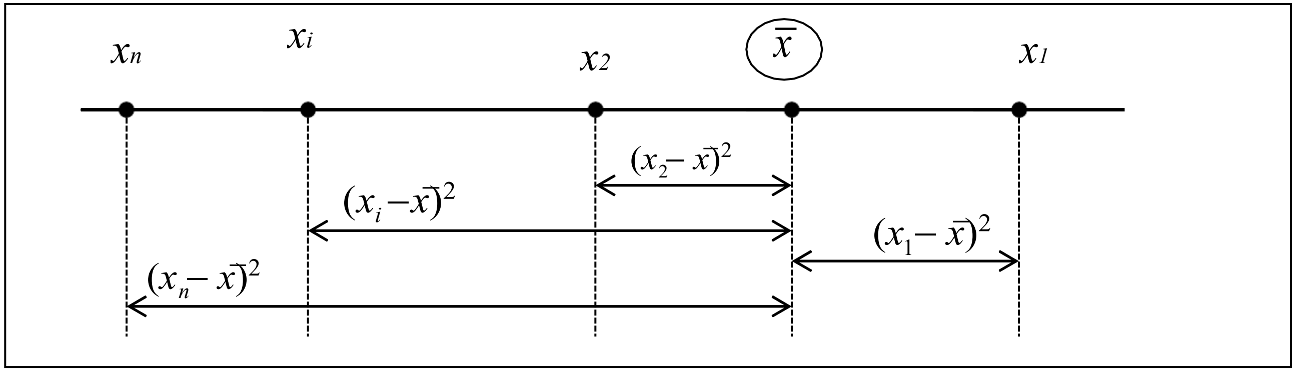
Variance: Is a measure which uses the mean as point of reference.

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

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

Population standard deviation

↓

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

Sample standard deviation

↓

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

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

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

- 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	$\mathrm{CV}_1 = \frac{s_1}{\bar{x}_1} \times 100(\%)$
Set2	$\bar{x}_2$	s_2	$\mathrm{CV}_2 = \frac{s_2}{\bar{x}_2} \times 100(\%)$

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).

<u>Example 2.5</u>

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

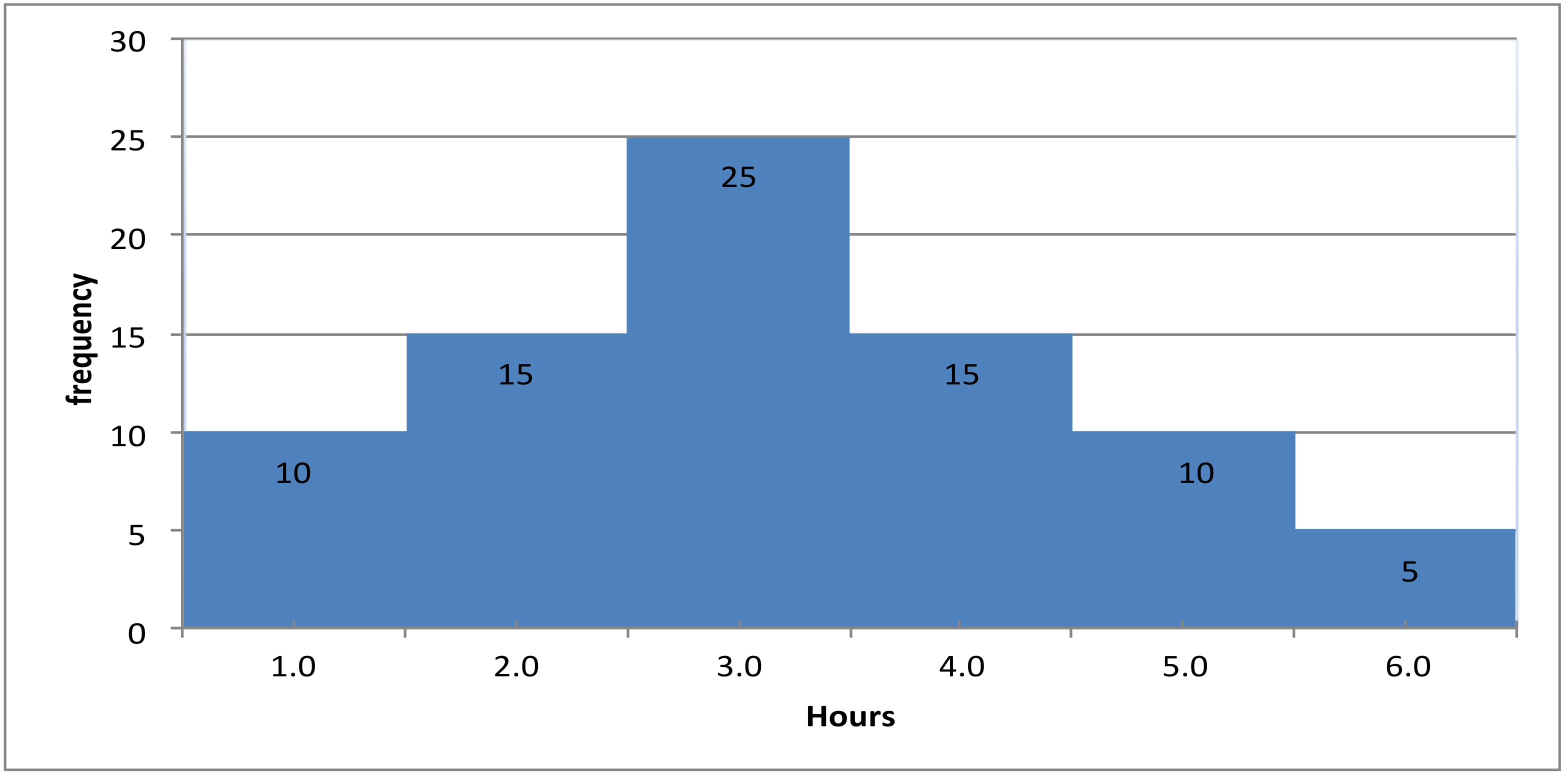
set1: on males aged 29: $\bar{x}_1$=66kg s_1=4.5kg $\Longrightarrow$ CV_1=(4.5/66)×100%=6.8%

set2: on males aged 10: $\bar{x}_2$=36kg s_1=4.5kg $\Longrightarrow$ CV_2=(4.5/36)×100%=12.5%

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).

