

Course: Biostatistics(PHBB6013)

Target: MPh Students

Course Instructors:

Awol Seid (MSc in Biostatistics)

Tewodros Getinet

(MSc, Ass. Prof. of Biostatistics)

Tolesa Diriba (MSc in Biostatistics)

Public Health Department

(SPHMMC)

Basic Biostatistics

Introduction

- **What is statistics?**
- **Statistics:** A field of study concerned with:
 - collection, organization, summarization, analysis, and interpretation of numerical data, &
 - the drawing of inferences about a body of data when only a small part of the data is observed.
- **Statistics** helps us use numbers to communicate ideas

- **Biostatistics:** The application of statistical methods to the fields of biological and medical sciences.
- Concerned with *interpretation* of biological data & the *communication* of information derived from these data
- Has central role in medical investigations

- The numbers must be presented in such a way that *valid interpretations* are possible
- *Statistics are everywhere* – just look at any newspaper or the current medical and public health literature.

Uses of biostatistics

- Provide methods of organizing information
- Assessment of health status
- Health program evaluation
- Resource allocation
- Magnitude of association
 - Strong vs weak association between exposure and outcome

Uses of biostatistics

- Assessing risk factors
 - Cause & effect relationship
- Evaluation of a new vaccine or drug
 - What can be concluded if the proportion of people free from the disease is greater among the vaccinated than the unvaccinated?
 - How effective is the vaccine (drug)?
 - Is the effect due to chance or some bias?
- Drawing of inferences
 - Information from sample to population

Types of Statistics

1. Descriptive statistics:

- Ways of organizing and summarizing data
- Helps to identify the general features and trends in a set of data and extracting useful information
- Also very important in conveying the final results of a study
- **Example:** tables, graphs, numerical summary measures

Types of Statistics

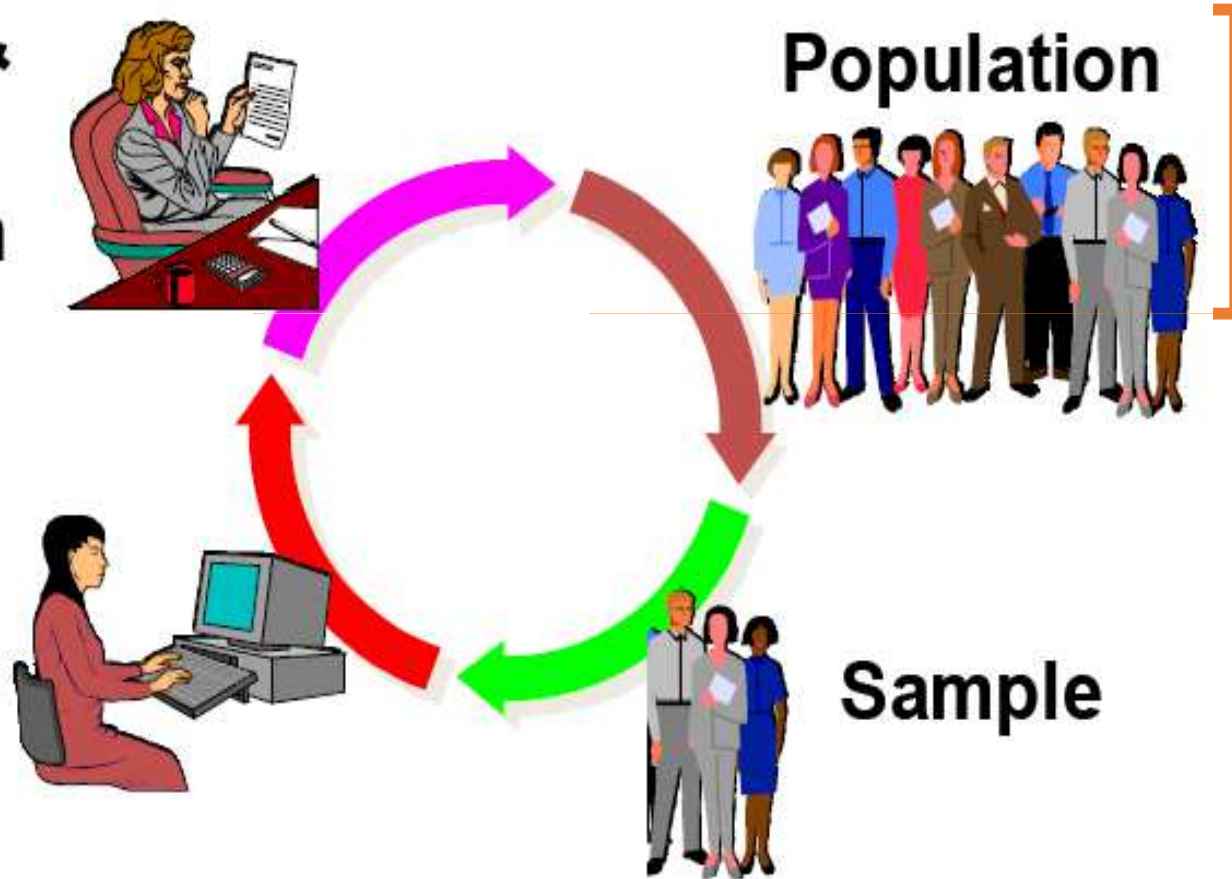
2. Inferential statistics:

- Methods used for drawing conclusions about a population based on the information obtained from a sample of observations drawn from that population
- **Example:** Principles of probability, estimation, confidence interval, comparison of two or more means or proportions, hypothesis testing, etc.

Inference Process

**Estimate &
test
population
parameter**

**Sample
statistic
(\bar{X})**

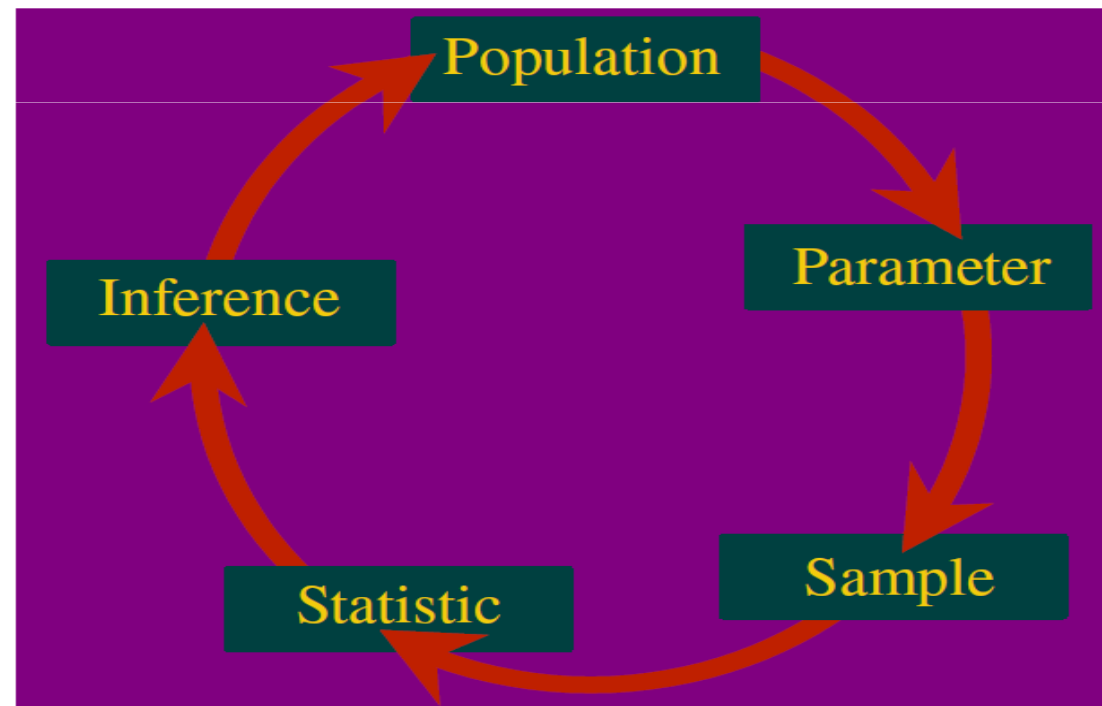


- **Parameter:** A descriptive measure computed from the data of a population.

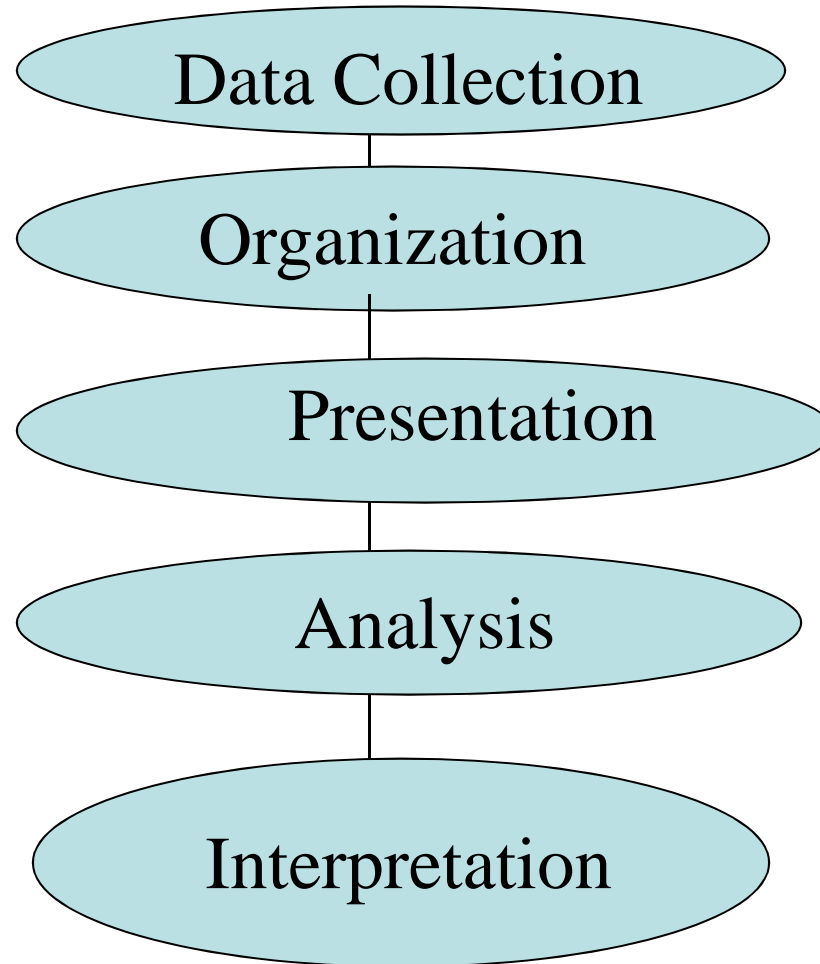
Example: the mean (μ) age of the population

- **Statistic:** A descriptive measure computed from the data of a sample.

Example: sample mean age (\bar{x})



Stages in Statistical Investigation



Data

- **Data** are numbers which can be measurements or can be obtained by counting
- The raw material for statistics
- Can be obtained from:
 - Routinely kept records, literature
 - Surveys
 - Counting
 - Experiments
 - Reports
 - Observation
 - Etc

Types of Data

1. **Primary data:** collected from the items or individual respondents directly by the researcher for the purpose of a study.
2. **Secondary data:** which had been collected by certain people or organization, & statistically treated and the information contained in it is used for other purpose by other people

Type of variables

- A **variable** is any characteristics, which can take on different values for different individuals or cases.

Example: age, blood pressure, enzyme level, heart beat per minute, etc.

- The main division is into two; **Qualitative** (or categorical) or **Quantitative** (or numerical) variables.
- **Qualitative variable:** a variable or characteristic which cannot be measured in quantitative (numeric) form but can only be identified by name or categories.

Example: Gender, Educational level, type of drug, stages of breast cancer (I, II, III, or IV).

- **Quantitative variable:** A quantitative variable is one that can be measured and expressed numerically and they can be of two types (**Discrete or Continuous**).

Discrete variables

Discrete data occur when the observations are integers that correspond with a count of some sort.

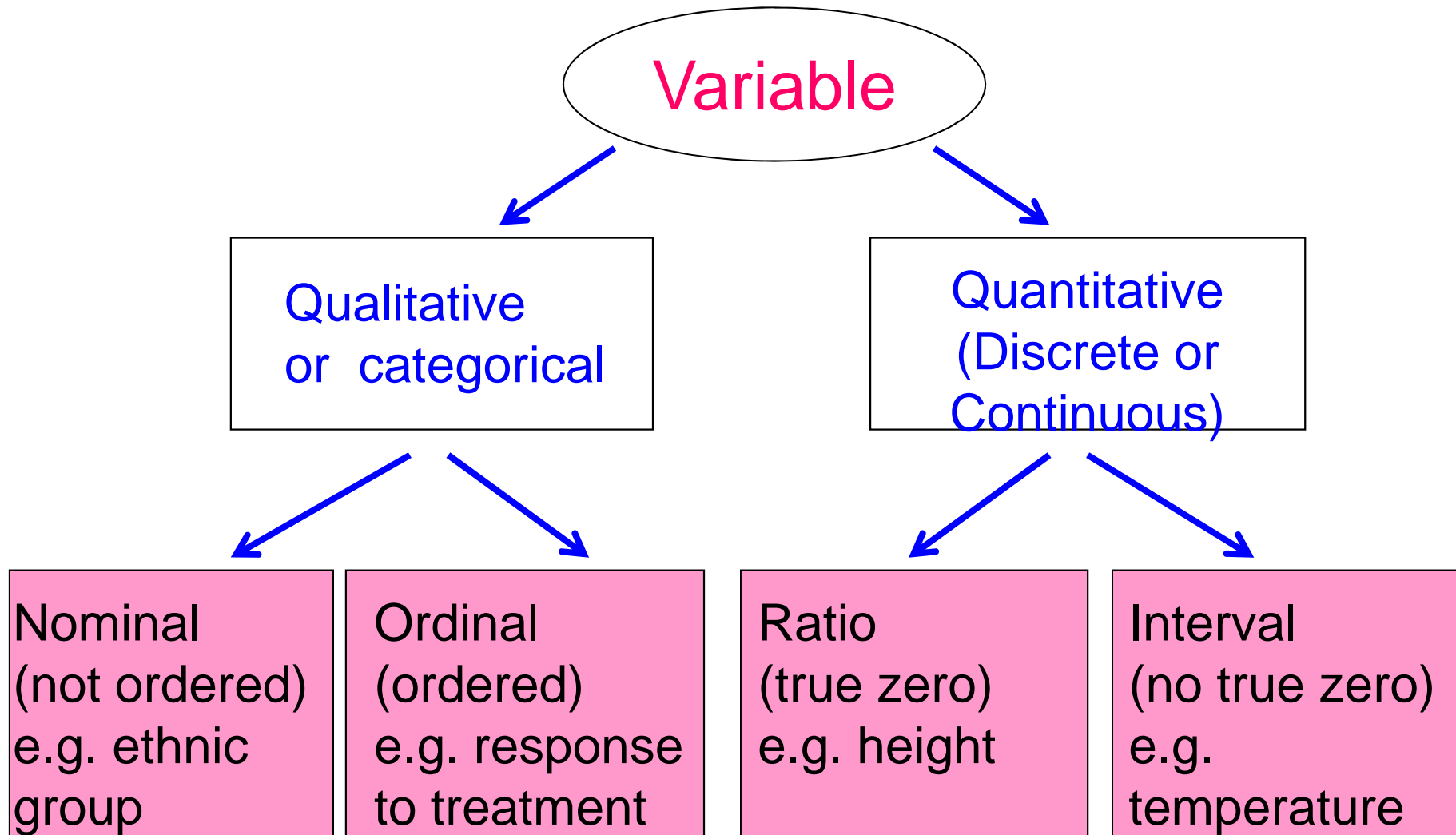
Examples : the number of bacteria colonies on a plate, the number of cells within a prescribed area upon microscopic examination, the number of heart beats within a specified time interval, etc.

Continuous variables

- ▶ For the given variable, each observation (data) theoretically falls somewhere along a continuum interval.
- ▶ One is not restricted, in principle, to particular values such as the integers of the discrete scale. The restricting factor is the degree of accuracy of the measuring instrument most clinical measurements.

Example: blood pressure, serum cholesterol level, height, weight, age etc. are on a numerical continuous scale.

Scales of Variables



Exercises

- Identify the type of data (nominal, ordinal, interval and ratio) represented by each of the following. Confirm your answers by giving your own examples.
 1. Blood group
 2. Job satisfaction index (1-5)
 3. Number of heart attacks per 1000 people
 4. Identification number
 5. The average weight gain of under five children's (with a special diet supplement) was 950 grams last month.

Methods of Data Organization and Presentation

Frequency Distributions (Tables)

- The actual summarization and organization of data starts from frequency distribution.
- **Frequency distribution:** *A table which has a list of each of the possible values that the data can assume along with the number of times each value occurs.*

- For **nominal and ordinal data**, frequency distributions are often used as a summary.
- Example:

Gender	Frequency	Relative Frequency
Male	30	65.2%
Female	16	34.8%
Total	46	100.0%

- The % of times that each value occurs, or the **relative frequency**, is often listed
- Tables make it easier to see how the data are distributed

- For continuous data, the values are grouped into non-overlapping intervals, usually of equal width if the range is maximum/ data is large.

Age	Frequency	Relative Frequency	Cum. Relative Freq.
<20	10	21.7%	21.7
20-55	25	54.4%	76.1
>55	11	23.9%	100%
Total	45	100%	

a) **Qualitative variable**: Count the number of cases in each category.

- **Example1**: The intensive care unit type of 25 patients entering ICU at a given hospital:

1. Medical
2. Surgical
3. Cardiac
4. Other

ICU Type	Frequency	Relative Frequency
Medical	12	0.48
Surgical	6	0.24
Cardiac	5	0.20
Other	2	0.08
Total	25	1.00

Example 2:

A study was conducted to assess the characteristics of a group of 234 smokers by collecting data on gender and other variables.

Gender, 1 = male, 2 = female

Gender	Frequency (n)	Relative Frequency
Male (1)	110	47.0%
Female (2)	124	53.0%
Total	234	100%

b) Quantitative variable:

- Select a set of continuous, non-overlapping intervals such that each value can be placed in one, and only one, of the intervals.
- The first consideration is how many intervals to include

Age	Frequency
15	1
19	2
29	1
31	1
34	1
39	1
52	2
53	2
45	2
65	1
71	1
74	1
75	2
76	2
77	1
78	1
79	1
84	1
85	1

For a continuous variable (e.g. – age), the frequency distribution of the individual ages is not so interesting.

Age Interval	Frequency
10-19	3
20-29	1
30-39	3
40-49	0
50-59	6
60-69	1
70-79	9
80-89	2
TOTAL	25

- We “see more” in frequencies of age values in “groupings”. Here, 10 year groupings make sense.
- Grouped data frequency distribution

To determine the number of class intervals and the corresponding width, we may use:

Sturge's rule:

$$K = 1 + 3.322(\log n)$$

$$W = \frac{L - S}{K}$$

where

K = number of class intervals

n = no. of observations

W = width of the class interval

L = the largest value

S = the smallest value

Example:

- Length of stay in time (hours) for 40 patients in a certain hospital:

23 24 18 43 20 36 24 26 23 21 42 15 19 20
22 14 13 10 19 27 29 22 44 28 34 32 23 39
21 31 16 28 41 18 12 27 15 21 25 36

$$K = 1 + 3.322 (\log 40) = 6.32 \approx 7$$

Maximum value = 44, Minimum value = 10

$$\text{Width} = (44-10)/7 = 4.86 \approx 5$$

Time (Hours)	Frequency	Relative Frequency	Cumulative Relative Frequency
10 -14	4	0.10	0.10
15 -19	7	0.175	0.275
20 -24	12	0.30	0.575
25 -29	7	0.175	0.75
30 -34	3	0.075	0.825
35-39	3	0.075	0.90
40-44	4	0.10	1.00
Total	40	1.00	

- **Cumulative frequencies:** When frequencies of two or more classes are added.
- **Cumulative relative frequency:** The percentage of the total number of observations that have a value either in that interval or below it.
- **Mid-point:** The value of the interval which lies midway between the lower and the upper limits of a class.

- **True limits:** Are those limits that make an interval of a continuous variable continuous in both directions
- Used for smoothening of the class intervals
- Subtract 0.5 from the lower and add it to the upper limit

Time (Hours)	True limit	Mid-point	Frequency
10-14	9.5 – 14.5	12	4
15-19	14.5 – 19.5	17	7
20-24	19.5 – 24.5	22	12
25-29	24.5 – 29.5	27	7
30-34	29.5 – 34.5	32	3
35-39	34.5 - 39.5	37	3
40-44	39.5-44.5	42	4
Total			40

Simple Frequency Distribution

- Primary and secondary cases of syphilis morbidity by age, 1989

Age group (years)	Cases	
	Number	Percent
0-14	230	0.5
15-19	4378	10.0
20-24	10405	23.6
25-29	9610	21.8
30-34	8648	19.6
35-44	6901	15.7
45-54	2631	6.0
>44	1278	2.9
Total	44081	100

Two Variable(Cross tab) Table

- Primary and secondary cases of syphilis morbidity by age and sex, 1989

Age group (years)	Number of cases		
	Male	Female	Total
0-14	40	190	230
15-19	1710	2668	4378
20-24	5120	5285	10405
25-29	5301	4306	9610
30-34	5537	3111	8648
35-44	5004	1897	6901
45-54	2144	487	2631
>44	1147	131	1278
Total	26006	18075	44081

Tables can also be used to present more than three or more variables.

Variable	Frequency (n)	Percent
Sex		
Male		
Female		
Age (yrs)		
15-19		
20-24		
25-29		
Religion		
Christian		
Muslim		
Occupation		
Student		
Farmer		
Merchant		

Diagrammatic Representation

- Pictorial representations of numerical data

Importance of diagrammatic representation:

1. Diagrams have greater attraction than mere figures.
2. They give quick overall impression of the data.
3. They have great memorizing value than mere figures.
4. They facilitate comparison
5. Used to understand patterns and trends

- Well designed graphs can be powerful means of communicating a great deal of information
- When graphs are poorly designed, they not only ineffectively convey message, but they are often misleading.

Specific types of diagrams and graphs include:

- Bar graph
- Pie chart



**Nominal, ordinal,
discrete data**

- Histogram
- Scatter plot
- Line graph
- Others

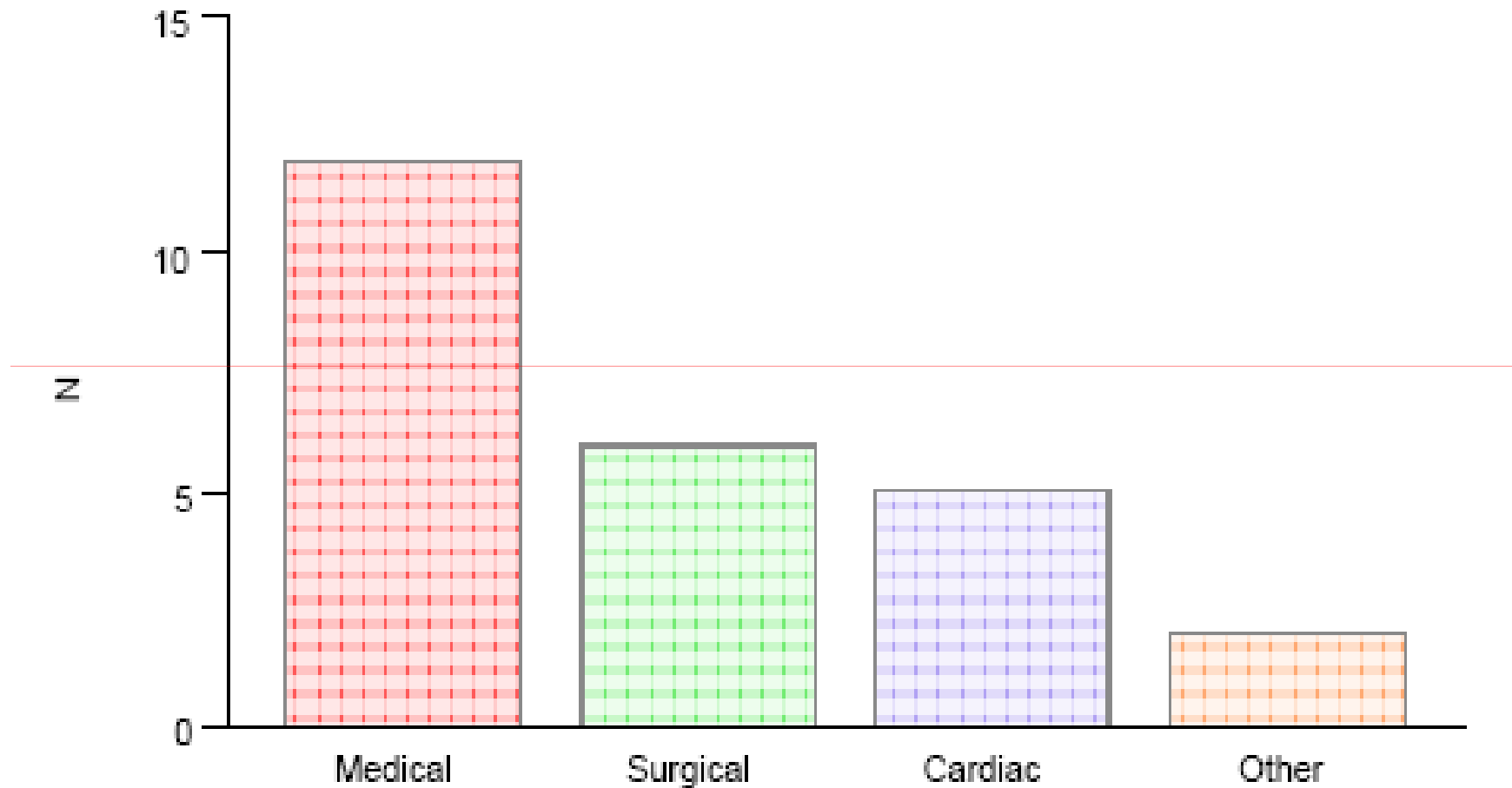


**Continuous
data**

1. Bar charts (or graphs)

- Categories are listed on the horizontal axis (X-axis)
- Frequencies or relative frequencies are represented on the Y-axis (ordinate)
- The height of each bar is proportional to the frequency or relative frequency of observations in that category

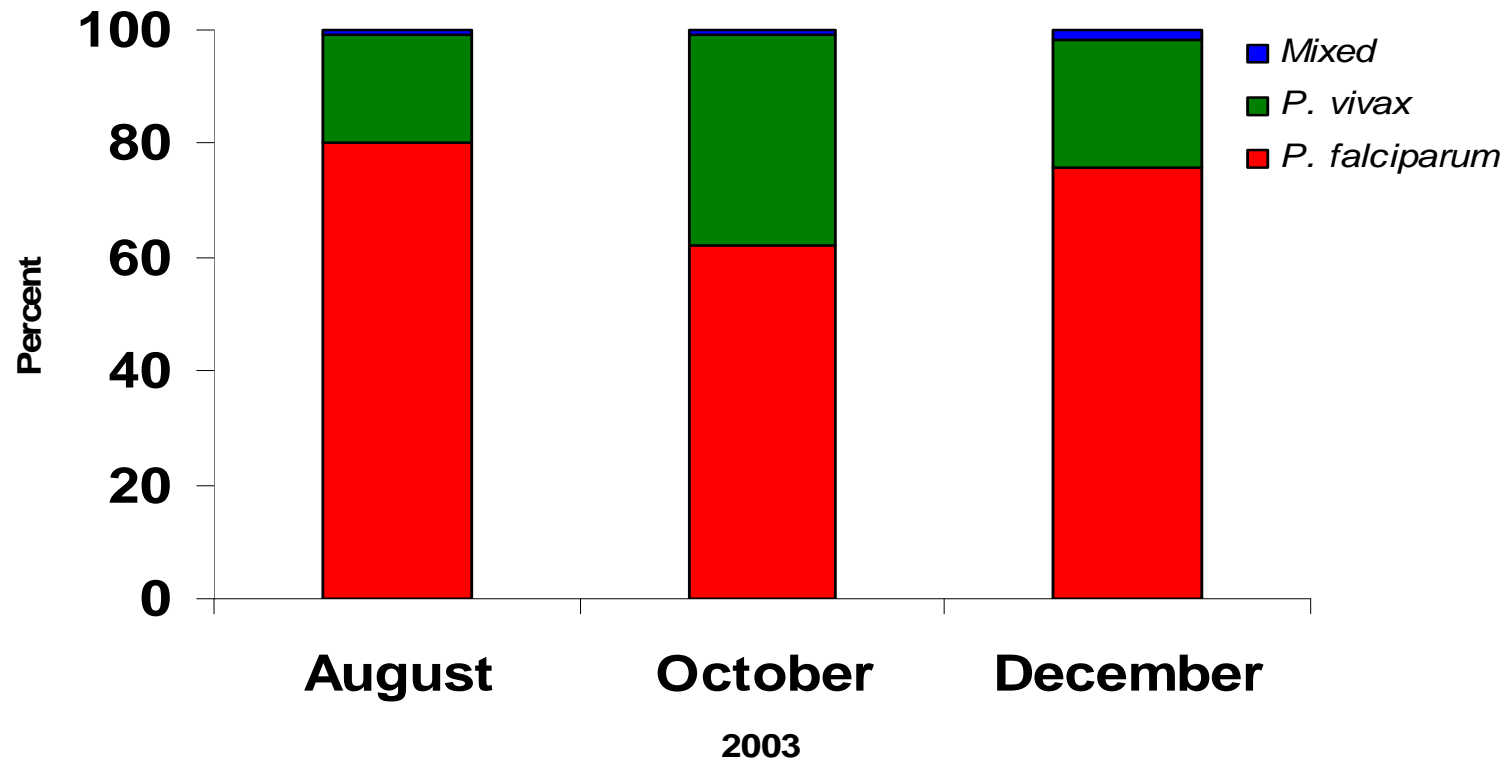
Simple Bar chart for the type of ICU for 25 patients



2. Sub-divided bar chart

- If there are different quantities forming the sub-divisions of the totals, simple bars may be sub-divided in the ratio of the various sub-divisions to exhibit the relationship of the parts to the whole.
- The order in which the components are shown in a “bar” is followed in all bars used in the diagram.
 - **Example:** Stacked and 100% Component bar charts

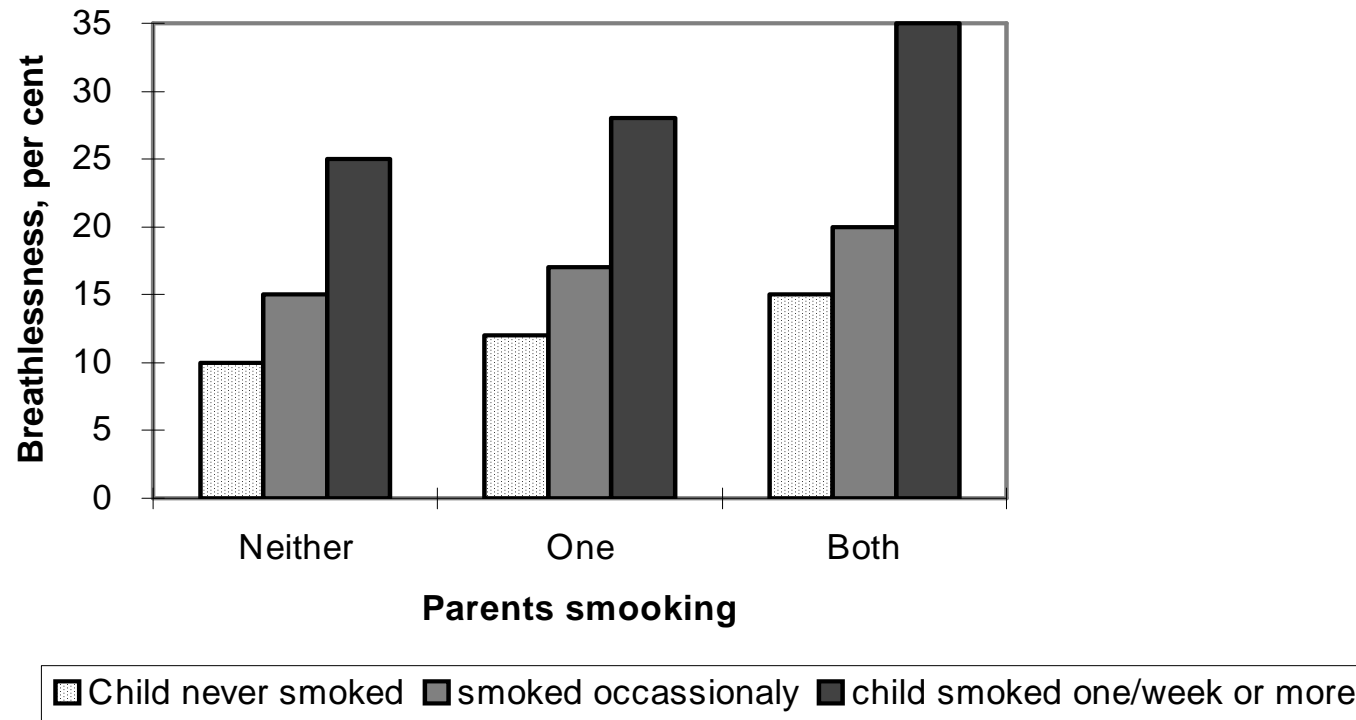
Example: Plasmodium species distribution for confirmed malaria cases, Zeway, 2003



3. Multiple bar graph

- Bar charts can be used to represent the relationships among more than two variables.
- The following figure shows the relationship between children's reports of breathlessness and cigarette smoking by themselves and their parents.

Prevalence of self reported breathlessness among school children, 1998



We can see from the graph quickly that the prevalence of the symptoms increases both with the child's smoking and with that of their parents.

