BIO401

PPT Slide Full BOOK

Admin:

Chaudhary Moazzam

* Zarva Chaudhary*

* Laiba Mahi *

Applied Biostatistics

Topic 1

Applied Biostatistics

Introduction to Biostatistics

Introduction to Biostatistics

Statistics is a field of study that is concerned with:

- The collection, organization, summarization, and analysis of data.
- The drawing of inferences about a body of data, when only a part of data is observed.

Introduction to Biostatistics

- Statistics also plays a vital role in all the phases of research starting from the planning to the policy making.

Introduction to Biostatistics

BioStatistics is

- The use of statistical tools for the data that is derived from biological sciences and medicine.
- Bio-statistical methods were used in the ancient civilizations like Ancient Greece, Ancient Rome, India and China.

Introduction to Biostatistics

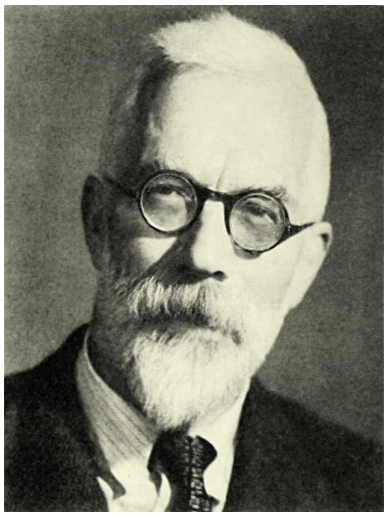
- One of the earliest known demographic / Bio-statistics study was conducted by John Graunt (1662). He developed life tables to warn off the onset and spread of Bubonic plague in London.

Introduction to Biostatistics

https://goo.gl/MVa1Jr

- The Lady with a Lamp (1820 – 1910)
- One of the early known Statistician who not only used statistical methodologies to detect the causes of deaths of British soldiers during Crimean War but also convinced Queen for evidence based policy making to provide better sanitary conditions for soldiers.

Introduction to Biostatistics



- Sir Ronald A. Fisher (1890-1962)
  - A Bio-Statistician
  - Evolutionary Biologist
  - Geneticist
- He developed Statistical Methodologies to combine Mendelian Genetics and natural selection.
- Famously known as "Father of Modern Statistical sciences".

https://goo.gl/XWdN6Z

Introduction to Biostatistics

Applications of Biostatistics:

- Public Health
- Pharmacology
- Epidemiology
- Medicine
- Genetics
- Genomics
- Proteomics
- Bioinformatics

etc etc

Introduction to Biostatistics

"To understand God's thoughts we must study statistics, for these are the measures of HIS purpose"

Florence Nightingale(1820-1910)

END

Applied Biostatistics

Topic # 2

Basic Terminology - I

Population:

Average person thinks of a population as collection of entities usually People.

In Statistics it is Defined as "an aggregate of entities for which we have an interest at a particular time." It can consists of Animals, machines, places or cells etc.

Basic Terminology - I

https://logofmaps.wordpress.com

Example:

If we are interested in knowing the weights of all the children enrolled in elementary schools of the Bahawalpur District.

Then our population will be all the the children enrolled in elementary schools of the Bahawalpur District.

Basic Terminology - I

Types of Population

1.Finite

If a population of values consists of a fixed number of these values, the population is said to be finite

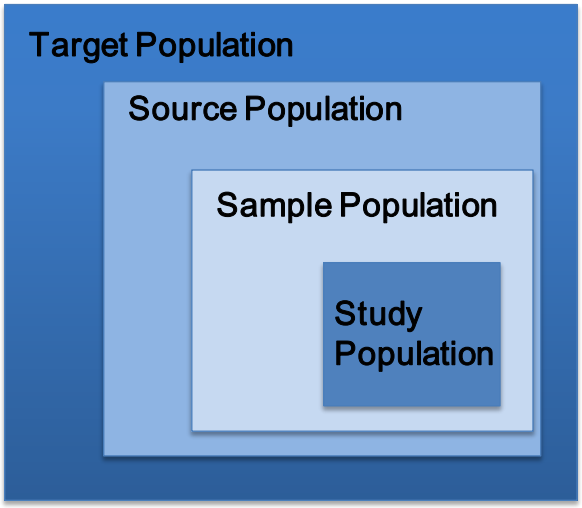
Basic Terminology - I

Types of Population

2.Infinite

If a population consists of an endless succession of values, the population is infinite one.

Basic Terminology - I



Target Population
The General Population that the study seeks to understand.

Source Population
The specific individuals from which a representative sample will be drawn.

Sample Population
Individuals asked to participate.

Study Population
Eligible participants.

Basic Terminology - I

Census

- A Census is a survey conducted on a full set of observations belonging to a given population.
- It is defined as “the complete enumeration of population of groups at a point in time with respect to well defined characteristics.

Basic Terminology - I

Population Parameter:

It is any summary number, like an average or percentage, that describes the entire population.

- These are the true values, which are usually unknown.

Basic Terminology - I

Population Parameter:

- These values are denoted by Greek letters. e.g.:
  - Population Mean μ (Mu)
  - Population Proportion π (Pi)

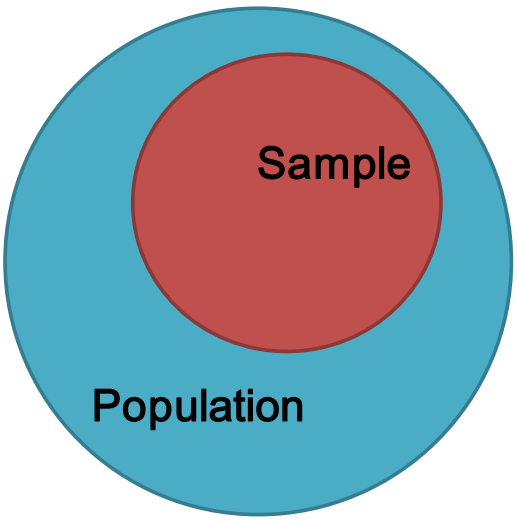
END

Applied Biostatistics

Basic Terminology - II

Basic Terminology - II

- It is not usually possible to study the whole population.
- Studying a population requires
  - A lot of time
  - resources



Basic Terminology - II

Sample is a representative part of a Population.

A sample survey is a study that obtains data from a subset of a population of our interest, in order to estimate population attributes.

Basic Terminology - II

Characteristics of a good sample

1. Representative.
2. It captures most of the variation in the population.
3. Obtained by using proper sampling methodology.

Basic Terminology - II

Characteristics of a good sample

4. Focus on objective
5. Informative with minimum use of resources.

Basic Terminology - II

Sample Statistic

Any summary number, like an average or percentage, that describes the sample.

- Sample statistics are random variables.
- These are the values which are usually denoted by Latin letters
  - Sample Standard Deviation (s)
  - Sample Proportion (p)

END

Applied Biostatistics

Topic # 4

Types of Statistics

Objectives of this course are twofold:

1. To Learn organize and summarize data.
2. To learn how to reach decisions about the large body of data by examining only a small part of it.

Types of Statistics

Types of Statistics

1. Descriptive Statistics
2. Inferential Statistics

Types of Statistics

Descriptive Statistics

These are the statistical methodologies which are used to Organize and Summarize data.

Together with simple graphics they form the basis for virtually every quantitative analysis of data.

Types of Statistics

Descriptive Statistics

Descriptive analysis is also known as Exploratory Data Analysis (EDA)

Types of Statistics

Inferential Statistics

These are statistical methodologies using which we reach a conclusion about a population on the basis of the information contained in the sample.

END

Applied Biostatistics

Topic 4

Data

The raw material of Statistics is data.
We may define data as figures. Figures result from the process of counting or from taking a measurement.

For example:

- When a hospital administrator counts the number of patients (counting).
- When a nurse weighs a patient (measurement)

Sources of Data:

We search for suitable data to serve as the raw material for our investigation.
Such data are available from one or more of the following sources:

1- Routinely kept records.

For example:

- Hospital medical records contain immense amounts of information on patients.
-Hospital accounting records contain a wealth of data on the facility's business
- activities.

2- External sources.

The data needed to answer a question may already exist in the form of published reports, commercially available data banks, or the research literature, i.e. someone else has already asked the same question.

3- Surveys:

The source may be a survey, if the data needed is about answering certain questions.

For example:

If the administrator of a clinic wishes to obtain information regarding the mode of transportation used by patients to visit the clinic,

then a survey may be conducted among patients to obtain this information.

4- Experiments.

Frequently the data needed to answer a question are available only as the result of an experiment.

For example:

If a nurse wishes to know which of several strategies is best for maximizing patient compliance,

she might conduct an experiment in which the different strategies of motivating compliance are tried with different patients.

Applied Biostatistics

Topic 5

Variables

It is a characteristic that takes on different values in different persons, places, or things.

For example:

- heart rate,
- the heights of adult males,
- the weights of preschool children,
- the ages of patients seen in a dental clinic.

Types of Variables

Quantitative Variables

It can be measured in the usual sense.

For example:

- the heights of adult males,
- the weights of preschool children,
- the ages of patients seen in a
- dental clinic.

Qualitative Variables

Many characteristics are not capable of being measured. Some of them can be ordered or ranked.

For example:

- classification of people into socio-economic groups,
- social classes based on income, education, etc.

Types of Quantitative Variables

A discrete variable

is characterized by gaps or interruptions in the values that it can assume.

For example:

- The number of daily admissions to a general hospital,
- The number of decayed, missing or filled teeth per child
- in an
- elementary
- school.

A continuous variable

can assume any value within a specified relevant interval of values assumed by the variable.

For example:

- Height,
- weight,
- skull circumference.

No matter how close together the observed heights of two people, we can find another person whose height falls somewhere in between.

Applied Biostatistics

Topic 6

Measurement Scales - I

- ***A Statistic:***

It is a descriptive measure computed from the data of a **sample**.

- ***A Parameter:***

It is a a descriptive measure computed from the data of a **population**.

Since it is difficult to measure a parameter from the population, a sample is drawn of size n, whose values are $\chi_1, \chi_2, \ldots, \chi_n$. From this data, we measure the statistic.

A measure of central tendency is a measure which indicates where the middle of the data is.

The three most commonly used measures of central tendency are:

The Mean, the Median, and the Mode.

The Mean:

It is the average of the data.

The Population Mean:

$\mu =$ which is usually unknown, then we use the

sample mean to estimate or approximate it.

The Sample Mean:

$=$

Example:

Here is a random sample of size 10 of ages, where

$\chi_1 = 42$, $\chi_2 = 28$, $\chi_3 = 28$, $\chi_4 = 61$, $\chi_5 = 31$,

$\chi_6 = 23$, $\chi_7 = 50$, $\chi_8 = 34$, $\chi_9 = 32$, $\chi_{10} = 37$.

$= (42 + 28 + \ldots + 37) / 10 = 36.6$

Properties of the Mean:

- **Uniqueness.** For a given set of data there is one and only one mean.
- **Simplicity.** It is easy to understand and to compute.
- **Affected by extreme values.** Since all values enter into the computation.

Example: Assume the values are 115, 110, 119, 117, 121 and 126. The mean = 118.

But assume that the values are 75, 75, 80, 80 and 280. The mean = 118, a value that is not representative of the set of data as a whole.

The Median:

When ordering the data, it is the observation that divide the set of observations into two equal parts such that half of the data are before it and the other are after it.

* If n is odd, the median will be the middle of observations. It will be the $(n+1)/2^{th}$ ordered observation.

When n = 11, then the median is the 6^{th} observation.

* If n is even, there are two middle observations. The median will be the mean of these two middle observations. It will be the $(n+1)/2^{th}$ ordered observation.

When n = 12, then the median is the 6.5^{th} observation, which is an observation halfway between the 6^{th} and 7^{th} ordered observation.

Example:

For the same random sample, the ordered observations will be as:

23, 28, 28, 31, 32, 34, 37, 42, 50, 61.

Since n = 10, then the median is the 5.5^{th} observation, i.e. = (32+34)/2 = 33.

Properties of the Median:

- **Uniqueness.** For a given set of data there is one and only one median.
- **Simplicity.** It is easy to calculate.
- **It is not affected by extreme values** as is the mean.

The Mode:

It is the value which occurs most **frequently.**

If all values are different there is **no mode**.

Sometimes, there are **more than one mode**.

Example:

For the same random sample, the value 28 is repeated two times, so it is the mode.

Properties of the Mode:

- Sometimes, it is not **unique.**
- It may be used for **describing qualitative data.**

Applied Biostatistics

Topic 8

Types of Statistics

Descriptive statistics are methods for organizing and summarizing data.
For example, tables or graphs are used to organize data, and descriptive values such as the average score are used to summarize data.
A descriptive value for a population is called a **parameter** and a descriptive value for a sample is called a **statistic**.

Inferential statistics are methods for using sample data to make general conclusions (inferences) about populations.

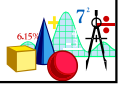
Because a sample is typically only a part of the whole population, sample data provide only limited information about the population. As a result, sample statistics are generally imperfect representatives of the corresponding population parameters.

Inference

Use a random sample to learn something about a larger population

Inference

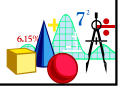
- **Two ways to make inference**
  - Estimation of parameters
    - * Point Estimation ($\bar{X}$ or p)
    - * Intervals Estimation
  - Hypothesis Testing



Statistic **Parameter**

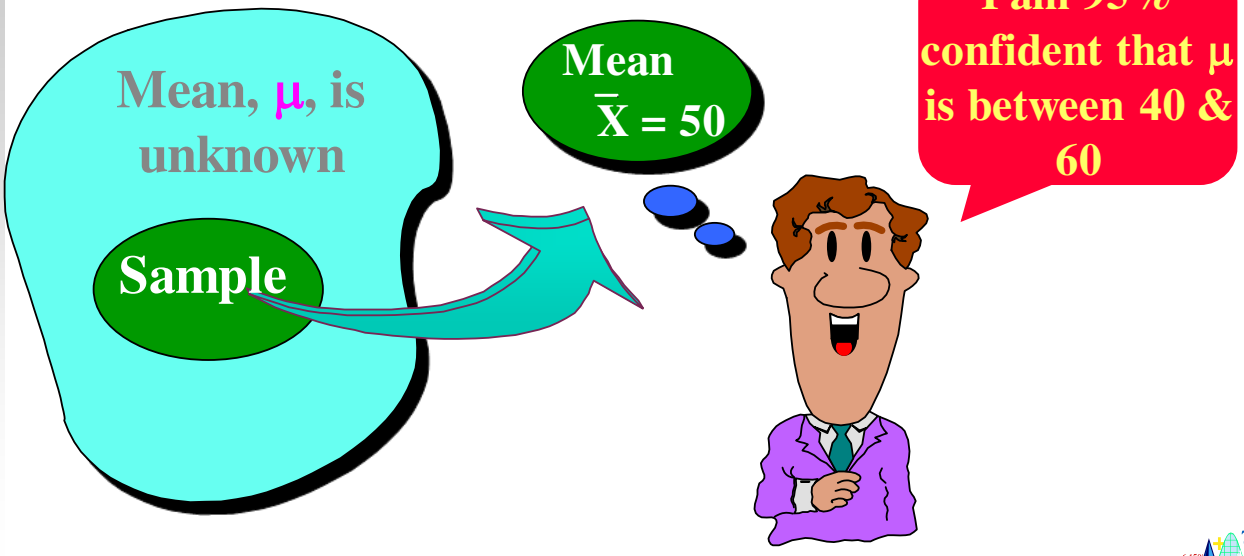
Mean:	$\bar{x}$	**estimates**	μ
Standard deviation:	**s**	**estimates**	σ
Proportion:	**p**	**estimates**	π