Example 2.4.1

For a sample of patients, we obtain the following graph for approximated hours spent without pain after a certain surgery



1) The **type of the graph** is:
a) Bar chart (b) polygon (c) histogram (d) line (e) curve

2) The **number of patients** stayed the **longest time** without pain is:
a) 10 (b) 15 (c) 6 (d) 5 (e) 80

3)The **percent of patients** spent 3.5 hours or more without pain is:
a)37.5% (b) 68.75% (c) 18.75% (d) 50% (e) 25%

4)The **lowest number of hours** spent without pain is:
a)10 (b) 1 (c) 0.5 (d) 5 (e) 25 (f) 6.5

5)What the approximate value of the **sample mean**
a)2.55 (b) 255 (c) 3 (d) 3.1875 (e) 40 (f) we can't find it

6)The **sample mode** equals
a)80 (b) 3 (c) 15 (d) 2,4 (e) 6 (f) we can't find it

The SPSS computer results of the age of patients in one of the Riyadh hospitals are given below

Find :

a) Variable name

a) The type of the variable

b) The mode

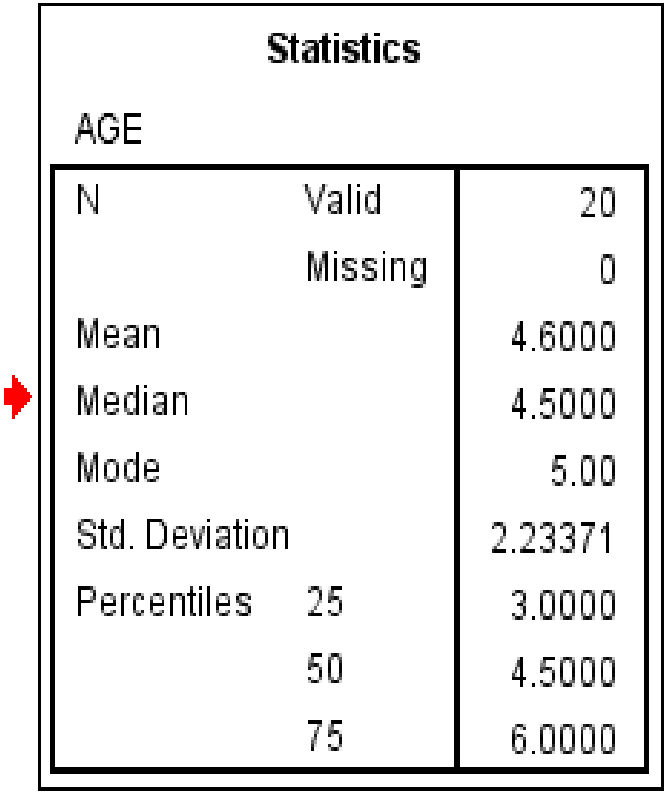
c) The mean age of the patients

d) The median age of the patients

e) The variance

f) Sample size

g) The coefficient of variation



Statistics

AGE

N	Valid	20
	Missing	0
Mean		4.6000
Median		4.5000
Mode		5.00
Std. Deviation		2.23371
Percentiles	25	3.0000
	50	4.5000
	75	6.0000

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

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

Notes

1. For any event A , $0 \le P(A) \le 1$ (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
$n(\Omega)=6$, $n(E_1)=3$, $n(E_2)=4$, $n(E_2)=3$

$P(E_1)=3/6=0.5$

$P(E_2)=4/6=0.667$

$P(E_3)=3/6=0.5$

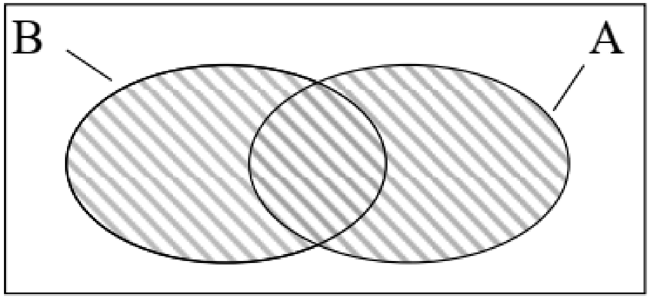
$P(E_4)=0$

Repaper that

- 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=$\{\}$= .

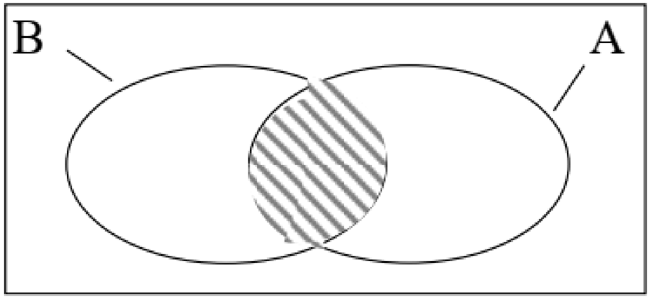
Relationships between events

- **Union** : $A \cup B$, consists of all those outcomes in A or in B or in both A and B



$A \cup B$

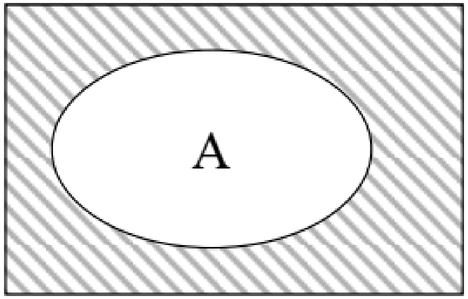
- **Intersection** : $A \cap B$, consists of all those outcomes in both A and B



$A \cap B$

- **Complement** : A^c (or A`)

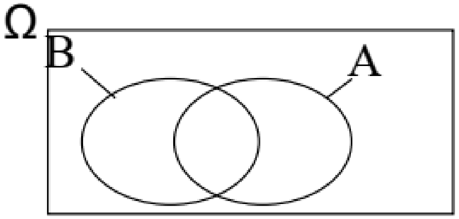
Consists of all outcomes that are in Ω but not in A