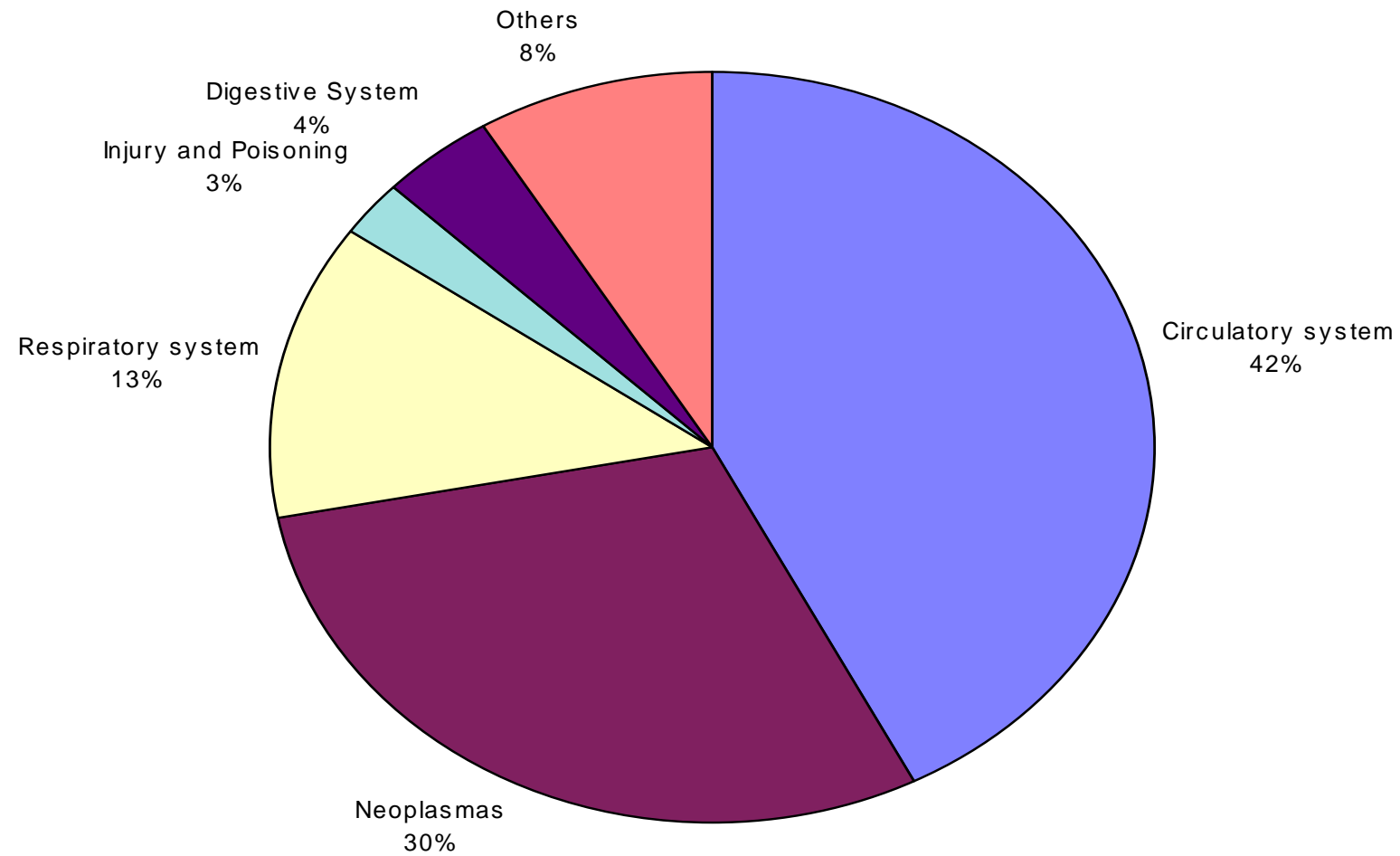
4. Pie chart

- Shows the relative frequency for each category by dividing a circle into sectors, the angles of which are proportional to the relative frequency.
- Used for a single categorical variable
- Use percentage distributions

Example: Distribution of deaths for females, in England
and Wales, 1989.

Cause of death	No. of death
Circulatory system	100 000
Neoplasm	70 000
Respiratory system	30 000
Injury and poisoning	6 000
Digestive system	10 000
Others	20 000
Total	236 000

Distribution fo cause of death for females, in England and Wales, 1989



5. Histogram

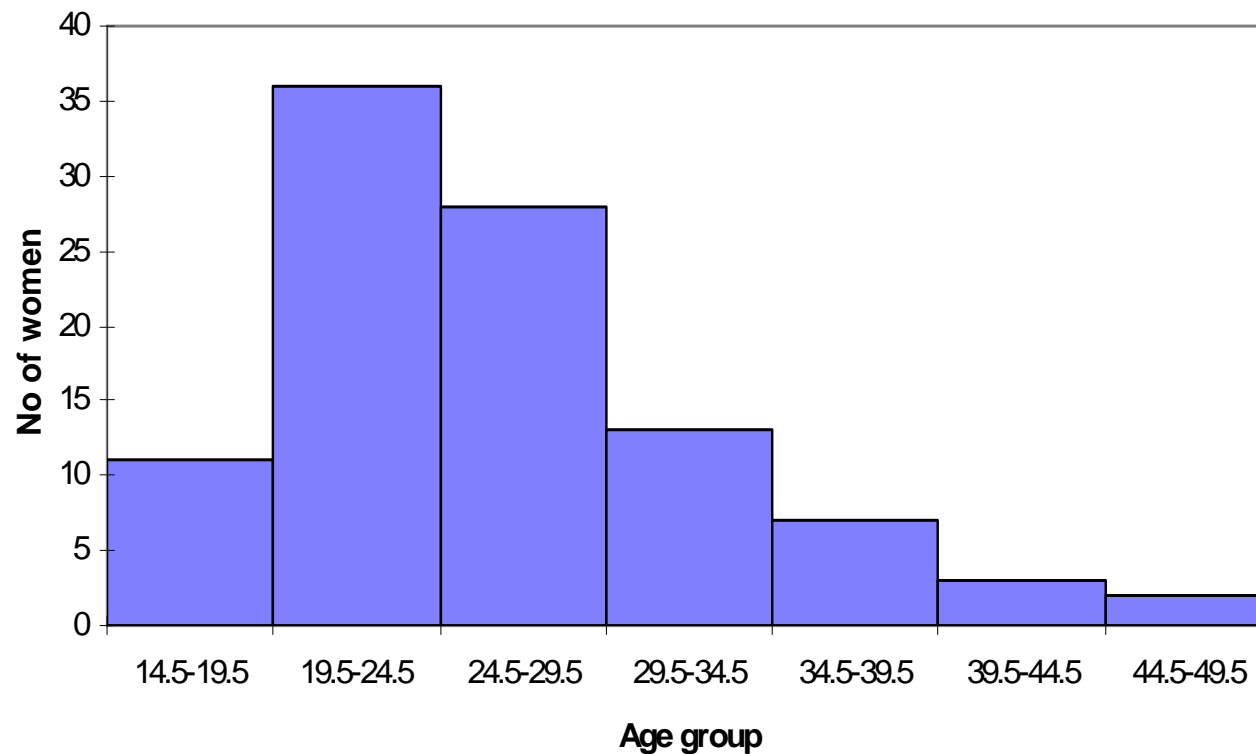
- Histograms are frequency distributions with continuous class intervals that have been turned into graphs.
- To construct a histogram, we draw the interval boundaries on a horizontal line and the frequencies on a vertical line.
- Non-overlapping intervals that cover all of the data values must be used.

- Bars are drawn over the intervals in such a way that the areas of the bars are all proportional in the same way to their interval frequencies.
- The area of each bar is proportional to the frequency of observations in the interval

Example: Distribution of the age of women at the time of marriage

Age group	15-19	20-24	25-29	30-34	35-39	40-44	45-49
Number	11	36	28	13	7	3	2

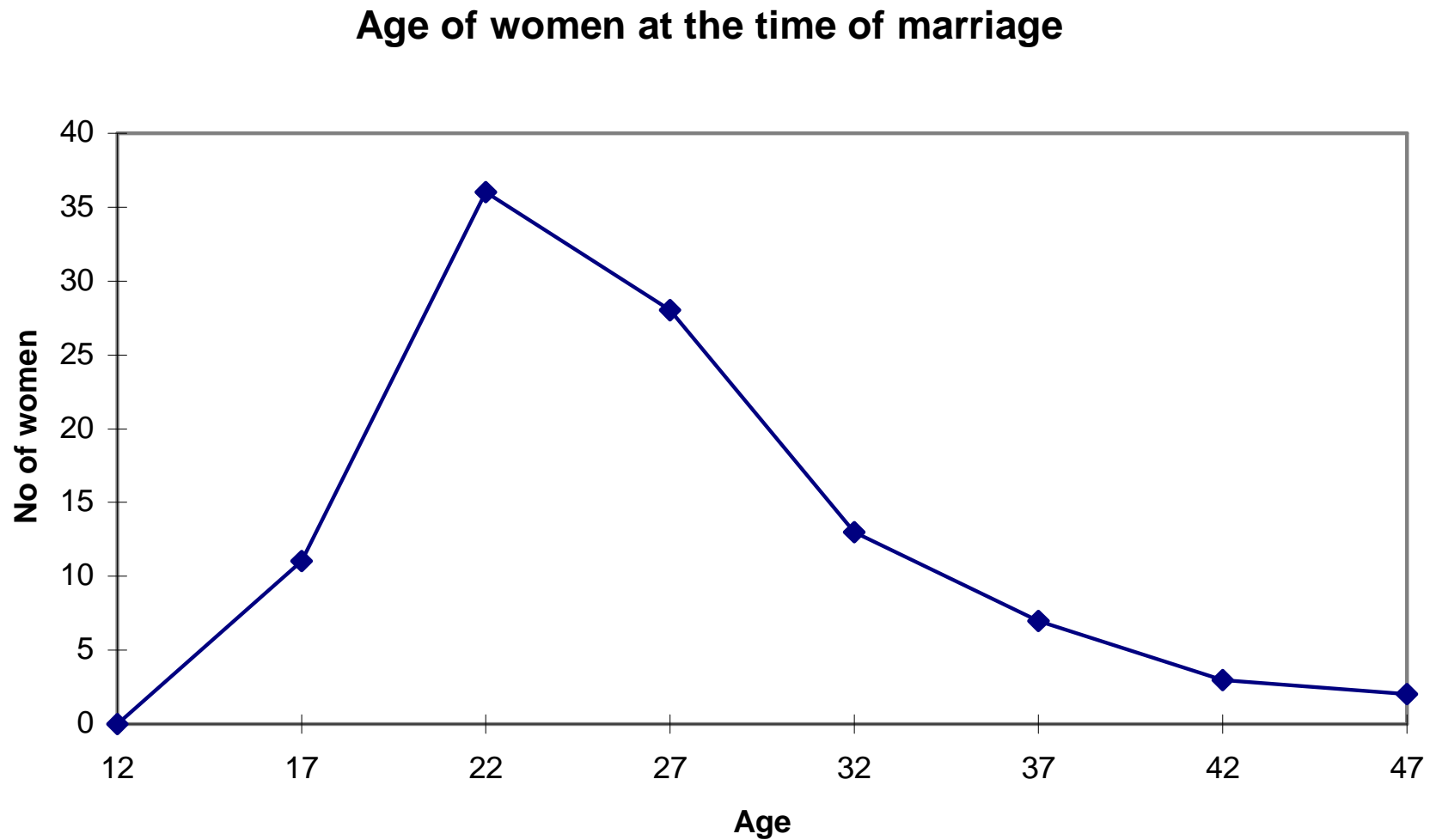
Age of women at the time of marriage



6. Frequency polygon

- A frequency distribution can be portrayed graphically in yet another way by means of a frequency polygon.
- To draw a frequency polygon we connect the mid-point of the tops of the cells of the histogram by a straight line.
- Useful when comparing two or more frequency distributions by drawing them on the same diagram

Frequency polygon for the ages of 2087 mothers with <5 children, Adami Tulu, 2003



7. Ogive Curve

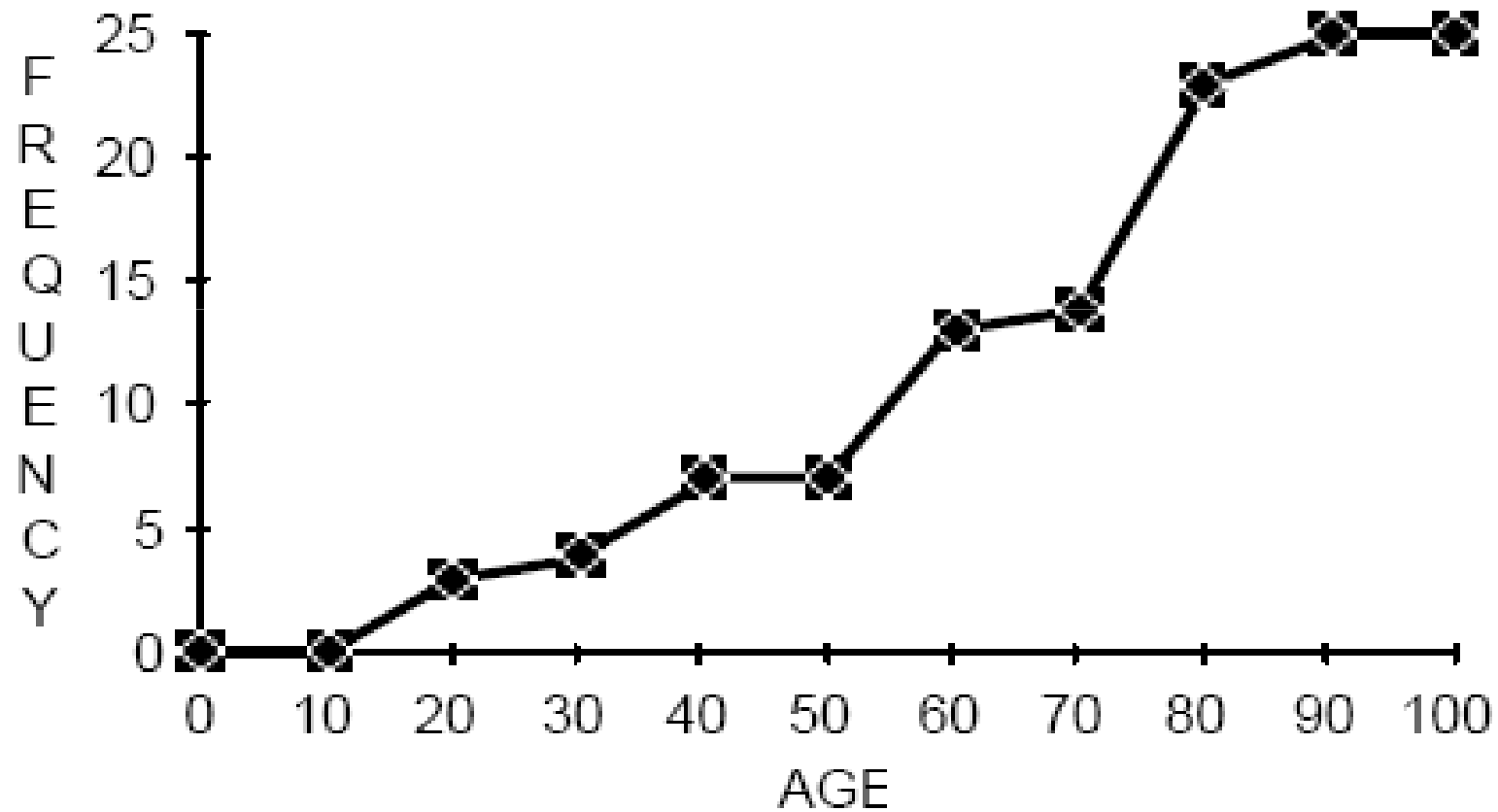
(The Cumulative Frequency Polygon)

- Some times it may be necessary to know the number of items whose values are more or less than a certain amount.
- We may, for example, be interested to know the no. of patients whose weight is <50 Kg or >60 Kg.
- To get this information it is necessary to have a 'cumulative' distribution.
- **Ogive curve** turns a cumulative frequency distribution in to graphs.
- Are much more common than frequency polygons

Cumulative Frequency and Cum. Rel. Freq. of Age of 25 ICU Patients

Age Interval	Frequency	Relative Frequency (%)	Cumulative frequency	Cumulative Rel. Freq. (%)
10-19	3	12	3	12
20-29	1	4	4	16
30-39	3	12	7	28
40-49	0	0	7	28
50-59	6	24	13	52
60-69	1	4	14	56
70-79	9	36	23	92
80-89	2	8	25	100
Total	25	100		

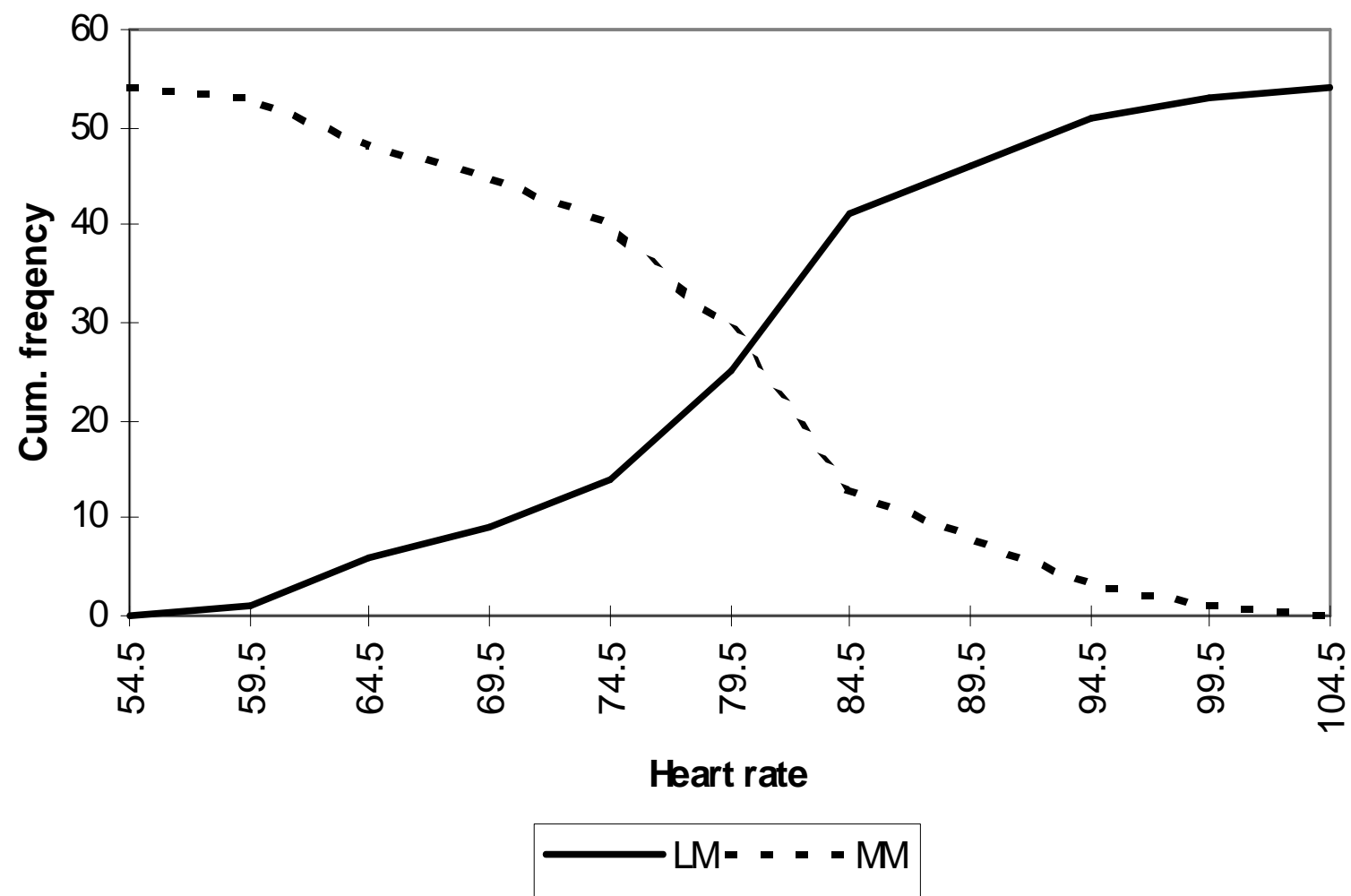
Cumulative frequency of 25 ICU patients



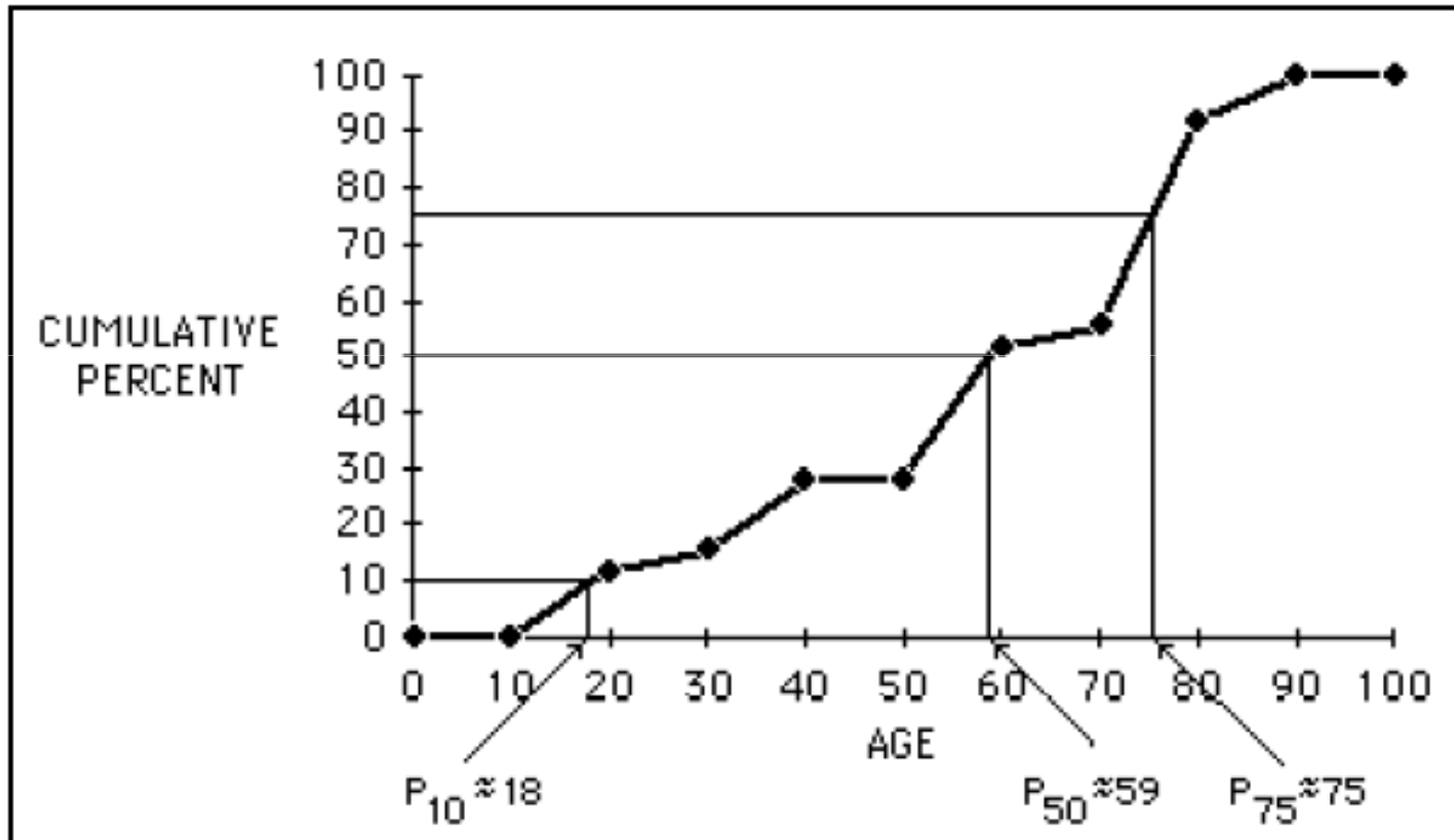
Example: Heart rate of patients admitted to hospital Y, 1998

Heart rate	No. of patients	Cumulative frequency Less than Method(LM)	Cumulative frequency More than Method(MM)
54.5-59.5	1	1	54
59.5-64.5	5	6	53
64.5-69.5	3	9	48
69.5-74.5	5	14	45
74.5-79.5	11	25	40
79.5-84.5	16	41	29
84.5-89.5	5	46	13
89.5-94.5	5	51	8
94.5-99.5	2	53	3
99.5-104.5	1	54	1

Heart rate of patients admitted in hospital Y, 1998



It is possible to estimate the values of percentiles from a cumulative frequency polygon.



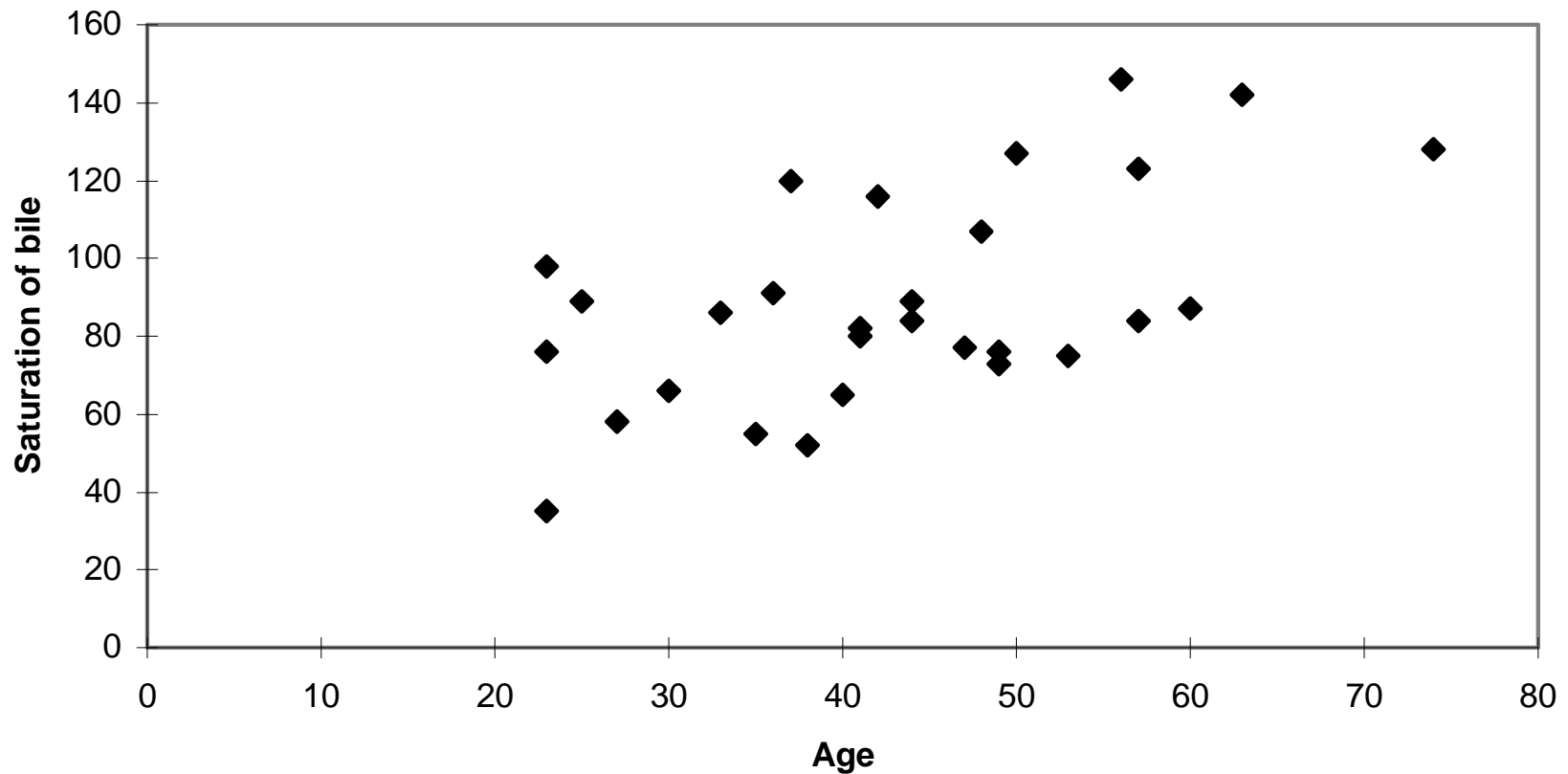
8. Scatter plot

- Most studies in medicine involve measuring more than one characteristic, and graphs displaying the relationship between two characteristics are common in literature.
- When both the variables are qualitative then we can use a multiple bar graph.

- For two quantitative variables we use bi-variate plots (also called scatter plots or scatter diagrams).
- In the study on percentage saturation of bile, information was collected on the age of each patient to see whether a relationship existed between the two measures.

- A scatter diagram is constructed by drawing X-and Y-axes.
- Each point represented by a point or dot(•) represents a pair of values measured for a single study subject

Age and percentage saturation of bile for women patients in hospital Z, 1998



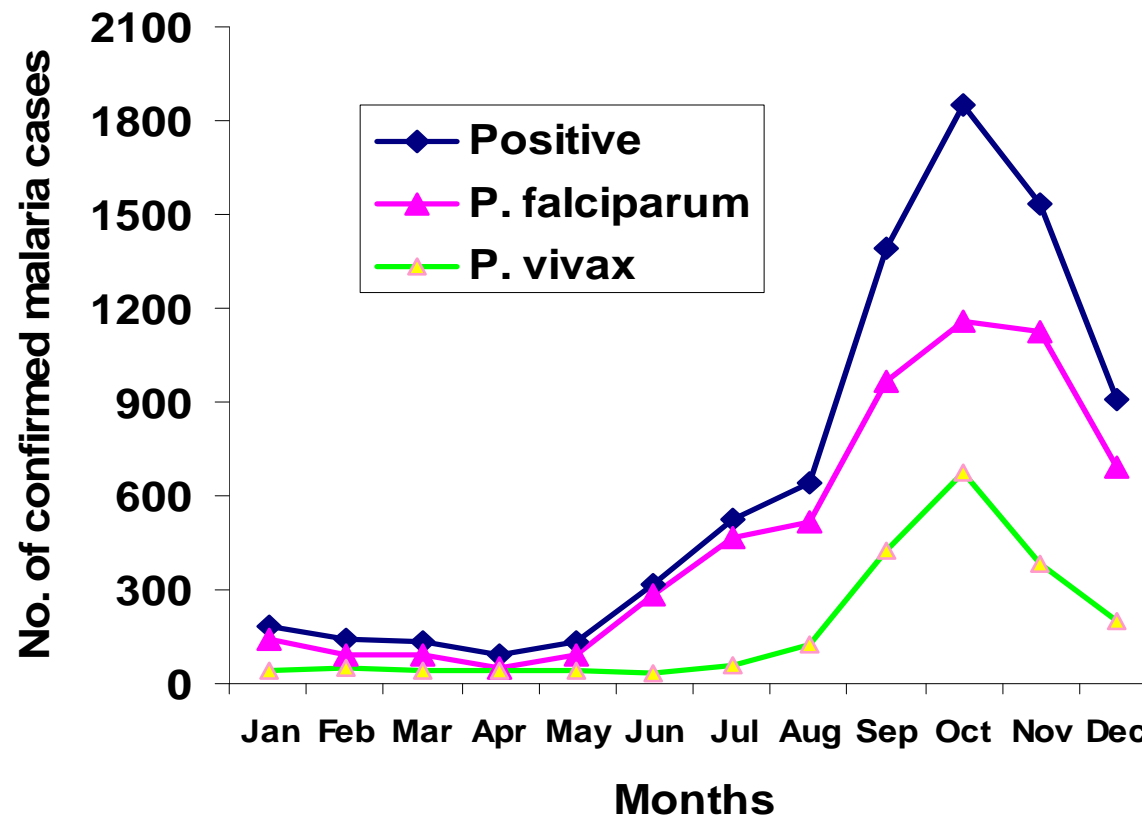
- The graph suggests the possibility of a positive relationship between age and percentage saturation of bile in women.

9. Line graph

- Useful for assessing the trend of particular situation overtime.
- Helps for monitoring the trend of epidemics.
- The time, in weeks, months or years, is marked along the horizontal axis, and

- Values of the quantity being studied is marked on the vertical axis.
- Values for each category are connected by continuous line.
- Sometimes two or more graphs are drawn on the same graph taking the same scale so that the plotted graphs are comparable.

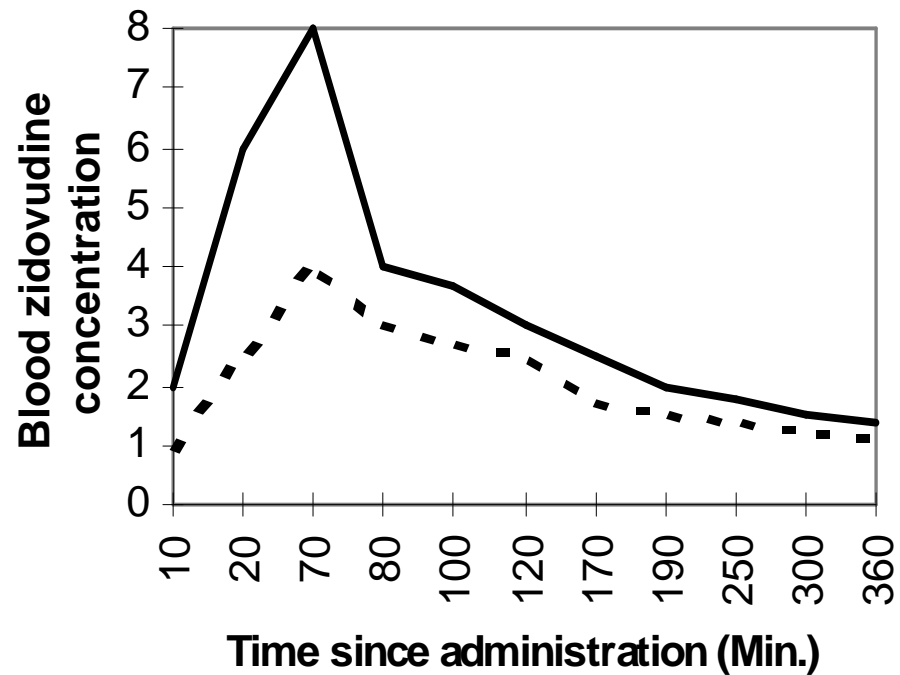
No. of microscopically confirmed malaria cases by species and month at Zeway malaria control unit, 2003



Line graph can be also used to depict the relationship between two continuous variables like that of scatter diagram.

- The following graph shows level of zidovudine (AZT) in the blood of AIDS patients at several times after administration of the drug, for with normal fat absorption and with fat mal absorption.