from sample **from entire population**



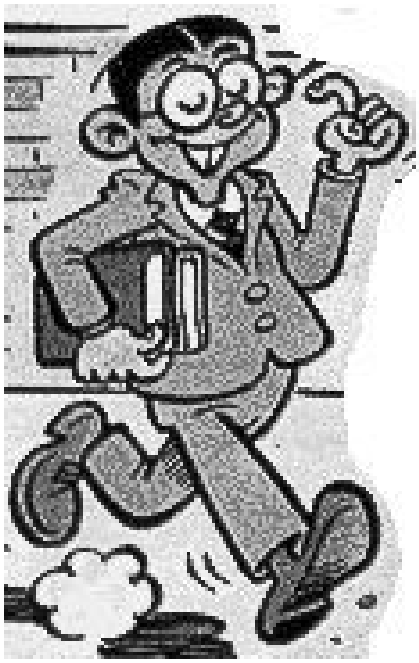
Estimation of parameters

Population Point estimate Interval estimate

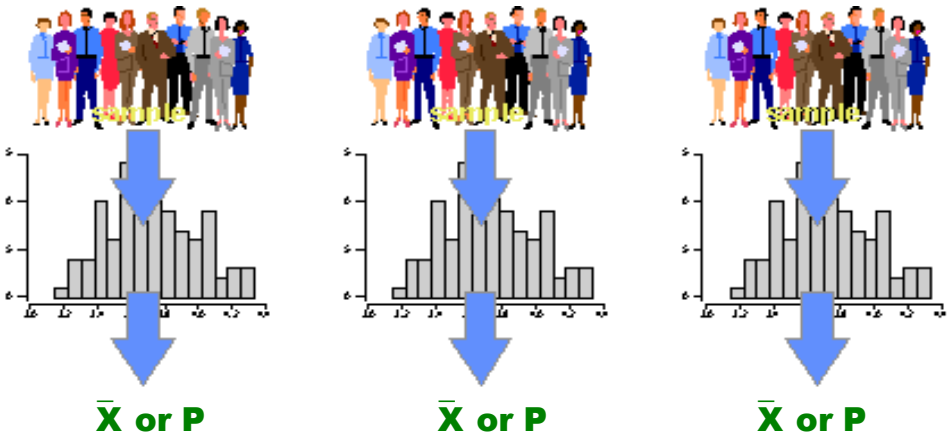


Parameter

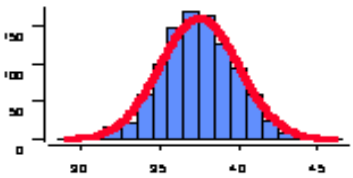
= Statistic ± Its Error



Sampling Distribution



The Sampling Distribution...



...is the distribution of a statistic across an infinite number of samples

Standard Error

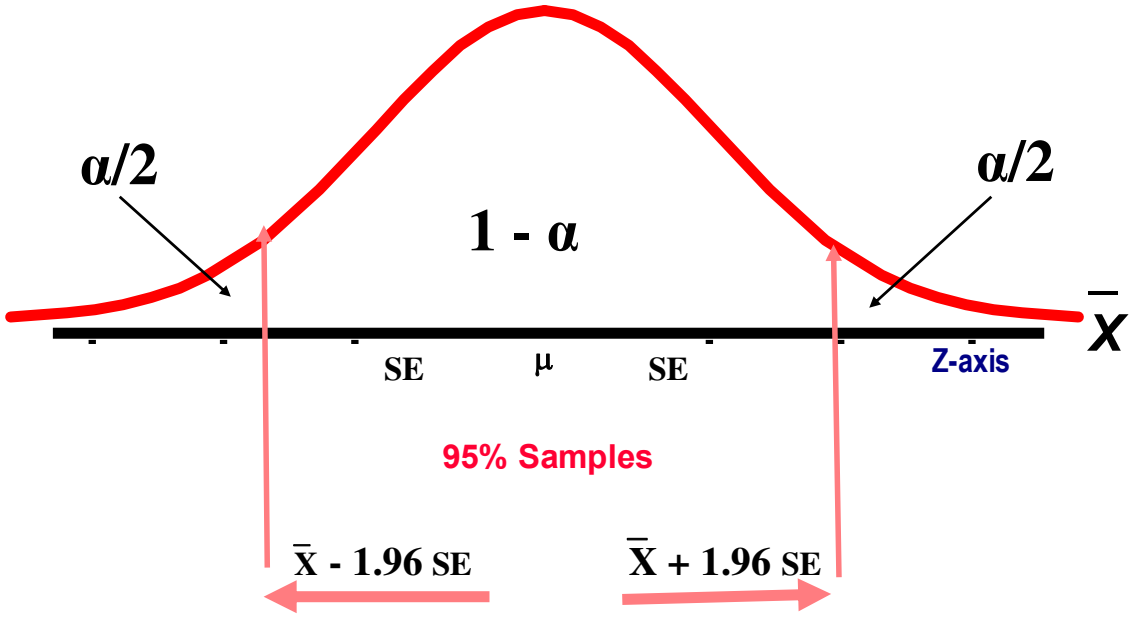
Quantitative Variable

$$SE\ (Mean) = \frac{S}{\sqrt{n}}$$

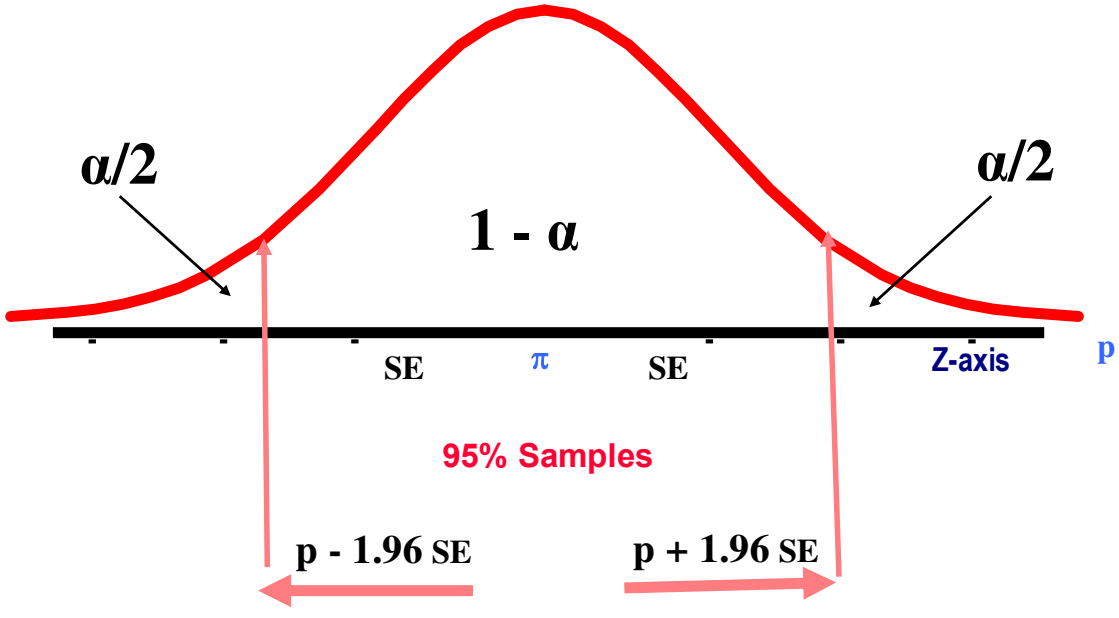
Qualitative Variable

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

Confidence Interval



Confidence Interval



Interpretation of CI

Probabilistic

In repeated sampling 100(1-□)% of all intervals around sample means will in the long run include □

Practical

We are 100(1-□)% confident that the single computed CI contains □

Example (Sample size≥30)

An epidemiologist studied the blood glucose level of a random sample of 100 patients. The mean was 170, with a SD of 10.

$SE = 10/10 = 1$

$\mu = \bar{X} \pm Z \times SE$

95%

Then CI:

$\mu = 170 \pm 1.96 \times 1 \quad 168.04 \leq \mu \geq 171.96$

Example (Proportion)

In a survey of 140 asthmatics, 35% had allergy to house dust. Construct the 95% CI for the population proportion.

$$\pi = p \pm Z\sqrt{\frac{P(1-p)}{n}} \qquad SE = \sqrt{\frac{0.35(1-0.35)}{140}} = 0.04$$

$$0.35 - 1.96 \times 0.04 \leq \pi \geq 0.35 + 1.96 \times 0.04$$

$$0.27 \leq \pi \geq 0.43$$

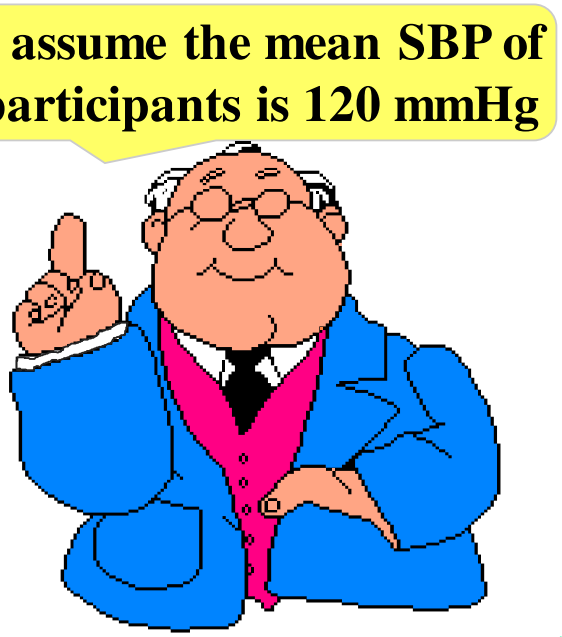
$$27\% \leq \pi \geq 43\%$$

Hypothesis testing

A statistical method that uses sample data to evaluate a hypothesis about a population parameter. It is intended to help researchers differentiate between real and random patterns in the data.

What is a Hypothesis?

An assumption about the population parameter.



Null & Alternative Hypotheses

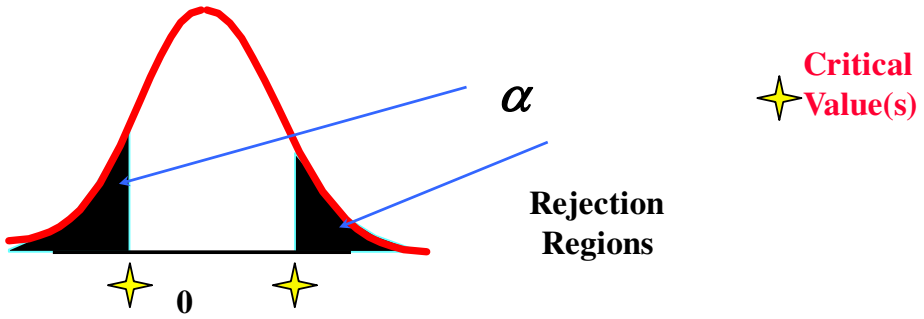
- H_0 Null Hypothesis states the Assumption to be tested e.g. SBP of participants = 120 (H_0: $\mu = 120$).

- H_1 Alternative Hypothesis is the opposite of the null hypothesis (SBP of participants ≠ 120 (H_1: $\mu \neq 120$). It may or may not be accepted and it is the hypothesis that is believed to be true by the researcher

Level of Significance, α

- Defines unlikely values of sample statistic if null hypothesis is true. Called rejection region of sampling distribution
- Typical values are 0.01, 0.05
- Selected by the Researcher at the Start
- Provides the Critical Value(s) of the Test

Level of Significance, a and the Rejection Region



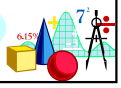
Result Possibilities

H_0: Innocent

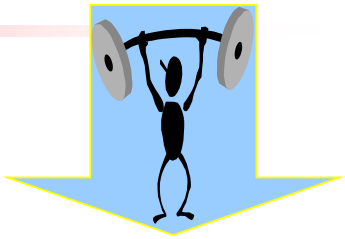
Jury Trial			Hypothesis Test		
	Actual Situation			Actual Situation	
Verdict	Innocent	Guilty	Decision	H_0 True	H_0 False
Innocent	Correct	Error	Accept H_0	$1 - \alpha$	Type II Error (β)
Guilty	Error	Correct	Reject H_0	Type I Error (α)	Power ($1 - \beta$)

False Positive

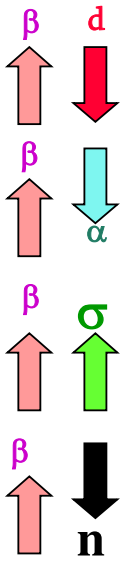
False Negative

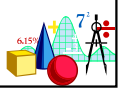


Factors Increasing Type II Error



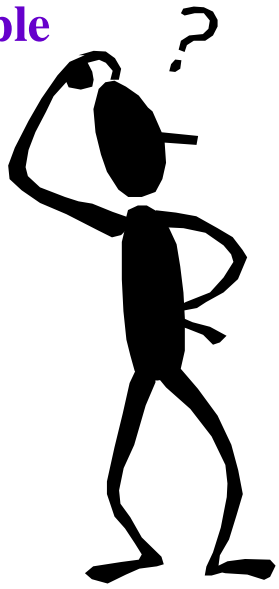
- True Value of Population Parameter
  - Increases When Difference Between Hypothesized Parameter & True Value Decreases
- Significance Level α
  - Increases When α Decreases
- Population Standard Deviation σ
  - Increases When σ Increases
- Sample Size n
  - Increases When n Decreases





p Value Test

- Probability of Obtaining a Test Statistic More Extreme ($\leq$ or $\geq$) than Actual Sample Value Given H_0 Is True
- Called Observed Level of Significance
- Used to Make Rejection Decision
  - If *p* value $\geq \alpha$, Do Not Reject H_0
  - If *p* value $< \alpha$, Reject H_0



Hypothesis Testing: Steps

Test the Assumption that the true mean SBP of participants is 120 mmHg.

State H_0	$H_0: \mu = 120$
State H_1	$H_1: \mu \neq 120$
Choose α	$\alpha = 0.05$
Choose *n*	n = 100
Choose Test:	*Z, t, X^2 Test (or p Value)*

Hypothesis Testing: Steps

Compute Test Statistic ***(or compute P value)***

Search for Critical Value

Make Statistical Decision rule

Express Decision

One sample-mean Test

- **Assumptions**
  - **Population is normally distributed**

- **t test statistic**

$$t = \frac{\text{sample mean} - \text{null value}}{\text{standard error}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

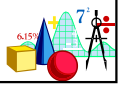
Example Normal Body Temperature

What is **normal body temperature**? Is it actually 37.6°C (on average)?