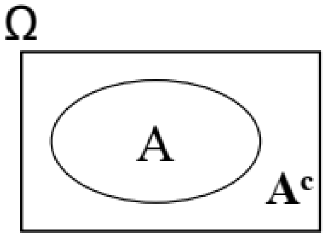
A^c

Notes:

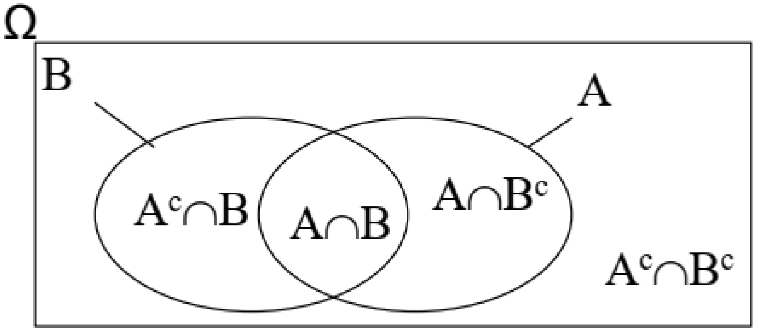
1- $n(A \cup B) = n(A) + n(B) - n(A \cap B)$
and hence
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$



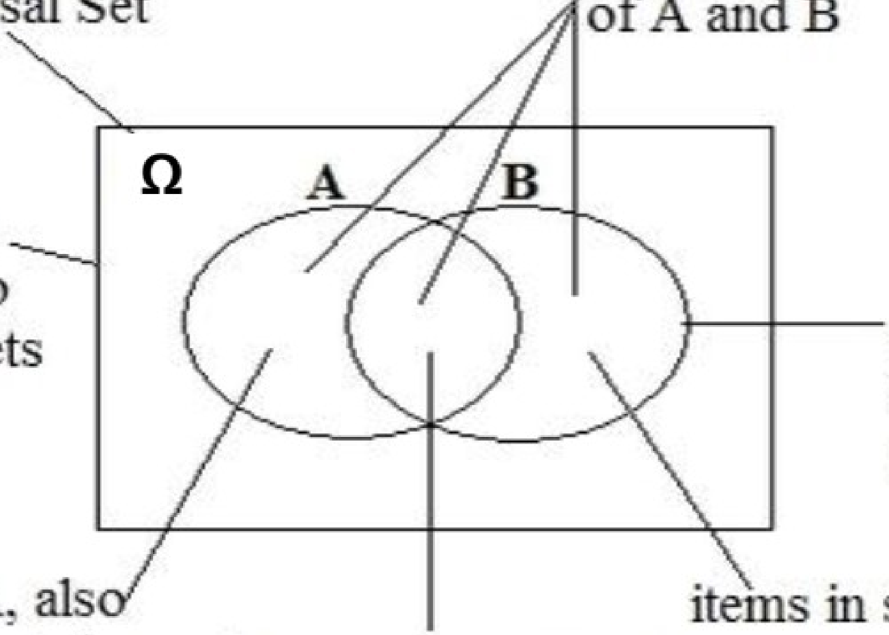
2. $n(A^c) = n(\Omega) - n(A)$
So that
$P(A^c) = 1 - P(A)$



Sets (events) can be represented by
Venn Diagram



A Venn Diagram:



The Universal Set

Union- elements of A and B

A rectangular box is used to enclose the sets of data

circles or ovals represent different sets

items in set A, also called "Elements" or "Memebers of the set"

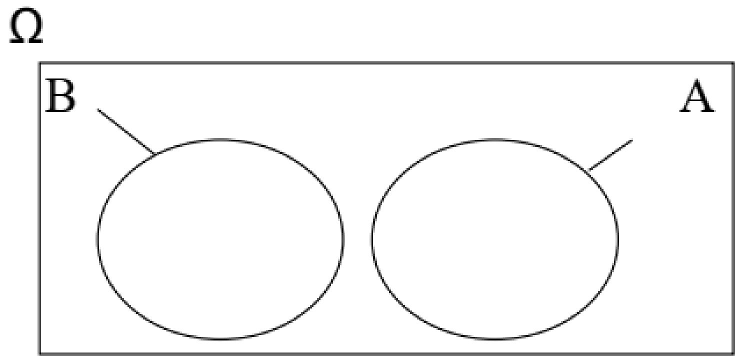
intersection- common elements of A and B (A∩B)

items in set B

Disjoint events

Two events A and B are said to be disjoint (mutually exclusive) if

$A \cap B = \phi$.



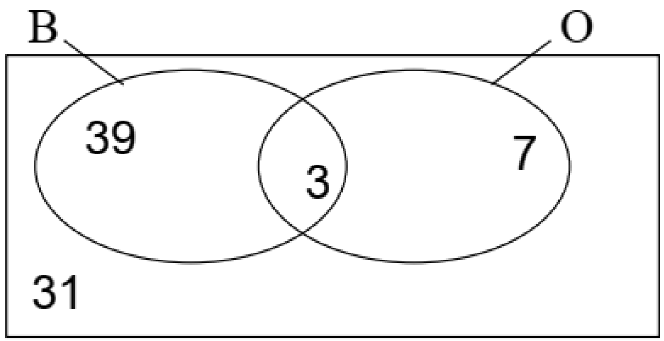
- In the case of disjoint events

$P(A \cap B)=0$

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

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

$P(B)=(3+39)/80=0.525$

- It is a boy and overweight

$P(B \cap O)= 3/80=0.0357$

- It is a boy or it is overweight

$P(B \cup O)= (39+3+7)/80=0.6125$

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

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

Notes:

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

Example

Referring to example 3.3 what is the probability that

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

$P(B \mid O)= P(B \cap O)/ P(O)= (3/80) / (10/ 80)=3/10=0.3$

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

$P(O^c \mid B^c)= P(B^c \cap O^c) / P(B^c) = (31/80) / [(7+31)/80]= 31/38= 0.716$

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

1-$P(A \cap B)=P(A)\,P(B)$

2-$P(A \mid B)= P(A)$ (Why?)

3- $P(B \mid A)= P(B)$ (Why?)