Response to administration of zidovudine in two groups of AIDS patients in hospital X, 1999

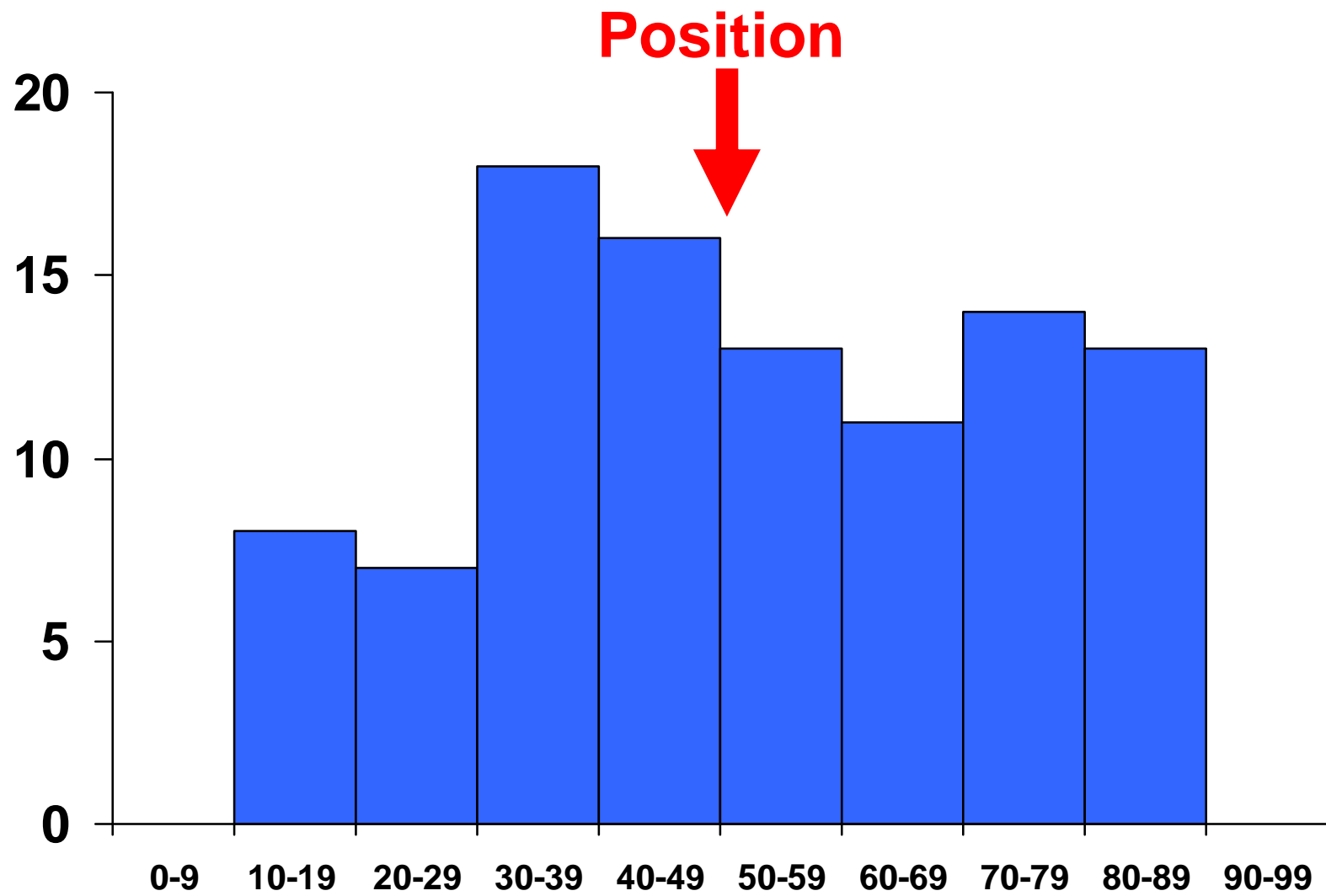


--- Fat malabsorption — Normal fat absorption

**Measures of Central Tendency
(MCT)
and
Measures of Dispersion
(MD)**

Measures of Central Tendency (MCT)

- the tendency of the statistical data to get concentrated at a certain value is called “**central tendency**”
- the various methods of determining the point about which the observations tend to concentrate are called **MCT**.
- the objective of calculating MCT is to determine a single figure which may be used to represent the whole data set.
- they facilitates comparison within one group or between groups of data.



Characteristics of a good MCT

1. It should be based on all the observations
2. It should not be affected by the extreme values
3. It should be as close to the maximum number of values as possible
4. It should have a definite value
5. It should not be subjected to complicated and tedious calculations
6. It should be capable of further algebraic treatment
7. It should be stable with regard to sampling fluctuation

- The most common measures of central tendency includes:
 - **Arithmetic Mean**
 - **Median**
 - **Mode**

Arithmetic Mean

Ungrouped Data

- the arithmetic mean is the "average" of the data set and by far the most widely used measure of central location
- is the sum of all the observations divided by the total number of observations(sample size)

$$\text{Mean} = \frac{\text{sum of values}}{\text{sample size}} = \frac{\Sigma (\text{values})}{n}$$

$$\begin{aligned}\bar{x} &= \frac{1}{n} \sum_{i=1}^n x_i \\ &= \frac{(x_1 + x_2 + x_3 + \dots + x_n)}{n}\end{aligned}$$

Example: The heart rates for $n=10$ patients were as follows (beats per minute): 167, 120, 150, 125, 150, 140, 40, 136, 120, 150. What is the arithmetic mean for the heart rate of these patients?

$$\begin{aligned}\bar{x} &= \frac{1}{10} \sum_{i=1}^{10} x_i \\ &= \frac{1298}{10} \\ &= 129.8 \text{ beats per minute}\end{aligned}$$

What does it imply?

Grouped Data

- Here the assumption is all values in a particular class interval are located at the mid point of the interval
- And the mean is calculated as follows;

$$\bar{X} = \frac{\sum_{i=1}^k m_i f_i}{\sum_{i=1}^k f_i}$$

where,

k = the number of class intervals

m_i = the mid-point of the i^{th} class interval

f_i = the frequency of the i^{th} class interval

Example: Compute the mean age at diagnosis of 169 cervical cancer patients; use the grouped frequency table below.

Mean=5810.5/169=34.48years*(What does it imply?)*

Class interval	Mid-point (mi)	Frequency (fi)	mifi
10-19	14.5	4	58.0
20-29	24.5	66	1617.0
30-39	34.5	47	1621.5
40-49	44.5	36	1602.0
50-59	54.5	12	654.0
60-69	64.5	4	258.0
Total	—	169	5810.5

Properties of the Arithmetic Mean.

- For a given set of data there is one and only one arithmetic mean (uniqueness).
- Easy to calculate and understand (simple).
- Influenced by each and every value in a data set
- Greatly affected by the extreme values
- In case of grouped data if any class interval is open, arithmetic mean can not be calculated.

Median

Ungrouped data

- The median is the value which divides the data set into two equal parts.
- If the number of observations is odd, the median will be the middle value when all values are arranged in order of magnitude.
- When the number of observations is even, the median will be located b/n two observations.
- In this case the median is the mean of these two observations, when all observations have been arranged in the order of their magnitude.

If the sample size n is ODD	median = $\frac{n+1}{2}$ th largest value
If the sample size n is EVEN	median = average of $\left(\left[\frac{n}{2}\right]th, \left[\frac{n+2}{2}\right]th\right)$ values

Example

- Data, from smallest to largest, are: 1, 1, 2, 3, 7, 8, 11, 12, 14, 19, 20
- The sample size, $n=11$
- Median is the $\frac{n+1}{2}$ th largest $= \frac{12}{2} = 6$ th largest value
- Thus, median value is $= 8$
- Five values are smaller than 8; five values are larger.

Example

- Data, from smallest to largest, are: 2, 5, 5, 6, 7, 10, 15, 21, 22, 23, 23, 25
- The sample size, $n=12$
- Median = average of $\frac{n}{2}$ th largest, $\frac{n+1}{2}$ th largest = average of 6th and 7th largest values
- Thus, median value is = average (10, 15) = 12.5

Grouped data

- In calculating the median from grouped data, we assume that the values within a class-interval are evenly distributed through the interval.
- The first step is to locate the class in which the median is located, using the following procedure.
- Find $n/2$ and see a class with a minimum cumulative frequency which contains $n/2$.
- Then, use the following formula;

$$\tilde{X} = L_m + \left(\frac{\frac{n}{2} - F_c}{f_m} \right) W$$

where,

L_m = lower class boundary of the median class

F_c = cumulative frequency of the class preceding the median class

f_m = frequency of the median class

W = width of the median class

n = total number of observations

Example: Compute the median age at diagnosis of 169 cervical cancer patients; use the grouped frequency table below.

$$n/2=169/2=84.5$$

Class interval	Mid-point (mi)	Frequency (fi)	Cum. freq
10-19	14.5	4	4
20-29	24.5	66	70
30-39	34.5	47	117
40-49	44.5	36	153
50-59	54.5	12	165
60-69	64.5	4	169
Total		169	

- $n/2 = 84.5$ = in the 3rd class (the median class is the 3rd class)
- Lower class boundary = 29.5,
- Frequency of the median class = 47
- $(n/2 - F_c) = 84.5 - 70 = 14.5$
- **Median = $29.5 + (14.5/47)10 = 32.58 \approx 33$**
- ***What does it imply?***

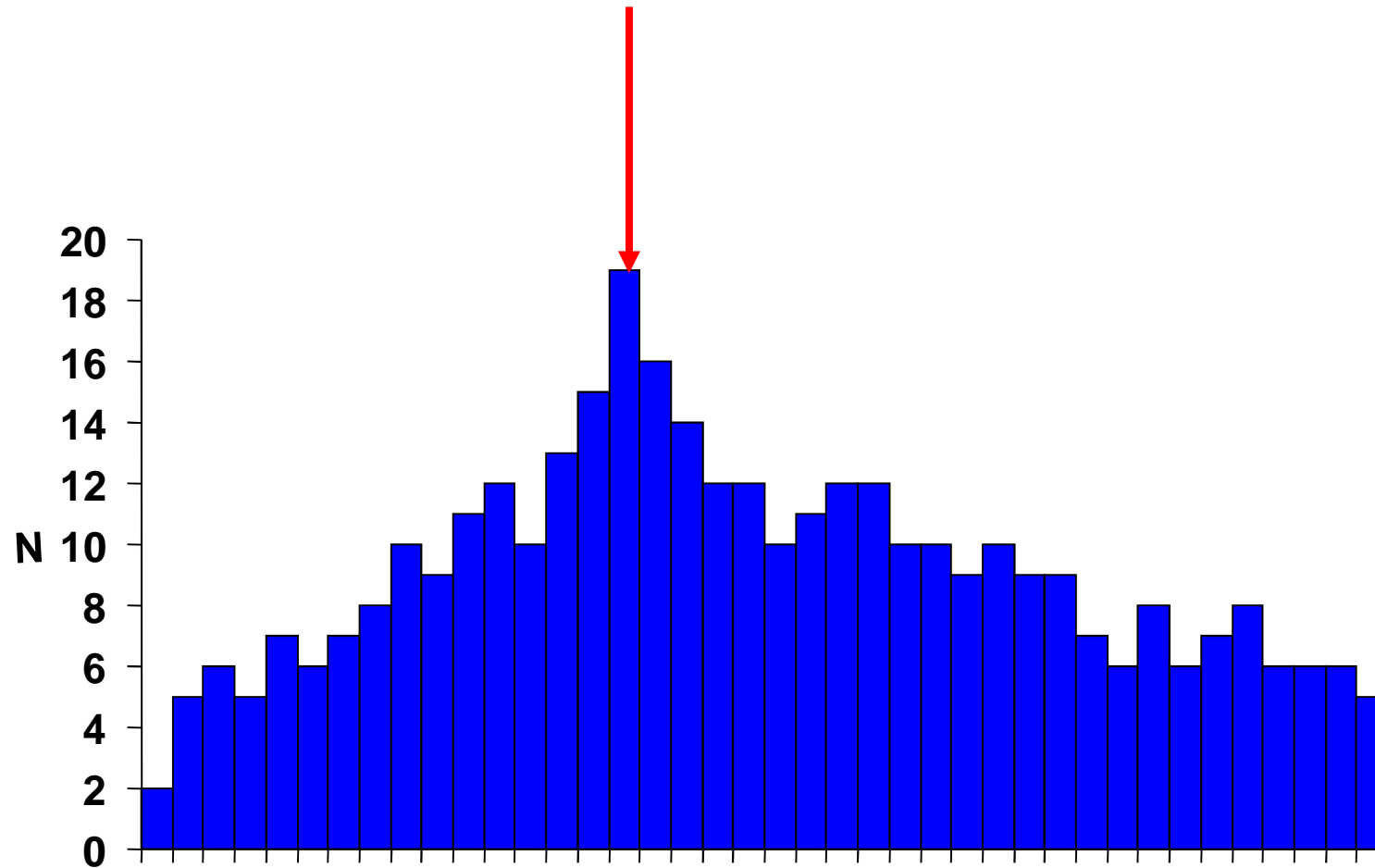
Properties of the median

- There is only one median for a given set of data (uniqueness)
- The median is easy to calculate
- Median is insensitive to very large or very small values
- Median can be calculated even in the case of open end intervals
- It is not based on all observations.

Mode

- Ungrouped data
- It is a value which occurs most frequently in a set of values.
- If all the values are different there is no mode, on the other hand, a set of values may have more than one mode.

Mode



- **Example**

- Data are: 1, 2, 3, 4, 4, 4, 4, 5, 5, 6
- Mode is 4 “Unimodal”

- **Example**

- Data are: 1, 2, 2, 2, 3, 4, 5, 5, 5, 6, 6, 8
- There are two modes – 2 & 5
- This distribution is said to be “bi-modal”

- **Example**

- Data are: 2.62, 2.75, 2.76, 2.86, 3.05, 3.12
- No mode, since all the values are different

Grouped data

- To find the mode of grouped data, we usually refer to the modal class, where the modal class is the class with the highest frequency.
- Once the modal class is/are identified it can be calculated using the following formula.

$$x = L_m + \left(\frac{\Delta_1}{\Delta_2 + \Delta_1} \right) * W$$

Where,

W = width of the modal class

L_m = lower class boundary of the modal class

f_1 = frequency of the class preceding the modal class

f_2 = frequency of the modal class

f_3 = frequency of the class succeeding the modal class

$$\Delta_1 = f_2 - f_1$$

$$\Delta_2 = f_2 - f_3$$

Example: Compute the most frequent age at diagnosis of 169 cervical cancer patients; use the grouped frequency table below.

Class interval	Mid-point (mi)	Frequency (fi)	Cum. freq
10-19	14.5	4	4
20-29	24.5	66	70
30-39	34.5	47	117
40-49	44.5	36	153
50-59	54.5	12	165
60-69	64.5	4	169
Total		169	

- the 2nd class is the modal class ($f_2 = 66$)
- $f_1 = 4$ and $f_3 = 47$
- Lower class boundary(L_m) = 19.5,
- Class width (W)= 10
- $= 84.5 - 70 = 14.5$
- $\Delta_1 = f_2 - f_1 = 62$
- $\Delta_2 = f_2 - f_3 = 19$
- **$Mode = 19.5 + (62/81) * 10 = 27.15 \approx 27$**
- ***What does it imply?***

Properties of mode

- It is not affected by extreme values
- It can be calculated for distributions with open end classes
- Often its value is not unique

Measures of Dispersion (MD)

Consider the following two sets of data:

A: 177 193 195 209 226

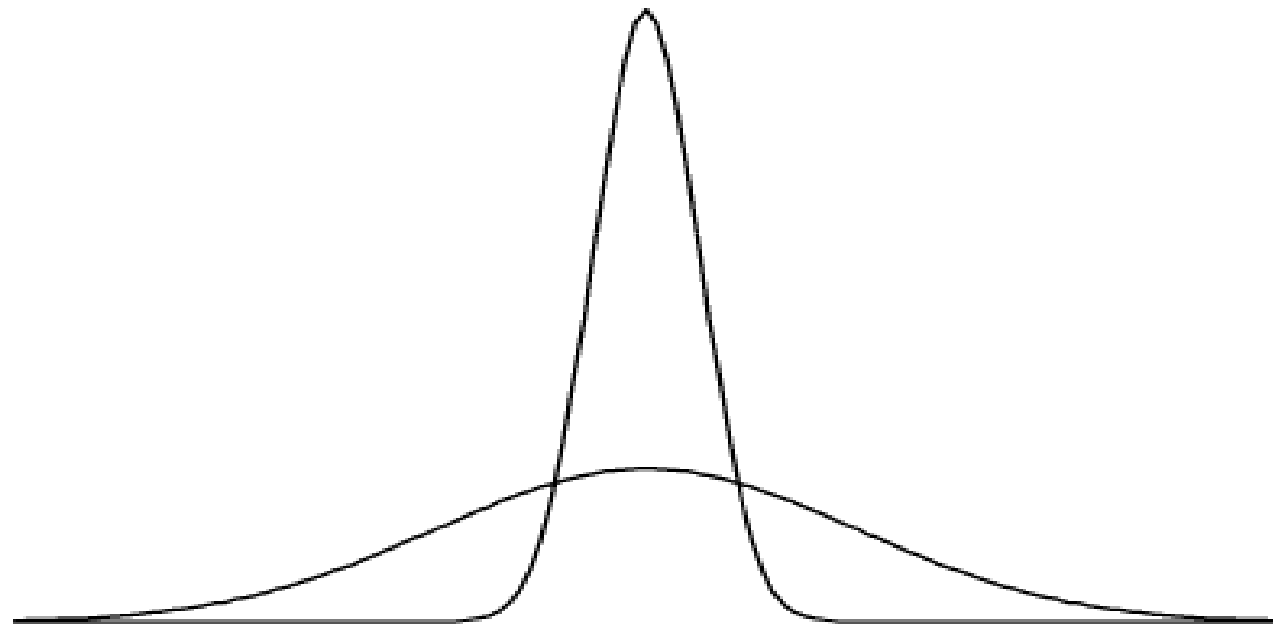
Mean = 200

B: 192 197 200 202 209

Mean = 200

Two or more sets may have the same mean and/or median but they may be quite different.

These two distributions have the same mean, median, and mode but they are different. What makes them different?



- MCT are not enough to give a clear understanding about the distribution of the data.
- We need to know something about the variability or spread of the values — whether they tend to be clustered close together, or spread out over a broad range
- This leads us to the concept of MD

- Measures that quantify the **variation or dispersion** of a set of data from its central location
- **Dispersion** refers to the variety exhibited by the values of the data.
- The amount may be small when the values are close together.
- *If all the values are the same, no dispersion*

Other synonymous term:

- “Measure of Variation”
- “Measure of Spread”
- “Measures of Scatter”

Common Measures of dispersion :

- Variance
- Standard deviation
- Standard error
- Coefficient of variation

Variance (σ^2 , s^2)

- The variance is the average of the squares of the deviations of each observation taken from their mean.
- It is squared because the sum of the deviations of the individual observations from their mean is always 0

$$0 = \sum (x_i - \bar{x})$$

- is used to measure the dispersion of values relative to their mean.

- When values are close to their mean the dispersion is smaller than scattered values.
 - **Population variance = σ^2**
 - **Sample variance = S^2**

Ungrouped data

- Let X_1, X_2, \dots, X_N be the measurement on N population units, then:

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

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

A sample variance is calculated for a sample of individual values (X_1, X_2, \dots, X_n) and uses the sample mean (\bar{X}) rather than the population mean (μ)

$$S^2 = \frac{\text{sum of (value - sample mean)}^2}{\text{sample size} - 1}$$

$$= \frac{\sum (\text{value} - \text{sample mean})^2}{n - 1}$$

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

Grouped data

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

where

m_i = the mid-point of the i^{th} class interval

f_i = the frequency of the i^{th} class interval

\bar{x} = the sample mean

k = the number of class intervals

Properties of Variance:

- The main disadvantage of variance is that its unit is the square of the unit of the original measurement values
- The variance gives more weight to the extreme values as compared to those which are near to mean value, because the difference is squared in variance.
- The drawbacks of variance are overcome by the standard deviation.

Standard deviation (σ , s)

- It is the square root of the variance.
- This produces a measure having the same scale as that of the individual values.

$$\sigma = \sqrt{\sigma^2} \text{ and } S = \sqrt{S^2}$$

- **Example:** Calculate the sample variance and SD using the data on survival times of $n=11$ patients after heart transplant surgery.

Patient Identifier, "i"	Survival (days), X_i	Mean for sample, \bar{X}	Deviation , $(X_i - \bar{X})$	Squared deviation $(X_i - \bar{X})^2$
1	135	161	-26	676
2	43	161	-118	13924
3	379	161	218	47524
4	32	161	-129	16641
5	47	161	-114	12996
6	228	161	67	4489
7	562	161	401	160801
8	49	161	-112	12544
9	59	161	-102	10404
10	147	161	-14	196
11	90	161	-71	5041
TOTAL	1771		0	285236

$$\blacklozenge \quad s^2 = \frac{\sum_{i=1}^{11} (X_i - \bar{X})^2}{n - 1} = \frac{285236}{10} = 28523.6 \text{ days}^2$$

$$\blacklozenge \quad s = \sqrt{s^2} = \sqrt{28523.6} = 168.89 \text{ days}$$

Example: Compute the variance and SD of the age at diagnosis of 169 cervical cancer patients

$$\text{Mean} = 5810.5 / 169 = 34.48 \text{ years}$$

$$S^2 = 20199.22 / 169 - 1 = 120.23 \text{ years}^2$$

$$SD = \sqrt{S^2} = \sqrt{120.23} = 10.96 \text{ years (What does it imply??)}$$

Class interval	(mi)	(fi)	(mi-Mean)	(mi-Mean) ²	(mi-Mean) ² fi
10-19	14.5	4	-19.98	399.20	1596.80
20-29	24.5	66	-9.98	99.60	6573.60
30-39	34.5	47	0.02	0.0004	0.0188
40-49	44.5	36	10.02	100.40	3614.40
50-59	54.5	12	20.02	400.80	4809.60
60-69	64.5	4	30.02	901.20	3604.80
Total		169		1901.20	20199.22

Properties of SD

- The SD has the advantage of being expressed in the same units of measurement as the mean
- However, if the units of measurements of variables of two data sets are not the same, then their variability can't be compared by comparing the values of SD.

SD Vs Standard Error (SE)

- **SD** is about the variability of individuals
- **SE** is used to describe the variability in the means of repeated samples taken from the same population

$$SE(\bar{X}) = \frac{SD}{\sqrt{n}}$$

- **Example:** *imagine 5,000 samples, each of the same size $n=11$. This would produce 5,000 sample means. The variability among the 5,000 sample means described using the SE, not the SD.*

Example: The heart transplant surgery

n=11, SD=168.89, Mean=161 days

$$SE(\bar{X}) = \frac{SD}{\sqrt{n}} = \frac{168.89}{\sqrt{11}} = 50.9$$

What does it imply??

Coefficient of variation (CV)

- When two data sets have different units of measurements, or their means differ sufficiently in size, the CV should be used as a measure of dispersion.
- It is the best measure to compare the variability of two series of sets of observations.
- Data with less coefficient of variation is considered more consistent.

- CV is the ratio of the SD to the mean multiplied by 100.

$$C V = \frac{S}{\bar{X}} \times 100$$

	SD	Mean	CV (%)
SBP	15mm	130mm	11.5
Cholesterol	40mg/dl	200mg/dl	20.0

- The data on Cholesterol is more variable/less consistent/less reliable than systolic blood pressure
- For single data 25% is the cutoff value.

Measure of relative standing (Z-Score)

- is the number of standard deviations that a given value X is below or above the mean
- $Z = \frac{xi - \bar{X}}{s}$ for the sample datasets
- $Z = \frac{xi - \mu}{\sigma}$ for population data sets
- Values above the mean have positive z-scores and values below the mean have negative Z-scores.
- It measures relative standing(performance)
- it is useful to transform a given data sets in to a new distribution

Observations with higher Z-score values reflect better performance (not always true)

Example: Two public health experts from two different areas were assessed for the time they have taken to accomplish a given task

PH Expert	Time	SD	Mean	Z-Score
A	14 hr	1hr	13hr	1
B	28 min.	4min.	20min.	2

- Expert A performs better relative to his group
- **Why??**

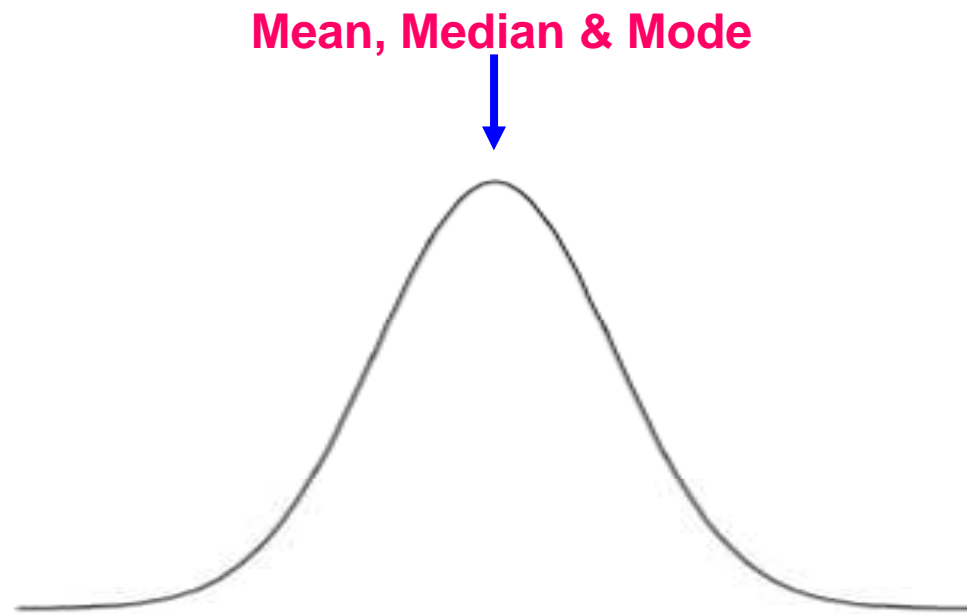
Properties of Z-Score

- The sum of Z-scores is always zero
- The mean of Z-score is zero
- The variance and standard deviation of z-score are equal to one

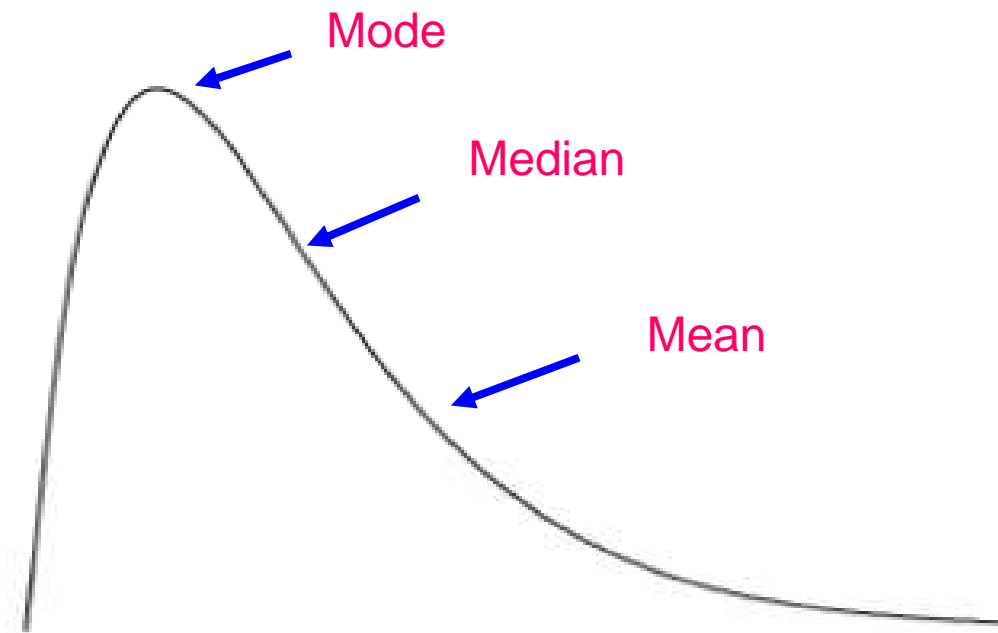
Measure of Shift and Size (Skewness and Kurtosis)

- **Skewness:** If extremely low or extremely high observations are present in a distribution, then the mean tends to shift towards those scores.
- It is the degree of symmetry or asymmetry of a distribution
- In a moderately skewed distribution;
$$\text{Mean} - \text{Mode} = 3(\text{Mean} - \text{Median})$$
- Based on the type of skewness, distributions can be:

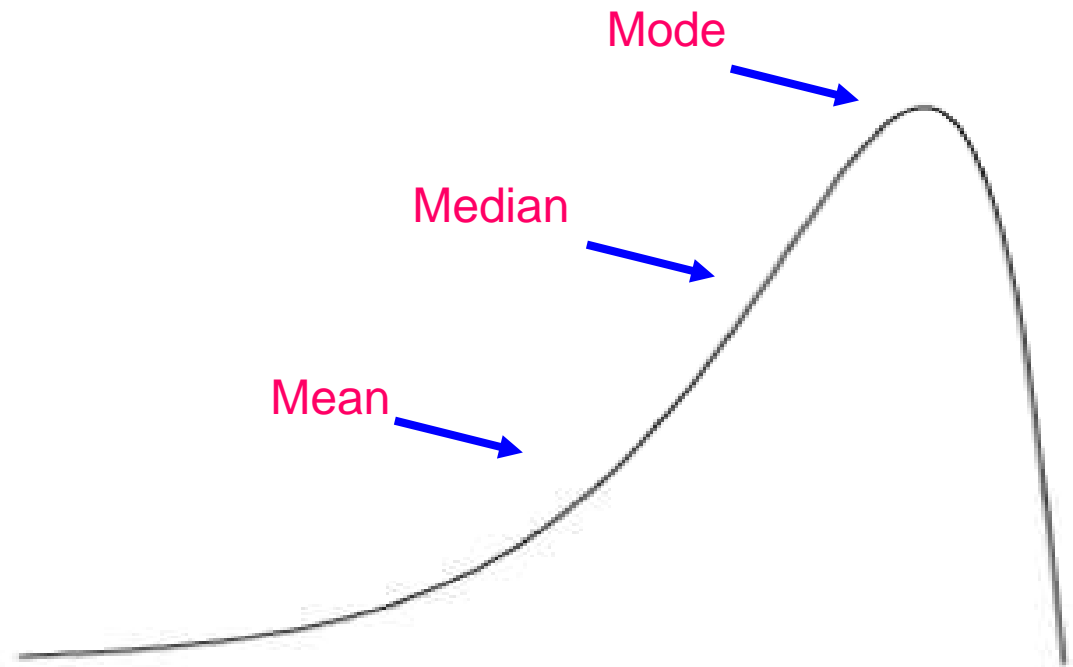
Symmetric distribution — Mean, median, and mode should all be approximately the same



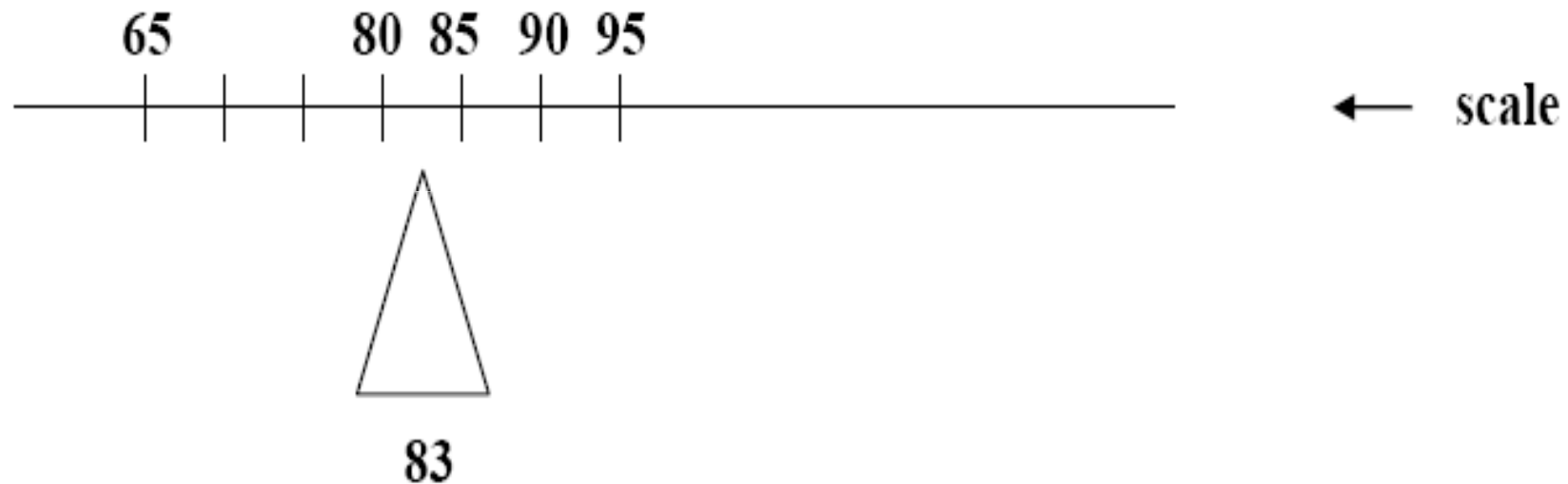
- **Skewed to the right (positively skewed):**
- Mean is greater than the mode
- majority of scores are at the left end of the curve and a few extreme large scores are scattered at the right end



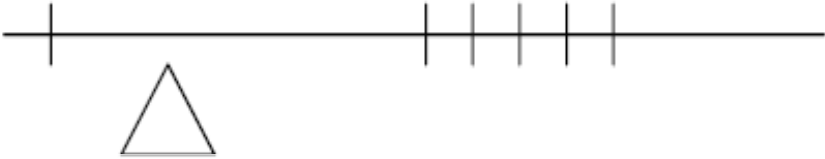
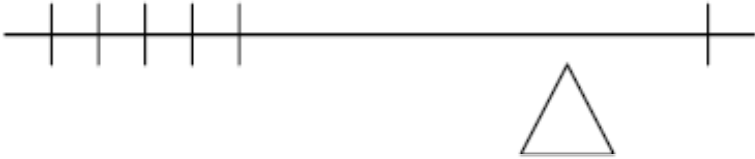
- **Skewed to the left (negatively skewed):**
- mean is less than the mode
- majority of scores are at the right end of the curve and a few small scores are scattered at the left end



The mean can be thought of as a
“balancing point”, “center of gravity”



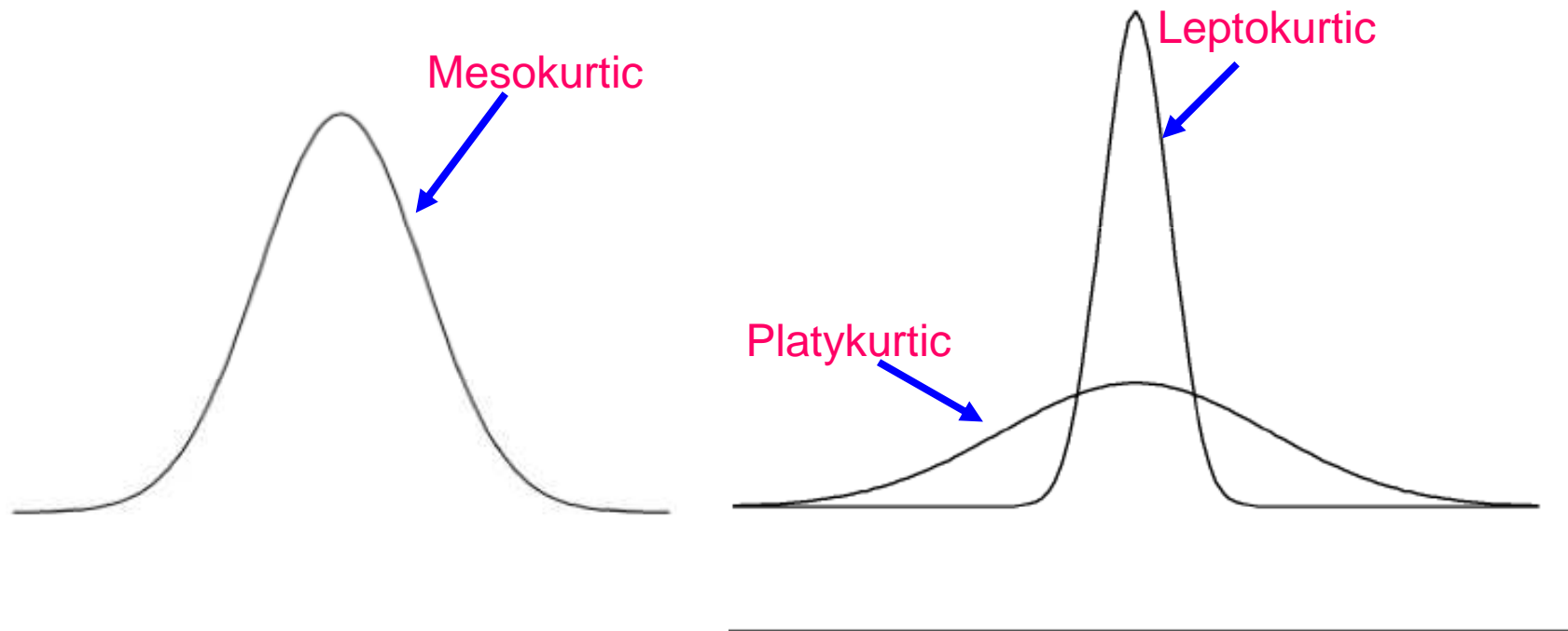
- When the data are skewed, the mean is “dragged” in the direction of the skewness

Negative Skewness (Left tail)	Positive Skewness (Right tail)
 <p>Mean is dragged left</p>	 <p>Mean is dragged right</p>

- *in this case, the mean is a poor measure of central location or does not reflect the center of the sample. Therefore, Median is recommended.*

- **Kurtosis:** is the degree of peakedness of a distribution.
- It is usually taken relative to a normal distribution.
- If a distribution is very peaked than a normal distribution then it is called **Leptokurtic** distribution
- if it is less peaked/flat than the normal curve it is called **Platykurtic**
- if it is moderate (normal) we call it **Mesokurtic.**

These three distributions have the same mean, median, and mode. But Variance (leptokurtic) < Variance (mesokurtic) < Variance (platykurtic)



Example: in a certain data coming from a moderately skewed distribution the mean = 74 and Mode = 60

- **What is the skewness type?**
- **Compute the median?**

Probability and Probability Distributions

Probability

- Chance of observing a particular outcome
- Likelihood of an event
- Assumes a “stochastic” or “random” process: i.e. the outcome is not predetermined - there is an element of chance
- An outcome is a specific result of a single trial of a probability experiment.
- Probability theory developed from the study of games of chance like dice and cards.
- A process like flipping a coin, rolling a die or drawing a card from a deck are **probability experiments**.