State the null and alternative hypotheses

H_0: $\mu = 37.6$°C

H_a: $\mu \neq 37.6$°C



Example Normal Body Temp (cont)

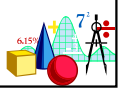
Data: random sample of $n = 18$ normal body temps

37.2	**36.8**	**38.0**	**37.6**	**37.2**	**36.8**	**37.4**	**38.7**	**37.2**
36.4	**36.6**	**37.4**	**37.0**	**38.2**	**37.6**	**36.1**	**36.2**	**37.5**

Summarize data with a test statistic

Variable	n	Mean	SD	SE	t	P
Temperature	18	37.22	0.68	0.161	2.38	0.029

$$t = \frac{\text{sample mean} - \text{null value}}{\text{standard error}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$



STUDENT'S t DISTRIBUTION TABLE

Degrees of freedom	Probability (p value)		
	0.10	0.05	0.01
1	6.314	12.706	63.657
5	2.015	2.571	4.032
10	1.813	2.228	3.169
17	1.740	2.110	2.898
20	1.725	2.086	2.845
24	1.711	2.064	2.797
25	1.708	2.060	2.787
∞	1.645	1.960	2.576

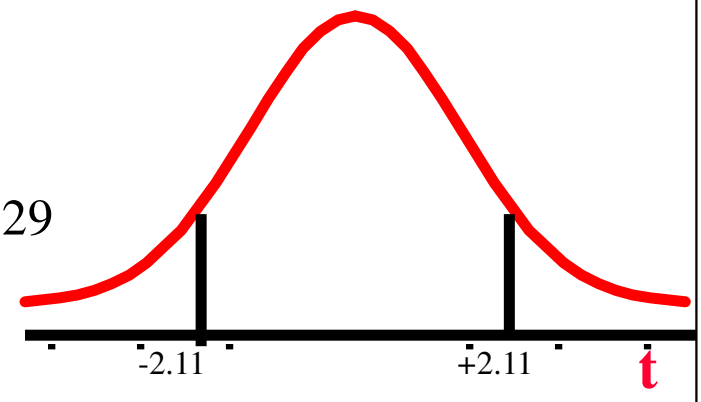
Example Normal Body Temp (cont)

Find the *p*-value

Df = n – 1 = 18 – 1 = 17

From SPSS: *p*-value = 0.029

From t Table: *p*-value is between 0.05 and 0.01.



Area to left of t = -2.11 equals area to right of t = +2.11.

The value t = 2.38 is between column headings 2.110& 2.898 in table, and for df =17, the *p*-values are 0.05 and 0.01.

Example Normal Body Temp (cont)

Decide whether or not the result is statistically significant based on the *p*-value

Using $\alpha = 0.05$ as the level of significance criterion, the results are **statistically significant** because 0.029 is less than 0.05. In other words, we can reject the null hypothesis.

Report the Conclusion

We can conclude, based on these data, that the mean temperature in the human population does not equal 37.6.

One-sample test for proportion

- Involves categorical variables
- Fraction or % of population in a category
- Sample proportion (p)
- Test is called Z test where:
- Z is computed value
- π is proportion in population (null hypothesis value)

$$p = \frac{X}{n} = \frac{number\ of\ successes}{sample\ size}$$

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

Critical Values: 1.96 at α=0.05

2.58 at α=0.01

Example

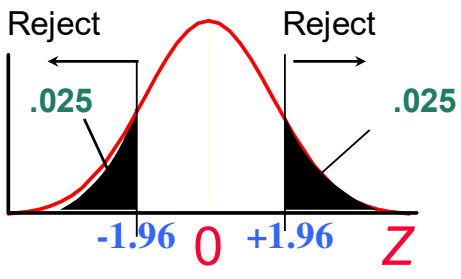
- In a survey of diabetics in a large city, it was found that 100 out of 400 have diabetic foot. Can we conclude that 20 percent of diabetics in the sampled population have diabetic foot.
- Test at the α =0.05 significance level.

Solution

H_o: $\pi = 0.20$

H_1: $\pi \neq 0.20$

$$Z = \frac{0.25 - 0.20}{\sqrt{\frac{0.20\,(1 - 0.20)}{400}}} = 2.50$$



Critical Value: 1.96

Decision:

We have sufficient evidence to reject the Ho value of 20%

We conclude that in the population of diabetic the proportion who have diabetic foot does not equal 0.20

Probability Sampling

Types of Probability Sampling Designs

- Simple random sampling
- Stratified sampling
- Systematic sampling
- Cluster (area) sampling
- Multistage sampling

Some Definitions

- N = the number of cases in the sampling frame
- n = the number of cases in the sample
- ${}_{N}C_{n}$ = the number of combinations (subsets) of n from N
- f = n/N = the sampling fraction

Simple Random Sampling

- Objective: Select n units out of N such that every ${}_{N}C_{n}$ has an equal chance.
- Procedure: Use table of random numbers, computer random number generator or mechanical device.
- Can sample with or without replacement.
- f=n/N is the sampling fraction.

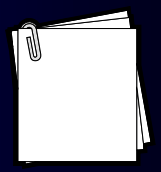
Simple Random Sampling

Example:

- Small service agency.
- Client assessment of quality of service.
- Get list of clients over past year.
- Draw a simple random sample of n/N.

Simple Random Sampling

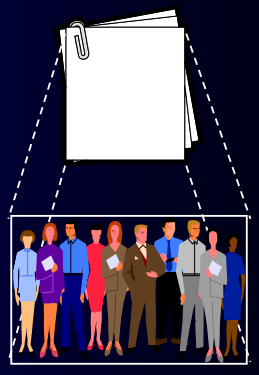
List of clients



Simple Random Sampling

List of clients

Random subsample



Stratified Random Sampling

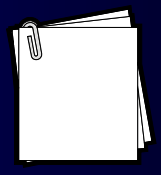
- Sometimes called "proportional" or "quota" random sampling.
- Objective: Population of N units divided into nonoverlapping strata N_1, N_2, N_3, ... N_i such that $N_1 + N_2 + ... + N_i = N$; then do simple random sample of n/N in each strata.

Stratified Sampling - Purposes:

- To insure representation of each strata, oversample smaller population groups.
- Administrative convenience -- field offices.
- Sampling problems may differ in each strata.
- Increase precision (lower variance) if strata are homogeneous within (like blocking).

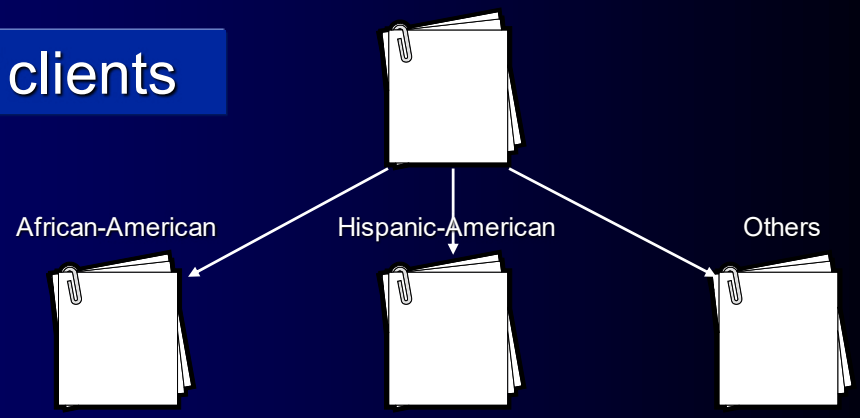
Stratified Random Sampling

List of clients



Stratified Random Sampling

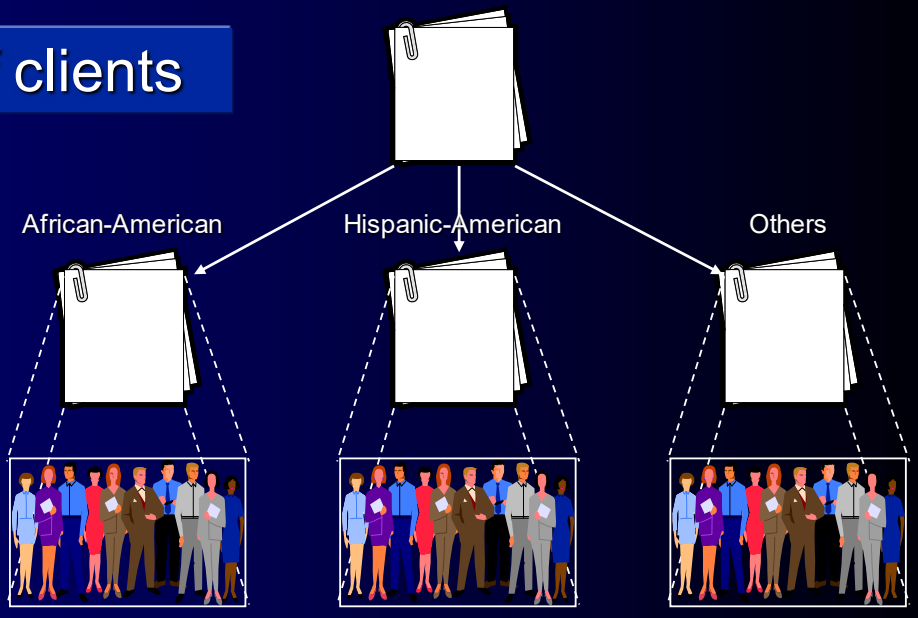
List of clients



Strata

Stratified Random Sampling

List of clients



Strata

Random subsamples of n/N

Proportionate vs. Disproportionate Stratified Random Sampling

- Proportionate: If sampling fraction is equal for each stratum
- Disproportionate: Unequal sampling fraction in each stratum
- Needed to enable better representation of smaller (minority groups)

Systematic Random Sampling

Procedure:

- Number units in population from 1 to N.
- Decide on the n that you want or need.
- N/n=k the interval size.
- Randomly select a number from 1 to k.
- Take every kth unit.

Systematic Random Sampling

- Assumes that the population is *randomly* ordered.
- Advantages: Easy; may be more precise than simple random sample.
- Example: The library (ACM) study.

Systematic Random Sampling

N = 100

1	26	51	76
2	27	52	77
3	28	53	78
4	29	54	79
5	30	55	80
6	31	56	81
7	32	57	82
8	33	58	83
9	34	59	84
10	35	60	85
11	36	61	86
12	37	62	87
13	38	63	88
14	39	64	89
15	40	65	90
16	41	66	91
17	42	67	92
18	43	68	93
19	44	69	94
20	45	70	95
21	46	71	96
22	47	72	97
23	48	73	98
24	49	74	99
25	50	75	100

Systematic Random Sampling

N = 100

Want n = 20

1	26	51	76
2	27	52	77
3	28	53	78
4	29	54	79
5	30	55	80
6	31	56	81
7	32	57	82
8	33	58	83
9	34	59	84
10	35	60	85
11	36	61	86
12	37	62	87
13	38	63	88
14	39	64	89
15	40	65	90
16	41	66	91
17	42	67	92
18	43	68	93
19	44	69	94
20	45	70	95
21	46	71	96
22	47	72	97
23	48	73	98
24	49	74	99
25	50	75	100

Systematic Random Sampling

N = 100

want n = 20

N/n = 5

1	26	51	76
2	27	52	77
3	28	53	78
4	29	54	79
5	30	55	80
6	31	56	81
7	32	57	82
8	33	58	83
9	34	59	84
10	35	60	85
11	36	61	86
12	37	62	87
13	38	63	88
14	39	64	89
15	40	65	90
16	41	66	91
17	42	67	92
18	43	68	93
19	44	69	94
20	45	70	95
21	46	71	96
22	47	72	97
23	48	73	98
24	49	74	99
25	50	75	100

Systematic Random Sampling

N = 100

Want n = 20

N/n = 5

Select a random number from 1-5: chose 4

1	26	51	76
2	27	52	77
3	28	53	78
4	29	54	79
5	30	55	80
6	31	56	81
7	32	57	82
8	33	58	83
9	34	59	84
10	35	60	85
11	36	61	86
12	37	62	87
13	38	63	88
14	39	64	89
15	40	65	90
16	41	66	91
17	42	67	92
18	43	68	93
19	44	69	94
20	45	70	95
21	46	71	96
22	47	72	97
23	48	73	98
24	49	74	99
25	50	75	100

Systematic Random Sampling

N = 100

Want n = 20

N/n = 5

Select a random number from 1-5: chose 4

Start with #4 and take every 5th unit

1	26	51	76
2	27	52	77
3	28	53	78
4	29	54	79
5	30	55	80
6	31	56	81
7	32	57	82
8	33	58	83
9	34	59	84
10	35	60	85
11	36	61	86
12	37	62	87
13	38	63	88
14	39	64	89
15	40	65	90
16	41	66	91
17	42	67	92
18	43	68	93
19	44	69	94
20	45	70	95
21	46	71	96
22	47	72	97
23	48	73	98
24	49	74	99
25	50	75	100

Cluster (Area) Random Sampling

Procedure:

- Divide population into clusters.
- Randomly sample clusters.
- Measure all units within sampled clusters.

Cluster (Area) Random Sampling

- Advantages: Administratively useful, especially when you have a wide geographic area to cover.

- Examples: Randomly sample from city blocks and measure all homes in selected blocks.

Multi-Stage Sampling

- Cluster (area) random sampling can be multi-stage.
- Any combinations of single-stage methods.

Multi-Stage Sampling

Example: Choosing students from schools

- Select all schools; then *sample* within schools.
- Sample schools; then measure *all* students.
- Sample schools; then *sample* students.

NON-PROBABILITY SAMPLING

Sampling

- Measuring a small portion of something and then making a general statement about the whole thing.
- Process of selecting a number of units for a study in such a way that the units represent the larger group from which they are selected.

Why We Need Sampling?

- Sampling makes possible the study of a large, (different characteristics) population.
- Sampling is for economy
- Sampling is for speed.
- Sampling is for accuracy.
- Sampling saves the sources of data from being all consumed.

General Types of Sampling

1. **Probability sampling**
2. **Non-probability sampling**

Non-probability sampling

- Unequal chance of being included in the sample (non-random)
- Non random or non - probability sampling refers to the sampling process in which, the samples are selected for a specific purpose with a pre-determined basis of selection.
- The sample is not a proportion of the population and there is no system in selecting the sample. The selection depends upon the situation.
- No assurance is given that each item has a chance of being included as a sample
- There is an assumption that there is an even distribution of characteristics within the population, believing that any sample would be representative.

Types of Non-Probability Sampling

- Judgment or purposive or deliberate sampling
- Convenience sampling
- Quota sampling
- Snow Ball Sampling

1. Judgment or purposive or deliberate sampling

- In this method, the sample selection is purely based on the judgment of the investigator or the researcher. This is because, the researcher may lack information regarding the population from which he has to collect the sample. Population characteristics or qualities may not be known, but sample has to be selected.

- In this method of sampling the choice of sample items depends primarily on the judgment of the researcher. In other words, the researcher determines and includes those items in the sample which he thinks are most typical of the universe with regard to the characteristics of research project.

- For example, suppose 100 boys are to be selected from a college with 1000 boys. If nothing is known about the students in this college, then the investigator may visit the college and choose the first 100 boys he meets. Or he may select 100 boys all belonging to III Year. Or he might select 25 boys from Commerce course, 25 from Science courses, 25 boys from Arts courses and 25 from Fine arts courses. Hence, when only the sample size is known, the investigator uses his discretion and select the sample.

The use of judgment sampling is justified by following premises:

- If there are a small number of sampling units is in the universe, judgment sampling enables inclusion of important units.
- Judgment stratification of population helps in obtaining a more representative sample in case research study wants to look into unknown traits of the population.
- Judgment sampling is a practical method to arrive at some solution to everyday business problems.