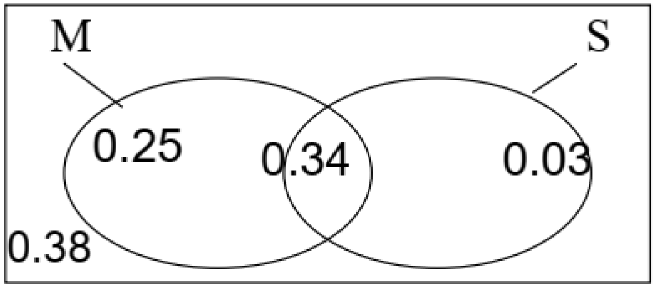
Example 3.4

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 ?

$P(M \cap S)=0.34$

2- Is a women?

$P(M^c)=$ 0.38+0.03= 0.41 (or $P(M^c)=1-P(M)=1-(0.25+0.34)=0.41$)

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

$P(M^c \cap S^c)=$ 0.38

4- Does not suffering from swollen leg?

$P(S^c)=$ 0.25+0.38= 0.63

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

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 the population mean (or the expected value of the random variable X) is

$$\boxed{\mu = \sum x \, P(X = x)}$$

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

$$\mu = \sum x \, P(X = x) = 1(0.1)+2(0.44)+3(0.3)+4(0.16)=2.52$$

-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

$P(X\le1)=0.1$

$P(X\le2)=P(X=1)+P(X=2)=0.1+0.44=0.54$

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

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

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- $\underline{P(X \leq a)}$ <u>is a cumulative distribution probability</u>

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

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 $\boxed{\mu=\Sigma x\, P(X=x)=n\,\pi}$

The variance $\boxed{\sigma^2=n\,\pi\,(1-\,\pi)}$

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***

$$\boxed{n=6 \,,\; \pi=0.15 \,,\;\; 1-\pi=0.85}$$

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

a - the probability distribution function is

$$\boxed{\begin{array}{l} P(X=x)=\binom{6}{x}0.15^{x}(0.85)^{6-x} \\ x=0,1,\ldots,6 \end{array}}$$

----

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

$$P(X=4)=\binom{6}{4}0.15^{4}(0.85)^{2}=(15)(0.15)^4(0.85)^2=0.00549$$

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

$$P(X=6)=\binom{6}{6}0.15^{6}(0.85)^{0}=0.15^{6}=0.00001$$

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

P(X=0)= $\binom{6}{0}0.15^0(0.85)^6 \;=0.85^6=0.37715$

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)]

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

$$=1-[\;0.37715\;+0.39933+\;0.17618\;]=1-0.95266\;=0.04734$$

F- the expected number of high blood pressure is $\mu=n\pi\;=6(0.15)=0.9$

and the variance is $\sigma^2=n\pi(1-\pi)\;=6(0.15)(0.85)=0.765$

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:

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

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 (الاصابات الخطيرة) in a particular factory in **a year**.
- The number of times a three year old child has an ear infection (عدوى الأذن) in **a year.**

- **Example 4.3:**

Suppose we are interested in the number of snake bite (لدغة الأفعى) 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. And the mean is 6*

Then X~ Poisson (6)

First note the following

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

$$\boxed{X\sim \text{Poisson } (6)}$$

- 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* $=(1/2)\lambda=3$

$$\boxed{Y\sim \text{Poisson } (3)}$$

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

$$\boxed{V\sim \text{Poisson } (12)}$$

1- The probability that the number of snake bites will be 7 in <u>*a year*</u> → λ=6

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

$$P(X=7)=\frac{e^{-6}6^{7}}{7!}=0.138$$

2- The probability that the number of cases will be less than 2 <u>in 6 months</u> → λ*=3

$$P(Y=y)=\frac{e^{-3}3^{y}}{y!}, y=0,1,2\dots$$

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

$$=\frac{e^{-3}3^{0}}{0!}+\frac{e^{-3}3^{1}}{1!} =0.0498+0.1494 \quad =0.1992$$

Remember
If X~ Poisson (λ)

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

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

3- The probability that the number of cases will be 13 in <u>2 years</u> → λ**=12

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

$$P(V=13)=\frac{e^{-12}12^{13}}{13!}=0.1056$$

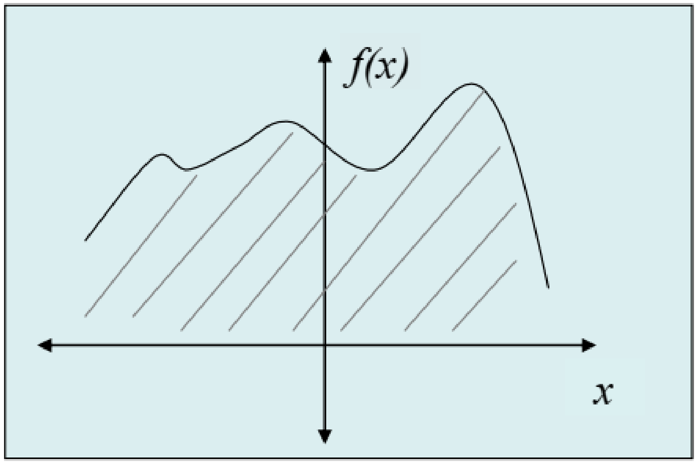
4- the expected number of snake bites in <u>a year</u>: $\mu=\lambda=6$ → λ=6

the variance of snake bites in <u>a year</u>: $\sigma^{2}=\lambda=6$

4.4 Probability Distribution of Continuous Random Variable

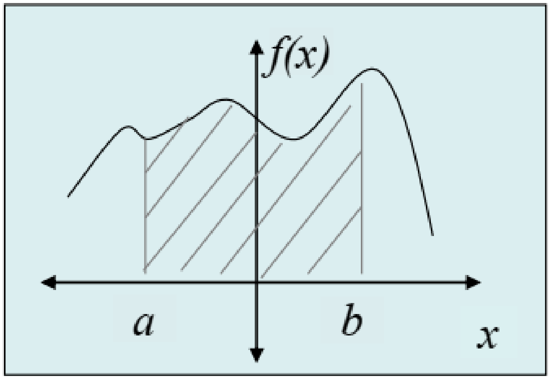
If X is a continuous random variable, then there exist a function *f(X)* called ***probability density function*** that has the following properties:

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

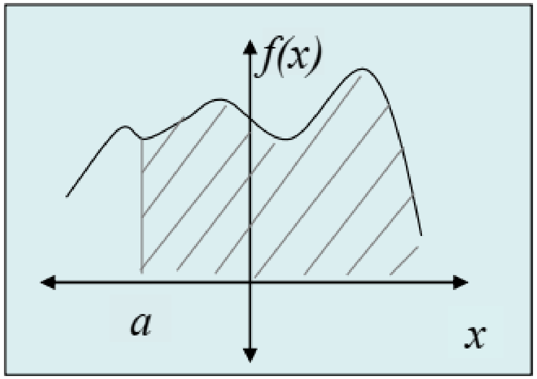


$$\text{area}= \int_{-\infty}^{\infty} f(x)dx = 1$$

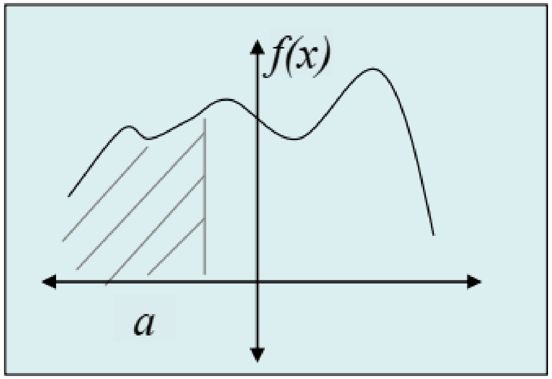
2- Probability of interval events are given by areas under the probability curve



P(a≤X≤b)= $\int_{a}^{b} f(x)dx$



P(X ≥ a)= $\int_{a}^{\infty} f(x)dx$



P(X≤ a)= $\int_{-\infty}^{a} f(x)dx$

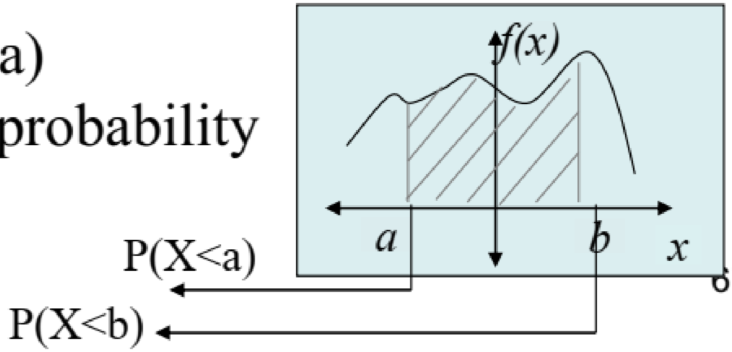
3- P(X=a)=0 (why?)

4-P(X≥a)=P(X>a) and P(X≤ a)=P(X<a)

7- P(X≤ a)= P(X<a) is the cumulative probability

5- P(X≥a)= 1- P(X≤a)

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



P(X<a)

P(X<b)

4.5 The Normal Distribution:

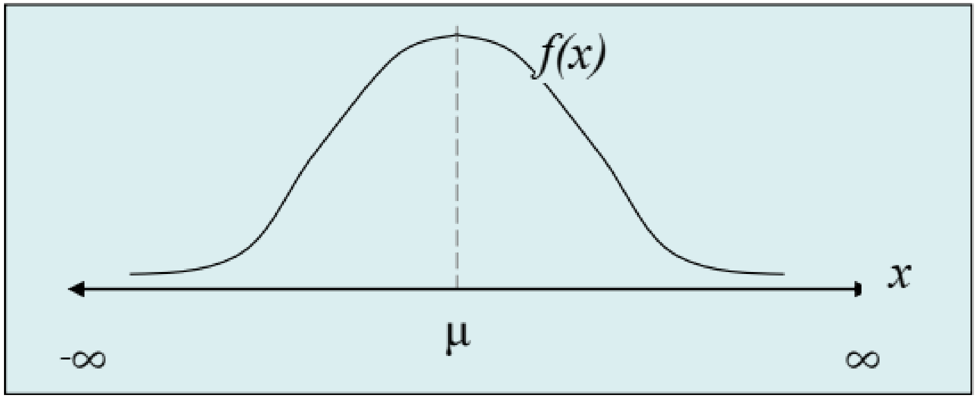
The normal distribution is one of the most important **continuous distribution** in statistics.

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

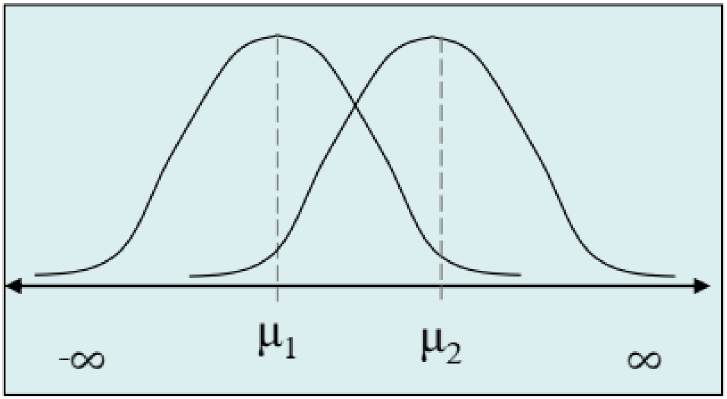
3- The graph of the density of a normal distribution has a bell shaped curve, that is symmetric about μ



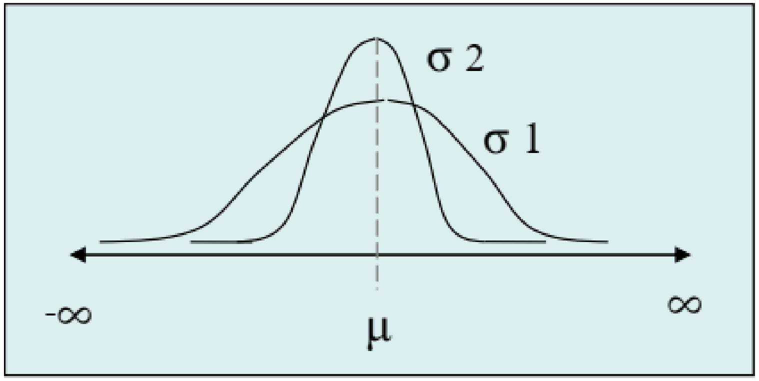
4- μ= mean=mode=median of the normal distribution.