Why Probability in Public Health and Medicine?

- Results are not certain
- Because medicine is an inexact science, physicians seldom predict an outcome with absolute certainty.
- **E.g.**, to formulate a diagnosis, a physician must rely on available diagnostic information about a patient
 - History and physical examination
 - Laboratory investigation, X-ray findings, ECG, etc

- Although no test result is absolutely accurate, it does affect the probability of the presence (or absence) of a disease (Sensitivity and specificity)
- understanding of probability is fundamental for quantifying the uncertainty that is inherent in the decision-making process
- Probability theory is a foundation for statistical inference, &
- Allows us to draw conclusions about a population of patients based on information obtained from a sample of patients drawn from that population

When can we talk about probability ?

- When dealing with a process that has an uncertain outcome
- *Experiment* = any process with an uncertain outcome
- When an experiment is performed, one and only one outcome is obtained

- **Event** = something that may happen or not when the experiment is performed
- An event either occurs or it does not occur
- Events are represented by uppercase letters such as *A*, *B*, and *C*

- **Probability of an Event E**

= a number between 0 and 1 representing the proportion of times that event E is expected to happen when the experiment is done over and over again under the same conditions

- Any event can be expressed as a subset of the set of all possible outcomes (S)

S = set of all possible outcomes

$$P(S) = 1$$

Two Categories of Probability

- Objective and Subjective Probabilities.
- *Objective probability*
 - 1) *Classical approach*
 - 2) *Frequentist approach*
 - 3) *Axiomatic approach*

Classical Probability

- uses sample space to determine the numerical probability that an event will happen
- it assumes that all outcomes in the sample space are equally likely to occur
- **Definition:** *If an event can occur in N mutually exclusive and equally likely ways, and if m of these possess a characteristic, E , the probability of the occurrence of $E = m/N$.*

$P(E)$ = the probability of $E = m/N$

$$\text{prob. of any event } E = \frac{\text{number of outcome in } E}{\text{total numbers of outcome in the sample space}}, p(E) = \frac{n(E)}{n(S)}$$

Example: Rolling a die -

- There are 6 possible outcomes: $= \{1, 2, 3, 4, 5, 6\}$.
 - Each is equally likely
- $P(i) = 1/6, i=1,2,\dots,6$.
 - $\text{SUM}(P(i)) = 1$
- what is the probability of 4 coming up?
 - $E=4, m = 1(\text{which is } 4) \text{ and } N = 6$
 - The probability of 4 coming up is
 - $P(4) = \text{the probability of } 4 = 1/6$

Relative Frequency Probability

- **Definition:** If a process is repeated a large number of times (n) under essentially identical conditions, in the long run the estimate will be closer to the true value.
- and if an event with the characteristic E occurs f times, the relative frequency of E ,

$$p(E) = \frac{\text{frequency for the class}}{\text{sum of the frequency in a distribution}} = \frac{f}{n}$$

Example: in a sample of 50 people, 21 are blood type O, 22 are blood type A, 5 are blood type B and 2 are blood type AB. Set up frequency distribution and find the following probabilities.

- a. Person that have blood type O
- b. Person that have blood type A or type B
- c. Person that have neither blood type A nor O
- d. Person not blood type AB.

Solution:

<u>Blood type</u>	<u>frequency</u>
A	22
B	5
AB	2
O	21
<u>Total</u>	<u>50</u>

a. $P(O) = f/n = 21/50 = 0.42$

b. $P(A \text{ or } B) = \frac{22}{50} + \frac{5}{50} = \frac{27}{50} = 0.54$

c. $P(\text{neither A nor O}) = \frac{5}{50} + \frac{2}{50} = \frac{7}{50} = 0.14$

d. $P(\text{not AB}) = 1 - P(AB) = 1 - \frac{2}{50} = \frac{48}{50} = 0.96$

- Since trials cannot be repeated an infinite number of times, theoretical probabilities are often estimated by **empirical probabilities** based on a finite amount of data

- ***Example:***

Of 158 people who attended a dinner party, 99 were ill.

$$***P (Illness) = 99/158 = 0.63 = 63\%.***$$

- In 1998, there were 2,500,000 registered live births; of these, 200,000 were LBW infants.
- Therefore, **$P (LBW) = 200,000/2,500,000 = 0.08$**

Subjective Probability

- uses a probability value based on an educated guess or estimate.
- a person or group makes an educated guess at a chance that an event will occur. This guess is based on the person's experience.

Example: an epidemiologist might say there is an 80% probability that an outbreak will occur in certain area.

- a physician may say that there is a 70% probability that a patient will cure.

- Someone says that he is 95% certain that a cure for AIDS will be discovered within 5 years, then he means that:

$$P(\text{discovery of cure for AIDS within 5 years}) = 95\% = 0.95$$

- ✓ Probabilities can be expressed as fraction, decimal or appropriate percentage.

Example: what is the probability of getting cure from a certain disease?

$\frac{1}{2}$, 0.5, 50%

Mutually Exclusive Events

- Two events A and B are *mutually exclusive* if they cannot both happen at the same time

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

- Example:
 - A coin toss cannot produce heads and tails simultaneously.
 - Weight of an individual can't be classified simultaneously as “underweight”, “normal”, “overweight”

Independent Events

- Two events A and B are *independent* if the probability of the first one happening is the same no matter how the second one turns out. OR. *The outcome of one event has no effect on the occurrence or non-occurrence of the other.*

$$P(A \cap B) = P(A) \times P(B) \text{ (Independent events)}$$

$$P(A \cap B) \neq P(A) \times P(B) \text{ (Dependent events)}$$

Example:

- The outcomes on the first and second coin tosses are independent

Intersection, and union

- The intersection of two events A and B , $A \cap B$, is the event that A and B happen simultaneously

$$P (A \text{ and } B) = P (A \cap B)$$

- Let A represent the event that a randomly selected newborn is LBW, and B the event that he or she is from a multiple birth
- The intersection of A and B is the event that the infant is both LBW and from a multiple birth

- The union of A and B , $A \cup B$, is the event that either A happens or B happens or they both happen simultaneously

$$P (A \text{ or } B) = P (A \cup B)$$

- In the example above, the union of A and B is the event that the newborn is either LBW or from a multiple birth, or both

Properties of Probability

1. The numerical value of a probability always lies between 0 and 1, inclusive.

$$0 \leq P(E) \leq 1$$

- ✓ *A value 0 means the event can not occur*
- ✓ *A value 1 means the event definitely will occur*
- ✓ *A value of 0.5 means that the probability that the event will occur is the same as the probability that it will not occur.*

2. The sum of the probabilities of all mutually exclusive outcomes is equal to 1.

$$P(E_1) + P(E_2) + \dots + P(E_n) = 1.$$

3. For two mutually exclusive events A and B,

$$P(A \text{ or } B) = P(A \cup B) = P(A) + P(B).$$

If not mutually exclusive:

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

4. The **complement** of an event A , denoted by \bar{A} or A^c , is the event that A does not occur

- Consists of all the outcomes in which event A does **NOT** occur

$$P(\bar{A}) = P(\text{not } A) = 1 - P(A)$$

- \bar{A} occurs only when A does not occur.
- These are complementary events.

- In the LBW example, the complement of A is the event that a newborn is not LBW
- In other words, A is the event that the child weighs 2500 grams at birth

$$P(\bar{A}) = 1 - P(A)$$

$$\begin{aligned} P(\text{not LBW}) &= 1 - P(\text{LBW}) \\ &= 1 - 0.076 \\ &= 0.924 \end{aligned}$$

Basic Probability Rules

1. Addition rule

- If events *A* and *B* are mutually exclusive:

$$P(A \text{ or } B) = P(A) + P(B)$$

$$P(A \text{ and } B) = 0$$

- More generally:

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

P(event A or event B occurs or they both occur)

Example: The probabilities below represent years of schooling completed by mothers of newborn infants

Mother's education	Probability
≤ 8 years	0.056
9 to 11 years	0.159
12 years	0.321
13 to 15 years	0.218
≥ 16 years	0.230
Not reported	0.016

- What is the probability that a mother has completed < 12 years of schooling?

$$P(\leq 8 \text{ years}) = 0.056 \text{ and}$$

$$P(9-11 \text{ years}) = 0.159$$

- Since these two events are mutually exclusive,

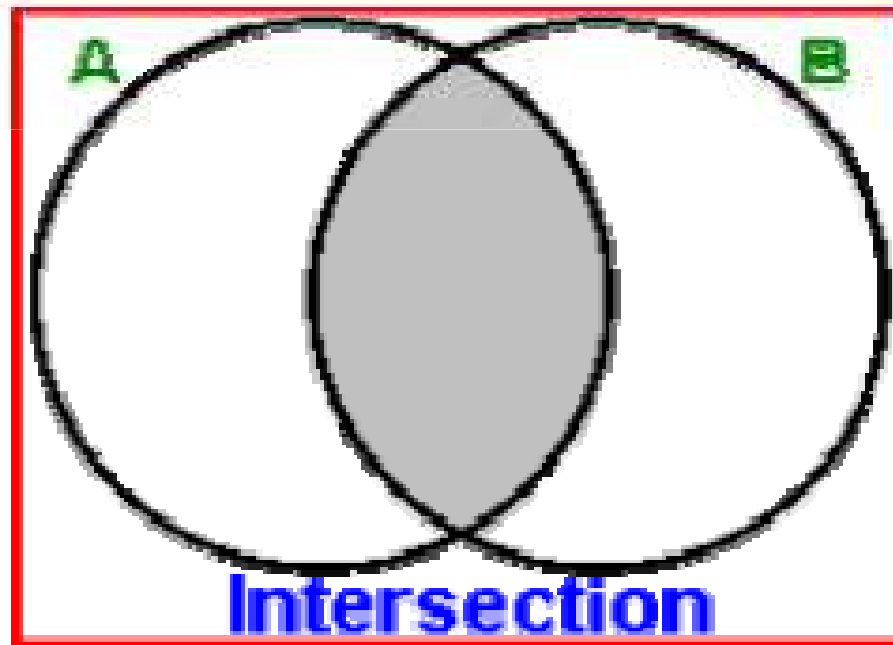
$$\begin{aligned} P(\leq 8 \text{ or } 9-11) &= P(\leq 8 \cup 9-11) \\ &= P(\leq 8) + P(9-11) \\ &= 0.056 + 0.159 \\ &= 0.215 \end{aligned}$$

- What is the probability that a mother has completed 12 or more years of schooling?

$$\begin{aligned}P(\geq 12) &= P(12 \text{ or } 13-15 \text{ or } \geq 16) \\&= P(12 \cup 13-15 \cup \geq 16) \\&= P(12) + P(13-15) + P(\geq 16) \\&= 0.321 + 0.218 + 0.230 \\&= 0.769\end{aligned}$$

If A and B are not mutually exclusive events,
then subtract the overlapping:

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



- **Example:** The following data are the results of electrocardiograms (ECGs) and radionuclide angiocardiograms (RAs) for 19 patients with post-traumatic myocardial confusions.
 - 7 patients developed both ECG and RA abnormality
 - 17 patients developed ECG abnormal
 - 9 patients developed RA abnormal

$$P(\text{ECG abnormal and RA abnormal}) = 7/19 = 0.37$$

$$P(\text{ECG abnormal or RA abnormal}) =$$

$$P(\text{ECG abnormal}) + P(\text{RA abnormal}) - P(\text{Both ECG and RA abnormal}) =$$

$$17/19 + 9/19 - 7/19 = 19/19 = 1.$$

Note: The problem is that the 7 patients whose ECGs and RAs are both abnormal are counted twice

2. Multiplication rule

– If A and B are independent events, then

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

– If dependent,

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

$P(A \text{ and } B)$ denotes the probability that **A** and **B** both occur at the same time.

Conditional Probability

- Refers to the probability of an event, given that another event is known to have occurred.
- “What happened first is assumed”
- The *conditional probability* that event B has occurred given that event A has already occurred is denoted $P(B|A)$ and is defined

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

- provided that $P(A) \neq 0$.

Example:

A study investigating the effect of prolonged exposure to bright light on retina damage in premature infants.

	Retinopathy YES	Retinopathy NO	TOTAL
Bright light	18	3	21
Reduced light	21	18	39
TOTAL	39	21	60

- The probability of developing retinopathy is:

$$\begin{aligned} P(\text{Retinopathy}) &= \frac{\text{No. of infants with retinopathy}}{\text{Total No. of infants}} \\ &= (18+21)/(21+39) \\ &= 0.65 \end{aligned}$$

- We want to compare the *probability of retinopathy*, given that the infant was *exposed to bright light*, with that the infant was *exposed to reduced light*.
- Exposure to bright light and exposure to reduced light are *conditioning events*, events we want to take into account when calculating *conditional probabilities*.

- The conditional probability of retinopathy, given exposure to bright light, is:
- $P(\text{Retinopathy/exposure to bright light}) = \frac{\text{No. of infants with retinopathy exposed to bright light}}{\text{No. of infants exposed to bright light}}$
 $= (18/60)/(21/60) = 18/21 = 0.86$

- $P(\text{Retinopathy/exposure to reduced light}) =$

$$\frac{\text{\# of infants with retinopathy exposed to reduced light}}{\text{No. of infants exposed to reduced light}}$$

$$= (21/60)/(39/60) = 21/39 = 0.54$$

- *The conditional probabilities suggest that premature infants exposed to bright light have a higher risk of retinopathy than premature infants exposed to reduced light.*

- For **independent events** A and B

$$P(A/B) = P(A).$$

- For **non-independent events** A and B

$$P(A \text{ and } B) = P(A/B) P(B)$$

(General Multiplication Rule)

Test for Independence

- Two events A and B are *independent* if:
- Two events A and B are *dependent* if

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

or

$$P(A \text{ and } B) = P(A) \cdot P(B)$$

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

or

$$P(A \text{ and } B) \neq P(A) \cdot P(B)$$

Example

- In a study of optic-nerve degeneration in Alzheimer's disease, postmortem examinations were conducted on 10 Alzheimer's patients. The following table shows the distribution of these patients according to sex and evidence of optic-nerve degeneration.
- Are the events “patients has optic-nerve degeneration” and “patient is female” independent for this sample of 10 patients?

Sex	Optic-nerve Degeneration	
	Present	Not Present
Female	4	1
Male	4	1

Solution

- $P(\text{Optic-nerve degeneration}/\text{Female}) =$

$$\frac{\text{No. of females with optic-nerve degeneration}}{\text{No. of females}}$$

$$= (4/10)/(5/10) = 4/5 = 0.80$$

$$P(\text{Optic-nerve degeneration}) = \frac{\text{No of patients with optic-nerve degeneration}}{\text{Total No. of patients}}$$

$$= 8/10 = 0.80$$

☞ The events are independent for this sample.

Exercise:

Culture and Gonodectin (GD) test results for
240 Urethral Discharge Specimens

GD Test Result	Culture Result		Total
	Gonorrhea	No Gonorrhea	
Positive	175	9	184
Negative	8	48	56
Total	183	57	240

1. What is the probability that a man has gonorrhea?
2. What is the probability that a man has a positive GD test?
3. What is the probability that a man has a positive GD test and gonorrhea?
4. What is the probability that a man has a negative GD test and does not have gonorrhea
5. What is the probability that a man with gonorrhea has a positive GD test?

6. What is the probability that a man does not have gonorrhea has a negative GD test?
7. What is the probability that a man does not have gonorrhea has a positive GD test?
8. What is the probability that a man with positive GD test has gonorrhea?

Probability Distributions

- **A probability distribution:** consists of a values a random variable can assume and the corresponding probabilities of the values.
- **Random Variable:** is a variable that takes a possible outcome and assigns a number to it, usually denoted by capital letters (X, Y, Z...).
- **Random variables** can be either discrete or continuous

- **A discrete random variable** are variables which can assume only a specific number of values. They have values that can be counted

Examples:

- Number of children in a family.
- Number of car accidents per week.
- Number of defective items in a laboratory store.
- Patient treatment status (Cure or Not cure).
- Death status (Dead or Alive)

- **A continuous random variable** are variables that can assume all values between any two give values. Are obtained from data that can be measured rather than counted

Examples:

- Height of patients at certain clinic.
- Mark of a student.
- Age at diagnosis.
- Length of hospital stay.

Examples: Flip a coin three times, let X be the number of heads in three tosses.

$S = \{ (HHH), (HHT), (HTH), (HTT), (THH), (THT), (TTH), (TTT) \}$

$$X(HHH) = 3$$

$$X(HHT) = X(HTH) = X(THH) = 2$$

$$X(THT) = X(TTH) = X(HTT) = 1$$

$$X(TTT) = 0$$

$$X = \{0, 1, 2, 3\}$$

X is random variable for an outcome Head

Example: Consider the experiment of tossing a coin three times. Let X be the number of heads. Construct the probability distribution of X .

Solution:

- First identify the possible value that X can assume.
- Calculate the probability of each possible distinct value of X and express X in the form of frequency distribution. We can also represent graphically.

$X=x$	0	1	2	3
$P(X=x)$	1/8	3/8	3/8	1/8

- Probability distribution denoted by P for discrete and f for continuous random variable.

Properties of probability distribution :

1. $0 \leq P(X) \leq 1$, if X is discrete

$0 \leq f(X) \leq 1$, if X is continuous

2. $\sum P(X=x) = 1$, If X is discrete

$\int f(x) dx = 1$, if X is continuous

Discrete Probability Distributions

- For a discrete random variable, the probability distribution specifies each of the possible outcomes of the random variable along with the probability that each will occur
- We represent a potential outcome of the random variable X by x

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

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

The following data shows the number of diagnostic services a patient receives. Is it a probability distribution?? Why??

x	$P(X = x)$
0	0.671
1	0.229
2	0.053
3	0.031
4	0.010
5	0.006

- What is the probability that a patient receives exactly 3 diagnostic services?

$$P(X=3) = 0.031$$

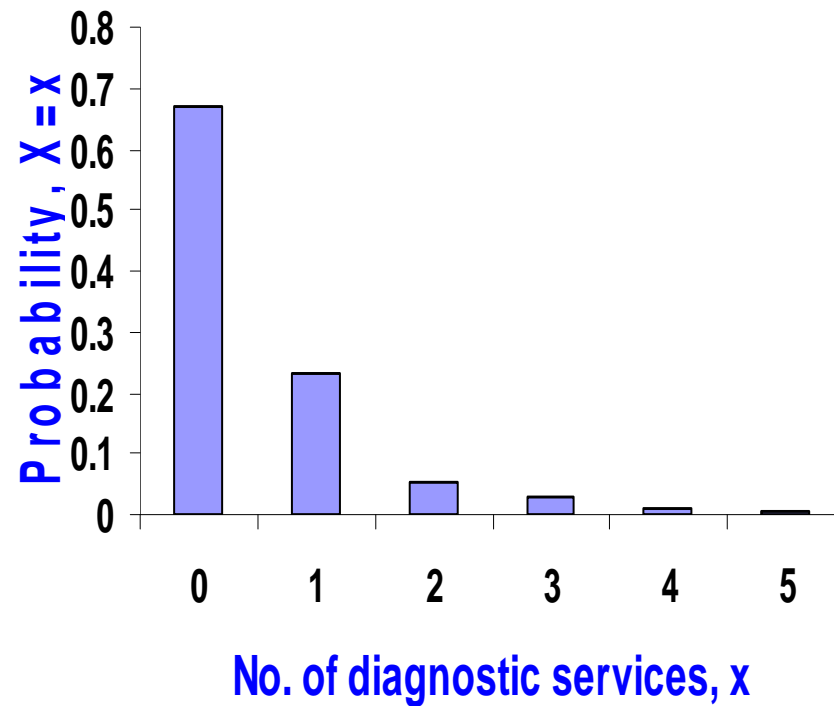
- What is the probability that a patient receives at most one diagnostic service?

$$\begin{aligned} P(X \leq 1) &= P(X = 0) + P(X = 1) \\ &= 0.671 + 0.229 \\ &= 0.900 \end{aligned}$$

- What is the probability that a patient receives at least four diagnostic services?

$$\begin{aligned}P(X \geq 4) &= P(X = 4) + P(X = 5) \\&= 0.010 + 0.006 \\&= 0.016\end{aligned}$$

Probability distributions can also be displayed using a graph



The Expected Value of a Discrete Random variable

- If a random variable is able to take on a large number of values, then a probability mass function might not be the most useful way to summarize its behavior
- Instead, measures of location and dispersion can be calculated

- The average value assumed by a random variable is called its **expected value**, or the **population mean**
- It is represented by $E(X)$ or μ
- *To obtain the expected value of a discrete random variable X , we multiply each possible outcome by its associated probability and sum all values with a probability greater than 0*

- For the diagnostic service data:

$$\begin{aligned}\text{Mean } (X) = E(X) &= 0(0.671) + 1(0.229) \\ &+ 2(0.053) \\ &+ 3(0.031) + 4(0.010) + 5(0.006) \\ &= 0.498 \approx 0.5\end{aligned}$$

- We would expect an average of 0.5 services for each visit

The Variance of a Discrete Random Variable

- The variance of a random variable X is called the **population variance** and is represented by $\text{Var}(X)$ or σ^2
- It quantifies the dispersion of the possible outcomes of X around the expected value μ

$$\begin{aligned}
\sigma^2 &= \sum (\mathbf{x_i} - \mu)^2 P(X = \mathbf{x_i}) = \sum (\mathbf{x_i} - E(X))^2 P(X = \mathbf{x_i}) \\
&= (0 - 0.5)^2(0.671) + (1 - 0.5)^2(0.229) \\
&\quad + (2 - 0.5)^2(0.053) + (3 - 0.5)^2(0.031) \\
&\quad + (4 - 0.5)^2(0.010) + (5 - 0.5)^2(0.006) \\
&= 0.782
\end{aligned}$$

$$\text{Standard deviation} = \sigma = \sqrt{0.782} = 0.884$$

- The most common discrete probability distributions are the

- Binomial distribution

- Poisson distribution.

Binomial Distribution

- It is one of the most widely encountered discrete probability distributions.
- Consider dichotomous (binary) random variable
- Is based on Bernoulli trial
 - When a single trial of an experiment can result in only **one of two mutually exclusive outcomes** (success or failure; dead or alive; sick or well, male or female)

Example:

- We are interested in determining whether a newborn infant will survive until his/her 70th birthday
- Let Y represent the survival status of the child at age 70 years
- $Y = 1$ if the child survives and $Y = 0$ if he/she does not

- The outcomes are mutually exclusive and exhaustive
- Suppose that 72% of infants born survive to age 70 years

$$P(Y = 1) = p = 0.72$$

$$P(Y = 0) = 1 - p = 0.28$$

The probability distribution of Y is

y	$P(Y = y)$
0	0.28
1	0.72

A binomial probability distribution occurs when the following requirements are met.

1. The procedure has a n identical trials.
2. The trials must be independent.
3. Each trial must have two possible outcomes.
4. The probability of success and failure must remain constant for each trial
 $[P(\text{success}) = p]$ and $[P(\text{failure}) = 1-p=q]$

- For a binomial experiment, the probability that outcome occurs exactly x times is:

- $P(X=x) = \binom{n}{x} P^x (1-P)^{(n-x)}$, $x = 0, 1, 2, \dots, n$.

$$= \frac{n!}{x!(n-x)!} P^x (1-P)^{(n-x)}$$

- ***n*** denotes the number of trials
- ***x*** denotes the number of successes in the *n* trials
- ***p*** denotes the probability of success
- ***q*** denotes the probability of failure (1- ***p***)

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

- Represents the number of ways of selecting *x* objects out of *n* where the order of selection does not matter.
- where $n! = n(n-1)(n-2)\dots(1)$, and $0! = 1! = 1$

Example:

- *Suppose we know that 40% of a certain population are cigarette smokers. If we take a random sample of 10 people from this population, what is the probability that we will have exactly 4 smokers in our sample?*

- If the probability that any individual in the population is a smoker to be $P=.40$, then the probability that $x=4$ smokers out of $n=10$ subjects selected is:

$$\begin{aligned}P(X=4) &= 10C4(0.4)^4(1-0.4)^{10-4} \\&= 10C4(0.4)^4(0.6)^6 = \\210(.0256)(.04666) \\&= 0.25\end{aligned}$$

- The probability of obtaining exactly 4 smokers in the sample is about 0.25.

Exercise

Each child born to a particular set of parents has a probability of 0.25 of having blood type O. If these parents have 5 children.

What is the probability that

- a. Exactly two of them have blood type O
- b. At most 2 have blood type O
- c. At least 4 have blood type O
- d. 2 do not have blood type O.

The Mean and Variance of a Binomial Distribution

- Once n and p are specified, we can compute the proportion of success,

$$p = x/n$$

- and the mean and variance of the distribution are given by :

$$E(X) = \mu = np, \sigma^2 = npq, \sigma = \sqrt{npq}$$

Example:

- 70% of a certain population has been immunized for polio. If a sample of size 50 is taken, what is the “expected total number”, in the sample who have been immunized?

$$\mu = np = 50(.70) = 35$$

- This tells us that “on the average” we expect to see 35 immunized subjects in a sample of 50 from this population.
- **Compute the standard deviation??**

The Poisson Distribution

- Is a **discrete probability distribution** used to model the number of occurrences of an event that takes place infrequently in **time** or **space**
- Applicable for **counts** of events over a given interval of time, for example:
 - number of patients arriving at an emergency department in a day
 - number of new cases of HIV diagnosed at a clinic in a month

- In such cases, we take a sample of days and observe the number of patients arriving at the emergency department on each day,
- or a sample of months and observe the number of new cases of HIV diagnosed at the clinic.
- *We are observing a count or number of events, rather than a yes/no or success/failure outcome for each subject or trial, as in the binomial.*

- *In theory, a random variable X is a count that can assume any integer value greater than or equal to 0*

- Suppose events happen *randomly* and *independently* in time at a constant rate. If events happen with rate λ events per unit time, the probability of x events happening in unit time is:

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

- where $x = 0, 1, 2, \dots, \infty$
- x is a potential outcome of X
- The constant λ (lambda) represents the rate at which the event occurs, or the expected number of events per unit time
- $e = 2.71828$

Example

- *The daily number of new registrations of cancer is 2.2 on average.*

What is the probability of

- a) Getting no new cases
- b) Getting 1 case
- c) Getting 2 cases
- d) Getting 3 cases
- e) Getting 4 cases

Solutions

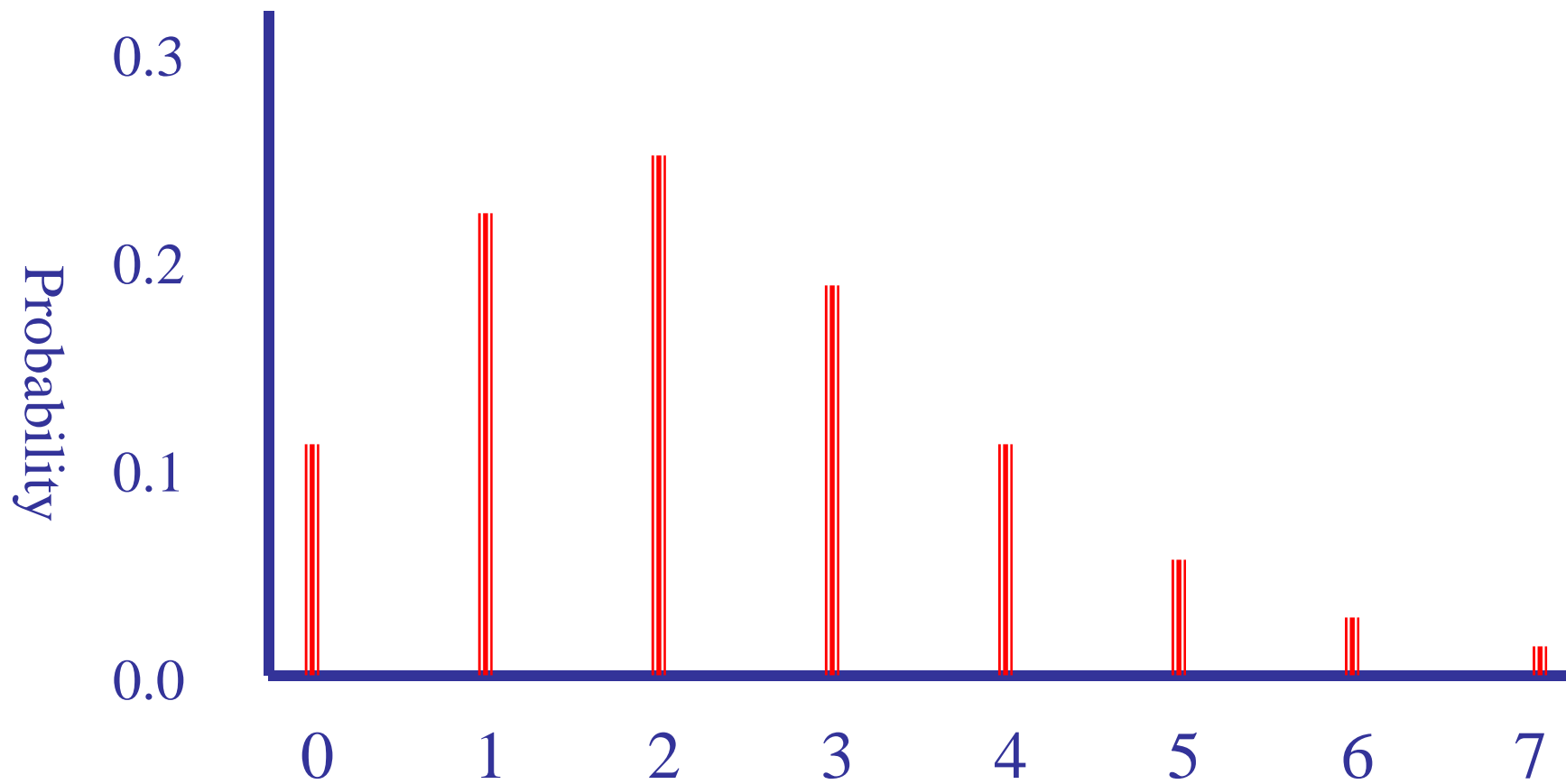
$$\text{a) } P(X = 0) = \frac{(2.2)^0 e^{-2.2}}{0!} = 0.111$$

$$\text{b) } P(X=1) = 0.244$$

$$\text{c) } P(X=2) = 0.268$$

$$\text{d) } P(X=3) = 0.197$$

$$\text{e) } P(X=4) = 0.108$$



Poisson distribution with mean 2.2

Characteristics

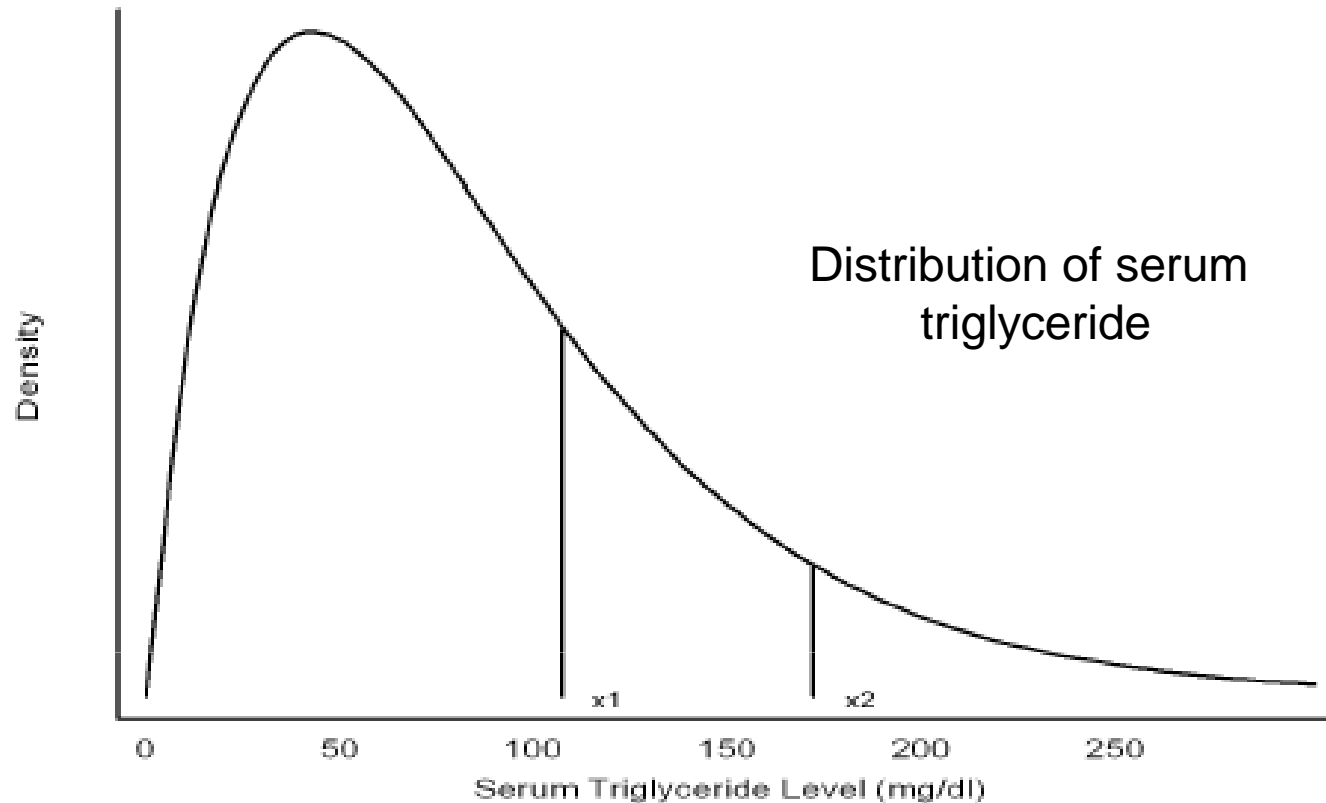
- It has no theoretical maximum value, but the probabilities tail off towards zero very quickly
- λ is the parameter of the Poisson distribution
- The mean is λ and the variance is also λ .