Limitations:

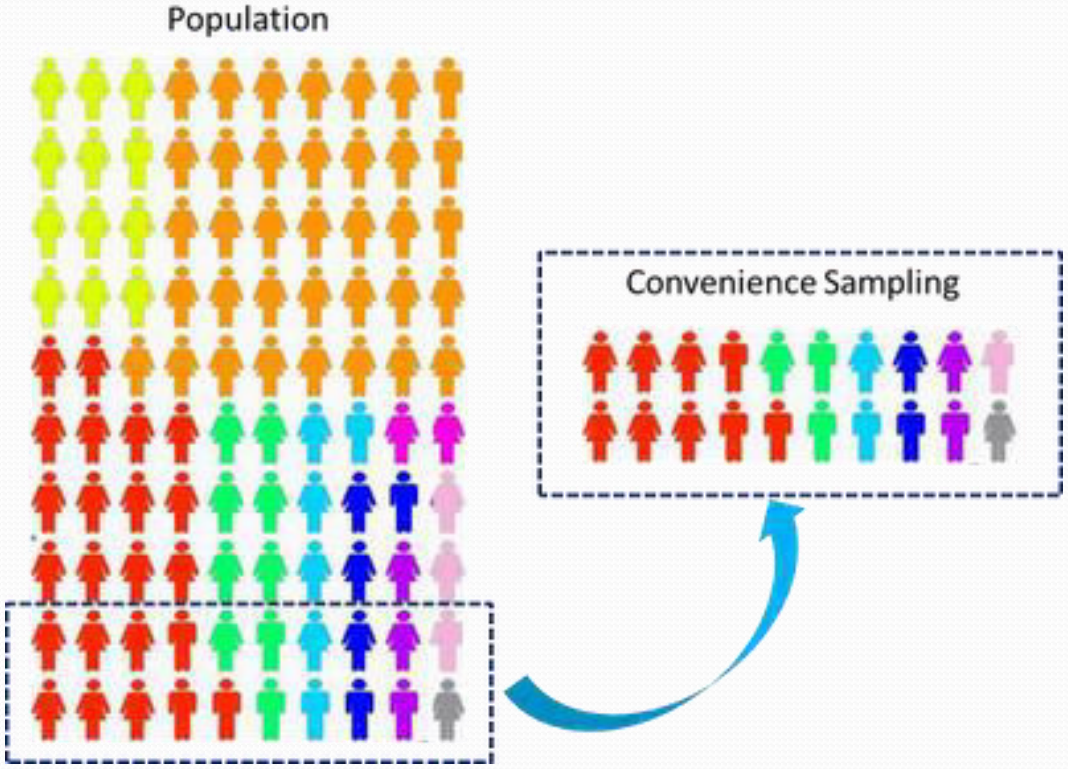
- The judgment sampling involves the risk that the researcher may establish conclusions by including those items in the sample which conform to his preconceived ideas.
- There is no objective way of evaluating the reliability of sample results.

2. Convenience sampling

- Convenience sampling is commonly known as unsystematic, accidental or opportunistic sampling. According to this procedure a sample is selected according to the convenience of the investigator.
- In this method of sampling the choice of sample items depends primarily on the judgment of the researcher. In other words, the researcher determines and includes those items in the sample which he thinks are most typical of the universe with regard to the characteristics of research project.
- A type of non probability sampling which involves the sample being drawn from that part of the population which is close to hand. That is, readily available and convenient.
- For example, suppose 100 car owners are to be selected. Then we may collect from the RTO's office the list of car owners and then make a selection of 100 from that to form the sample.

A convenience sampling may be used in the following cases:

- i) When universe is not well defined,
- ii) When sampling unit is not clear, and
- iii) When complete list of the source is not available.



3. QUOTA SAMPLING

- In this method, the sample size is determined first and then quota is fixed for various categories of population, which is followed while selecting the sample.
- In this method the quota has to be determined in advance and intimated to the investigator. The quota for each segment of the population may be fixed at random or with a specific basis. Normally such a sampling method does not ensure representativeness of the population.
- Example: - Suppose we want to select 100 students, then we might say that the sample should be according to the quota given below : Boys 50%, Girls 50% Then among the boys, 20% college students, 40% plus two students, 30% high school students and 10% elementary school students. A different or the same quota may be fixed for the girls.

MERITS OF QUOTA SAMPLING

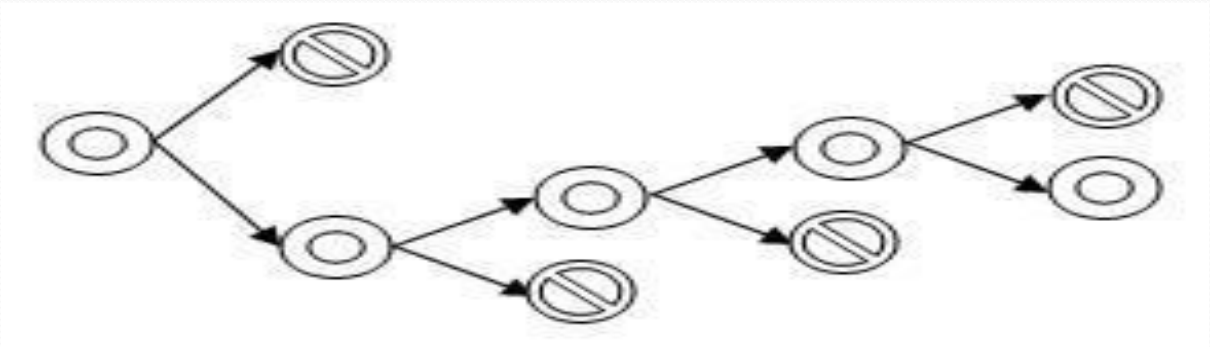
- Reduces cost of preparing sample and field work, since ultimate units can be selected so that they are close together.
- Introduces some stratification effect.

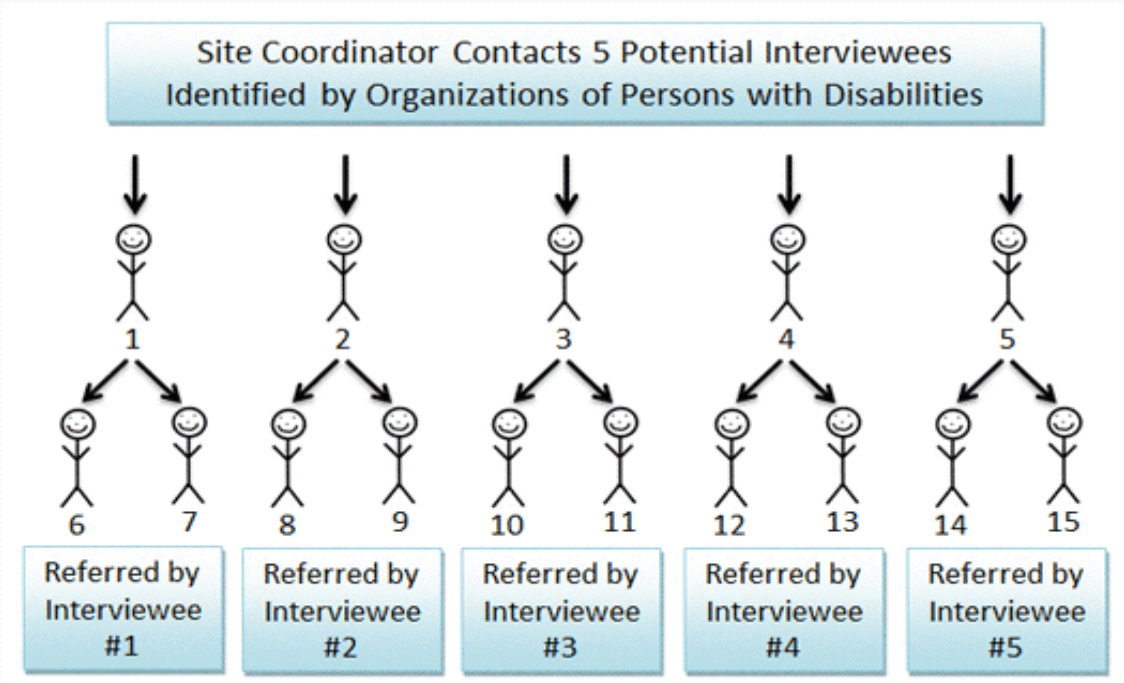
DEMERITS OF QUOTA SAMPLING

- Introduces bias of investigator is not involved at any stage, the errors of the method cannot be estimated by statistical procedures.
- Since random sampling is not involved at any stage, the errors of the method cannot be estimated by statistical procedures. Quota sampling is most commonly used in marketing survey and election polls.

4. SNOWBALL SAMPLING

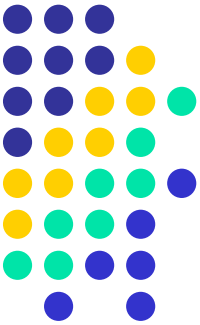
- It refers to Identifying someone who meets the criteria for inclusion in the study.
- Selection of additional respondents is based on referrals from the initial respondents.





Describing Data

Charts and Graphs



Lecture Objectives

You should be able to:

1. Define **Basic Terms**
2. Recognize **Types of Data** and **Data Scales**
3. **Draw appropriate graphs** based on type of data and type of analysis desired.
4. **Interpret** the graphs

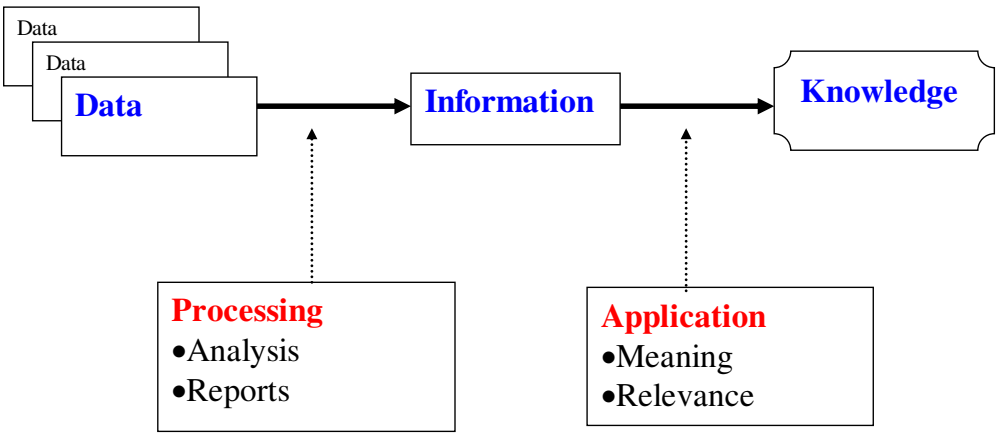
Basic Terms

1. Data, Information, and Knowledge
2. Populations and Samples
3. Variables and Observations

Types of Data:

1. Categorical and Numerical
2. Cross Sectional and Time Ordered

Data, Information, and Knowledge



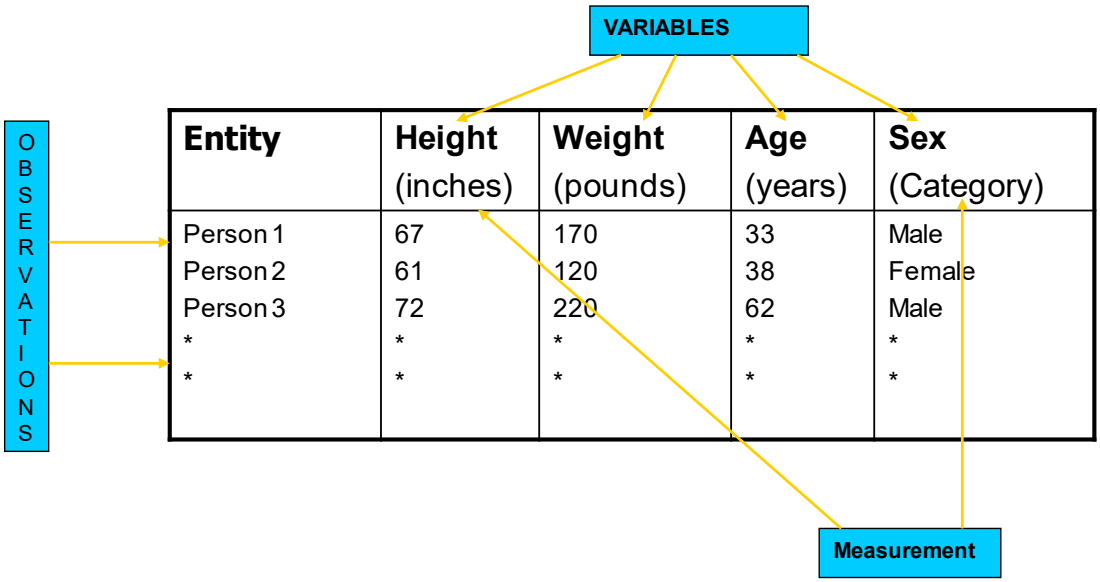
Populations and Samples

Population: Collection of all possible entities of interest
Described by Parameters

Sample: Subset of collection
Described by Statistics

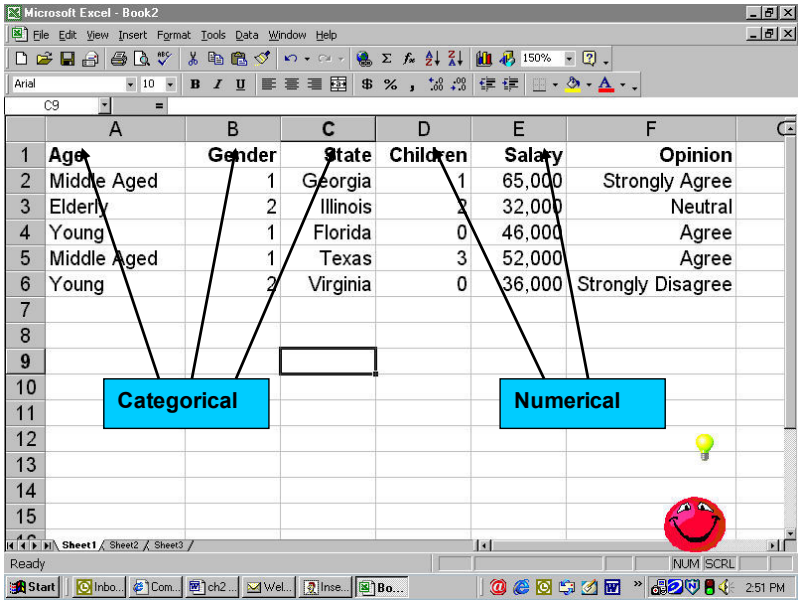
Statistical Inference
Art and science of using samples to make conclusions about populations.

Variables and Observations



Entity	Height (inches)	Weight (pounds)	Age (years)	Sex (Category)
Person 1	67	170	33	Male
Person 2	61	120	38	Female
Person 3	72	220	62	Male
*	*	*	*	*
*	*	*	*	*

Types of Data: Categorical and Numerical



	A	B	C	D	E	F
1	Age	Gender	State	Children	Salary	Opinion
2	Middle Aged	1	Georgia	1	65,000	Strongly Agree
3	Elderly	2	Illinois	2	32,000	Neutral
4	Young	1	Florida	0	46,000	Agree
5	Middle Aged	1	Texas	3	52,000	Agree
6	Young	2	Virginia	0	36,000	Strongly Disagree

Data Scales

Data are generally classified into four types:

1. **Nominal** – Categorical data
2. **Ordinal** – shows ranks, intervals may vary
3. **Interval** – intervals are constant, arbitrary 0
4. **Ratio** – Numeric data with a 'real' 0 value.

Ordinal, Interval and Ratio scales are all Numeric data.

Types of Data: Time Series and Cross-sectional

Month	Population (Millions)
1900	56
1910	58
1920	60
1930	65
1940	76
1950	84
1960	95
1970	120

Variable(s) over time

	1970		
	Population	GDP	Gender
Country	(Millions)	$ Billion	Ratio
USA	160	575	0.998
China	800	155	1.105
India	600		
Nigeria	100		
Japan	120		
Canada	30		

Variable(s) at one point in time across multiple entities (countries in this case)

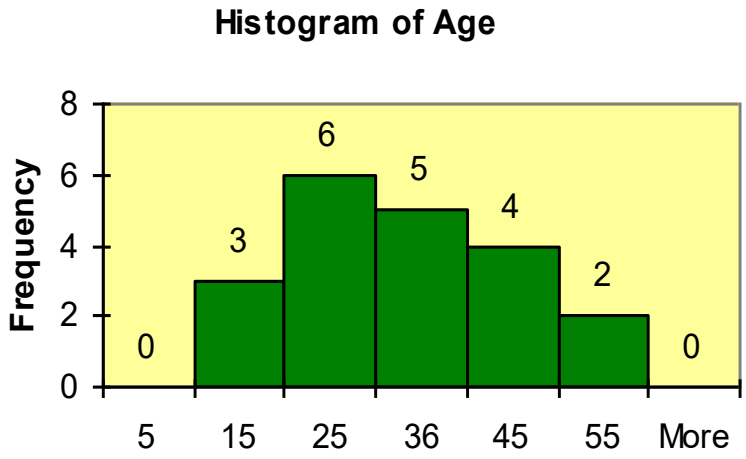
Numeric Data (Interval or Ratio): Frequency Tables

A Frequency Table showing a classification of the AGE of attendees at an event.