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

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



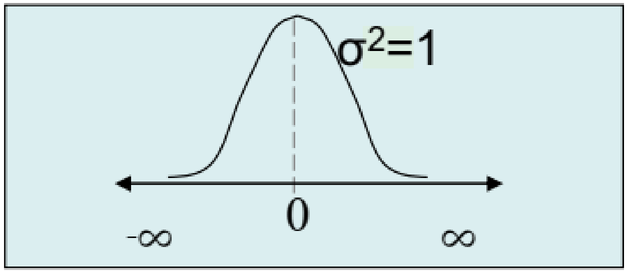
$\mu_1 < \mu_2$



$\sigma_1 > \sigma_2$

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 $\boxed{Z=\frac{X-\mu}{\sigma} \sim N(0, 1).}$

Notes

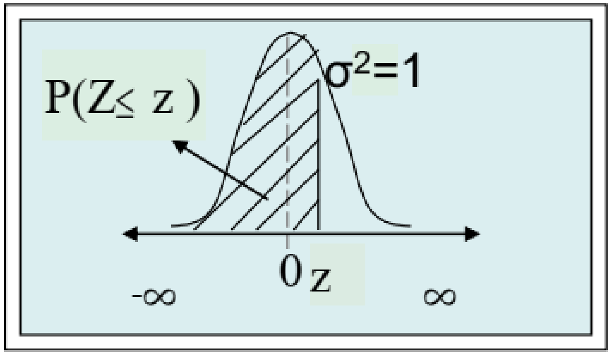
- The probability $A= P(Z\leq z)$ is the area to the left of z under the standard normal curve.

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

Calculating probabilities from Normal (0,1)

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

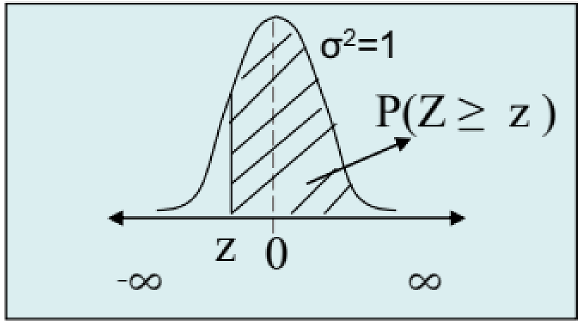
(the area under the curve to the left of z)



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

$\uparrow$ From the table

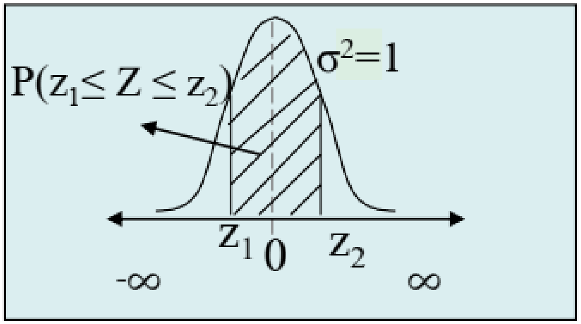
(the area under the curve to the right of z)



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

$\uparrow$ From the $\uparrow$ 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 $\sigma=\sqrt{0.81}=0.9$

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

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

$= 1- 0.1335=0.8665$.

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

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

$= 0.1335- 0= 0.1335$.

d) P(X=13)=0

Result(1)

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

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

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

Central Limit Theorem

Let X_1, X_2, …,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

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

(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

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

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).

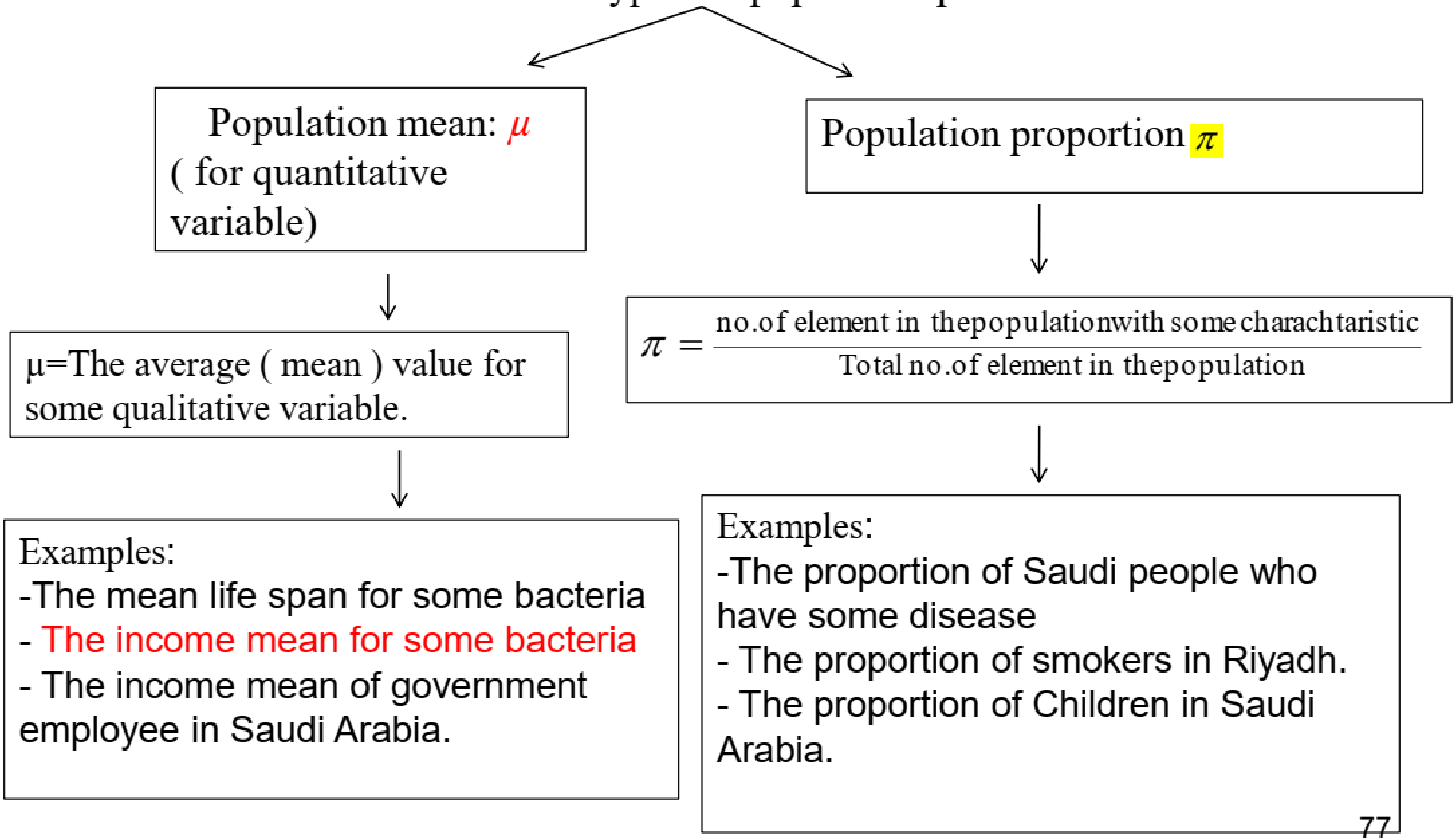
There are two general areas of statistical inference:

- **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 for some bacteria
- The income mean of government employee in Saudi Arabia.

Population proportion π

$$\pi = \frac{\text{no.of element in thepopulationwith some charachtaristic}}{\text{Total no.of element in thepopulation}}$$

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**.
- $\boxed{\bar{x} \text{ is a point estimate of } \mu}$

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

----

2) Interval Estimation (Confidence Interval:C.I) of μ

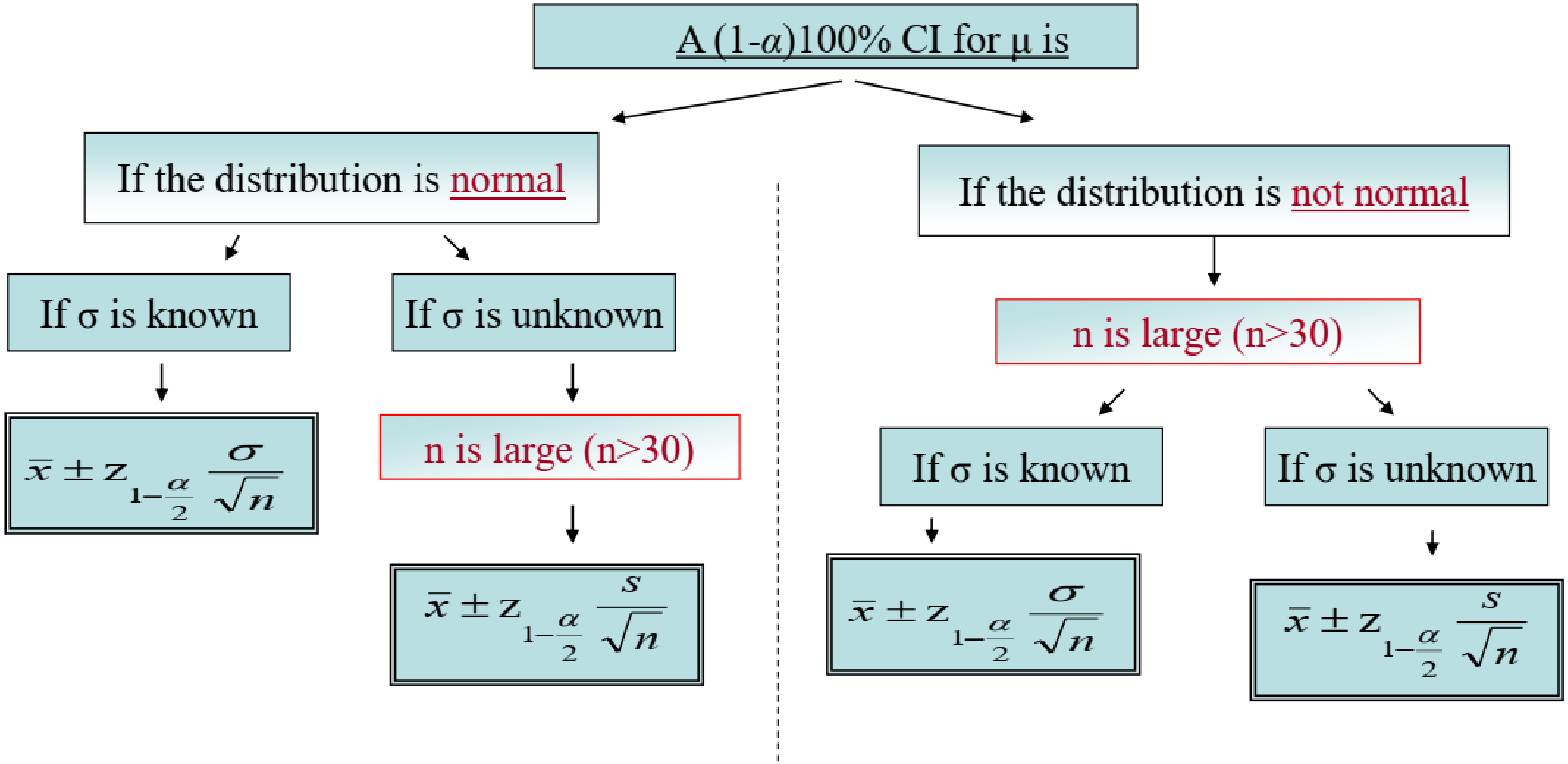
- Definition: (1-α)100% Confidence Interval:

(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).

$1-\alpha$: the confidence coefficient.

L: Lower limit of the confidence interval.

U : upper limit of the confidence interval.



A $(1-\alpha)100\%$ CI for μ is

If the distribution is normal

If σ is known

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

If σ is unknown

n is large (n>30)

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

If the distribution is not normal

n is large (n>30)

If σ is known

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

If σ is unknown

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

Note: The C.I $\bar{x} \pm z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ 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) of (variable) for the (population) is between L and U.