Exercise:

- In a given geographical area, cases of tetanus are reported at a rate of $\lambda = 4.5/\text{month}$
- What is the probability that 0 cases of tetanus will be reported in a given month?
- What is the probability at least 1 cases of tetanus will be reported in a given month?
- Find the mean and variance of cases of tetanus will be reported in a given month

Continuous Probability Distributions

- A continuous random variable X can take on any value in a specified interval or range
- The probability distribution of X is represented by a smooth curve called a **probability density function**



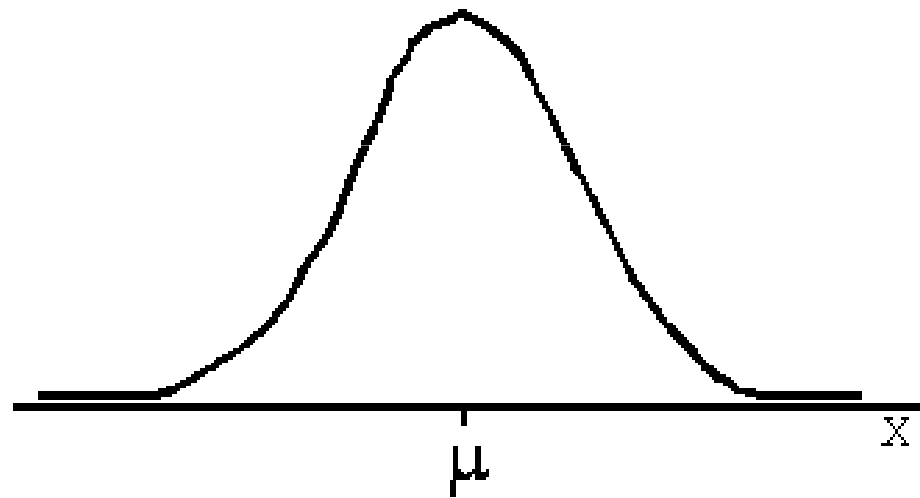
- The area under the smooth curve is equal to 1
- The area under the curve between any two points x_1 and x_2 is the probability that X takes a value between x_1 and x_2

- Instead of assigning probabilities to specific outcomes of the random variable X , probabilities are assigned to ranges of values
- The probability associated with any one particular value is equal to 0
- Therefore, $P(X=x) = 0$
- Also, $P(X \geq x) = P(X > x)$

The Normal distribution

- The ND is the most important probability distribution in statistics
- Frequently called the “Gaussian distribution” or **bell-shape** curve.
- Variables such as blood pressure, weight, height, serum cholesterol level, and IQ score — are approximately normally distributed

A random variable is said to have a normal distribution if it has a probability distribution that is symmetric and *bell-shaped*



- The ND is vital to statistical work, most estimation procedures and hypothesis tests underlie ND
- The concept of “probability of $X=x$ ” in the discrete probability distribution is replaced by the “probability density function $f(x)$ ”
- The ND is also an approximating distribution to other distributions (e.g., binomial)

- A random variable X is said to follow ND, if and only if, its probability density function is:

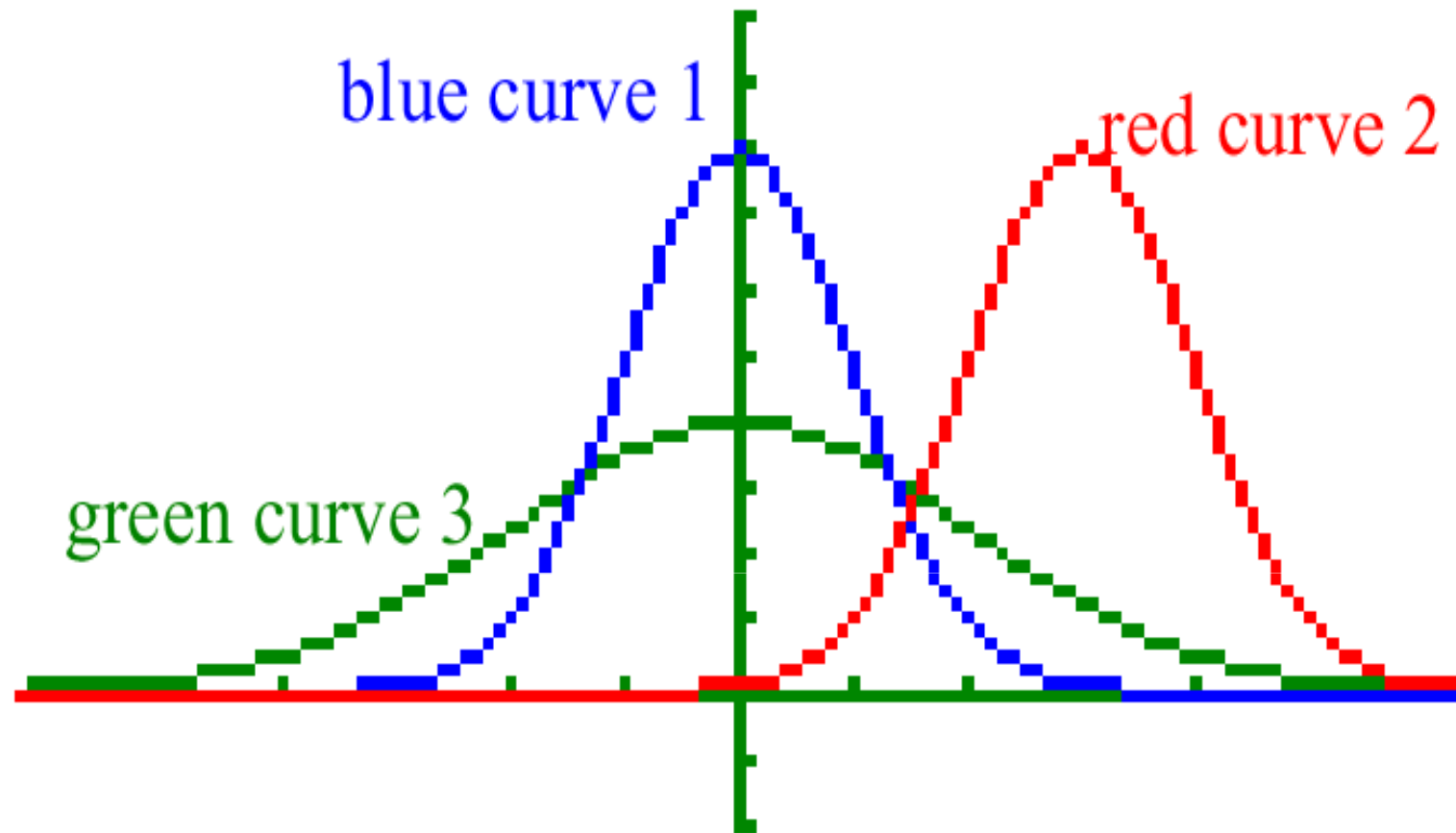
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty < x < \infty.$$

- ✓ π (pi) = 3.14159
- ✓ $e = 2.71828$, x = Value of X
- ✓ Range of possible values of X : $-\infty$ to $+\infty$
- ✓ μ = Expected value of X (“average”)
- ✓ σ^2 = Variance of X .
- ✓ μ and σ are the parameters of the normal distribution — they completely define its shape

The normal distribution with mean μ and variance σ^2 is represented by $N(\mu, \sigma^2)$

To find $P(X \leq x)$, we would have to draw the probability density function of $N(\mu, \sigma^2)$ and determine the area to the left of x

1. **The mean μ tells you about location -**
 - Increase μ - Location shifts right
 - Decrease μ – Location shifts left
 - Shape is unchanged
2. **The variance σ^2 tells you about narrowness or flatness of the bell -**
 - Increase σ^2 - Bell flattens. Extreme values are more likely
 - Decrease σ^2 - Bell narrows. Extreme values are less likely
 - Location is unchanged



$$\mu_1 = \mu_3 < \mu_2 \quad \text{but} \quad \sigma_1 = \sigma_2 < \sigma_3$$

Properties of the Normal Distribution

1. It is symmetrical about its mean, μ .
2. The mean, the median and mode are almost equal. It is unimodal.
3. The total area under the curve about the x-axis is 1 square unit.
4. The curve never touches the x-axis.
5. As the value of σ increases, the curve becomes more and more flat and vice versa.

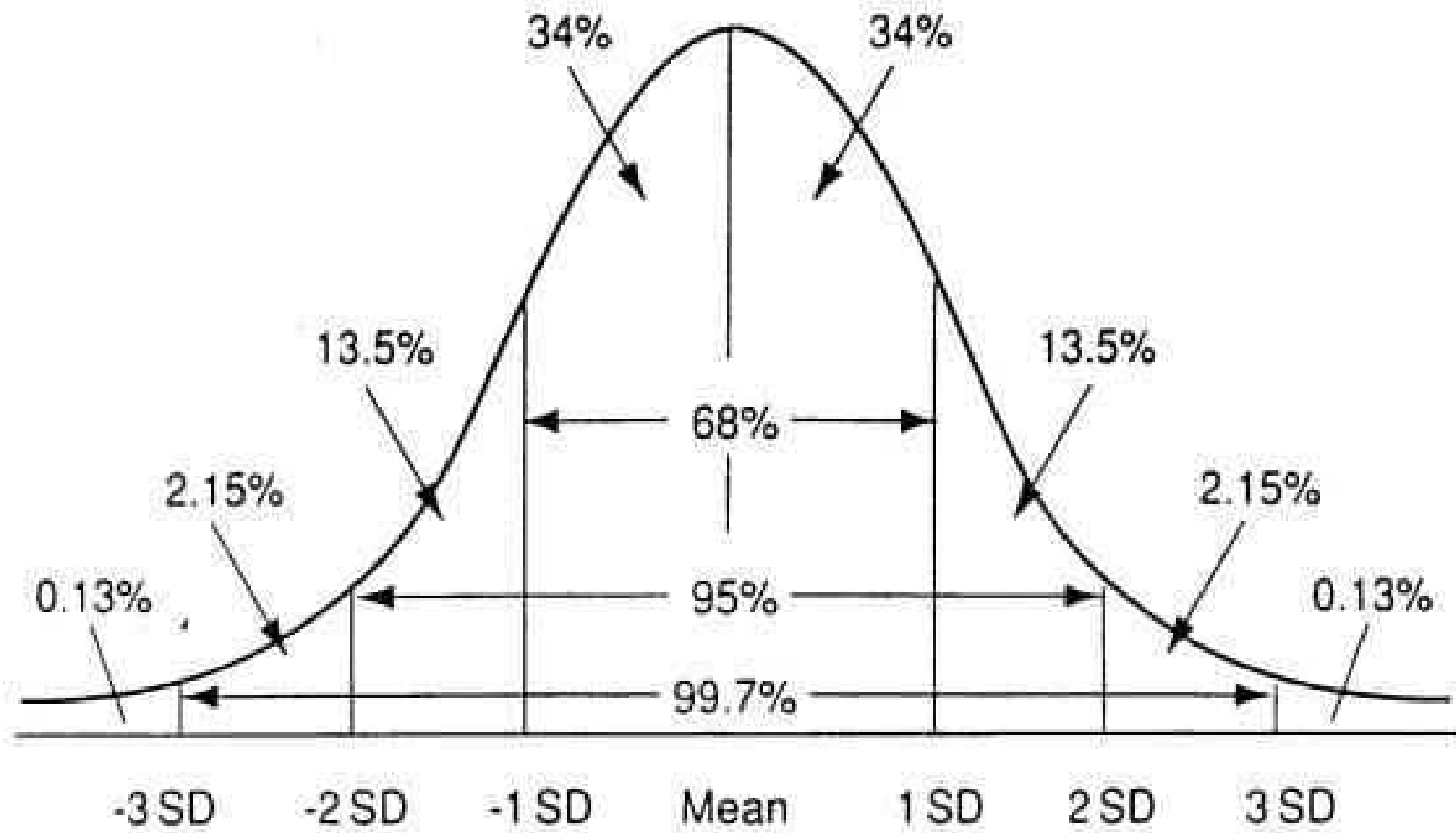
6. Perpendiculars of:

$\mu \pm \text{SD}$ contain about 68%;

$\mu \pm 2 \text{ SD}$ contain about 95%;

$\mu \pm 3 \text{ SD}$ contain about 99.7%
of the area under the curve.

7. The distribution is completely determined
by the parameters μ and σ .



- We have different normal distributions depending on the values of μ and σ^2 .
- We cannot tabulate every possible distribution
- Tabulated normal probability calculations are available only for the ND with $\mu = 0$ and $\sigma^2=1$.

Standard Normal Distribution

- It is a normal distribution that has a mean equal to 0 and a SD equal to 1, and is denoted by $N(0, 1)$.
- The main idea is to standardize all the data that is given by using Z-scores.
- These Z-scores can then be used to find the area (and thus the probability) under the normal curve.

Z - Transformation

- If a random variable $X \sim N(\mu, \sigma)$ then we can transform it to a SND with the help of Z-transformation

$$Z = \frac{x - \mu}{\sigma}$$

- Z represents the Z-score for a given x value

Some Useful Tips

$$\Pr [Z \leq -z] = \Pr [Z \geq +z]$$

$$\Pr [Z < z] + \Pr [Z \geq +z] = 1$$

$$\Pr [Z \geq 0] = .5$$

$$\Pr [Z \leq 0] = .5$$

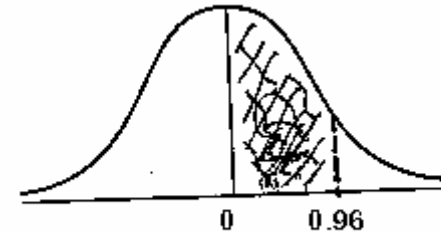
Examples:

1. Find the area under the standard normal distribution which lies

a) Between $Z=0$ and $Z=0.96$

Solution:

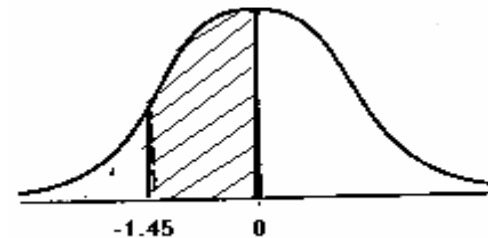
$$\text{Area} = P(0 < Z < 0.96) = 0.3315$$



b) between $Z = -1.45$ and $Z = 0$

Solution:

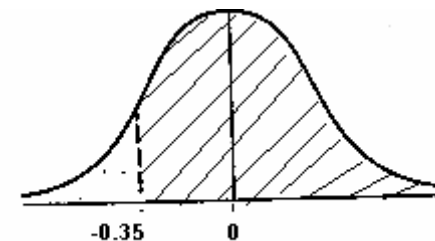
$$\begin{aligned}\text{Area} &= P(-1.45 < Z < 0) \\ &= P(0 < Z < 1.45) \\ &= 0.4265\end{aligned}$$



c) To the right of $Z = -0.35$

Solution:

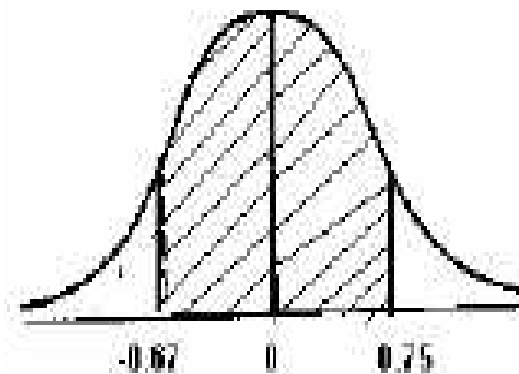
$$\begin{aligned}\text{Area} &= P(Z > -0.35) = P(-0.35 < Z < 0) + P(Z > 0) \\ &= P(0 < Z < 0.35) + P(Z > 0) \\ &= 0.1368 + 0.50 = 0.6368\end{aligned}$$



e) Between $Z = -0.67$ and $Z = 0.75$

Solution:

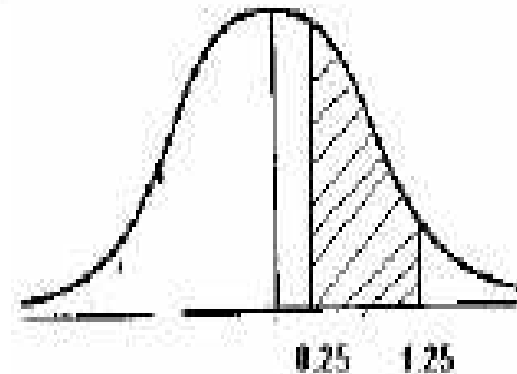
$$\begin{aligned} \text{Area} &= P(-0.67 < Z < 0.75) \\ &= P(-0.67 < Z < 0) + P(0 < Z < 0.75) \\ &= P(0 < Z < 0.67) + P(0 < Z < 0.75) \\ &= 0.2486 + 0.2734 = 0.5220 \end{aligned}$$



f) Between $Z = 0.25$ and $Z = 1.25$

Solution:

$$\begin{aligned} \text{Area} &= P(0.25 < Z < 1.25) \\ &= P(0 < Z < 1.25) - P(0 < Z < 0.25) \\ &= 0.3934 - 0.0987 = 0.2957 \end{aligned}$$



2. Find the value of Z if

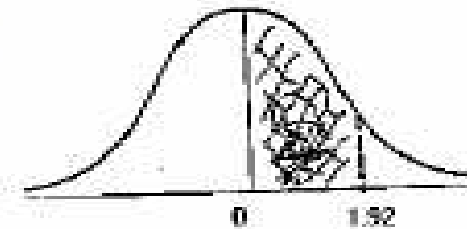
a) The normal curve area between 0 and z(positive) is 0.4726

Solution

$P(0 < Z < z) = 0.4726$ and from table

$$P(0 < Z < 1.92) = 0.4726$$

$$\Leftrightarrow z = 1.92 \dots \text{uniqueness of Area.}$$



b) The area to the left of z is 0.9868

Solution

$$P(Z < z) = 0.9868$$

$$= P(Z < 0) + P(0 < Z < z)$$

$$= 0.50 + P(0 < Z < z)$$

$$\Rightarrow P(0 < Z < z) = 0.9868 - 0.50 = 0.4868$$

and from table

$$P(0 < Z < 2.2) = 0.4868$$

$$\Leftrightarrow z = 2.2$$

Exercise

1. Compute $P(-1 \leq Z \leq 1.5)$

Ans: 0.7745

2. Find the area under the SND from 0 to 1.45

Ans: 0.4265

3. Compute $P(-1.66 < Z < 2.85)$

Ans: 0.9493

Applications of the Normal Distribution

- The ND is used as a model to study many different variables.
- The ND can be used to answer probability questions about continuous random variables.
- Following the model of the ND, a given value of x must be converted to a z score before it can be looked up in the z table.

Example:

- The diastolic blood pressures of males 35–44 years of age are normally distributed with $\mu = 80$ mm Hg and $\sigma^2 = 144$ mm Hg²
 $\sigma = 12$ mm Hg
- Therefore, a DBP of $80+12 = 92$ mm Hg lies 1 SD above the mean
- Let individuals with BP above 95 mm Hg are considered to be hypertensive

- a. What is the probability that a randomly selected male has a BP above 95 mm Hg?

$$\begin{aligned}P(X > 95) &= P\left(\frac{X - 80}{12} > \frac{95 - 80}{12}\right) \\&= P(Z > 1.25) \\&= 0.1056\end{aligned}$$

- Approximately 10.6% of this population would be classified as hypertensive

b. What is the probability that a randomly selected male has a DBP above 110 mm Hg?

$$Z = \frac{110 - 80}{12} = 2.50$$

$$P(Z > 2.50) = 0.0062$$

- Approximately 0.6% of the population has a DBP above 110 mm Hg

c. What is the probability that a randomly selected male has a DBP below 60 mm Hg?

$$Z = \frac{60 - 80}{12} = -1.67$$

$$P(Z < -1.67) = 0.0475$$

- Approximately 4.8% of the population has a DBP below 60 mm Hg

d. What value of DBP cuts off the upper 5% of this population?

- Looking at the table, the value $Z = 1.645$ cuts off an area of 0.05 in the upper tail
- We want the value of X that corresponds to $Z = 1.645$

$$Z = \frac{X - \mu}{\sigma}$$
$$1.645 = \frac{X - \mu}{\sigma}, \quad X = 99.7$$

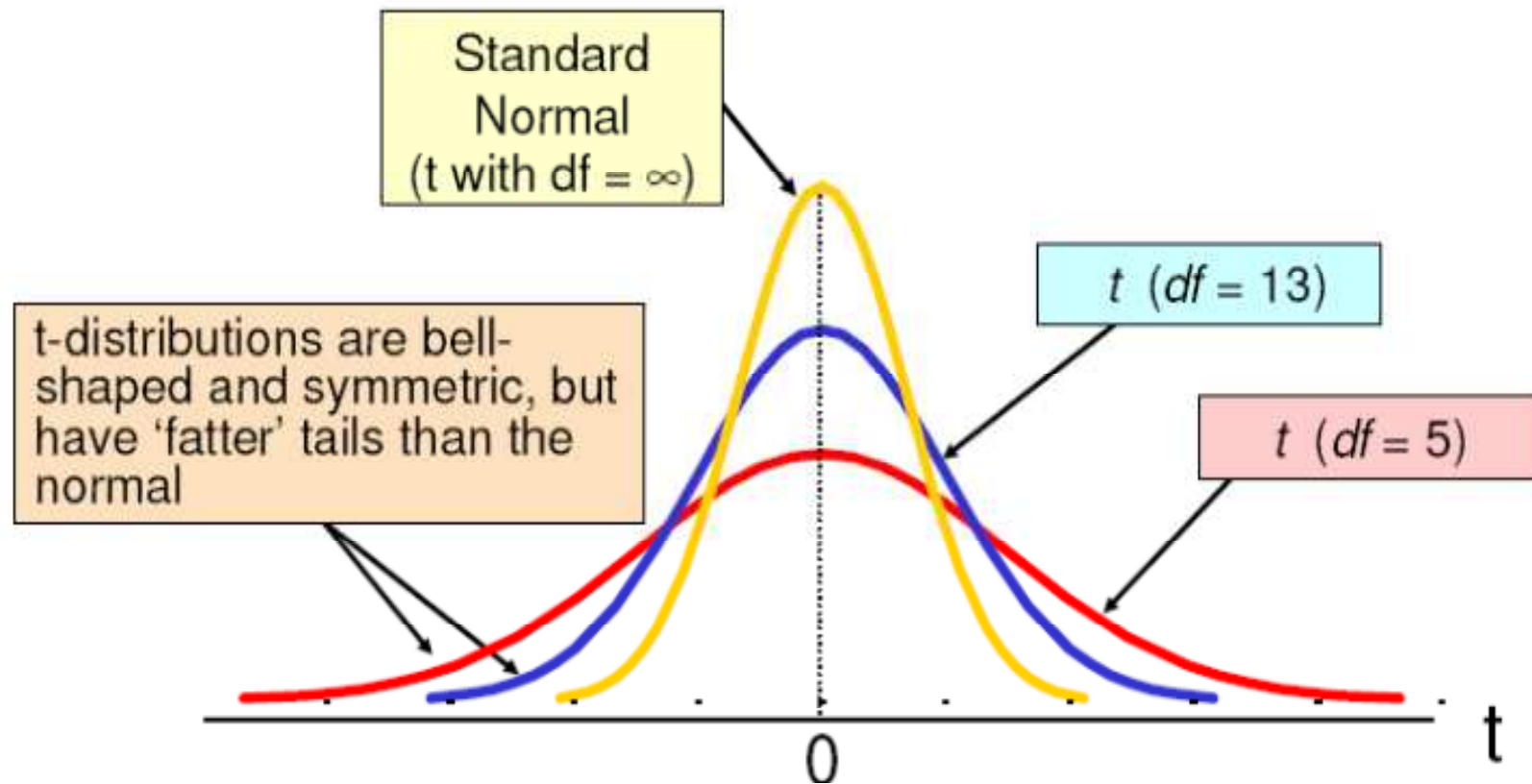
- Approximately 5% of the men in this population have a DBP greater than 99.7 mm Hg

Exercise:

- The blood glucose level of a healthy individuals is normally distributed with $\mu = 80$ mg/dL and $\sigma = 4.8$ mg/dL
- What is the probability that a randomly selected individual has blood glucose level above 130 mg/dL?
- What is the probability that a randomly selected individual has blood glucose level below 70 mg/dL?
- What is the probability that a randomly selected individual has blood glucose level between 70 and 130 mg/dL?

Student's t Distribution

- Bell Shaped
- Symmetric about zero (the mean)
- Flatter than the Normal $(0,1)$. This means
 - The variability of a t is greater than that of a Z that is normal $(0,1)$
 - Thus, there is more area under the tails and less at center
 - Because variability is greater, resulting confidence intervals will be wider.



- Note: t approaches z as n increases

Sample Size and Sampling

❖ Focus

- ✓ To answer the following three questions
 - ❖ What is the group of people from which we want to draw the sample?
 - ❖ How many people do we need in our sample?
 - ❖ How will this people be selected?

The Group of People

❖ Population

- ✓ Is a complete set of elements that possess some common characteristic
- ✓ Should be clearly and explicitly defined in terms of place, time, and other relevant criteria

❖ Target Population

- ✓ The entire group of people or objects to which the researcher wishes to generalize the study findings
 - ❖ All people with AIDS

❖ Source Population

- ✓ A subset of the target population
- ✓ May be limited to region, state, city, or institution
 - ❖ All people with AIDS in A.A.

❖ Study Population

- ✓ A subset of the source population
- ✓ Population from which the sample actually was drawn and about which a conclusion can be made.
 - ❖ All people with AIDS in A.A. satisfying the inclusion criteria

Sample Size

(How many?)

- **Sample Size:** *The number of study subjects selected to represent a given study population.*

Sample Size Determination

- ✓ In planning of any investigation we must decide how many people needed to be studied in order to answer the research questions
- ✓ If the sample is too small we may fail to detect important effects, or may estimate effects too imprecisely
- ✓ If the study is too large then we will waste resources, but we can increase precision, we may also detect unimportant effect/ difference
- ✓ In general, it is much better to increase the accuracy of data collection than to increase the sample size after a certain point
- ✓ Too small is the worst
- ✓ The feasible sample size is determined by the availability of resources to collect information and to analyze it

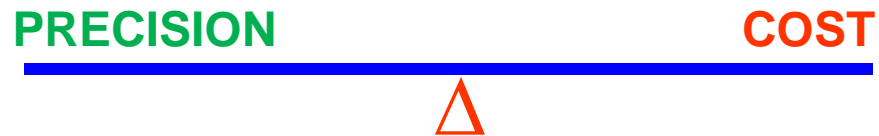
Sample size determination depends on the:

- Objective of the study
- Design of the study
 - Descriptive/Analytic
- Degree of precision required for generalization
- Clinically important differences to detect
- Degree of confidence with which to conclude
- Power

Common question:

- *“How many subjects should I study?”*
- Too small sample = Waste of time and resources
= Results have no practical use
- Too large sample = Waste of resources
= Precision will increase

When deciding on sample size:



↑ Sample size = ↑ Precision = ↑ Cost

Points for Consideration

1. Sample size estimates might need to be adjusted to compensate for non-response rate, patient dropout or loss to follow-up, etc.
2. If sampling is from a finite population of size $N(\leq 10,000)$, then:

$$n = \frac{n_0}{\left(1 + \frac{n_0}{N}\right)}$$

where n_0 is the sample from an infinite population. When N is large in comparison to n , (i.e., $n/N \leq 0.05$), the finite population correction may be ignored.

3. Design effect for complex sampling designs. Common values: multiply n by 2, 3, ...5.

Sample size for single sample

- A. Sample size for estimating a single population mean
- B. Sample size to estimate a single population proportion

A. Sample size for estimating a single population mean

$$n = \frac{(z_{\alpha/2})^2 \cdot \sigma^2}{d^2}$$

Example:

Suppose that for a certain group of cancer patients, we are interested in estimating the mean age at diagnosis. We would like a 95% CI of 5 years wide. If the population SD is 12 years, how large should our sample be?

$$n = \frac{(z_{\alpha/2})^2 \cdot \sigma^2}{d^2} = \frac{(1.96)^2 (144)}{(2.5)^2} = 88.5 \approx 89$$

Final sample size???

But the population σ^2 is most of the time unknown

As a result, it has to be estimated from:

- *Pilot or preliminary sample:*
 - Select a pilot sample and estimate σ^2 with the sample variance, s^2
- *Previous or similar studies*

B. Sample size to estimate a single population proportion

$$n = \frac{(z_{\alpha/2})^2 \cdot pq}{d^2}$$

Example:

Suppose that you are interested to know the proportion of infants who breastfed >18 months of age in a rural area. Suppose that in a similar area, the proportion (p) of breastfed infants was found to be 0.20. What sample size is required to estimate the true proportion within $\pm 3\%$ points with 95% confidence. Let $p=0.20$, $d=0.03$, $\alpha=5\%$

$$n = \frac{(z_{\alpha/2})^2 \cdot pq}{d^2} = \frac{(1.96)^2 (0.2)(0.8)}{(0.03)^2} = 683$$

Final sample size???

But the population P is most of the time unknown

As a result, it has to be estimated from:

- *Pilot or preliminary sample:*
 - Select a pilot sample and estimate P with the sample variance, p
- *Previous or similar studies*
- *Take 50 % if there is no literature or pilot not conducted*

2. Sample Size: Two Samples

- A. Estimation of the difference between two population means
- B. Estimation of the difference between two population proportions

A. Sample size for estimating a difference in two means

- Aim: Estimate $\mu_1 - \mu_2$

$$n = \frac{z^2 \cdot (\sigma_1^2 + \sigma_2^2)}{d^2}$$

If equal sample size in both groups is required

B. Sample size for estimating a difference in two proportions

- **Aim:** Estimate $p_1 - p_2$

$$n = \frac{z^2 \cdot (p_1 q_1 + p_2 q_2)}{d^2}$$

If equal sample size in both groups is required

But the population **Values** are most of the time unknown

As a result, they has to be estimated from:

- *Pilot or preliminary sample:*
 - Select a pilot sample and estimate the population values
- *Previous or similar studies*

3. Sample Size Based on Hypothesis Testing

- **Type I error (α)** = The probability of rejecting H_0 when it is true

$$\alpha = P(\text{reject } H_0 \mid H_0 \text{ true})$$

- **Type II error (β)** = The probability of not rejecting H_0 when it is false

$$\beta = P(\text{do not reject } H_0 \mid H_0 \text{ false})$$

- **Power (1- β)** = the probability H_0 is rejected given that it is false
= P (rejecting H_0 / H_0 is false)
- If the power of a test is low, then there is little chance of detecting a difference even if one really exists
- **Power** is an important part of the design of a study

- Power $(1 - \beta) = 50\%$, $Z_\beta = 0.00$
- Power $(1 - \beta) = 75\%$, $Z_\beta = 0.67$
- Power $(1 - \beta) = 80\%$, $Z_\beta = 0.84$
- Power $(1 - \beta) = 90\%$, $Z_\beta = 1.28$
- Most of the studies recommend power of at least 80%.

A. Comparison between two means (Equal sample sizes)

$$n_1 = n_2 = \frac{(z_{\alpha/2} + z_{\beta})^2 (\sigma_1^2 + \sigma_2^2)}{\Delta^2}$$

$$\Delta = |\mu_1 - \mu_2|$$

The means and variances of the two respective groups are (μ_1, σ_1^2) and (μ_2, σ_2^2) .

Note: This formula is quite general, and applies to comparative cross sectional, case-control as well as cohort studies .

Example

1. Determine the sample sizes required to detect a difference of 5 mm in mean blood pressure between individuals receiving placebo and those receiving drug with $\alpha = 5\%$ and power of 0.80
- Assume $\sigma_1 = \sigma_2 = 15$ mm in each group.
 - We are interested in testing:

$$H_0: \mu_1 - \mu_2 = 5, H_A: \mu_1 - \mu_2 \neq 5$$

$$n = \frac{(z_\alpha + z_\beta)^2 (\sigma_1^2 + \sigma_2^2)}{\Delta^2} = \frac{(1.96 + 0.84)^2 (15^2 + 15^2)}{5^2} = 141.1$$

- We would need 142 individuals in each group

Final sample size???