Class	Frequency	Relative Frequency	Percent
10 to 20	3	0.15	15
20 to 30	6	0.30	30
30 to 40	5	0.25	25
40 to 50	4	0.20	20
50 to 60	2	0.10	10
	20	**1.00**	**100**

Frequency Histograms

A graphical display of distribution of frequencies



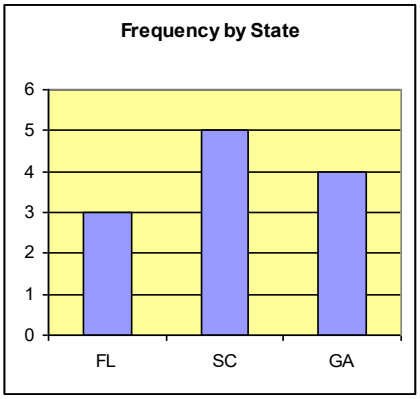
Developing Frequency Tables and Histograms

1. **Sort Raw Data** in Ascending Order:
   12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35,
   37, 38, 41, 43, 44, 46, 53, 58
2. **Find Range**: 58 - 12 = 46
3. **Select Number of Classes**: 5 (usually between 5 and 15)
4. **Compute Class Interval** (width): 10 (range/classes = 46/5 then round up)
5. **Determine Class Boundaries** (limits): 10, 20, 30, 40, 50
6. **Compute Class Midpoints**: 15, 25, 35, 45, 55
7. **Count Observations & Assign to Classes**

Categorical Data: Bar Charts

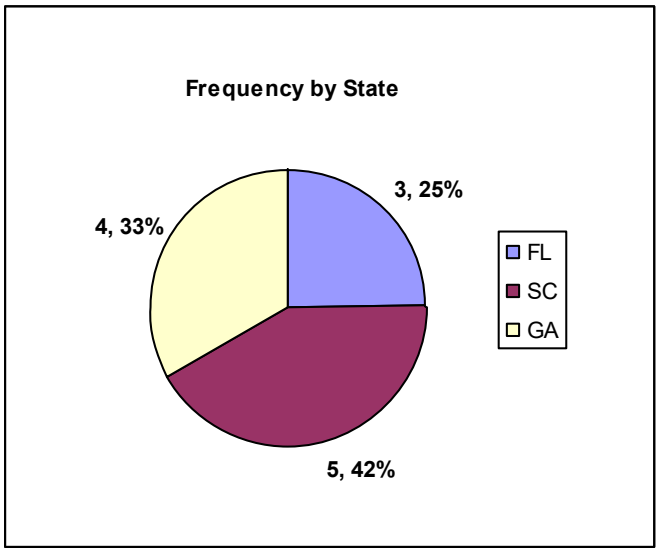
Obs	Age	Gender	State	Salary
1	25	M	FL	25
2	28	F	SC	36
3	31	M	GA	44
4	35	F	GA	38
5	36	M	SC	56
6	38	F	FL	68
7	42	M	SC	79
8	51	F	FL	64
9	55	M	GA	88
10	61	F	FL	71
11	62	M	GA	92
12	65	F	SC	54

State	Freq
FL	3
SC	5
GA	4



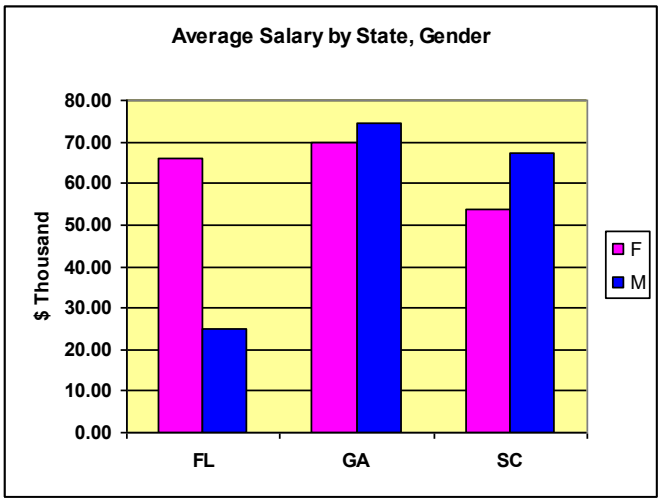
Categorical Data: Pie Charts

State	Freq
FL	3
SC	5
GA	4

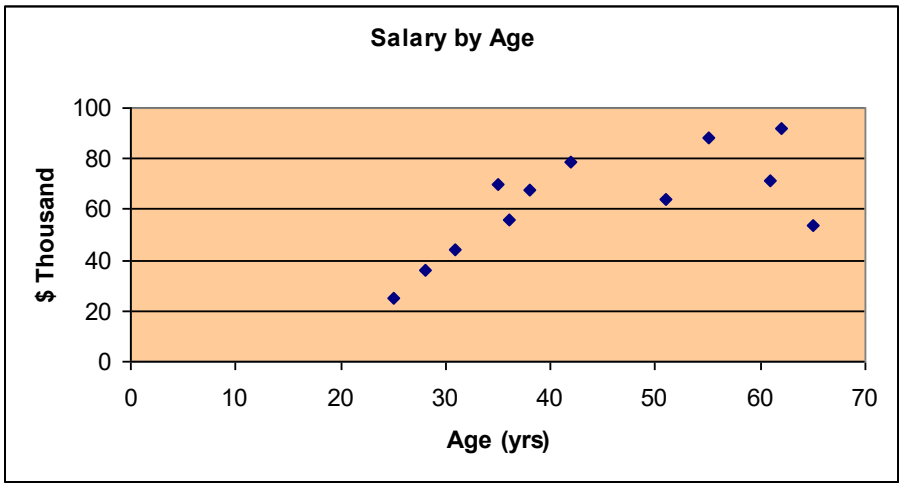


Numeric Data by Category

	F	M
FL	66.00	25.00
GA	70.00	74.67
SC	53.67	67.50

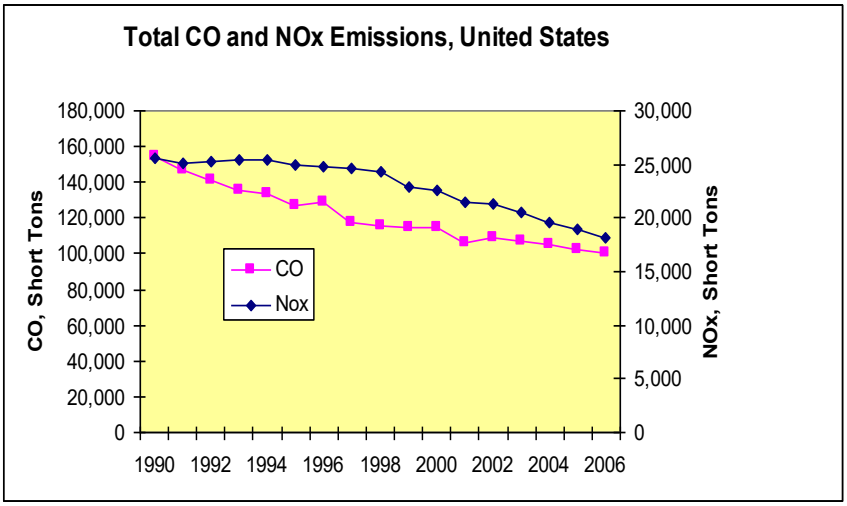


Bivariate Numerical Data
Scatter Plot



Two variables, different units

Year	CO	Nox
1990	154,188	25,527
1991	147,128	25,180
1992	140,895	25,261
1993	135,902	25,356
1994	133,558	25,350
1995	126,778	24,955
1996	128,859	24,786
1997	117,911	24,706
1998	115,380	24,347
1999	114,541	22,843
2000	114,465	22,599
2001	106,263	21,546
2002	109,235	21,277
2003	107,062	20,476
2004	104,892	19,564
2005	102,721	18,947
2006	100,552	18,226



Source:
http://www.epa.gov/ttn/chief/trends/trends06/nationaltier1upto2006basedon2002finalv2.1.xls

Chapter Summary

Categorization: Bar, Pie charts
Distribution: Stem and Leaf, Histogram, Box Plot
Relationships: Scatter Plots, Line Charts
Multivariate: Spider Plots, Maps, Bubble Charts

Contingency Tables

- Chapters Seven, Sixteen, and Eighteen
- Chapter Seven
  - Definition of Contingency Tables
  - Basic Statistics
  - SPSS program (Crosstabulation)
- Chapter Sixteen
  - Basic Probability Theory Concepts
  - Test of Hypothesis of Independence

Contingency Tables (continued)

- Chapter Eighteen
  - Measures of Association
  - For nominal variables
  - For ordinal variables

Basic Empirical Situation

- Unit of data.
- Two nominal scales measured for each unit.
  - Example: interview study, sex of respondent, variable such as whether or not subject has a cellular telephone.
  - Objective is to compare males and females with respect to what fraction have cellular telephones.

Crosstabulation of Data

- Prepare a data file for study.
  - One record per subject.
  - Three variables per record: subject ID, sex of subject, and indicator variable of whether subject has cellular telephone.
- SPSS analysis
  - Statistics, summarize, crosstabs
- Basic information is the contingency table.

Two Common Situations

- Hypothesized causal relation between variables.
- No hypothesized causal relation.

Hypothesized Causal Relation

- Classification of variables
  - Independent variable is one hypothesized to be cause. Example: sex of respondent.
  - Dependent variable is hypothesized to be the effect. Example: whether or not subject has cellular telephone.
- Format convention
  - Columns to categories of independent variable
  - Rows to categories of dependent variable

Association Study

- No hypothesized causal mechanism.
  - Whether or not subject above median on verbal SAT and whether or not above median on quantitative SAT.
- No convention about assigning variables to rows and columns.

Contingency Table

- One column for each value of the column variable; C is the number of columns.
- One row for each value of the row variable; R is the number of rows.
- R x C contingency table.

Contingency Table

- Each entry is the OBSERVED COUNT O(i,j) of the number of units having the (i,j) contingency.
- Column of marginal totals.
- Row of marginal totals.

Example Contingency Table (Hypothetical)

Own Cell Telephone	Male	Female	Total
Yes	60	80	140
No	140	120	260
Total	200	200	400

Example Contingency Table (Hypothetical)

- Entry 60 in the upper left hand corner means that there were 60 male respondents who owned a cellular telephone.
- ASSUME marginal totals are known:
- THEN, knowing entry of 60 means that you can deduce all other entries.
- This 2 x 2 table has one degree of freedom.
- R x C table has (R-1)(C-1) degrees of freedom.

Row and Column Percentages

- Natural to use percentages rather than raw counts.
  - Remember that you want to use these numbers for comparison purposes.
  - The term "rate" is often used to refer to a percentage or probability.
- Can ask for column percentages, row percentages, or both.
  - Percentage in the direction of the independent variable (usually the column).

Relation of Percentages to Probabilities

- ASSUME that the column variable is the independent variable.
- THEN the column percentages are estimates of the conditional probabilities given the setting of the independent variable.
- The basic questions revolve around whether or not the conditional distributions are the same for all settings of the independent variable.

Bar Charts

- Graphical means of presenting data.
- SPSS analysis
  - Graphs, bar chart.
- Can use either count scale or percentage scale (prefer percentage scale).
- Can have bars side by side or stacked.

Generalization of the R x C contingency table

- Can have three or more variables to classify each subject. These are called “layers”.
  - In example, can add whether respondent is student in college or student in high school.

Chapter Sixteen: Comparing Observed and Expected Counts

- Basic hypothesis
- Definitions of expected counts.
- Chi-squared test of independence.

Basic Hypothesis

- ASSUME column variable is the independent variable.
- Hypothesis is independence.
- That is, the conditional distribution in any column is the same as the conditional distribution in any other column.

Expected Count

- Basic idea is proportional allocation of observations in a column based on column total.
- Expected count in (i, j) contingency = E(i,j)= total number in column j *total number in row i/total number in table.
- Expected count need not be an integer; one expected count for each contingency.

Residual

- Residual in (i,j) contingency = observed count in (i,j) contingency - expected count in (i,j) contingency.
- That is, R(i,j)= O(i,j)-E(i,j)
- One residual for each contingency.

Pearson Chi-squared Component

- Chi-squared component for (i, j) contingency =C(i,j)= (Residual in (i, j) contingency)^2/expected count in (i, j) contingency.
- C(i,j)=(R(i,j))^2 / E(i,j)

Assessing Pearson Component

- Rough guides on whether the (i, j) contingency has an excessively large chi-squared component C(i,j):
  - the observed significance level of 3.84 is about 0.05.
  - Of 6.63 is about 0.01.
  - Of 10.83 is 0.001.

Pearson Chi-Squared Test

- Sum C(i,j) over all contingencies.
- Pearson chi-squared test has (R-1)(C-1) degrees of freedom.
- Under null hypothesis
  - Expected value of chi-square equals its degrees of freedom.
  - Variance is twice its degrees of freedom

Special Case of 2 x 2 Contingency Table

Status of Row Var	Column On	Column Off	Total
On	A	B	A+B
Off	C	D	C+D
Total	A+C	B+D	N

Chi-squared test for a 2x2 table

- 1 degree of freedom [(R-1)(C-1)=1]
- Value of chi-squared test is given by
- $N(AD-BC)^2/[(A+B)(C+D)(A+C)(B+D)]$
- There is a correction for continuity

Computer Output for Chi-Squared Tests

- Output gives value of test.
- Asymptotic significance level (p-value)
- Four types of test
  - Pearson chi-squared
  - Pearson chi-squared with continuity correction
  - Likelihood ratio test (theoretically strong test)
  - Fisher's exact test (most accepted, if given.

Example Problem Set

- The independent variable is whether or not the subject reported using marijuana at time 3 in a study (time 3 is roughly in later high school). The dependent variable is whether or not the subject reported using marijuana at time 4 in a study (time 4 is roughly in middle college or beginning independent living). The contingency table is on the next slide.

Marijuana Use at Time 4 by Marijuana Use at Time 3

Use at time 4	No use at time 3	Used at time 3	Total
No use at time 4	120	9	129
Used at time 4	95	142	237
Total	215	151	366

Example Question 1

- Which of the following conclusions is correct about the test of the null hypothesis that the distribution of whether or not a subject uses marijuana at time 3 is independent of whether the subject uses marijuana at time 4?

- Usual options.

Solution to question 1

- Find the significance level in the chi-square test output. Pearson chi-square (without and with continuity correction), likelihood ratio, and Fisher's exact had significance levels of 0.000.
- Option A (reject at the 0.001 level of significance) is the correct choice.

Example Question 2

- How many degrees of freedom does the contingency table describing this output have?
- Solution: (R-1)(C-1)=(2-1)(2-1)=1.

Example Question 3

- Specify how the expected count of 97.8 for subject's who did use marijuana at time 3 and time 4 was calculated?
- Solution:
- Total number using at time 3 was 151.
- Total number using at time 4 was 237.
- Total N was 366.
- Expected Count=151*237/366.