μ

Example 5.1:

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

α/2=0.025

1- α/2=0.975

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

b) α=0.1

$\alpha/2=0.05$

$1- \alpha/2=0.95$

$\boxed{Z_{0.95} = 1.645}$

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

Example 5.2: On 123 patient of diabetic ketoacidosis (الحماض الكيتوني السكري) 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) $\Rightarrow$ the 90% CI for μ is given by

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

$90\% = (1- \alpha)100\% \quad \Rightarrow 1- \alpha=0.9$

$\alpha=0.1 \quad \Rightarrow \quad \alpha/2=0.05 \quad \Rightarrow \quad 1-\alpha/2=0.95$

$Z_{0.95} = 1.645$

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

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

$$= (26.2 - (1645)\frac{3.3}{\sqrt{123}},\ 26.2 + (1645)\frac{3.3}{\sqrt{123}})$$

$$= (25.71,\ 26.69)$$

Interpretation: We are 90 % confident that the mean blood glucose level of diabetic ketoacidosis patient in Saudi Arabia is between 25.71 and 26.69

Exercises

Q1: Suppose that we are interested in making some statistical inferences about the mean μ of normal population with standard deviation 0.2 . Suppose that a random sample of size n = 49 from this population gave the sample mean4.5

The distribution of is

(a) N(0,1) (b) t(48) (c)N(μ,$(0.02857)^2$) (d)N(μ,2.0)

A good point estimate for μ is

(a) 4.5 (b) 2 (c) 2.5 (d) 7 (e) 1.125

Assumptions is

(a) Normal, σ known (b) Normal, σ unknown (c)not Normal, σ known
(d) not Normal, σ unknown

(4)A 95% confident interval for μ is

(a) (3.44,5.56) (b) (3.34,5.66) (c) (4.444, 4.556)
(d) (3.94,5.05) (e) (3.04,5.96)

Q2:An electronics company wanted to estimate in monthly operating expenses riyals (μ) . Assume that the population variance equals 0.584 .
Suppose that a random sample of size 49 is taken and found that the sample mean equals 5.47 . Find

Point estimate for μ
The distribution of the sample mean is
The assumptions ?
A 90% confident interval for μ.

Q3:The random variable X, representing the lifespan of a certain light bulb is distributed normally with mean of 400 hours ,and standard deviation of 10 hours.
-What is the probability that a particular light bulb will last for more than 380 hours ?
-What is the probability that a particular light bulb will last for exactly 399 hours ?
-What is the probability that a particular light bulb will last for between 380 and 420 hours ?
The mean is ……..
The variance is…..
The standard deviation ……

Q4: The tensile of a certain type of thread is approximately normally distributed with standard deviation of 6.8 Kg. A sample of 20 pieces of the thread has an average strength of 72.8 Kg. Then

A point estimate of the population mean of tensile strength μ is

(a)72.8 (b) 20 (c) 6.8 (d) 46.24 (e) none of these

A 98% Confident interval for mean of tensile strength μ,the lower bound equal to :

(a)68.45 (b) 69.26 (c) 71.44 (d) 69.68 (e) none of these

A 98% Confident interval for mean of tensile strength μ,the upper bound equal to :

(a)74.16 (b) 77.15 (c) 75.92 (d) 76.34 (e) none of these

5.3: Estimation of Population Proportion π

- **Recall that, the population proportion**

$$\pi = \frac{\text{no.of element in thepopulationwith some charachtaristic}}{\text{Total no.of element in thepopulation}}$$

N

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

$$p = \frac{\text{no. of elementin thesample with some charachtristic}}{\text{Totalno. of elementin thesample}}$$

n

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

$$\boxed{p \approx \mathrm{N}(\pi, \pi(1-\pi)/n).}$$

and hence

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

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

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

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 (خوف) of dental care in Riyadh, 22% of 347 adults said they would hesitate (تردد) 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= 347 , p= 22%=0.22,

np=(347)(0.22)=76.34 >5 and n(1-p)=(437)(0.78)=270.66>5

1- point estimation of π is p=0.22

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

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

$$=\left(0.22-(1.96)\sqrt{\frac{0.22(0.78)}{347}},0.22+(1.96)\sqrt{\frac{0.22(0.78)}{347}}\right)$$

$$=(0.22-(1.96)(0.0222379),0.22+(1.96)(0.0222379))$$

$$=(0.176,0.264)$$

<u>Interpretation</u>: we are 95% confident that the true proportion of adult in Riyadh who hesitate to take a dental appointment is between 0.176 and 0.264 .

Exercises

Q1: A random sample of 200 students from a certain school showed that 15 students smoke. let π be the proportion of smokers in the school.

- Find a point estimate for π
- Find 95% confidence interval for π

Q2. A researcher was interested in making some statistical inferences about the proportion of females (π) among the students of a certain university. A random sample of 500 students showed that 150 students are female.

1. A good point estimate for$\pi$is

(A) 0.31 (B) 0.30 (C) 0.29 (D) 0.25 (E) 0.27

1.The lower limit of a 90% confidence interval for $\pi$is

(A) 0.2363 (B) 0.2463 (C) 0.2963 (D) 0.2063 (E) 0.2663

1.The upper limit of a 90% confidence interval for $\pi$is

(A) 0.3337 (B) 0.3137 (C) 0.3637 (D) 0.2937 (E) 0.3537

Q3. In a random sample of 500 homes in a certain city, it is found that 114 are heated by oil. Let π be the proportion of homes in this city that are heated by oil.

1. Find a point estimate for π.
2. Construct a 98% confidence interval for π.

Q4. In a study involved 1200 car drivers, it was found that 50 car drivers do not use seat belt.

- A point estimate for the proportion of car drivers who do not use seat belt is:
  (A) 50 (B) 0.0417 (C) 0.9583 (D) 1150 (E) None of these

- The lower limit of a 95% confidence interval of the proportion of car drivers not using seat belt is
  (A) 0.0322 (B) 0.0416 (C) 0 .0304 (D) –0.3500 (E) None of these

- The upper limit of a 95% confidence interval of the proportion of car drivers not using seat belt is
  (A) 0.0417 (B) 0.0530 (C) 0.0512 (D) 0.4333 (E) None of these

Q5. A study was conducted to make some inferences about the proportion of female employees (π) in a certain hospital. A random sample gave the following data:

- Calculate a point estimate (p) for the proportion of female employees (π).
- Construct a 90% confidence interval for p.

Sample size	250
Number of females	120