B. Comparison between two means (Unequal sample sizes)

$$n_1 = \frac{(r + 1) \sigma^2 (Z_\beta + Z_{\alpha/2})^2}{r \text{ difference}^2}$$

$$r = n_2/n_1$$

Example: Suppose we anticipate twice as many non OC users as OC users entering the study. Determine the sample size to achieve an 80% power in the study using $\alpha=0.05$. $r = 2$.

$$n_1 = 1.5 * ((15.34^2 + 18.23^2)(1.96 + 0.84)^2 / (5.42)^2)$$

= 228 OC users

and $n_2 = 2(228) = 456$ non-OC users.

Final sample size???

C. Comparison between two proportions (Equal sample sizes)

$$n = \frac{2(\bar{p})(1 - \bar{p})(Z_{\beta} + Z_{\alpha/2})^2}{(p_1 - p_2)^2}$$

$$\bar{p} = \frac{p_1 + p_2}{2}$$

Let $p_1=0.35$, $p_2=0.25$, and $\Delta=p_1-p_2=0.35-0.25=0.10$

We would need approximately 330 subjects in each group. Final Sample Size???

Note: This formula is quite general, and applies to comparative cross sectional, case-control as well as cohort studies .

D. Comparison between two proportions (Unequal sample sizes)

$$n = \frac{r+1}{r} \frac{\bar{p}(1-\bar{p})(Z_{\beta} + Z_{\alpha/2})^2}{(p_1 - p_2)^2}$$

Where

$$r = n_2/n_1$$

$$\bar{p} = (p_1 + r p_2) / (1 + r)$$

$$\bar{q} = 1 - \bar{p}.$$

Note: This formula is quite general, and applies to comparative cross sectional, case-control as well as cohort studies .

Example

- A study is proposed to study the link b/n anticoagulant therapy and bleeding. Patients are divided into two groups: as cases and controls. Suppose that 5% of cases and 22% of controls are anticipated to have exposure for the treatment. How large sample should such a study be to have an 80% chance of finding a significance association at a ratio of 1:2 for cases and control at $\alpha = 5\%$.

Solution

$$p_1=0.05, p_2=0.22, \quad \bar{p} = (0.05+2*0.22)/(1+2) = 0.16, \\ \Delta = 0.22-0.05 = 0.17$$

We would need approximately 165 subjects in total

$n_1 = 55$ cases and $n_2 = 2(55) = 110$ control.

Final Sample Size??

Comparison between two proportions (Equal sample sizes)

$$n_1 = n_2 = \frac{\left(z_{\alpha/2} \sqrt{2\bar{p}\bar{q}} + z_{\beta} \sqrt{p_1q_1 + p_2q_2}\right)^2}{\Delta^2}$$

Where

$$\bar{p} = \frac{p_1 + p_2}{2} \quad \bar{q} = 1 - \bar{p}.$$

$$\Delta = p_1 - p_2$$

Note: This formula is quite general, and applies to cohort case-control as well as comparative Cross sectional Studies.

- Let $p_1=0.35$, $p_2=0.25$, and $\Delta=p_1-p_2=0.35-0.25=0.10$

$$n_1 = n_2 = \frac{\left(1.96\sqrt{2(0.30)(.70)} + 0.84\sqrt{(0.35)(0.65) + (0.25)(0.75)}\right)^2}{(0.10)^2}$$

$$n_1 = n_2 = 328.1$$

- We would need approximately **329** subjects in each group
- Final Sample Size??

Comparison between two proportions (Unequal sample sizes)

$$n_1 = \frac{\left[Z_{\alpha/2} \sqrt{\bar{p}\bar{q}} (1 + 1/\lambda) + Z_{\beta} \sqrt{p_1 q_1 + p_2 q_2 / \lambda} \right]^2}{\Delta^2}$$

$$\bar{q} = 1 - \bar{p}.$$

Where $n_2 = n_1 \lambda$, $\bar{p} = (p_1 + \lambda p_2) / (1 + \lambda)$

Example

- A study is proposed to study the effect of a new anticoagulant therapy. Patients are to be randomly divided into two groups: one receives the anticoagulant, and the other placebo. The groups are then followed for the incidence of major bleeding events over 3 years. Suppose that 5% of treated patients and 22% of controls are anticipated to experience a major event over 3 years. How large sample should such a study be to have an 80% chance of finding a significance difference at a ratio of 1:2 for treated and control at $\alpha = 5\%$.

Solution

$$p_1=0.05, p_2=0.22, \bar{p} = (0.05+2*0.22)/(1+2) = 0.16$$

$$q_1=0.95, q_2=0.78, \Delta = 0.22-0.05 = 0.17$$

We would need approximately 147 subjects in total

$n_1=49$ treated and **$n_2 = 2(49) = 98$** control.

$$n_1 = \frac{[1.96\sqrt{0.16*0.84(1+1/2)} + 0.84\sqrt{0.05*0.95+0.22*0.78/2}]^2}{(0.17)^2}$$

$$= 49$$

$$n_2 = 2*n_1 = 2*49 = 98$$

We would approximately need **147** subjects

Final Sample Size??

But the population **Values** are most of the time unknown

As a result, they has to be estimated from:

- *Pilot or preliminary sample:*
 - Select a pilot sample and estimate the population values
- *Previous or similar studies*

Home Take Messages

- Sample size calculations covered in this section are based on single outcome from a study participant
- And are not used for;
 - Survival analysis

<http://www.sample-size.net/sample-size-survival-analysis/>

- Repeated measure / longitudinal studies
 - Diagnostic studies
 - Count data
- Epi Info and other software's or online calculators can be used

Sampling (How?)

- ✓ Is the technique or procedure of selecting a sample from a population
- ✓ Before selection we should answer the following three questions
 - ❖ What is the group of people from which we want to draw the sample?
 - ❖ How many people do we need in our sample?
 - ❖ How will this people be selected?

Advantages of Sampling

- ✓ Cost - sampling saves time, labor and money.
- ✓ Quality of data - more time and effort can be spent on getting reliable data on each individual included in the sample.

Definition of Some Basic terms

❖ Sampling Unit

- ✓ The unit of selection in the sampling process
 - ❖ Districts, Kebeles, HH, Persons...

❖ Study Unit

- ✓ The unit on which the observations will be collected
- ✓ Each and every element in your investigation
- ✓ They are not necessarily the same as the sampling unit
 - ❖ Districts, Kebeles, HH, Persons...

❖ Sample Design

- ✓ The scheme (plan) for selecting the sampling units from the study population

❖ Sampling Frame

- ✓ The list of each and every element in the study population from which the sample to be selected

Methods of Sampling

- ✓ An important issue influencing the choice of the most appropriate sampling method is whether a sampling frame is available, that is, a listing of all the units that compose the study population.
- ✓ Probability and Non-probability sampling techniques

Non-Probability Sampling Methods

❖ Convenience Sampling

- ✓ Study units are selected as they are available at the time of data collection, for the convenience of the investigator

❖ Quota Sampling

- ✓ is a method that insures that a certain number of sample units from different categories with specific characteristics appear in the sample

❖ Purposeful Sampling

- ✓ Strategies for qualitative studies when focusing on a limited number of informants
- ✓ Study units will be selected by the judgment of the investigator if he/she thinks they are representatives of the population under investigation

❖ **Snow-ball sampling:** If the first sample is selected by the investigator then the next individual is selected by the first from the target group and it goes in similar fashion until the determined sample size has been gained .

- ✓ This method is appropriate for critical, more of personal, rare and other subjects.
- ✓ For instance; individuals like commercial sex workers, homosexuals, etc.

N.B The above sampling methods do not claim to be representative of the entire population

❖ Probability Sampling Methods

- ✓ They involve **random selection** procedures to ensure that each unit of the sample is chosen on the basis of chance
- ✓ All units of the study population should have an equal or at least a known chance of being included in the sample
- ✓ There are different techniques under it

❖ Simple Random Sampling (SRS)

- ✓ This is the most basic scheme of random sampling.
- ✓ To select a simple random sample you need to:
 - Make a numbered list of all the units in the population from which you want to draw a sample
 - Decide on the size of the sample
 - Select the required number of sampling units, using a “lottery” method or a table of random numbers

❖ Systematic Random Sampling

- ✓ Individuals are chosen at regular intervals from the sampling frame
- ✓ The sampling interval can be determined as $K = N/n$



❖ Stratified Random Sampling

- ✓ It is appropriate when the distribution of the characteristic to be studied is strongly affected by certain variable (heterogeneous population).
- ✓ The population is first divided into groups (strata) according to a characteristic of interest (eg, sex, geographic area, prevalence of disease, economical status, etc.).
- ✓ A separate sample is then taken independently from each stratum, by simple random or systematic sampling.
- ✓ **Proportional allocation** - if quota is given based on the size of the strata.
- ✓ **Equal allocation** - if equal quota is given for each strata.

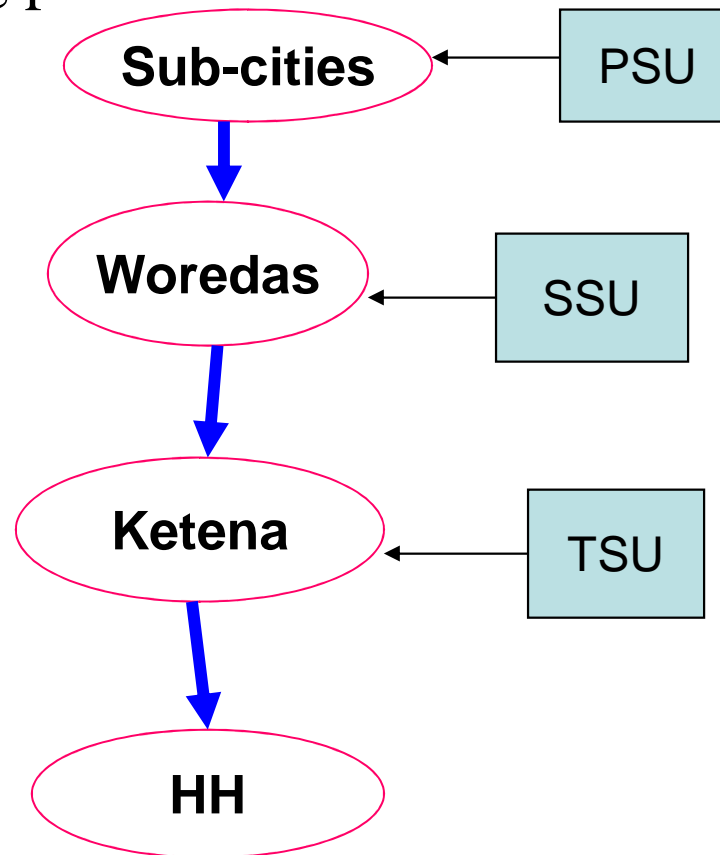


❖ Cluster Random Sampling

- ✓ When a list of groupings of study units is available (e.g. villages, districts etc.) or can be easily compiled, a number of these groupings can be randomly selected
- ✓ The process of assuming a group of study units as single for simplicity

❖ . **Multi-Stage Random Sampling**

- ✓ This method is appropriate when the population is large and widely scattered
- ✓ The number of stages of sampling is the number of times a sampling procedure is carried out



❖ Probability Proportional to Size (PPS) Sampling

- ✓ It is most useful when the sampling units vary considerably in size
- ✓ To assures that those in larger sites have the same probability of getting into the sample as those in smaller sites, and vice verse
- ✓ But the same number of units has to be sampled from each cluster

Statistical Estimation

- Up until this point, we have assumed that the values of the parameters of a probability distribution are known.
- In the real world, the values of these population parameters are usually not known
- Instead, we must try to say something about the way in which a random variable is distributed using the information contained in a sample of observations

- The process of drawing conclusions about an entire population based on the data in a sample is known as statistical inference.
- Methods of inference usually fall into one of two broad categories: estimation or hypothesis testing.
- For now, we will focus on using the observations in a sample to estimate a population parameter

Estimation

- **Is concerned with estimating the values of specific population parameters based on sample statistics.**
- **is about using information in a sample to make estimates of the characteristics (parameters) of the source population.**

Example

- A sample survey revealed:
 - *Proportion of smokers among a certain group of population aged 15 to 24.*
 - *Mean of SBP among sampled population*
 - *Prevalence of HIV-positive among people involved in the study*

 The next question is what can we predict about the characteristics of the population from which the sample was drawn

Estimation, Estimator & Estimate

- ♣ **Estimation** is the computation of a statistic from sample data, often yielding a value that is an approximation (guess) of its target, an unknown true population parameter value.
- ♣ The statistic itself is called an **estimator** and can be of two types - **point or interval**.
- ♣ The value or values that the estimator assumes are called **estimate**.

- Two methods of estimation are commonly used: point estimation and interval estimation
- Point estimation involves the calculation of a single number to estimate the population parameter
- Interval estimation specifies a range of reasonable values for the parameter

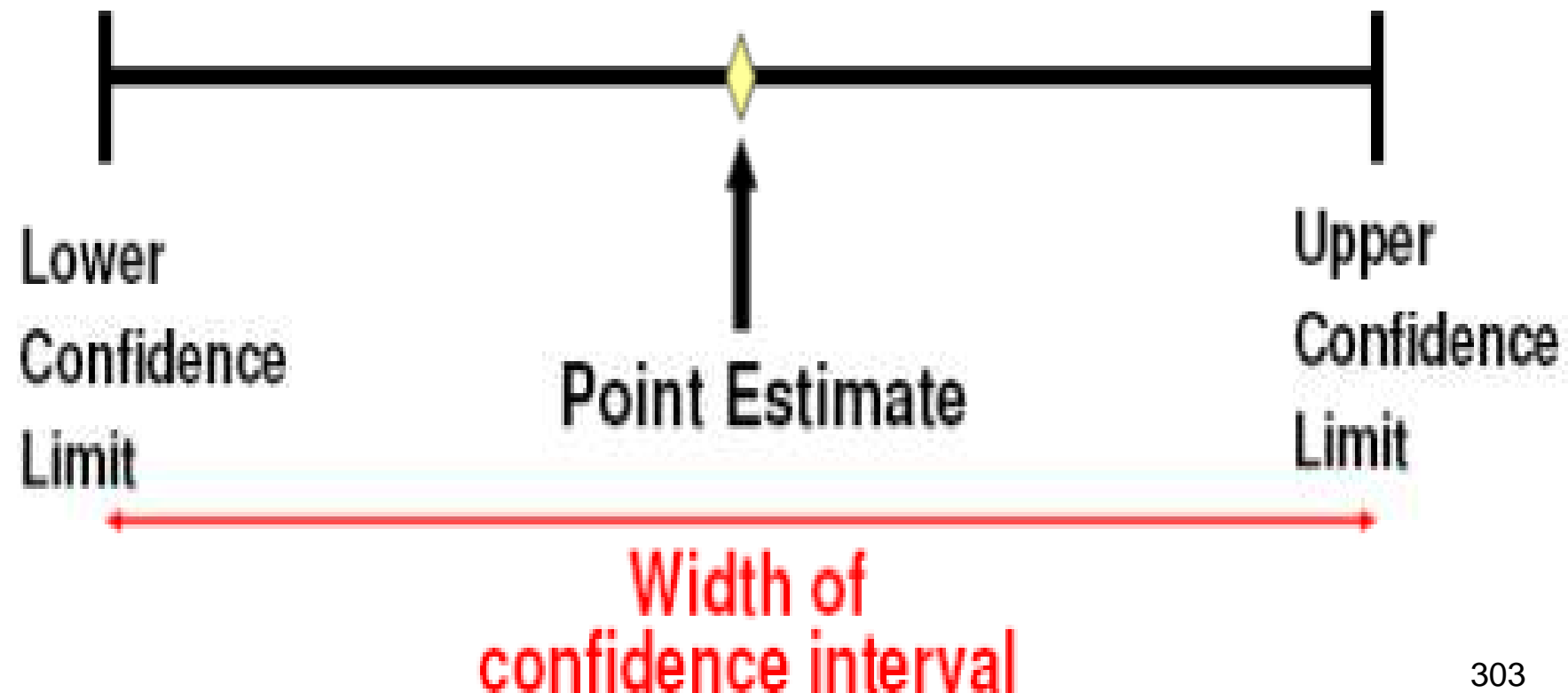
Point versus Interval Estimators

- ♣ An estimator that represents a "single best guess" is called a point estimator.
- ♣ When the estimate is of the form of a "range of plausible values", it is called an interval estimator.
- ♣ Thus,
 - A point estimate is of the form: [Value],
 - Whereas, an interval estimate is of the form: [lower limit, upper limit]

Example -

The sample mean \bar{X}_n , calculated using data in a sample of size n , is a point estimator of the population mean μ . If $\bar{X}_n = 10$, the value 10 is called a point estimate of the population mean μ .

- A **point estimate** is a single number,
- a **confidence interval** provides additional information about variability



Properties of good estimators

A. Unbiased

♣ A statistic is said to be an unbiased estimator if its expected value is equal to the estimated parameter.

Sample mean (\bar{X}) is an unbiased estimator of population mean.

$$E (\bar{X}) = \mu$$

B. Consistent

- ♣ A statistic is said to be a consistent estimator if its value close to the parameter as the sample size grows larger.

C. Relatively Efficient

- ♣ A statistic is said to be an efficient estimator if its variance is smaller.

Illustration for data from a normal distribution

1. The unbiased estimators are the sample mean \bar{X} and median \tilde{X}
2. $\text{variance} [\bar{X}] < \text{variance} [\tilde{X}]$

Choose the sample mean \bar{X} . It is the minimum variance unbiased estimator.

For a random sample of data from a normal probability distribution, \bar{X} is the minimum variance unbiased estimator of the population mean μ .

Estimating the Sampling Error

- Any estimates derived from samples are subject to the sampling error.
- This comes from the fact that only a part of the population was observed, instead of the whole.
- A different samples could have come up with different results. The amount of variation that exists among the estimates from the different possible samples is the sampling error.

- The set of sample means in repeated random samples of size n from a given population has variance σ^2/n .
- The standard deviation of this set of sample means is σ/\sqrt{n} and is referred to as the standard error of the mean (sem) or the standard error.
- The sem is estimated by s/\sqrt{n} if σ is unknown.

- *The sampling error is dependent on on sample size (n), the variability of individual sample points (σ), sampling and estimation methods.*
- **As n increases, the sample mean (\bar{X}) and the sample variance s^2 approach the values of the true population parameters, μ and σ^2 , respectively.**

Example

- *Suppose that the mean \pm sd of DBP on 20 old males is 78.5 ± 10.3 mm Hg.*
 1. What is our best estimate of μ ?
 2. What is the sem?
 3. Compare the sem with the sd.

- The following table gives the se for mean of DBP for different sample sizes.

n	sem
1	10.3
20	2.3
100	1.0

- Our best estimate of μ is 78.5.
- The sem of this estimate is $10.3/\sqrt{20} = 2.3$
- The sem (2.3) is much smaller than sd (10.3).

1. Point Estimate

- A single numerical value used to estimate the corresponding population parameter.

Sample Statistics are Estimators of Population Parameters	
Sample mean, \bar{X}	μ
Sample variance, S^2	σ^2
Sample proportion, \hat{p} or $\hat{\pi}$	p or π
Sample Odds Ratio, \hat{OR}	OR
Sample Relative Risk, \hat{RR}	RR
Sample correlation coefficient, r	ρ

2. Interval Estimation

- **Interval estimation** specifies a range of reasonable values for the population parameter based on a point estimate.
- **A confidence interval is a particular type of interval estimator.**

Confidence Intervals

- Give a plausible range of values of the estimate likely to include the “true” (population) value with a given confidence level.
- An interval estimate provides more information about a population characteristic than does a point estimate
- Such interval estimates are called confidence intervals.

- CIs also give information about the precision of an estimate.
- How much uncertainty is associated with a point estimate of a population parameter?
- When sampling variability is high, the CI will be wide to reflect the uncertainty of the observation.
- Wider CIs indicate less certainty.

- CIs can also answer the question of *whether or not* an association exists or a treatment is beneficial or harmful. (analogous to p-values...)

e.g., if the CI of an odds ratio includes the value 1.0 we cannot be confident that exposure is associated with disease.

- **A CI in general:**
 - Takes into consideration variation in sample statistics from sample to sample
 - Based on observation from 1 sample
 - Gives information about closeness to unknown population parameters
 - Stated in terms of level of confidence
 - Never 100% sure

General Formula:

The general formula for all CIs is:

The value of the statistic in my sample
(eg., mean, odds ratio, etc.)

point estimate \pm (measure of how confident we want to be) \times (standard error)

From a Z table or a t table, depending
on the sampling distribution of the
statistic (Critical value).

Standard error of the statistic.

Lower limit = Point Estimate - (Critical Value) x (Standard Error)

Upper limit = Point Estimate + (Critical Value) x (Standard Error)

- **A wide interval suggests imprecision of estimation.**
- **Narrow CI widths reflects large sample size or low variability or both.**
- ***Note: Measure of how confident we want to be = critical value = confidence coefficient***

Confidence Level

- Confidence Level
 - Confidence in which the interval will contain the unknown population parameter
- A percentage (less than 100%)
 - Example: 95%
- Also written $(1 - \alpha) = .95$

Definition: 95% CI

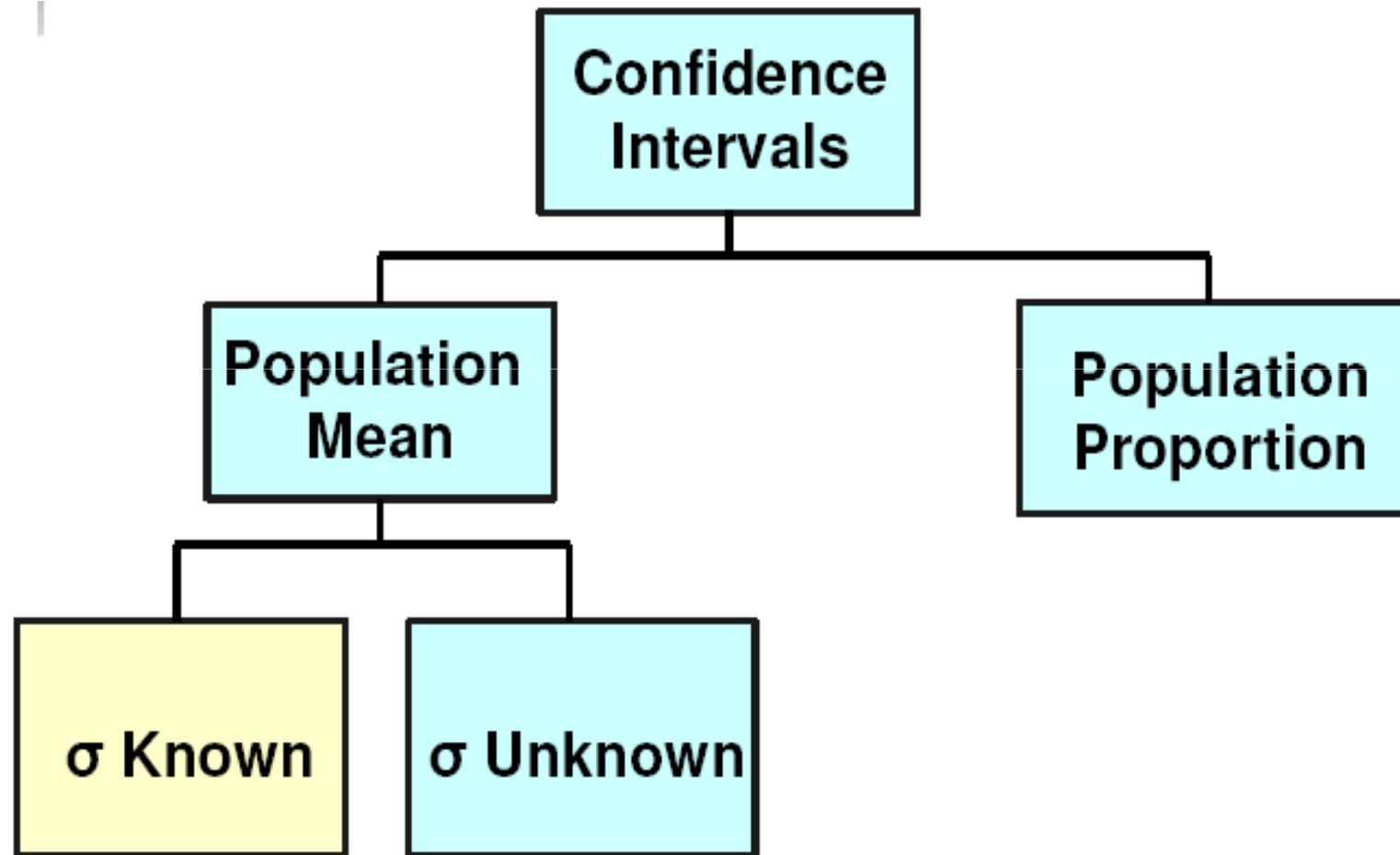
1. Probabilistic interpretation:

- If all possible random samples (an infinite number) of a given sample size (e.g. 10 or 100) were obtained and if each were used to obtain its own CI, then 95% of all such CIs would contain the unknown population parameter; the remaining 5% would not.
- It is incorrect to say “*There is a 95% probability that the CI contains the unknown population parameter*”.

2. Practical interpretation

- When sampling is from a normally distributed population with known standard deviation, we are $100(1-\alpha)$ [e.g., 95%] confident that the single computed interval contains the unknown population parameter.

Estimation for Single Population



1. CI for a Single Population Mean (normally distributed)

A. Known variance or large sample size

- **There are 3 elements to a CI:**
 1. Point estimate
 2. SE of the point estimate
 3. Confidence coefficient
- **Consider the task of computing a CI estimate of μ for a population distribution that is normal with σ *known*.**
- **Available are data from a random sample of size = n .**

■ Assumptions

- *Population standard deviation (σ) is known*
- *Sample size is large ($n \geq 30$)*
- *Population normally distributed*
- **Use Z- distribution**

A $100(1-\alpha)\%$ C.I. for μ is:

$$\bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

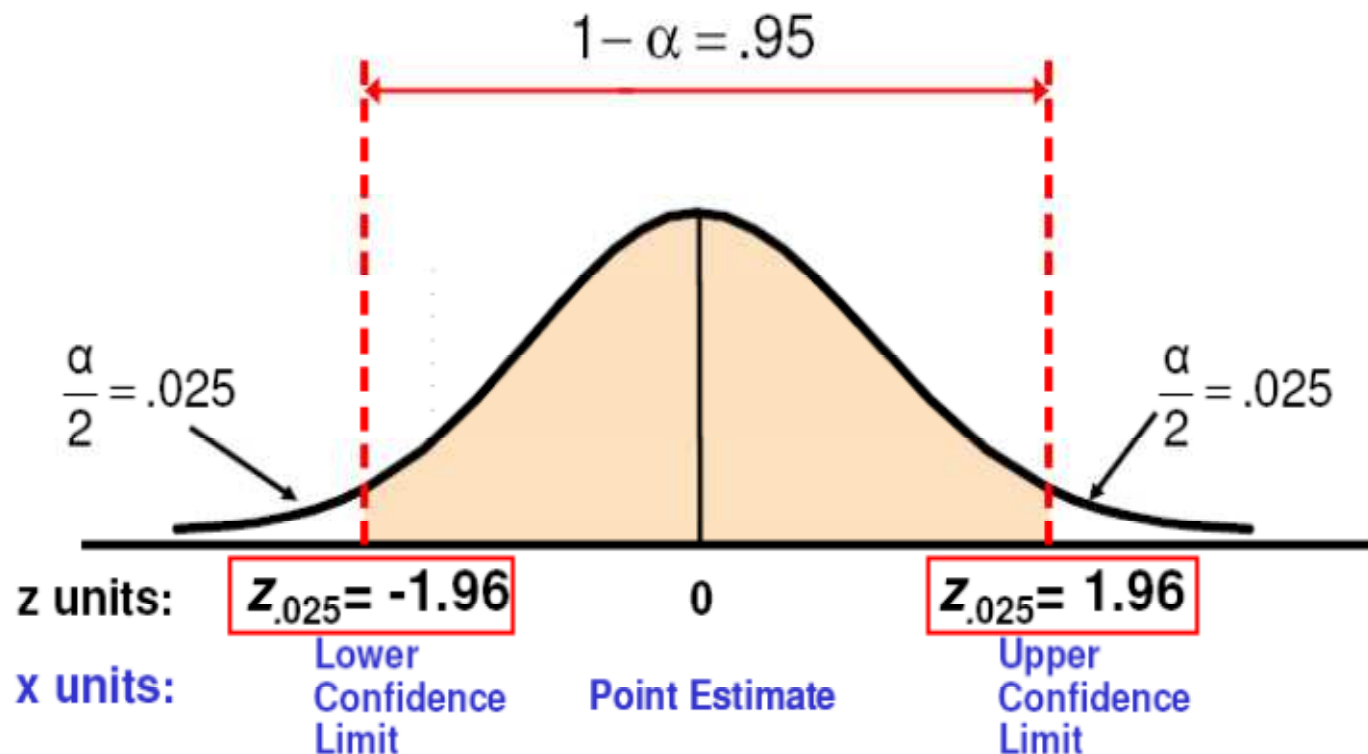
- α is to be chosen by the researcher, most common values of α are 0.05, 0.01 and 0.1.

1. The Point Estimate of μ is the Sample Mean \bar{X}
2. The Standard Error of \bar{X}_n is σ/\sqrt{n}
3. Commonly used CLs are 90%, 95%, and 99%

Confidence Level ((1- α) 100%)	Confidence Coefficient ($Z_{\alpha/2}$)
99%	2.576
95%	1.960
90%	1.645
80%	1.282

Finding the Critical Value

- Consider a 95% confidence interval: $z_{\alpha/2} = \pm 1.96$

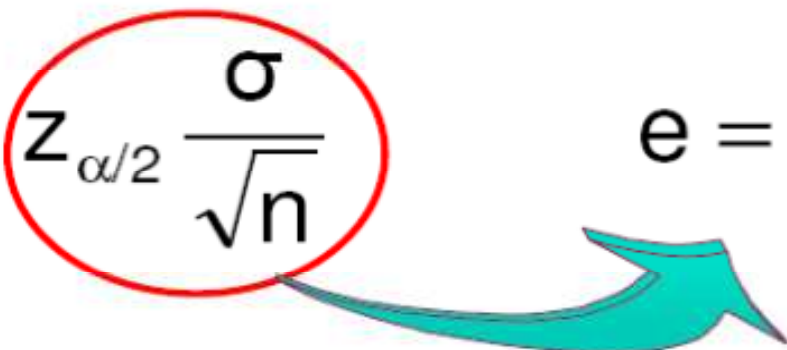


Margin of Error

(Precision of the estimate)

- **Margin of Error (e):** the amount added and subtracted to the point estimate to form the confidence interval

Example: Margin of error for estimating μ , σ known:

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

$$e = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Factors Affecting Margin of Error

$$e = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

- The CI for mean and margin of error is determined by n , σ or s , and α .
 - *As n increases, the size of the CI decreases.*
 - *As σ or s increases, the size of the CI increases.*
 - *As the confidence level increases (α decreases), the size of the CI increases.*

Example:

1. **Waiting times (in hours) at a particular hospital are believed to be approximately normally distributed with a variance of 2.25 hr.**
 - a. A sample of 20 outpatients revealed a mean waiting time of 1.52 hours. Construct the 95% CI for the estimate of the population mean.
 - b. Suppose that the mean of 1.52 hours had resulted from a sample of 32 patients. Find the 95% CI.
 - c. What effect does larger sample size have on the CI?

a.

$$1.52 \pm 1.96 \frac{\sqrt{2.25}}{\sqrt{20}} = 1.52 \pm 1.96(.33) \\ = 1.52 \pm .65 = (.87, 2.17)$$

- ***We are 95% confident that the true mean waiting time is between 0.87 and 2.17 hrs.***
- ***Although the true mean may or may not be in this interval, 95% of the intervals formed in this manner will contain the true mean.***
- ***An incorrect interpretation is that there is 95% probability that this interval contains the true population mean.***

b.

$$1.52 \pm 1.96 \frac{\sqrt{2.25}}{\sqrt{32}} = 1.52 \pm 1.96(.27) \\ = 1.52 \pm .53 = (.99, 2.05)$$

c. The larger the sample size makes the CI narrower (more precision).

- When constructing CIs, it has been assumed that the standard deviation of the underlying population, σ , is known
- What if σ is not known?
- In practice, if the population mean μ is unknown, then the standard deviation, σ , is probably unknown as well.
- In this case, the SE of the population can be replaced by the SE of the sample if the sample size is large enough ($n \geq 30$). With large sample size, we assume a normal distribution

- *Example: It was found that a sample of 35 patients were 17.2 minutes late for appointments, on the average, with SD of 8 minutes. What is the 90% CI for μ ? Ans: (15.0, 19.4).*
- Since the sample size is fairly large (≥ 30) and the population SD is unknown, we assume the distribution of sample mean to be normally distributed based on the CLT and the sample SD to replace population σ .

B. Unknown variance

(and small sample size, $n < 30$)

- What if the σ for the underlying population is unknown and the sample size is small?
- As an alternative we use *Student's t distribution*.

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

$$t\text{-score} = t_{DF=n-1} = \frac{\bar{X} - \mu}{s/\sqrt{n}} \text{ is distributed Student's } t \text{ with degrees of freedom} = (n-1)$$

■ Assumptions

- ***Population standard deviation (σ) is unknown***
- ***Sample size is small ($n < 30$)***
- ***Population normally distributed***
- ***If population is not normal, use CLT***
 - Use Student's t Distribution
 - Confidence Interval Estimate

$$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

What happens to CI as sample gets larger?

$$\bar{x} \pm Z \left(\frac{s}{\sqrt{n}} \right)$$

$$\bar{x} \pm t \left(\frac{s}{\sqrt{n}} \right)$$

For large samples:

Z and t values become almost identical, so CIs are almost identical.

Degrees of Freedom (df)

df = Number of observations that are free to vary after sample mean has been calculated

$$df = n - 1$$

Example: Suppose the mean of 3 numbers is 8.0

Let $x_1 = 7$
Let $x_2 = 8$
What is x_3 ?