Chapter 2
Describing Data: Graphs and Tables

Basic Concepts in Data Analysis

Data, Information, and Knowledge

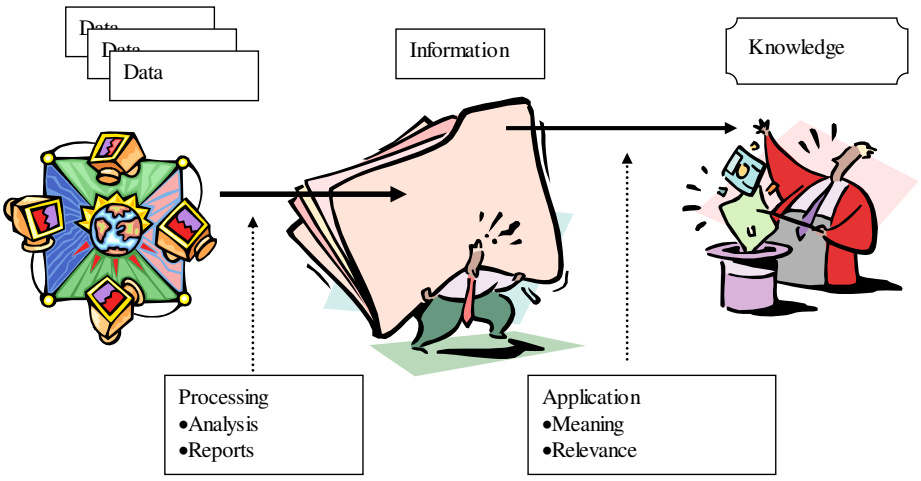
Populations and Samples

Variables and Observations

Types of Data: Categorical and Numerical

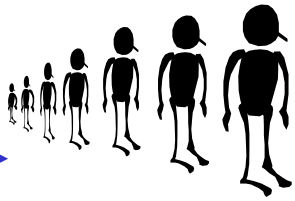
Types of Data: Cross Sectional and Time Ordered

Data, Information, and Knowledge



Populations and Samples

Statistical Inference



Sample: Subset of collection of all possible entities (observation units)

Data on sample is what is available.

KNOWN
Statistics are used to describe samples. These can vary across samples.

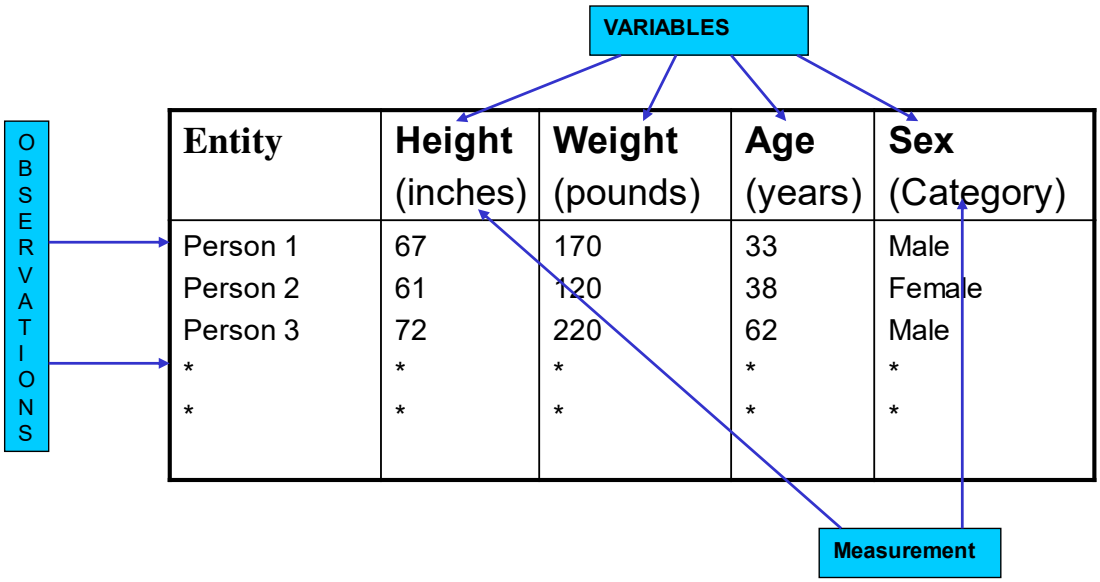
***Statistical Inference** is the art and science of drawing inferences/ conclusions about a population of interest.*

Population: Collection of all possible entities (observation units)

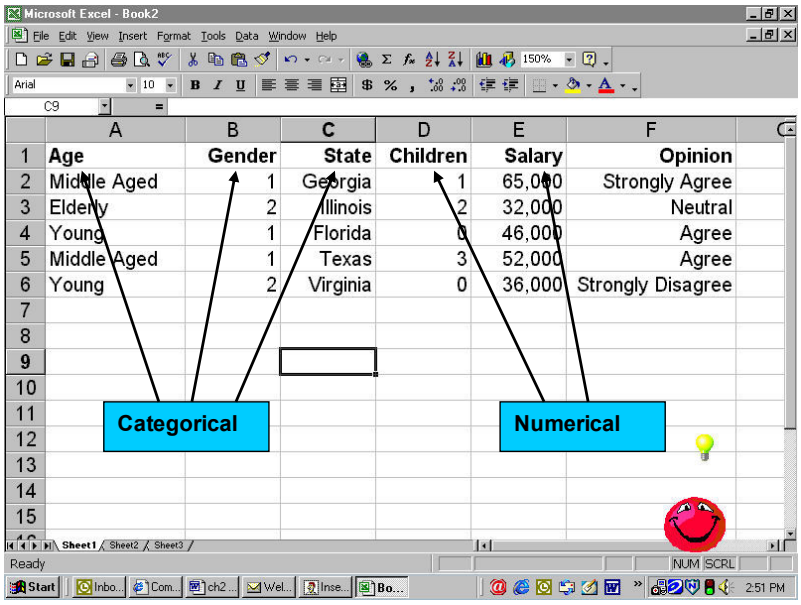
Data on the whole population is usually not available.

UN**KNOWN**
Parameters are used to describe populations. These are constants for a population.

Variables and Observations



Types of Data: Categorical and Numerical



	A	B	C	D	E	F
1	Age	Gender	State	Children	Salary	Opinion
2	Middle Aged	1	Georgia	1	65,000	Strongly Agree
3	Elderly	2	Illinois	2	32,000	Neutral
4	Young	1	Florida	0	46,000	Agree
5	Middle Aged	1	Texas	3	52,000	Agree
6	Young	2	Virginia	0	36,000	Strongly Disagree

Types of Data: Cross-sectional and Time Ordered

Period	Plant 1	Plant 2	Plant 3	Plant 4
Jan	← Cross Sectional Data →			
Feb				
Mar				
Apr				
May				
Jun				
July				

Time Ordered Data

Questions

What was the absenteeism at Plant 1 in Jan. 1998?

Was the annual absenteeism the same for all plants?

Was absenteeism stable at plant 1 during 1998?

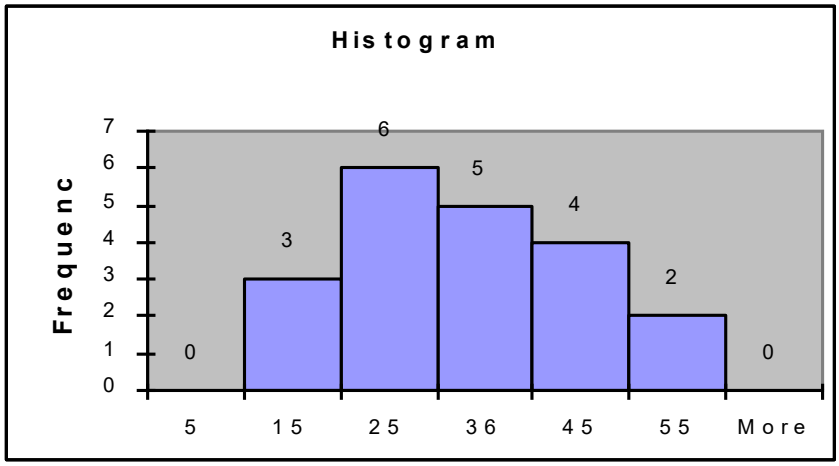
Frequency Tables

A Frequency Table showing a classification of the AGE of attendees at an event.

Class	Frequency	Relative Frequency	Percentage
10 but under 20	3	.15	15
20 but under 30	6	.30	30
30 but under 40	5	.25	25
40 but under 50	4	.20	20
50 but under 60	2	.10	10
Total	20	1	100

Frequency Histograms

A graphical display of distribution of frequencies



Developing Frequency Tables and Histograms

Sort Raw Data in Ascending Order:

12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58

Find Range: 58 - 12 = 46

Select Number of Classes: 5 (usually between 5 and 15)

Compute Class Interval (width): 10 (range/classes = 46/5 then round up)

Determine Class Boundaries (limits): 10, 20, 30, 40, 50

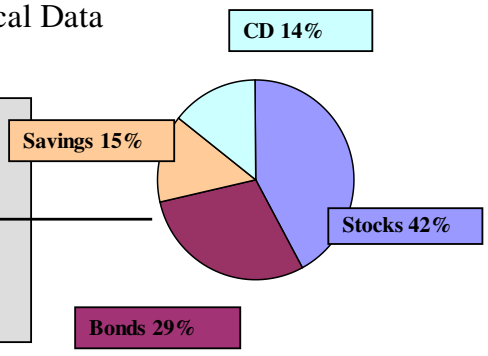
Compute Class Midpoints: 15, 25, 35, 45, 55

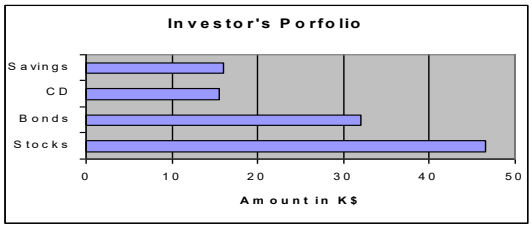
Count Observations & Assign to Classes

Bar and Pie Charts

Displaying Categorical Data

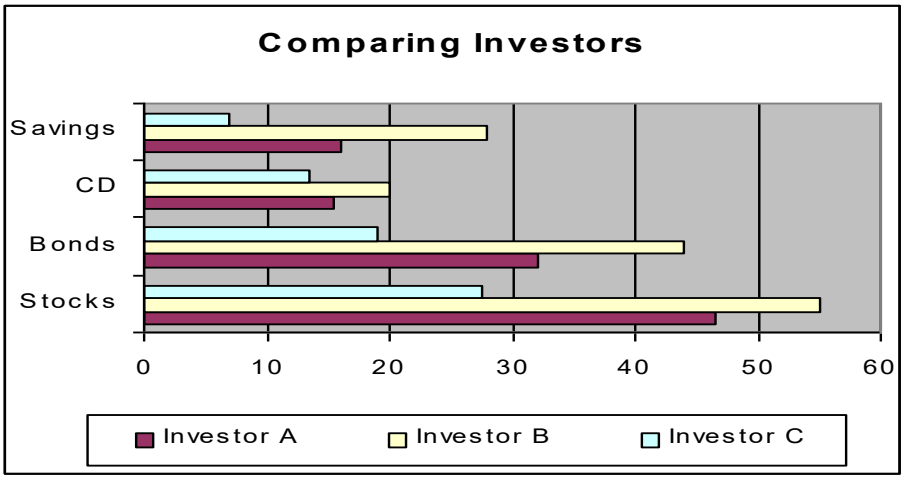
Investment Category	Amount	Percentage (in thousands $)
Stocks	46.5	42.27
Bonds	32	29.09
CD	15.5	14.09
Savings	16	14.55
Total	**110**	**100**





Side by Side Chart

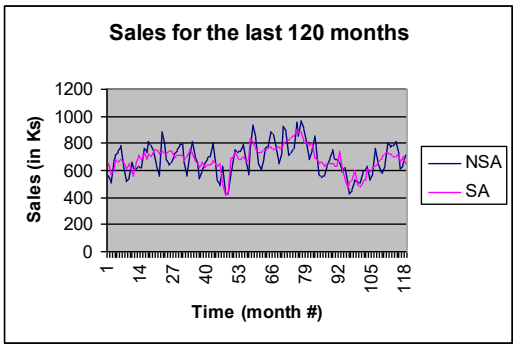
Displaying Categorical Bivariate Data: Contingency Tables and Side-by-Side Charts



Scatter Plot for bivariate numerical data

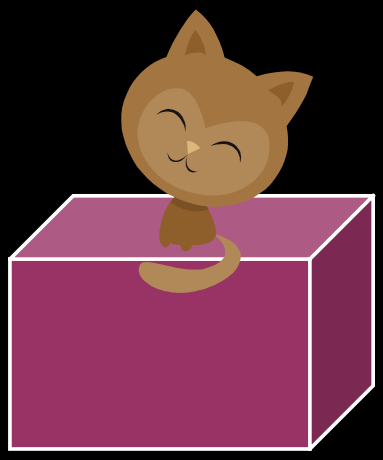
Shows relationship between two variables.

Can one be used to predict the other?



Time-Series and Regression Analysis are used to predict one variable's value based on the other. Correlation analyses is used to measure the strength of linear relationship among two variables.

Box-and-Whisker Plots



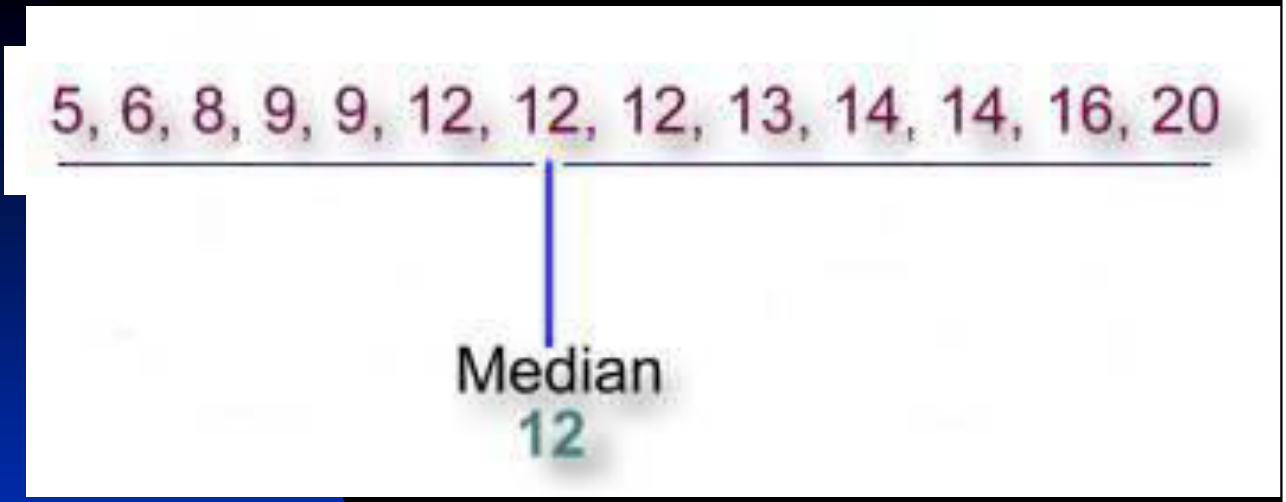
Step 1 – Order Numbers

12, 13, 5, 8, 9, 20, 16, 14, 14, 6, 9, 12, 12

1. Order the set of numbers from least to greatest

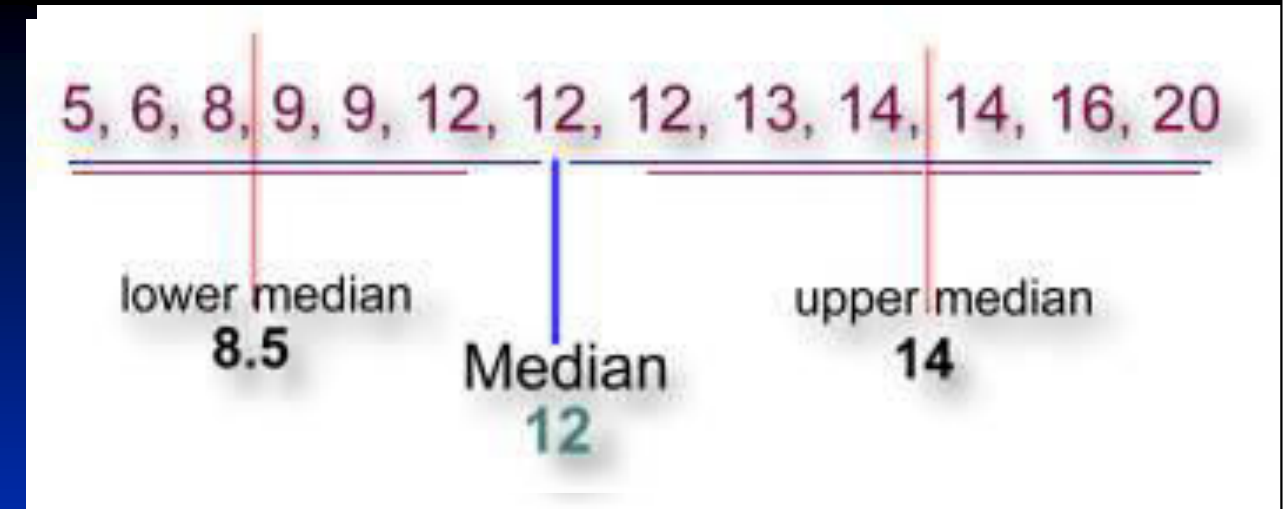
5, 6, 8, 9, 9, 12, 12, 12, 13, 14, 14, 16, 20

Step 2 – Find the Median



2. Find the **median**. The median is the middle number. If the data has two middle numbers, find the mean of the two numbers. What is the median?