If the mean of these three values is 8.0, then x_3 **must be 9** (i.e., x_3 is not free to vary)

Here, $n = 3$, so degrees of freedom = $n - 1 = 3 - 1 = 2$

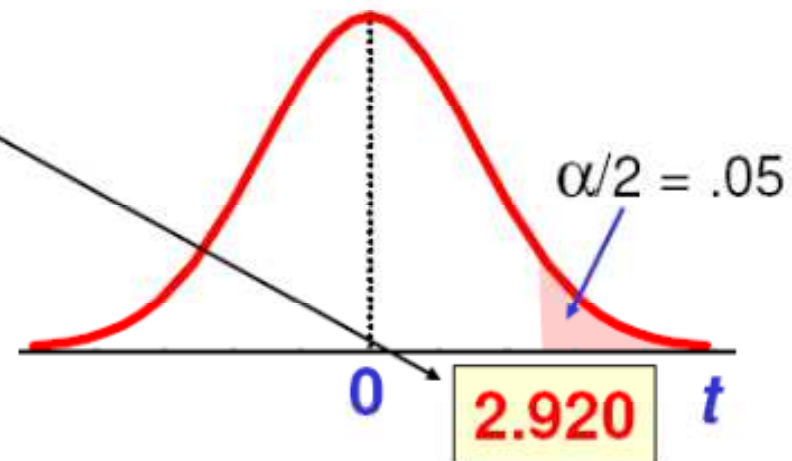
(2 values can be any numbers, but the third is not free to vary for a given mean)

Student's t Table

Upper Tail Area			
df	.25	.10	.05
1	1.000	3.078	6.314
2	0.817	1.886	2.920
3	0.765	1.638	2.353

Let: $n = 3$
 $df = n - 1 = 2$
 $\alpha = .10$
 $\alpha/2 = .05$

The body of the table contains t values, not probabilities



t distribution values

- With comparison to the Z value

Confidence Level	t (10 d.f.)	t (20 d.f.)	t (30 d.f.)	z
.80	1.372	1.325	1.310	1.28
.90	1.812	1.725	1.697	1.64
.95	2.228	2.086	2.042	1.96
.99	3.169	2.845	2.750	2.57

Note: $t \rightarrow z$ as n increases

Example

A random sample of size $n=20$ durations (minutes) of cardiac bypass surgeries has a mean duration of $\bar{X} = 267$ minutes, and variance $s^2 = 36,700$ minutes². Assuming the underlying distribution is normal with unknown variance, construct a 90% CI estimate of the unknown true mean, μ .

- Standard error = $\frac{S}{\sqrt{n}} = \frac{\sqrt{36,700}}{\sqrt{20}} = 42.7$ minutes
- t-value at 90% CL at 19 df = 1.729

Putting this all together –

$$\begin{aligned}\text{Lower limit} &= (\text{point estimate}) - (\text{conf coeff.}) (\text{SE of point estimate}) \\ &= 267 - (1.729)(42.7) \\ &= 193.17\end{aligned}$$

$$\begin{aligned}\text{Upper limit} &= (\text{point estimate}) + (\text{conf coeff}) (\text{SE of point estimate}) \\ &= 267 + (1.729)(42.7) \\ &= 340.83\end{aligned}$$

Thus, a 90% confidence interval for the true mean duration of surgery is (193.2, 340.8) minutes.

Exercise

- Compute a 95% CI for the mean birth weight based on $n = 10$, sample mean = 116.9 and $s = 21.70$.
- From the t Table, $t_{9, 0.975} = 2.262$
- Ans: (101.4, 132.4)

2. CIs for single population proportion, p

- An interval estimate for the population proportion (p) can be calculated by adding an allowance for uncertainty to the sample proportion (\bar{p})
- Is based on three elements of CI.
 - **Point estimate**
 - **SE of point estimate**
 - **Confidence coefficient**

- Recall that the distribution of the sample proportion is approximately normal if the sample size is large, with standard deviation

$$\sigma_p = \sqrt{\frac{p(1-p)}{n}}$$

- We will estimate this with sample data:

$$s_{\bar{p}} = \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

- Upper and lower confidence limits for the population proportion are calculated with the formula

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

- where
 - z is the standard normal value for the level of confidence desired
 - \bar{p} is the sample proportion
 - n is the sample size

Lower limit = Point Estimate - (Critical Value) x (Standard Error of Estimate)

Upper limit = Point Estimate + (Critical Value) x (Standard Error of Estimate)

Hence,

$$\left(\hat{p} - 1.96 \sqrt{\frac{\hat{p} \hat{q}}{n}}, \hat{p} + 1.96 \sqrt{\frac{\hat{p} \hat{q}}{n}} \right)$$

is an approximate 95% CI for the true proportion p .

Example 1

- A random sample of 100 people shows that 25 are left-handed. Form a 95% CI for the true proportion of left-handers.

1. $\bar{p} = 25/100 = .25$

2. $S_{\bar{p}} = \sqrt{\bar{p}(1-\bar{p})/n} = \sqrt{.25(.75)/n} = .0433$

3. $.25 \pm 1.96 (.0433)$

0.1651 0.3349

Interpretation

- We are 95% confident that the true percentage of left-handers in the population is between
16.51% and 33.49%.
- Although this range may or may not contain the true proportion, 95% of intervals formed from samples of size 100 in this manner will contain the true proportion.

Changing the sample size

- Increases in the sample size reduce the width of the confidence interval.

Example:

- If the sample size in the above example is doubled to 200, and if 50 are left-handed in the sample, then the interval is still centered at .25, but the width shrinks to

.1931

Example 2

- *It was found that 28.1% of 153 cervical-cancer cases had never had a Pap smear prior to the time of case's diagnosis. Calculate a 95% CI for the percentage of cervical-cancer cases who never had a Pap test.*
- A 95% CI is given by

$$\begin{aligned}\hat{p} \pm 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}} &= .281 \pm 1.96\sqrt{\frac{.281(.719)}{153}} \\ &= .281 \pm .071 = (.210, .352).\end{aligned}$$

Example 3

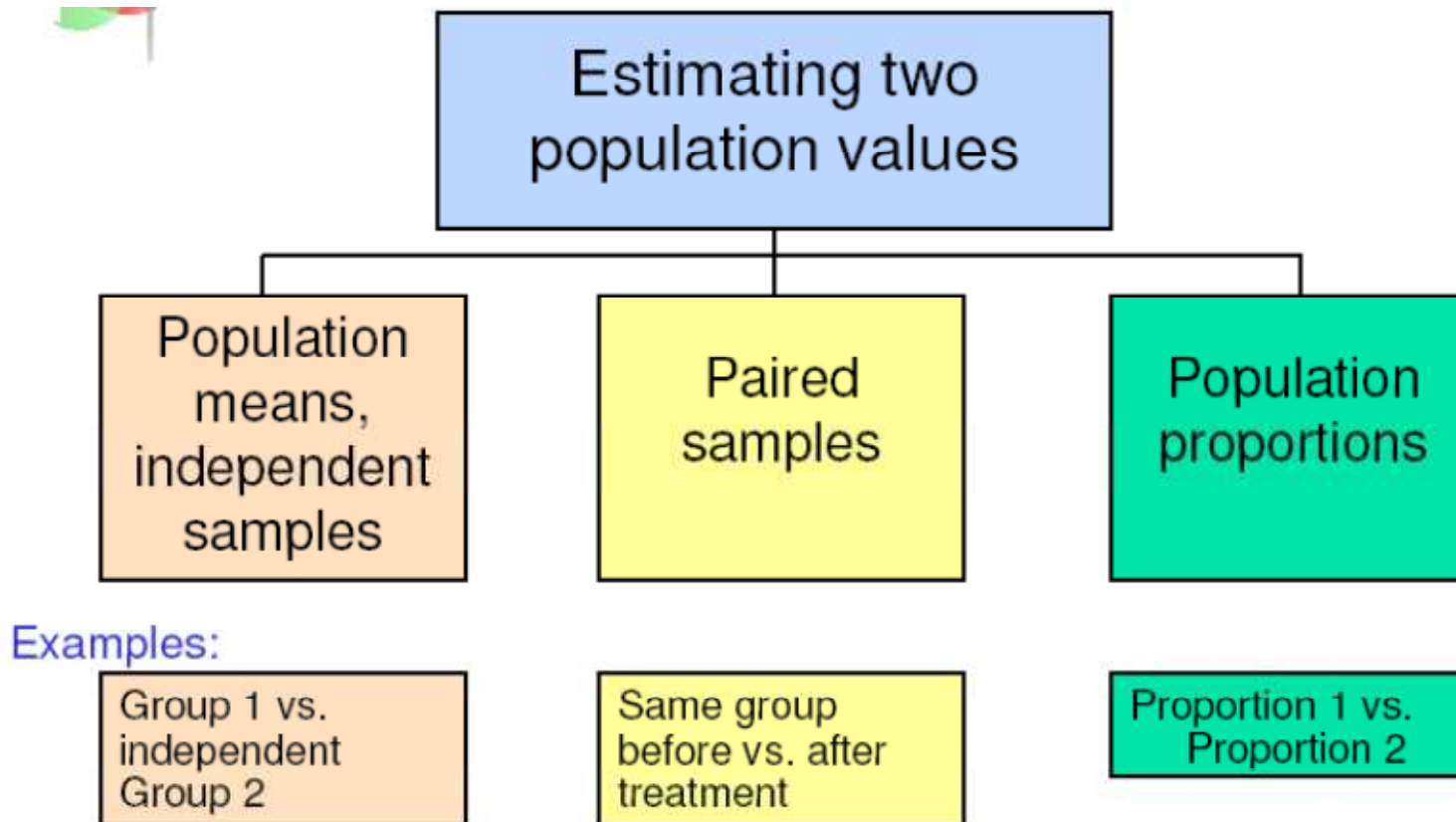
- Suppose that among 10,000 female operating-room nurses, 60 women have developed breast cancer over five years. Find the 95% for p based on point estimate.
- Point estimate = $60/10,000 = 0.006$
- The 95% CI for p is given by the interval:

$$\left(\hat{p} - 1.96 \sqrt{\frac{\hat{p} \hat{q}}{n}}, \hat{p} + 1.96 \sqrt{\frac{\hat{p} \hat{q}}{n}} \right)$$

- The 95% CI for p is:

$$\begin{aligned} \left(.006 - 1.96 \sqrt{\frac{.006(.994)}{10,000}}, .006 + 1.96 \sqrt{\frac{.006(.994)}{10,000}} \right) &= (.006 - .0015, .006 + .0015) \\ &= (.0045, .0075) \end{aligned}$$

Estimation for Two Populations



3. CI for the difference between population means (normally distributed)

A. Known variances (2 independent samples)

- When σ_1 and σ_2 are known and both populations are normal or both sample sizes are at least 30, the test statistic is a z-value...

...and the standard error of $\bar{x}_1 - \bar{x}_2$ is

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

The point estimate for the difference is

$$\bar{X}_1 - \bar{X}_2$$

The confidence interval for

$\mu_1 - \mu_2$ is:

$$\left(\bar{X}_1 - \bar{X}_2 \right) \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Assumptions

- Samples are randomly and independently drawn
- Population distributions are normal or both sample sizes are ≥ 30
- Population standard deviations are known

Illustration

- A researcher performs a drug trial involving two independent groups.
 - A **control** group is treated with a placebo while, separately;
 - The **intervention** group is treated with an active agent.
 - Interest is in a comparison of the **mean control** response with the **mean intervention** response under the assumption that the responses are independent.

Examples

- **We are interested in the similarity of the two groups.**

1) Is mean blood pressure the same for males and females?

2) Is body mass index (BMI) similar for breast cancer cases versus non-cancer patients?

3) Is length of stay (LOS) for patients in hospital “A” the same as that for similar patients in hospital “B”?

- Thus, evidence of similarity of the two groups is reflected in a difference between means that is “near” zero.
- Focus is on $[\mu_{\text{Group 1}} - \mu_{\text{Group 2}}]$

Example

- *Researchers are interested in the difference between serum uric acid levels in patients with and without Down's syndrome.*
- Patients without Down's syndrome
 - $n=12$, sample mean=4.5 mg/100ml, $\sigma^2=1.0$
- Patients with Down's syndrome
 - $n=15$, sample mean=3.4 mg/100ml, $\sigma^2=1.5$
- Calculate the 95% CI.
- $SE = 0.43$, 95% CI = $1.1 \pm 1.96 (0.43) = (0.26, 1.94)$
- WE are 95% confident that the true difference between the two population means is between 0.26 and 1.94.

B. Unknown variances (Independent samples)

I. Population variances equal (large sample)

- **Assumptions:**
 - *Samples are randomly and independently drawn*
 - *Both of the sample sizes are ≥ 30*
 - *Population standard deviations are unknown*

Forming confidence estimates:

- Use sample standard deviation s to estimate σ , and
- the test statistic is a z-value

The confidence interval for
 $\mu_1 - \mu_2$ is:

$$\left(\bar{x}_1 - \bar{x}_2\right) \pm z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Example

- The mean CD4 + cells for 112 men with HIV infection was 401.8 with a SD of 226.4. For 75 men without HIV, the mean and SD were 828.2 and 274.9, respectively. Calculate a 99% CI for the difference between population means.
- SE of the difference b/n two means = 38.28
- 99% CI = $426.4 \pm 2.58 (38.28)$
= (327.6, 525.2)

II. Population variances equal (small sample)

- **Assumptions:**
 - Populations are normally distributed
 - *The populations have equal variances*
 - *Samples are independent*
 - *One or both sample sizes are <30*
 - *Population standard deviations are unknown*

* If $0.5 \leq s_1^2/s_2^2 \leq 2$ then we assume that the population variances are equal.

Forming confidence estimates:

- The population variances are assumed equal, so use the two sample standard deviations and pool them to estimate σ
- The test statistic is a t value with $(n_1 + n_2 - 2)$ degrees of freedom
- The pooled estimate (s_p^2) is the weighted average of the two sample variances.

- The pooled standard deviation is :

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

- The standard error of the estimate is given by:

$$s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

The confidence interval for
 $\mu_1 - \mu_2$ is:

$$\left(\bar{x}_1 - \bar{x}_2 \right) \pm t_{\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Where $t_{\alpha/2}$ has $(n_1 + n_2 - 2)$ d.f.,

and

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Example 1

- A study was conducted to compare the serum iron levels of children with cystic fibrosis to those of healthy children. Serum iron levels were measured for random samples of $n_1 = 9$ healthy children and $n_2 = 13$ children with cystic fibrosis.

For the 9 healthy children, $\bar{x}_1 = 18.9 \mu\text{mol/l}$ and $s_1 = 5.9 \mu\text{mol/l}$

For the 13 children with cystic fibrosis, $\bar{x}_2 = 11.9 \mu\text{mol/l}$ and $s_2 = 6.3 \mu\text{mol/l}$

- The two underlying populations of serum iron levels are independent and normally distributed.

For now, assume that $\sigma_1^2 = \sigma_2^2 = \sigma^2$

$$\begin{aligned}s_p^2 &= \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \\&= \frac{(9 - 1)(5.9)^2 + (13 - 1)(6.3)^2}{9 + 13 - 2} \\&= 37.74\end{aligned}$$

The difference in sample means $\bar{x}_1 - \bar{x}_2 = 7.0$ can be used as a point estimate for the true difference in population means $\mu_1 - \mu_2$

We could also construct a confidence interval for $\mu_1 - \mu_2$

A t-value at 95% CL with 20 df is 2.086

$$(\bar{x}_1 - \bar{x}_2) \pm 2.086 \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

or

$$(18.9 - 11.9) \pm 2.086 \sqrt{37.74 \left(\frac{1}{9} + \frac{1}{13} \right)}$$

or

$$(1.4, 12.6)$$

is a 95% confidence interval for $\mu_1 - \mu_2$

Example 2

- Birth weights of children born to 14 heavy smokers (group 1) and to 15 non-smokers (group 2) were sampled from live births at a large teaching hospital. For the heavy smokers, sample mean = 3.17 kg, SD = 0.46 and for non-smokers, sample mean = 3.63 kg and SD = 0.36.
- $S_p = 0.4121$, $SE = 0.1531$, t-value at 27 df = 2.05
- 95% CI = (0.14, 0.77)

III. Population variances unequal (small sample)

- **The confidence interval for $\mu_1 - \mu_2$ is:**

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2, d'} \sqrt{\left(\frac{s_1^2}{n_1}\right) + \left(\frac{s_2^2}{n_2}\right)}$$

- Where the degree of freedom (d') is given by: Welch -Satterthwaite approximation

$$d' = \frac{\left[(s_1^2/n_1) + (s_2^2/n_2) \right]^2}{\left[\frac{(s_1^2/n_1)^2}{(n_1-1)} + \frac{(s_2^2/n_2)^2}{(n_2-1)} \right]}$$

Round d' down to the nearest integer — this is d''

Example

For the tuberculosis meningitis example, a random sample of $n_1 = 37$ HIV infected patients has mean age at diagnosis $\bar{x}_1 = 27.9$ years and standard deviation $s_1 = 5.6$ years

A sample of $n_2 = 19$ uninfected patients has mean age at diagnosis $\bar{x}_2 = 38.8$ years and standard deviation $s_2 = 21.7$ years

$$\begin{aligned}
 d' &= \frac{\left[(s_1^2/n_1) + (s_2^2/n_2)\right]^2}{\left[\frac{(s_1^2/n_1)^2}{(n_1-1)} + \frac{(s_2^2/n_2)^2}{(n_2-1)}\right]} \\
 &= \frac{\left[(5.6^2/37) + (21.7^2/19)\right]^2}{\left[\frac{(5.6^2/37)^2}{(37-1)} + \frac{(21.7^2/19)^2}{(19-1)}\right]} \\
 &= 19.24
 \end{aligned}$$

and

$$d'' = 19$$

- For a t distribution with 19 df at 95% CL t -value is 2.093.
- Therefore, a 95% confidence interval would take the form

$$(\bar{x}_1 - \bar{x}_2) \pm 2.093 \sqrt{\left(\frac{s_1^2}{n_1}\right) + \left(\frac{s_2^2}{n_2}\right)}$$

- Using the data from two samples of patients with tuberculosis meningitis, the 95% CI for $\mu_1 - \mu_2$ is

$$(27.9 - 38.8) \pm 2.093 \sqrt{\left(\frac{5.6^2}{37}\right) + \left(\frac{21.7^2}{19}\right)}$$

or

$$(-21.5, -0.3)$$

C. Paired Samples

- Tests Means of 2 **Related** Populations
 - △ Paired or matched samples
 - △ Repeated measures (before/after)
 - △ Use **difference** between paired values:
$$\mathbf{d = x_1 - x_2}$$
- Eliminates variation among subjects
- Assumptions:
 - Both populations are normally distributed,
 - Or, if not normal, use large samples.

Paired Data

- **Paired data arises when each individual in a sample is measured twice.**
- **Measurement might be "pre/post", "before/after", "right/left", "parent/child", etc.**

Examples of paired data

- 1) Blood pressure prior to and following treatment,
 - 2) Number of cigarettes smoked per week measured prior to and following participation in a smoking cessation program,
 - 3) Number of sex partners in the month prior to and in the month following an HIV education campaign.
- Notice in each of these examples that the two occasions of measurement are linked by virtue of the two measurements being made on the same individual.
 - Longitudinal or follow-up study

Paired differences

- If two measurements of the same phenomenon (eg. blood pressure, # cigarettes/week, etc) X and Y are measured on an individual and if each is normally distributed, then their difference is also distributed normal.
- The interest in the difference between two measurements

The i^{th} paired difference is d_i , where

$$d_i = x_{1i} - x_{2i}$$

The point estimate for the population mean paired difference is \bar{d} :

$$\bar{d} = \frac{\sum_{i=1}^n d_i}{n}$$

The sample standard deviation is

$$s_d = \sqrt{\frac{\sum_{i=1}^n (d_i - \bar{d})^2}{n-1}}$$

n is the number of pairs in the paired sample

The confidence interval for \bar{d} is

$$\bar{d} \pm t_{\alpha/2} \frac{s_d}{\sqrt{n}}$$

- Where $t_{\alpha/2}$ is with $n-1$ df.

Example

- Ten hypertensive patients are screened at a neighborhood health clinic and are given methyl dopa, a strong antihypertensive medication for their condition. They are asked to come back 1 week later and have their blood pressures measured again. Suppose the initial and follow-up SBPs (mm Hg) of the patients are given below.

Patient number	Initial SBP	Follow-up SBP
1	200.0	188.0
2	194.0	212.0
3	236.0	186.0
4	163.0	150.0
5	240.0	200.0
6	225.0	222.0
7	203.0	190.0
8	180.0	154.0
9	177.0	180.0
10	240.0	225.0

1. What is the mean and Sd of the difference?
2. What is the standard error of the mean?
3. Assume that the difference is normally distributed, construct a 95% CI for μ .

Answer

- We have the following data and summary statistics

Patient number	$d_i(I - F)$	d_i^2
1	12.0	144
2	-18.0	324
3	50.0	2500
4	13.0	169
5	40.0	1600
6	3.0	9
7	13.0	169
8	26.0	676
9	-3.0	9
10	15.0	225
Total	151.0	5825

$$\bar{d} = \frac{\sum d_i}{n} = \frac{151}{10} = 15.1$$

$$s = \sqrt{\frac{\sum (d_i - \bar{d})^2}{n-1}} = \sqrt{\frac{3544.9}{9}} = \sqrt{393.878} = 19.85$$

$$se = \frac{s}{\sqrt{n}} = \frac{19.85}{\sqrt{10}} = 6.28$$

A 95% CI for μ is provided by

$$\begin{aligned} & \left(\bar{d} - \frac{t_{9, .975}s}{\sqrt{n}}, \bar{d} + \frac{t_{9, .975}s}{\sqrt{n}} \right) \\ &= [15.1 - 2.262(6.28), 15.1 + 2.262(6.28)] \\ &= (15.1 - 14.2, 15.1 + 14.2) = (0.9, 29.3). \end{aligned}$$

4. Two Population Proportions

- We are often interested in comparing proportions from 2 populations:
 - Is the incidence of disease A the same in two populations?
 - Patients are treated with either drug D, or with placebo. Is the proportion “improved” the same in both groups?

Goal: Form a confidence interval for or test a hypothesis about the difference between two population proportions, $p_1 - p_2$

Assumptions:

$$n_1 p_1 \geq 5 \quad , \quad n_1 (1 - p_1) \geq 5$$

$$n_2 p_2 \geq 5 \quad , \quad n_2 (1 - p_2) \geq 5$$

The point estimate for the difference is $\bar{p}_1 - \bar{p}_2$

Confidence Interval for Two Population Proportions

- SE of the difference =

$$\sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \frac{\bar{p}_2(1-\bar{p}_2)}{n_2}}$$

- The confidence interval for $p_1 - p_2$ is:

$$(\bar{p}_1 - \bar{p}_2) \pm z_{\alpha/2} \sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \frac{\bar{p}_2(1-\bar{p}_2)}{n_2}}$$

The following formula is also equally used

- An approximate 95% confidence interval takes the form

$$(\hat{p}_1 - \hat{p}_2) \pm 1.96 \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$$

Example

- In a clinical trial for a new drug to treat hypertension, $n_1 = 50$ patients were randomly assigned to receive the new drug, and $n_2 = 50$ patients to receive a placebo. 34 of the patients receiving the drug showed improvement, while 15 of those receiving placebo showed improvement.
- Compute a 95% CI estimate for the difference between proportions improved.

- $p_1 = 34/50 = 0.68$, $p_2 = 15/50 = 0.30$
- The point estimate for the difference is:

$$= [0.68 - 0.30] = 0.38$$
- SE of the difference =
$$\sqrt{\frac{.68(.32)}{50} + \frac{.30(.70)}{50}} = 0.0925$$
- 95% CI
 - Lower = (point estimate) - $(Z_{\alpha/2})$ (SE)

$$= 0.38 - (1.96)(0.0925) = 0.20$$
 - Upper = (point estimate) + $(Z_{\alpha/2})$ (SE)

$$= 0.38 + (1.96)(0.0925) = 0.56$$
- 95% CI = (0.20, 0.56)

Hypothesis Testing

- **The majority of statistical analyses involve comparison, most obviously between treatments or procedures or between groups of subjects.**
- **Hypotheses are formulated, experiments are performed, and results are evaluated for their consistency (non-consistency) using with a hypothesis.**
- **Hypothesis Testing (HT) provides an objective framework for making decisions using probabilistic methods**

- The purpose of HT is to aid the clinician, researcher or administrator in reaching a decision (conclusion) concerning a population by examining a sample from that population.

Hypothesis

- Is a statement about one or more populations
- Is a claim (assumption) about a population parameter
- Is frequently concerned with the parameters of the population about which the statement is made.
- Is a formal scientific process that accounts for statistical uncertainty

- **Two types of hypothesis:**

- ❖ ***Research hypothesis:***

- Is the supposition or conjecture that motivates the research. It may be proposed after numerous repeated observation
- It leads directly to statistical hypotheses.

- ❖ ***Statistical hypothesis:***

- Stated in such a way that they can be evaluated by using appropriate statistical technique.

Examples of Research Hypotheses

Population Mean

- The average length of stay of patients admitted to the hospital is five days
- The mean birthweight of babies delivered by mothers with low SES is lower than those from higher SES.
- Etc

Population Proportion

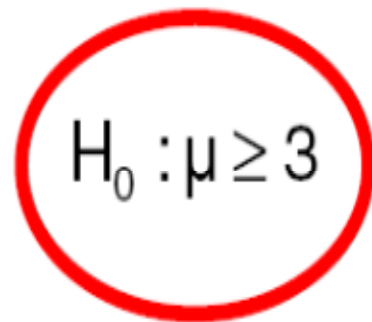
- The proportion of adult smokers in Addis Ababa is believed to be $p = 0.40$
- The prevalence of HIV among non-married adults is higher than that in married adults
- Etc

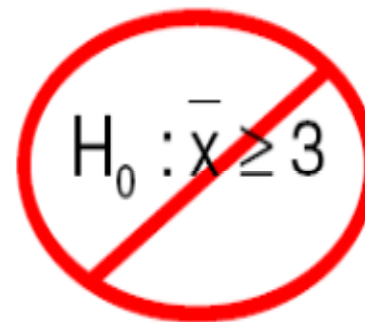
Types of Hypothesis (statistical)

1. The Null Hypothesis, H_0

- Is a statement claiming that there is no difference between the hypothesized value and the population value.
 - (The effect of interest is zero = no difference)
- States the assumption (hypothesis) to be tested

- H_0 is a statement of agreement (or no difference)
- H_0 is always about a population parameter, not about a sample statistic


$$H_0 : \mu \geq 3$$


$$H_0 : \bar{x} \geq 3$$

- Begin with the assumption that the H_0 is true
 - Similar to the notion of innocent until proven guilty
- Always contains “=” sign
- May or may not be rejected

2. The Alternative Hypothesis, H_A

- Is a statement of what we will believe is true if our sample data causes us to reject H_0 .
- Is generally the hypothesis that is believed (or needs to be supported) by the researcher.
- Is a statement that disagrees (opposes) with H_0
- (The effect of interest is not zero= difference)
- Never contains “=” sign
- May or may not be accepted

Steps in Hypothesis Testing

1. Formulate the appropriate statistical hypotheses clearly

- Specify H_0 and H_A

$$H_0: \mu = \mu_0$$

$$H_0: \mu = \mu_0$$

$$H_0: \mu = \mu_0$$

$$H_A: \mu \neq \mu_0$$

$$H_A: \mu > \mu_0$$

$$H_A: \mu < \mu_0$$

two-tailed

one-tailed

one-tailed

2. Set up a suitable significance level:

- The level of significance is the probability of rejecting the true null hypothesis.
- It indicates the level of significance that signifies the probability of computing type I error.

- It is usually denoted by α and should be specified before any samples are drawn.
- The level of significance is arbitrarily chosen small numbers usually 0.05, 0.01...

		Condition of Null Hypothesis	
		True	False
Possible Action	Fail to reject H_0	Correct action	Type II error
	Reject H_0	Type I error	Correct action

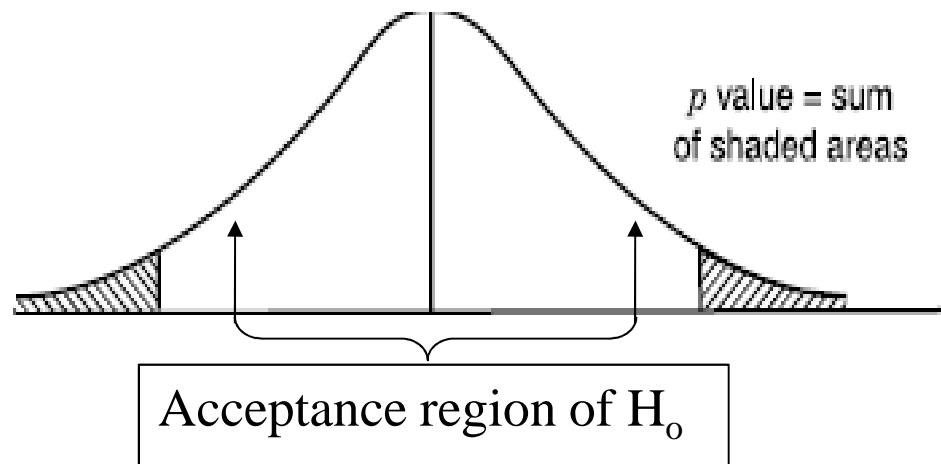
3. Decide on the appropriate test statistic for the hypothesis. E.g., One population

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

OR

$$t_{n-1} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

4. Determine the critical region: it is the area that indicates the rejection region of the hypothesis.



5. **Doing computation:** it is the right way of computing the test statistic and other results from the sample. Then we need to see whether sample result falls in the rejection region or in acceptance regions.

6. **Making decision:** finally we draw statistical conclusions. A statistical decision comprises either accepting the null hypothesis or rejecting it.

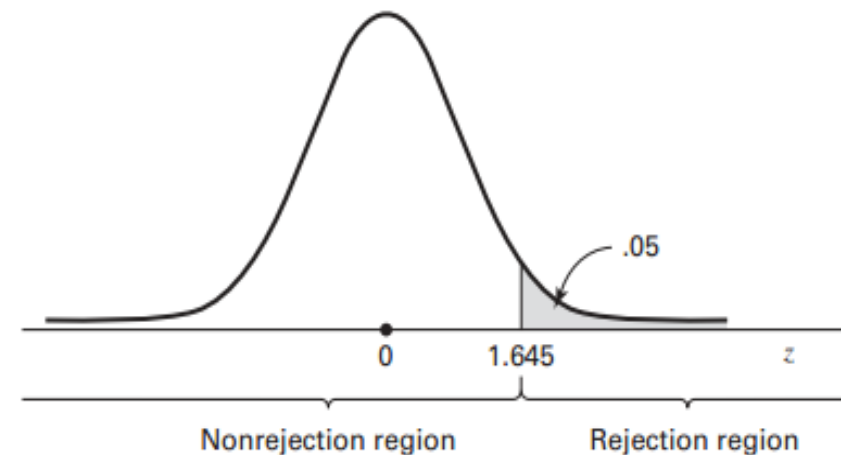
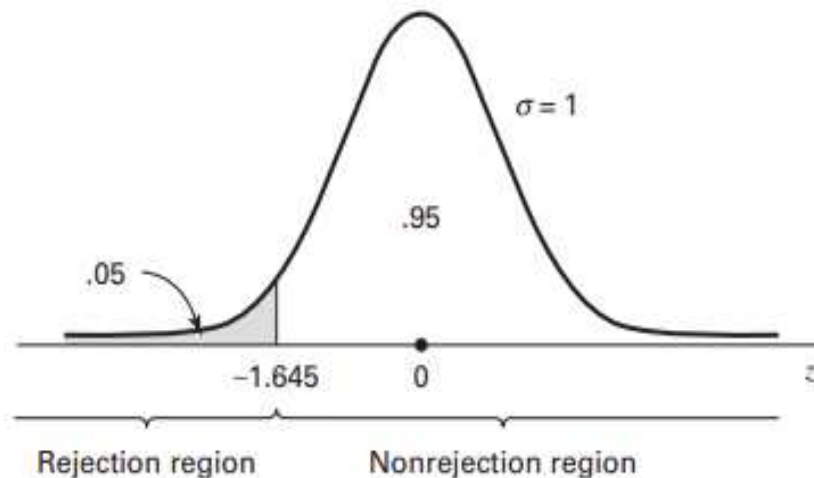
Remark:

1. from z or t table, if the calculated value is greater than tabulated value, the null hypothesis is rejected, i.e. the statistical results are significant.

2. The p-value

- p-Value is the probability of obtaining values of a test statistic as extreme as that observed if the null hypothesis is true.
- The p-value for a test of a hypothesis is the smallest value of α for which the null hypothesis is accepted or rejected.
- When p-value is below the cut off level (α), say 0.05, the result is called statistically significant; when above 0.05 it is called not significant.
- Reject H_0 if *P-value* $< \alpha$ or Accept H_0 if *P-value* $> \alpha$

- In a *one tail test*, the rejection region is at one end of the distribution or the other.
- **Decision rules** $Z_{\text{cal}} > Z_{\text{tab}}$ or $Z_{\text{cal}} < -Z_{\text{tab}}$ reject H_0
- In a *two tail test*, the rejection region is split between the two tails.
- **Decision rules** $|Z_{\text{cal}}| > Z_{\text{tab}}$ reject H_0
- Which one is used depends on the way the alternative hypothesis is written.
- The same is true for t-test also



Rules for Stating Statistical Hypotheses

1. One population

- Indication of equality (either $=$, \leq or \geq) must appear in H_0 .

$$H_0: \mu = \mu_0, \quad H_A: \mu \neq \mu_0$$

$$H_0: P = P_0, \quad H_A: P \neq P_0$$

- Can we conclude that a certain population mean is

– not 30?

$$H_0: \mu = 30 \quad \text{and} \quad H_A: \mu \neq 30$$

– greater than 50?

$$H_0: \mu = 50 \qquad H_A: \mu > 50$$

- Can we conclude that the proportion of patients with leukemia who survive more than six years is not 60%?

$$H_0: P = 0.6 \quad H_A: P \neq 0.6$$

2. Two populations

$$H_0: \mu_1 = \mu_2$$

$$H_A: \mu_1 \neq \mu_2$$

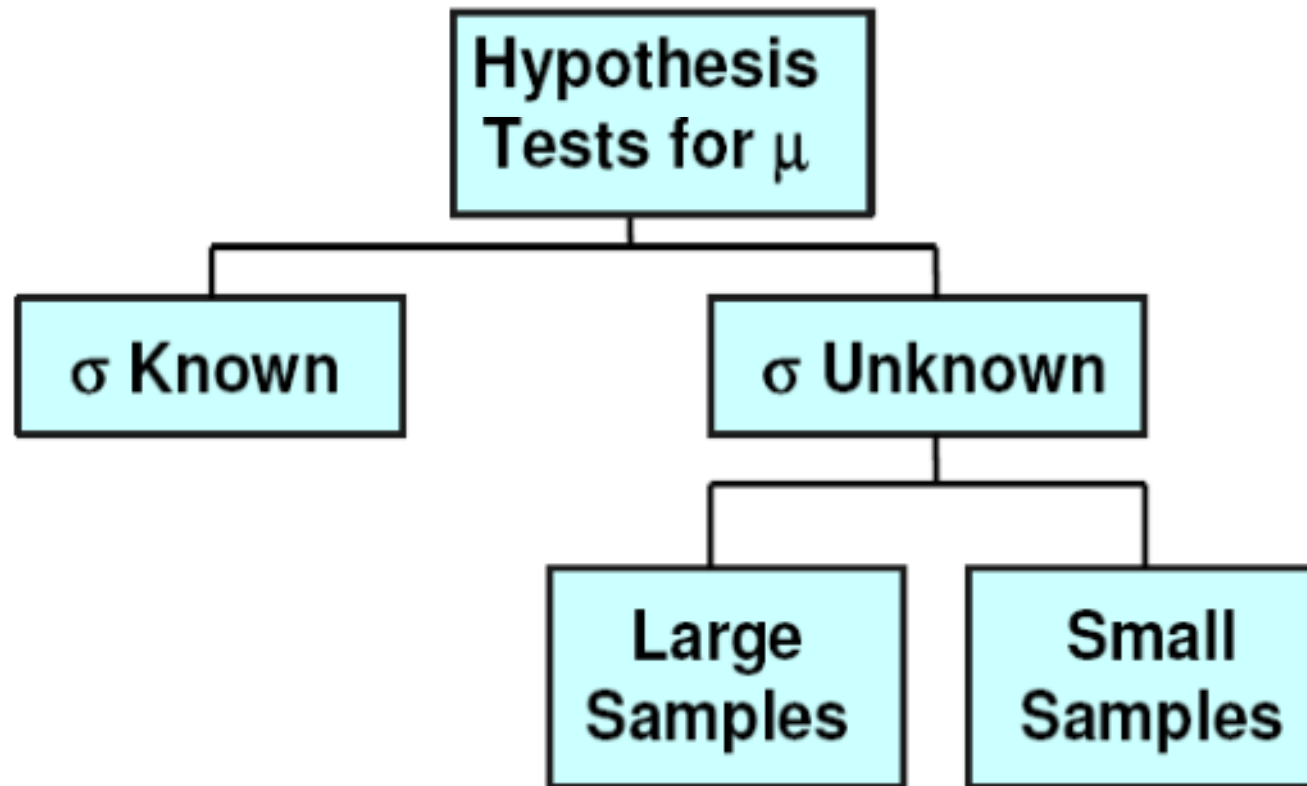
$$H_0: P_1 = P_2$$

$$H_A: P_1 \neq P_2$$

In summary,

1. What you hope to conclude should be placed in the H_A .
2. The H_0 should have a statement of equality, $=$.
3. The H_0 is the hypothesis that is tested
4. The H_0 and H_A are complementary.

1. Hypothesis Testing of a Single Mean (Normally Distributed)



1.1 Known Variance

The test statistic is:

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Example: Two-Tailed Test

1. A simple random sample of 10 people from a certain population has a mean age of 27. Can we conclude that the mean age of the population is not 30? The population variance is 20. Let $\alpha = .05$.

A. Data

$n = 10$, sample mean = 27, $\sigma^2 = 20$, $\alpha = 0.05$

B. Assumptions

Simple random sample

Normally distributed population

C. Hypotheses

$$H_0: \mu = 30$$

$$H_A: \mu \neq 30$$

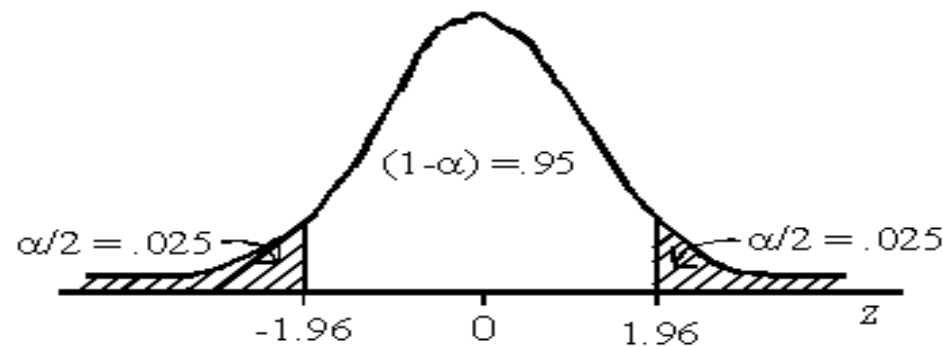
D. *Test statistic*

As the population variance is known, we use Z as the test statistic.

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$

E. Decision Rule

- Reject H_0 if the Z_{cal} value falls in the rejection region.
- Don't reject H_0 if the Z_{cal} value falls in the non-rejection region.
- Because of the structure of H_0 it is a two tail test. Therefore, reject H_0 if $Z_{cal} < -1.96$ or $Z_{cal} > 1.96$ or $|Z_{cal}| > 1.96$



F. Calculation of test statistic

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$
$$z = \frac{27 - 30}{\sqrt{20/10}} = \frac{-3}{1.4142} = -2.12$$

G. Statistical decision

We reject the H_0 because $Z = -2.12$ is in the rejection region ($-2.12 < -1.96$). The value is significant at 5%.

H. Conclusion

We conclude that μ is not 30. $P\text{-value} = 0.0340 < 0.05$

A Z value of -2.12 corresponds to an area of 0.0170. Since there are two parts to the rejection region in a two tail test, the P-value is twice this which is .0340.

Hypothesis test using confidence interval

- A problem like the above example can also be solved using a confidence interval.
- A confidence interval will show that the calculated value of Z does not fall within the boundaries of the interval. However, it will not give a probability.
- Confidence interval

$$\bar{x} \pm z \sigma \sqrt{n}$$

$$27 \pm 1.96 \sqrt{20/10}$$

$$27 \pm 1.96 (1.4142)$$

$$(24.228, 29.772)$$

Example: One -Tailed Test

- A simple random sample of 10 people from a certain population has a mean age of 27. Can we conclude that the mean age of the population is less than 30? The population variance is known to be 20. Let $\alpha = 0.05$.
- **Data**
 $n = 10$, sample mean = 27, $\sigma^2 = 20$, $\alpha = 0.05$
- **Hypotheses**
 $H_0: \mu = 30$, $H_A: \mu < 30$

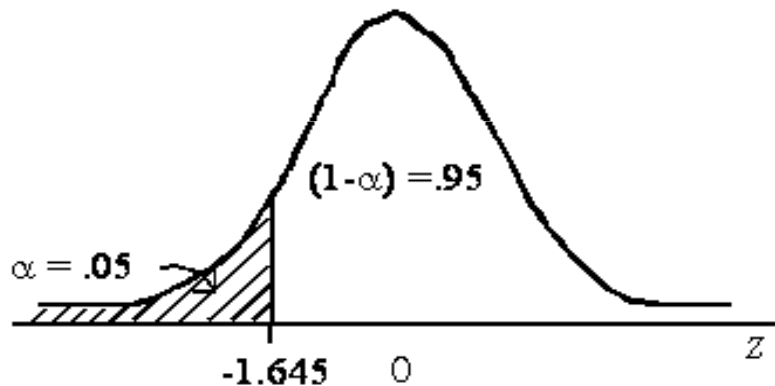
- **Test statistic**

$$z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

$$z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

$$z = \frac{27 - 30}{\sqrt{20/10}} = \frac{-3}{1.4142} = -2.12$$

- **Rejection Region**



Lower tail test

- *With $\alpha = 0.05$ and the inequality, we have the entire rejection region at the left. The critical value will be $Z_{tab} = -1.645$. Reject H_0 if $Z_{cal} < -1.645$.*

- **Statistical decision**
 - We reject the H_0 because $-2.12 < -1.645$.
- **Conclusion**
 - We conclude that $\mu < 30$.
 - $p = .0170$ this time because it is only a one tail test and not a two tail test.

1.2 Unknown Variance

- In most practical applications the standard deviation of the underlying population is not known
- In this case, σ can be estimated by the sample standard deviation s .
- If the underlying population is normally distributed and $n < 30$, then the test statistic is:

$$t_{n-1} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Example: Two-Tailed Test

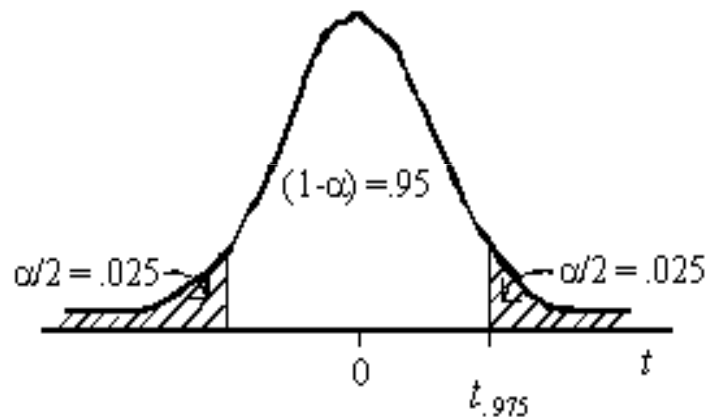
- A simple random sample of 14 people from a certain population gives a sample mean body mass index (BMI) of 30.5 and sd of 10.64. Can we conclude that the BMI is not 35 at α 5%?
- $H_0: \mu = 35, H_A: \mu \neq 35$
- **Test statistic**

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

- If the assumptions are correct and H_0 is true, the test statistic follows Student's t distribution with 13 degrees of freedom.

- **Decision rule**

- We have a two tailed test. With $\alpha = 0.05$ it means that each tail is 0.025. The critical t_{tab} values with 13 df are -2.1604 and 2.1604.
- We reject H_0 if the $t_{cal} < -2.1604$ or $t_{cal} > 2.1604$.



$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

$$t = \frac{30.5 - 35}{10.639/\sqrt{14}} = \frac{4.5}{2.8434} = -1.58$$

- Do not reject H_0 because -1.58 is not in the rejection region. Based on the data of the sample, it is possible that $\mu = 35$. *P-value* = 0.1375

Sampling from a population that is not normally distributed

- Here, we do not know if the population displays a normal distribution.
- However, with a large sample size, we know from the Central Limit Theorem that the sampling distribution of the population is distributed normally.
- With a large sample ($n \geq 30$), we can use Z as the test statistic calculated using the sample sd.

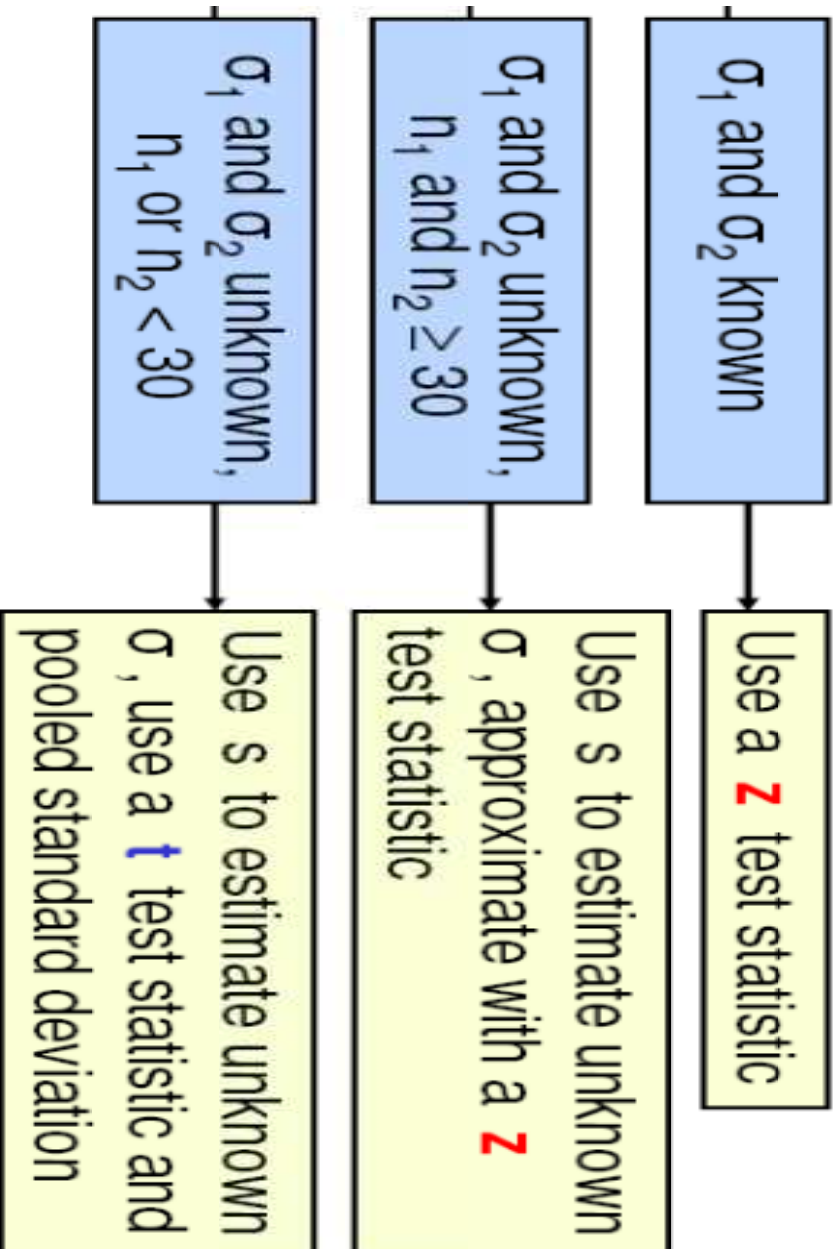
$$Z = \frac{\bar{X} - \mu_0}{s / \sqrt{n}}$$

2. Hypothesis Testing about the Difference Between Two Population Means

(Normally Distributed)

- When studying one-sample tests for a continuous random variable, the unknown mean μ of a single population was compared to some known value μ_0 .
- We are usually interested in comparing the means of two different populations when the values of both means are unknown

Two Sample Means, Independent Samples



2.1 Known Variances

(Independent Samples)

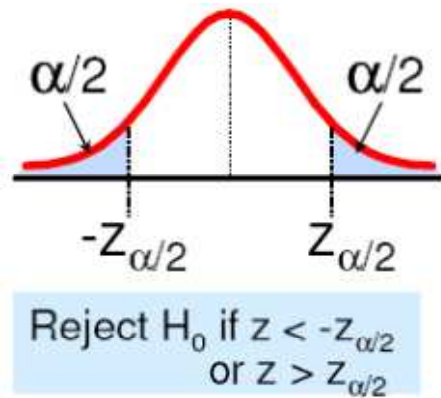
- When two independent samples are drawn from a normally distributed population with known variance, the test statistic for testing the H_0 of equal population means is:

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Example:

- Researchers wish to know a difference in mean serum uric acid (SUA) levels between normal individuals and individuals with Down's syndrome. The means SUA levels on 12 individuals with Down's syndrome and 15 normal individuals are 4.5 and 3.4 mg/100 ml, respectively. with variances. ($\sigma^2=1$, $\sigma^2=1.5$, respectively). Is there a difference between the means of both groups at α 5%?
- **Hypotheses:**
 $H_0: \mu_1 - \mu_2 = 0$ or $H_0: \mu_1 = \mu_2$
 $H_A: \mu_1 - \mu_2 \neq 0$ or $H_A: \mu_1 \neq \mu_2$

- With $\alpha = 0.05$, the critical values of Z_{tab} are -1.96 and +1.96. We reject H_0 if $Z_{\text{cal}} < -1.96$ or $Z_{\text{cal}} > +1.96$.



$$z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

$$z = \frac{(4.5 - 3.4) - 0}{\sqrt{1/12 + 1.5/15}} = \frac{1.1}{.4282} = 2.57$$

- Reject H_0 because $2.57 > 1.96$.
- From these data, it can be concluded that the population means are not equal. A 95% CI would give the same conclusion. *P-value* = 0.01.

2.2 Unknown Variances

i. Equal variances (Independent samples)

- With equal population variances, we can obtain a pooled value from the sample variances.
- The test statistic for $\mu_1 - \mu_2$ is:

• Where t has $(n_1 + n_2 - 2)$ df., and

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2[(1/n_1) + (1/n_2)]}} \quad s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Example:

- We wish to know if we may conclude, at the 95% confidence level, that smokers, in general, have greater lung damage than do non-smokers.

$$\text{Smokers: } \bar{x}_1 = 17.5 \quad n_1 = 16 \quad s_1^2 = 4.4752$$

$$\text{Non-Smokers: } \bar{x}_2 = 12.4 \quad n_2 = 9 \quad s_2^2 = 4.8492$$

$$\alpha = .05$$

- Calculation of Pooled Variance**

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$s_p^2 = \frac{(15)(4.4711) + (8)(4.8492)}{16 + 9 - 2}$$

$$s_p^2 = \frac{299.86 + 188.12}{23}$$

$$s_p^2 = 21.2165$$

- **Hypotheses:**

$$H_0: \mu_1 - \mu_2 = 0, \quad H_A: \mu_1 > \mu_2$$

- With $\alpha = 0.05$ and $df = 23$, the critical value of t_{tab} is 1.7139. We reject H_0 if $t_{cal} > 1.7139$.

- **Test statistic**

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{s_p^2/n_1 + s_p^2/n_2}}$$

$$t = \frac{(17.5 - 12.4) - 0}{\sqrt{21.2165/16 + 21.2165/9}} = \frac{5.1}{1.92} = 2.6563$$

- Reject H_0 because $2.6563 > 1.7139$. On the basis of the data, we conclude that $\mu_1 > \mu_2$.

ii. Unequal variances (Independent samples)

- We are still interested in testing

$$H_0 : \mu_1 = \mu_2 \text{ vs } H_A : \mu_1 \neq \mu_2$$

- The test statistic used is:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}}$$

- To compute a test statistic, we simply substitute s_1^2 for σ_1^2 and s_2^2 for σ_2^2 .

- Where the degree of freedom (d') is given by:

$$d' = \frac{\left[(s_1^2/n_1) + (s_2^2/n_2) \right]^2}{\left[\frac{(s_1^2/n_1)^2}{(n_1-1)} + \frac{(s_2^2/n_2)^2}{(n_2-1)} \right]}$$

Round d' down to the nearest integer — this is d''

- If $t_{cal} > t_{d'', \alpha/2}$ or $t_{cal} < -t_{d'', \alpha/2}$ then reject H_0 .

Example:

- Suppose we want to compare the characteristics of tuberculosis meningitis for patients infected with HIV and those not infected with HIV. In particular, we are interested in comparing age at diagnosis. A random sample of $n_1 = 37$ HIV infected patients has mean age at diagnosis $\bar{X}_1 = 27.9$ years and $s_1 = 5.6$ years. A sample of $n_2 = 19$ uninfected patients has mean age at diagnosis $\bar{X}_2 = 38.8$ years and $s_2 = 21.7$ years

- The test statistic is:

$$\begin{aligned} t &= \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}} \\ &= \frac{(27.9 - 38.8) - 0}{\sqrt{(5.6^2/37) + (21.7^2/19)}} \\ &= -2.15 \end{aligned}$$

- Note that

$$\begin{aligned}
 d' &= \frac{\left[(s_1^2/n_1) + (s_2^2/n_2) \right]^2}{\left[\frac{(s_1^2/n_1)^2}{(n_1-1)} + \frac{(s_2^2/n_2)^2}{(n_2-1)} \right]} \\
 &= \frac{\left[(5.6^2/37) + (21.7^2/19) \right]^2}{\left[\frac{(5.6^2/37)^2}{(37-1)} + \frac{(21.7^2/19)^2}{(19-1)} \right]} \\
 &= 19.24
 \end{aligned}$$

- And

$$d'' = 19$$

- For a t distribution with 19 df, the area to the left of $-2.15(t_{\text{cal}})$ is between 0.01 and 0.025
- Therefore, $0.02 < p < 0.05$ ($t_{\text{cal}}(-2.15) < -t_{19,0.025} (-2.093)$)
- For a test conducted at $\alpha = 0.05$, H_0 is rejected
- We conclude that among patients diagnosed with tuberculosis meningitis, those who are infected with HIV tend to be younger than those who are not

Sampling from populations that are not normally distributed

- In this situation, the results of the CLT may be employed if sample sizes are large (≥ 30).
- If the population variances are known, they are used; but if unknown, the sample variances based on large sample sizes are used as estimates.

The test statistic for
 $\mu_1 - \mu_2$ is:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Hypothesis Testing for Paired Samples

- Two samples are paired when each data point of the first sample is matched and is related to a unique data point of the second sample.
- Tests means of 2 related populations
 - Paired or matched samples
 - Repeated measures (before/after)
 - Longitudinal or follow-up study

- Assumptions:
 - Both populations are normally distributed
 - Or, if not normal, use large samples

The Paired t Test

The test statistic for \bar{d} is

$$t = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}}$$

Where $t_{\alpha/2}$ has $n - 1$ d.f.

and s_d is:

$$s_d = \sqrt{\frac{\sum_{i=1}^n (d_i - \bar{d})^2}{n-1}}$$

n is the number of pairs in the paired sample
 s_d = Sample standard deviation

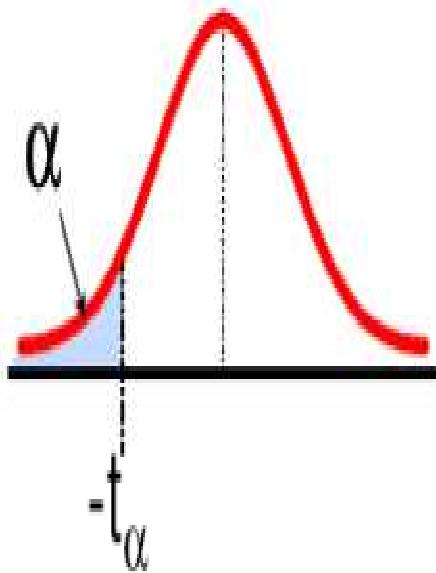
The i^{th} paired difference is d_i , where

$$d_i = x_{1i} - x_{2i}$$

The point estimate for the population mean paired difference is \bar{d} :

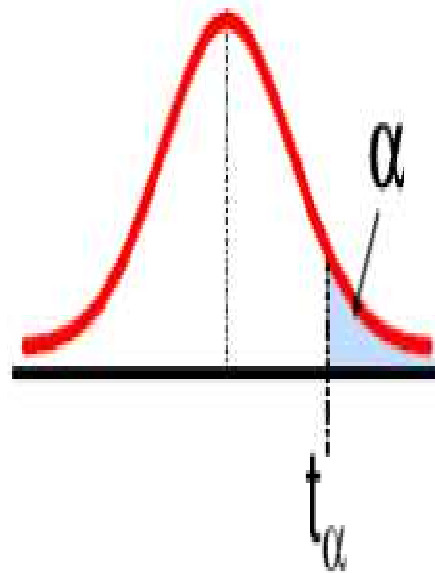
$$\bar{d} = \frac{\sum_{i=1}^n d_i}{n}$$

$$H_A: \mu_d < 0$$



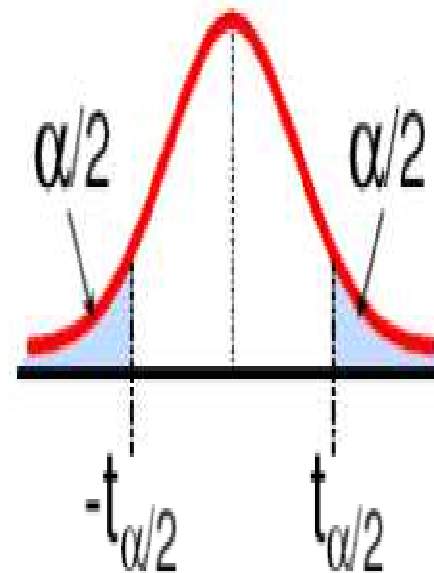
Reject H_0 if $t < -t_\alpha$

$$H_A: \mu_d > 0$$



Reject H_0 if $t > t_\alpha$

$$H_A: \mu_d \neq 0$$



Reject H_0 if $t < -t_{\alpha/2}$
or $t > t_{\alpha/2}$

Example:

- The following data show the SBP levels (mm Hg) in 10 women while not using (baseline) and while using (follow-up) oral contraceptives. Can we conclude that there is a difference between mean baseline and follow-up SBP at $\alpha 5\%$? $d_i = \text{baseline} - \text{follow-up}$

i	SBP (baseline)	SBP (follow-up)	d_i
1	115	128	13
2	112	115	3
3	107	106	-1
4	119	128	9
5	115	122	7
6	138	145	7
7	126	132	6
8	105	109	4
9	104	102	-2
10	115	117	2

$$\bar{d} = (13 + 3 + \dots + 2)/10 = 4.80$$

$$S_d^2 = [(13-4.8)^2 + \dots + (2-4.8)^2]/9 = 20.844$$

$$S_d = \sqrt{20.844} = 4.566$$

$$t_{cal} = 4.80/(4.566/\sqrt{10}) = 4.80/1.44 = 3.32$$

- From the Table, $t_{9, 0.025} = 2.262$
- Since $t_{cal} (= 3.32) > t_{9, \alpha/2} (2.262)$ H_0 is rejected
- P-value is between 0.001 and 0.01
- Since 3.32 falls in the rejection region, there is a significance difference between the population means SBP while not using and using OC use.

Hypothesis Tests for Proportions

- Involves categorical values
- Two possible outcomes
 - “Success” (possesses a certain characteristic)
 - “Failure” (does not possess that characteristic)
- Fraction or proportion of population in the “success” category is denoted by p

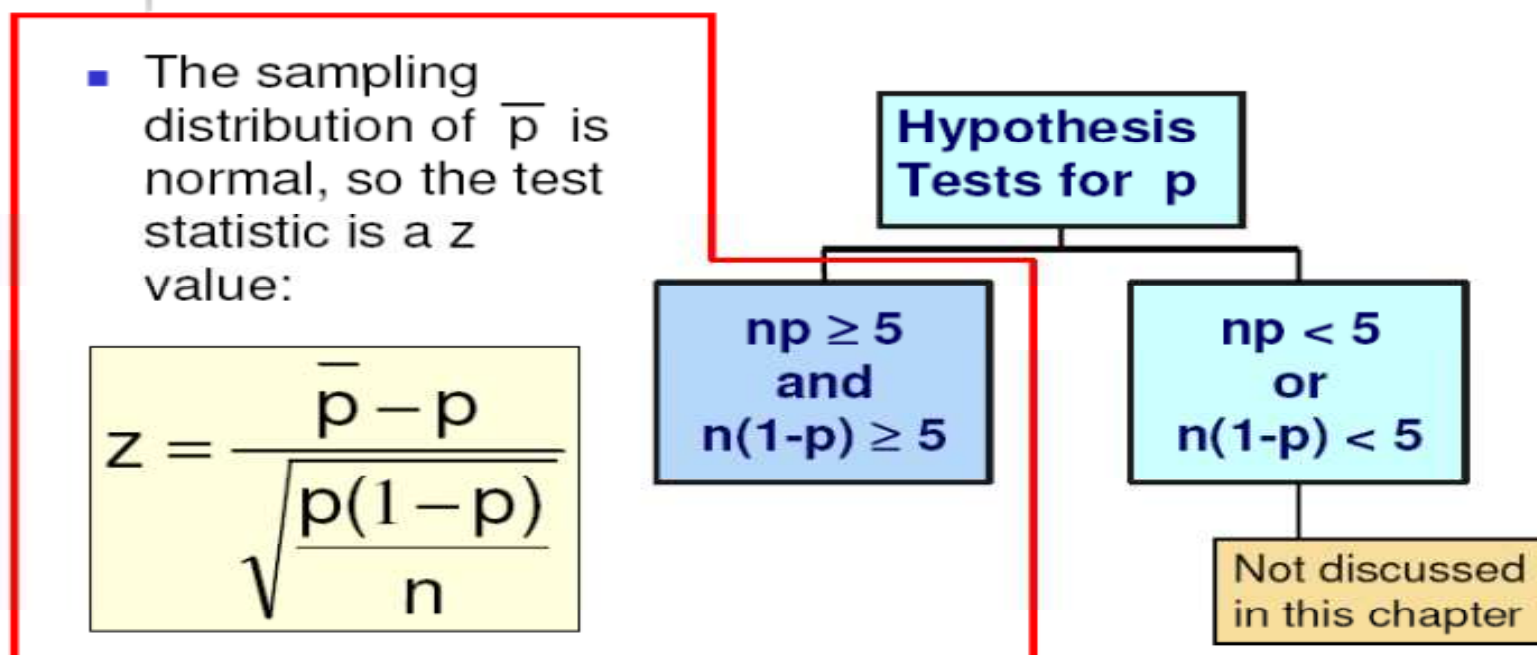
Proportions

- Sample proportion in the success category is denoted by \bar{p}

- $$\bar{p} = \frac{x}{n} = \frac{\text{number of successes in sample}}{\text{sample size}}$$

3. Hypothesis Testing about a Single Population Proportion

(Normal Approximation to Binomial Distribution)



- When both np and $n(1-p)$ are at least 5, \bar{p} can be approximated by a normal distribution with mean and standard deviation

■

$$\mu_{\bar{p}} = p$$

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}}$$

Example

- We are interested in the probability of developing asthma over a given one-year period for children 0 to 4 years of age whose mothers smoke in the home. In the general population of 0 to 4-year-olds, the annual incidence of asthma is 1.4%. If 10 cases of asthma are observed over a single year in a sample of 500 children whose mothers smoke, can we conclude that this is different from the underlying probability of $p_0 = 0.014$? $\alpha = 5\%$

$$H_0 : p = 0.014$$

$$H_A : p \neq 0.014$$

- The test statistic is given by:

$$z = \frac{\bar{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

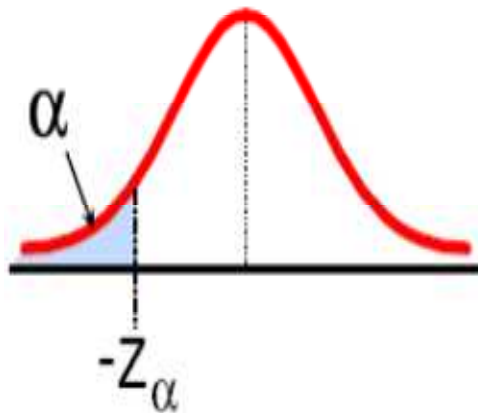
$$\begin{aligned}\hat{p} &= \frac{10}{500} \\ &= 0.02\end{aligned}$$

$$\begin{aligned}z &= \frac{0.02 - 0.014}{\sqrt{(0.014)(0.986)/500}} \\ &= 1.14\end{aligned}$$

- The critical value of $Z_{\alpha/2}$ at $\alpha=5\%$ is ± 1.96 .
- Don't reject H_0 since $Z (=1.14)$ in the non-rejection region between ± 1.96 .
- **P-value = 0.2542**
- We do not have sufficient evidence to conclude that the probability of developing asthma for children whose mothers smoke in the home is different from the probability in the general population

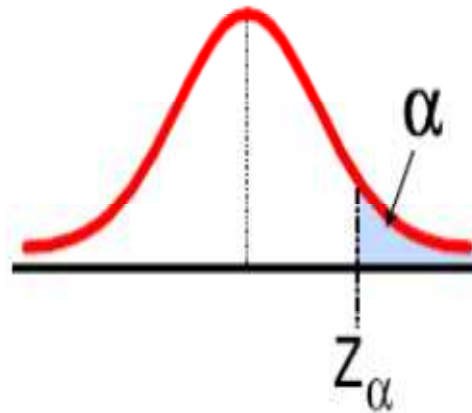
4. Hypothesis Tests about the Difference Between Two Population Proportions

$$H_A: p_1 - p_2 < 0$$



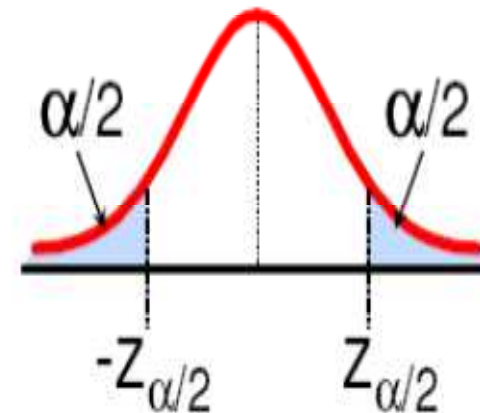
Reject H_0 if $z < -z_\alpha$

$$H_A: p_1 - p_2 > 0$$



Reject H_0 if $z > z_\alpha$

$$H_A: p_1 - p_2 \neq 0$$



Reject H_0 if $z < -z_{\alpha/2}$
or $z > z_{\alpha/2}$

Since we begin by assuming the null hypothesis is true, we assume $p_1 = p_2$ and pool the two \bar{p} estimates

The pooled estimate for the overall proportion is:

$$\bar{p} = \frac{n_1 \bar{p}_1 + n_2 \bar{p}_2}{n_1 + n_2} = \frac{X_1 + X_2}{n_1 + n_2}$$

Where X_1 = the observed number of events in the first sample and X_2 = the observed number of events in the second sample

The test statistic for
 $p_1 - p_2$ is:

$$z = \frac{(\bar{p}_1 - \bar{p}_2) - (p_1 - p_2)}{\sqrt{\bar{p}(1 - \bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Example

- A study was conducted to investigate the possible cause of gastroenteritis outbreak following a lunch served in a high school cafeteria. Among the 225 students who ate the sandwiches, 109 became ill. While, among the 38 students who did not eat the sandwiches, 4 became ill. Is there a significant difference between the two groups at $\alpha = 5\%$.

- We wish to test

$H_0: p_1 = p_2$ against the alternative

$H_A: p_1 \neq p_2$

$$\begin{aligned}
 \hat{p}_1 &= \frac{x_1}{n_1} \\
 &= \frac{109}{225} \\
 &= 48.4\%
 \end{aligned}$$

- Assume that the sample sizes are large enough, and the normal approximation to the binomial distribution is valid.
- If the H_0 is true, then $p_1 = p_2 = p$

$$\bar{p} = \frac{n_1 \bar{p}_1 + n_2 \bar{p}_2}{n_1 + n_2} = \frac{x_1 + x_2}{n_1 + n_2}$$

$$\begin{aligned}\bar{p} &= \frac{109 + 4}{225 + 38} \\ &= 0.430\end{aligned}\quad \begin{aligned}Z &= \frac{0.484 - 0.105}{\sqrt{(.430)(.570) \left(\frac{1}{225} + \frac{1}{38} \right)}} \\ &= 4.36\end{aligned}$$

The area under the standard normal curve to the right of 4.36 is less than 0.0001. Therefore, $p < 0.0002$. We reject H_0 at the 0.05 level. (4.36 > 1.96)

The proportion of students who became ill differs in the two groups; those who ate the prepared sandwiches were more likely to develop gastroenteritis.