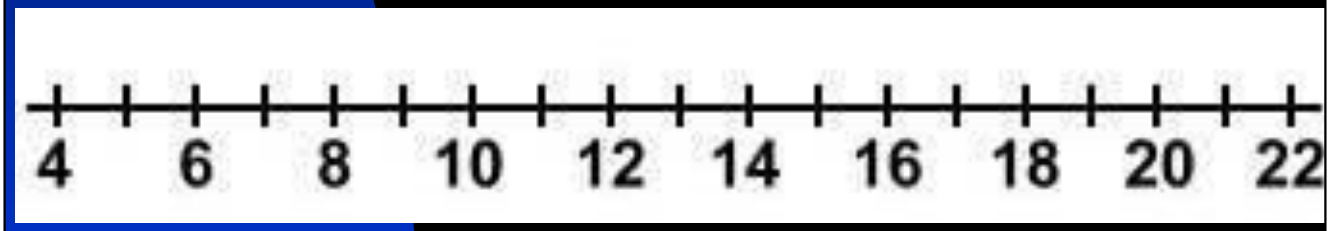
Step 3 – Upper & Lower Quartiles



3. Find the **lower and upper medians or quartiles**. These are the middle numbers on each side of the median. What are they?

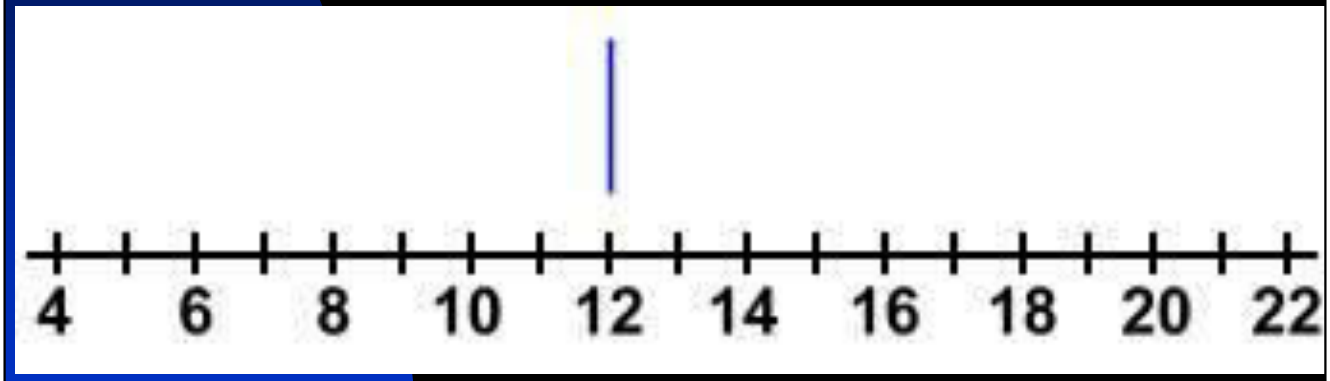
Step 4 – Draw a Number Line

Now you are ready to construct the actual box & whisker graph. First you will need to draw an ordinary number line that extends far enough in both directions to include all the numbers in your data:



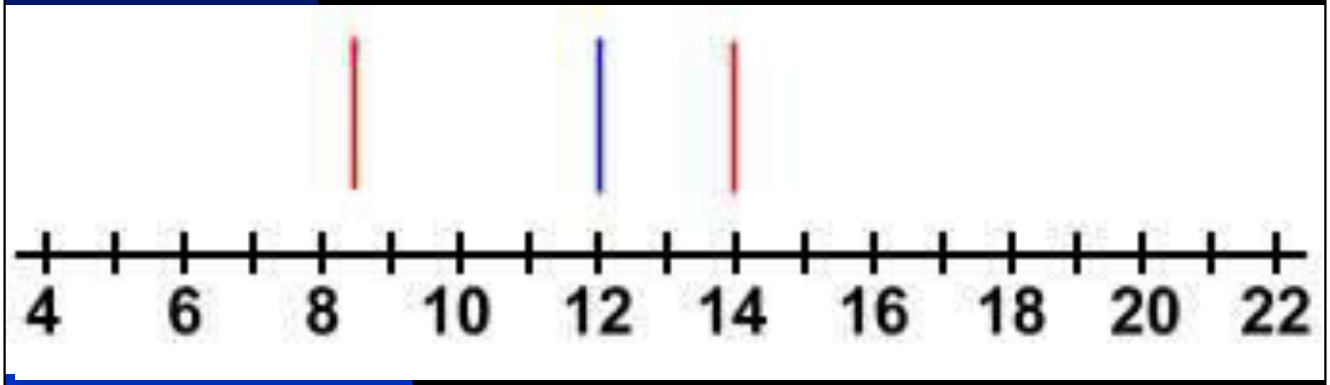
Step 5 – Draw the Parts

Locate the main **median 12** using a vertical line just above your number line:



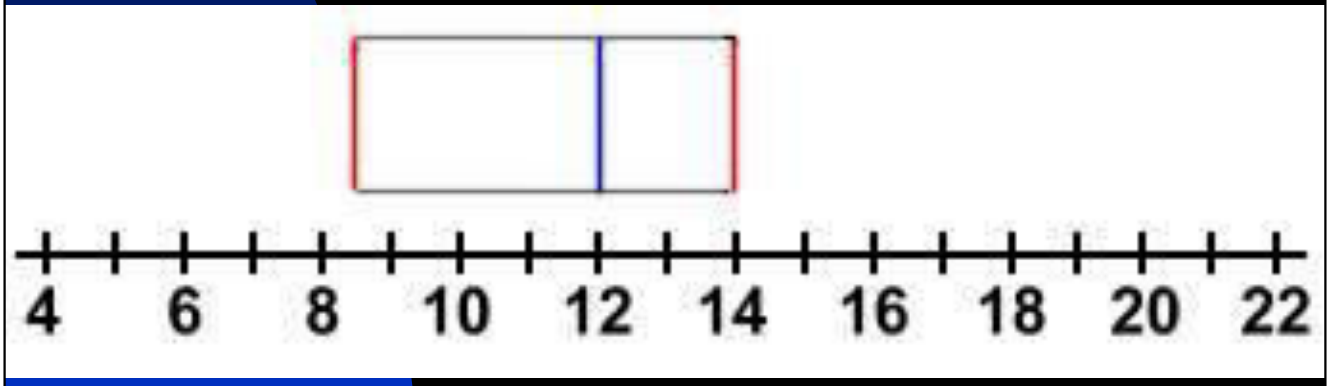
Step 5 – Draw the Parts

Locate the **lower median 8.5** and the **upper median 14** with similar vertical lines:



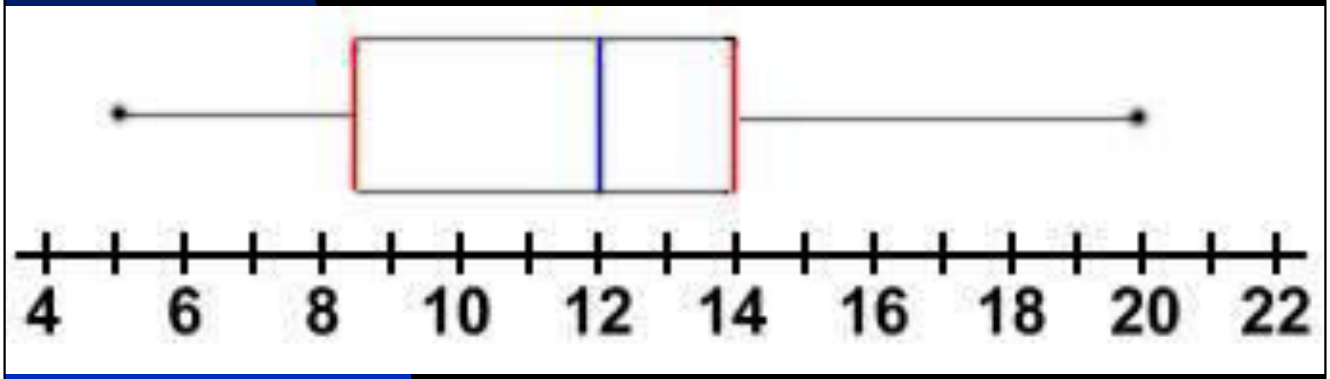
Step 5 – Draw the Parts

- Next, draw a box using the lower and upper median lines as endpoints:

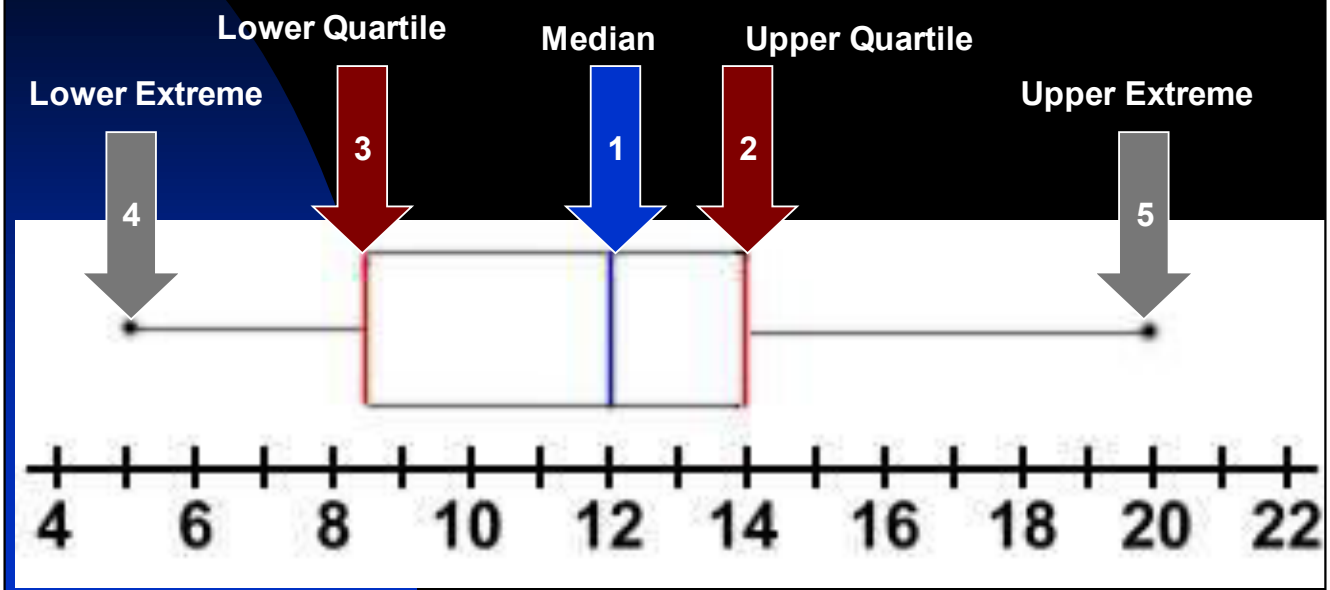


Step 5 – Draw the Parts

Finally, the **whiskers** extend out to the data's smallest number **5** and largest number **20**:



Step 6 - Label the Parts of a Box-and-Whisker Plot



Name the parts of a Box-and-Whisker Plot

Interquartile Range

The interquartile range is the difference between the upper quartile and the lower quartile.

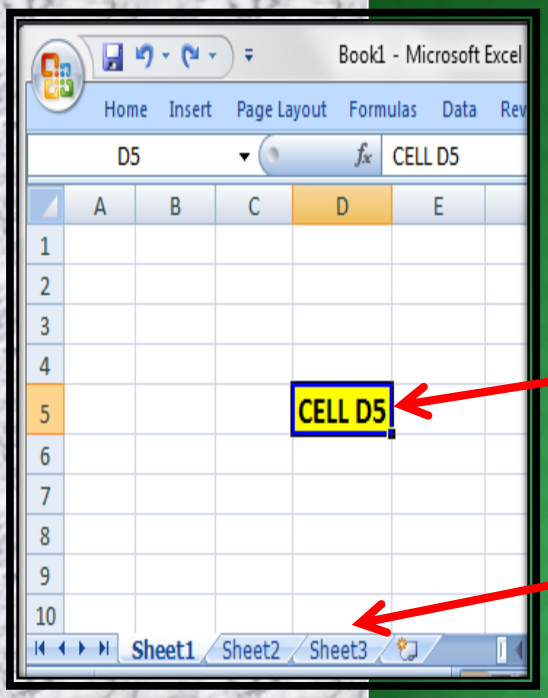
$$14 - 8.5 = 5.5$$

INDEX

INTRODUCTION TO MS-EXCEL

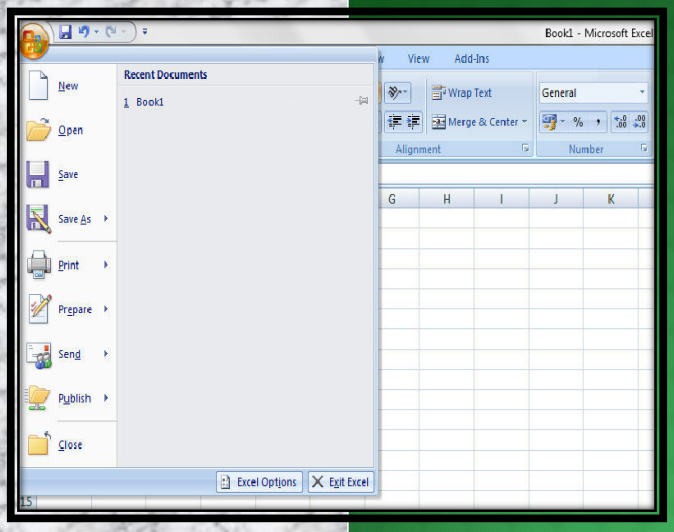
- Excel is a computer program used to create electronic spreadsheets.
- Within excel user can organize data ,create chart and perform calculations.
- Excel is a convenient program because it allow user to create large spreadsheets, reference information, and it allows for better storage of information.
- Excels operates like other Microsoft(MS) office programs and has many of the same functions and shortcuts of other MS programs.

OVERVIEW OF EXCEL



- Microsoft excel consists of workbooks. Within each workbook, there is an infinite number of worksheets.
- Each worksheet contains **Columns and Rows.**
- Where a column and a row intersect is called a **cell.** For e.g. cell **D5** is located where column **D** and row **5** meet.
- The tabs at the bottom of the screen represent different worksheets within a workbook. You can use the scrolling buttons on the left to bring other worksheets into view.

OFFICE BUTTON

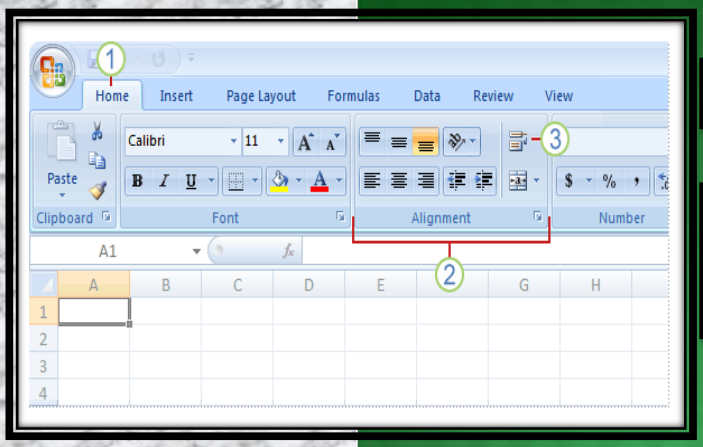


OFFICE BUTTON CONTAINS..

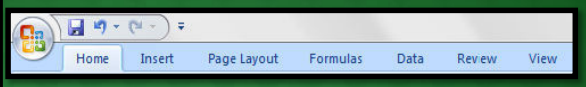
- NEW-TO OPEN NEW WORKBOOK. (CTRL+N)
- OPEN-TO OPEN EXISTING DOCUMENT (CTRL+O)
- SAVE-TO SAVE A DOCUMENT. (CTRL+S)
- SAVE AS-TO SAVE COPY DOCUMENT. (F12)
- PRINT-TO PRINT A DOCUMENT. (CTRL+P)
- PREPARE-TO PREPARE DOCUMENT FOR DISTRIBUTION.
- SEND-TO SEND A COPY OF DOCUMENT TO OTHER PEOPLE.
- PUBLISH-TO DISTRIBUTE DOCUMENT TO OTHER PEOPLE.
- CLOSE-TO CLOSE A DOCUMENT (CTRL+W).

RIBBONS

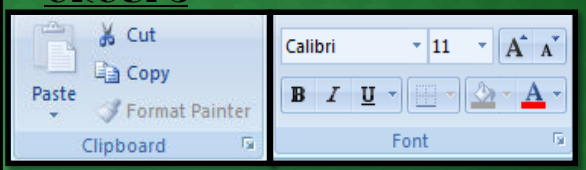
THE THREE PARTS OF THE RIBBON ARE



TABS



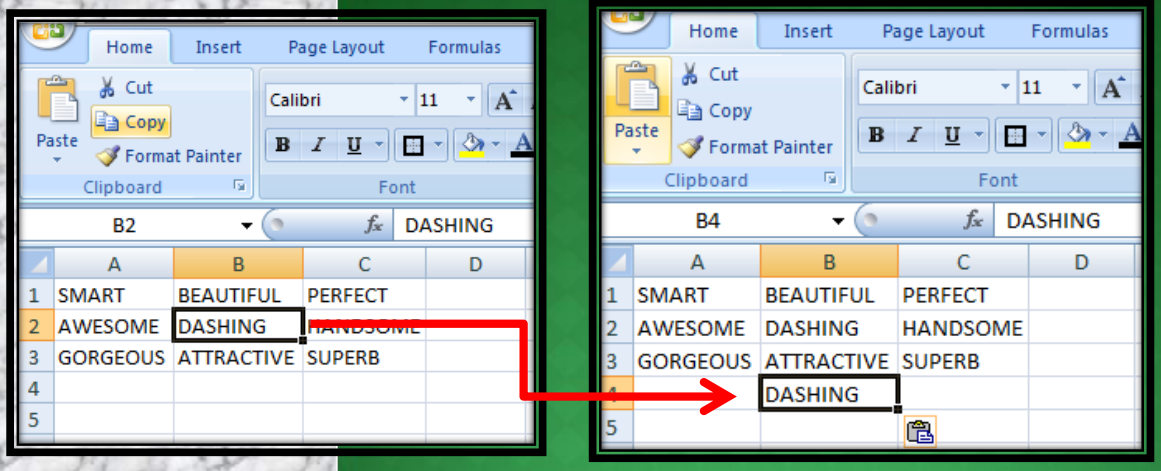
GROUPS



COMMANDS

1. **TABS:** THERE ARE SEVEN TABS ACROSS THE TOP OF THE EXCEL WINDOW.
2. **GROUPS**: GROUPS ARE SETS OF RELATED COMMANDS, DISPLAYED ON TABS.
3. **COMMANDS**: A COMMAND IS A BUTTON, A MENU OR A BOX WHERE YOU ENTER INFORMATION.

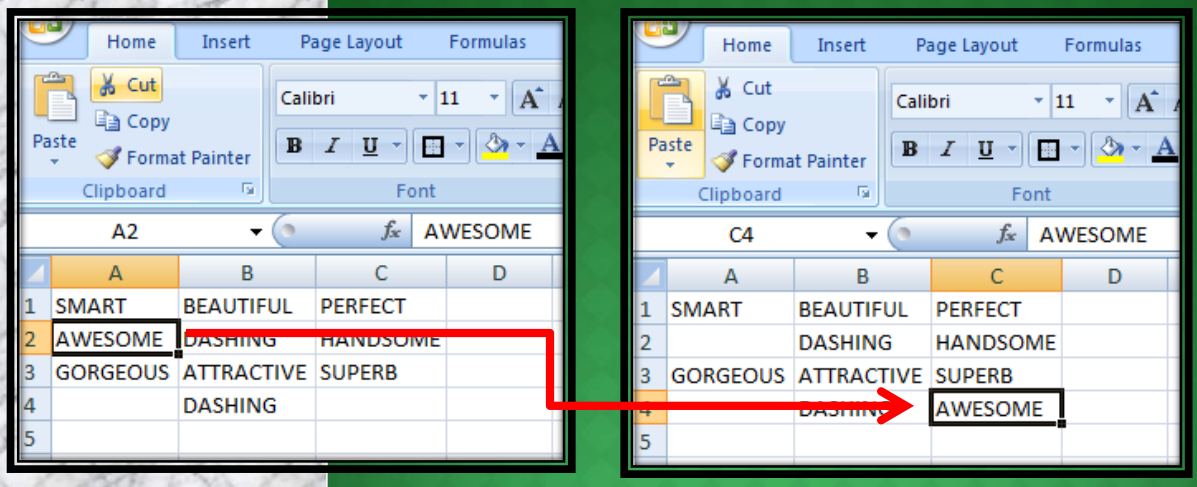
WORKING WITH CELLS



TO COPY AND PASTE CONTENTS:

Select the **cell or cells** you wish to **copy**.
Click the **Copy** command in the Clipboard group on the Home tab.
Select the **cell or cells** where you want to **paste** the information.
Click the **Paste** command.
The copied information will now appear in the new cells.

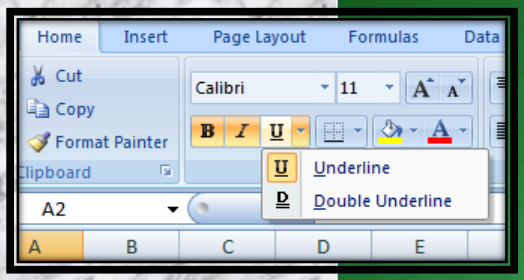
WORKING WITH CELLS

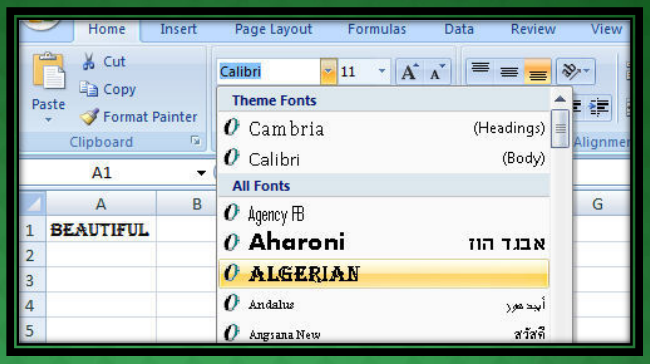


To Cut and Paste Cell Contents:

Select the **cell or cells** you wish to **cut**.
Click the **Cut** command in the Clipboard group on the Home tab.
Select the **cell or cells** where you want to **paste** the information.
Click the **Paste** command.
The cut information will be removed and **now appear** in the new **cells.**

FORMATTING TEXT





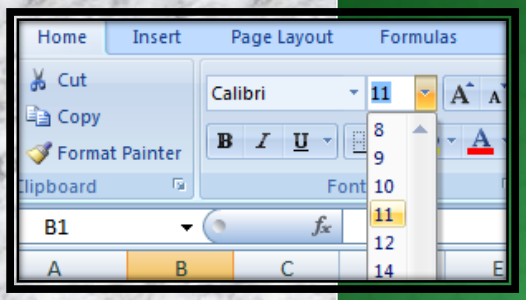
TO FORMAT TEXT IN BOLD, ITALICS OR UNDERLINE:

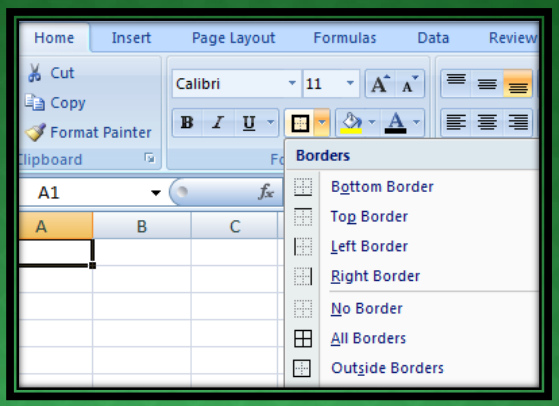
Left-click a **cell** to select it or drag your cursor over the text in the formula bar to select it.
Click the **Bold, Italics or underline** command.

TO CHANGE THE FONT STYLE:

Select the cell or cells you want to format.
Left-click the **drop-down arrow** next to the **Font Style** box on the Home tab.
Select a **font style** from the list.

FORMATTING TEXT





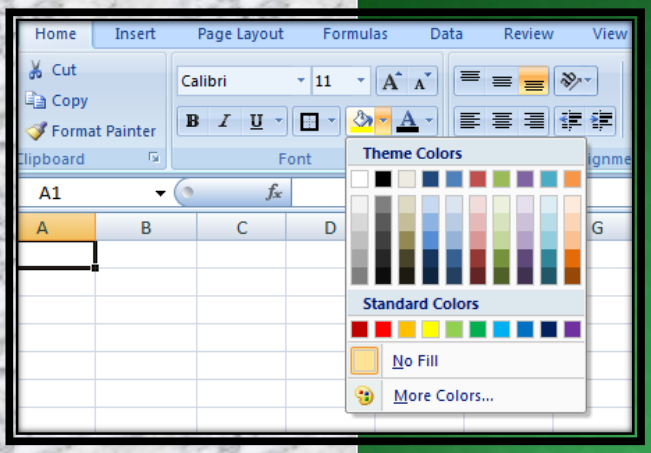
TO CHANGE THE FONT SIZE:

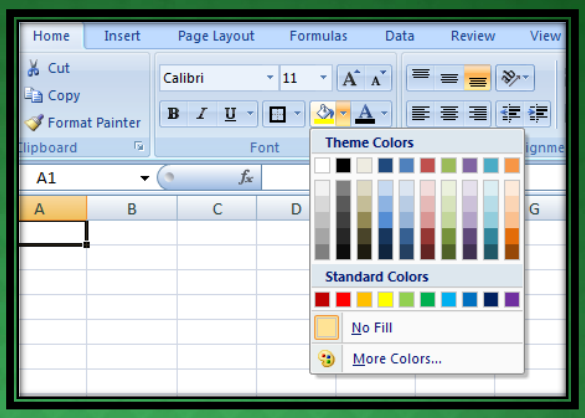
Select the cell or cells you want to format.
Left-click the **drop-down arrow** next to the **Font Size** box on the Home tab.
Select a **font size** from the list.

TO ADD A BORDER:

Select the cell or cells you want to format.
Click the **drop-down arrow** next to the **Borders** command on the Home tab. A menu will appear with border options.

FORMATTING TEXT





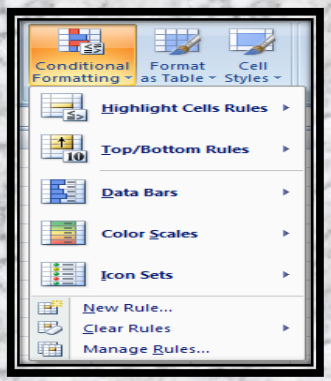
TO CHANGE THE TEXT COLOUR:

Select the cell or cells you want to format.
Left-click the **drop-down arrow** next to the **Text Color** command. A color palette will appear.
Select a color from the palette.

TO ADD A FILL COLOUR:

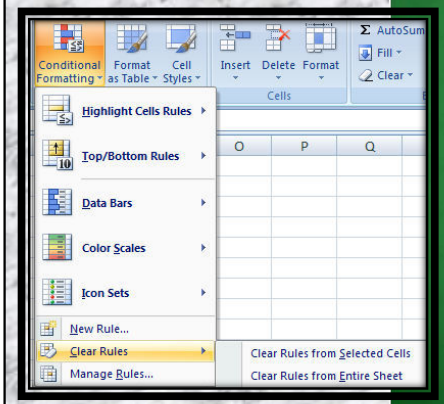
Select the cell or cells you want to format.
Click the **Fill command**. A color palette will appear.
Select a color from the palette.

CONDITIONAL FORMATTING



TO APPLY CONDITIONAL FORMATTING:

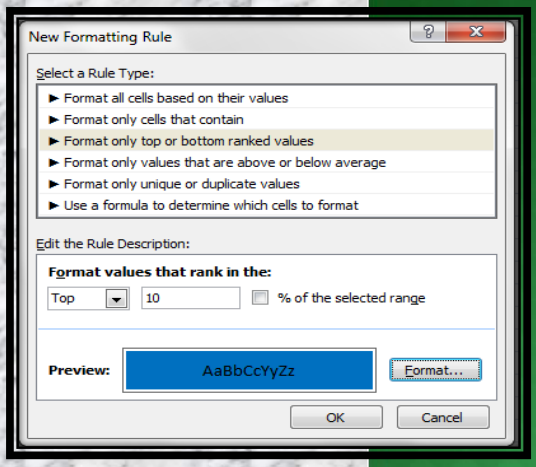
Select the cells you would like to format.
Select the **Home** tab.
Locate the **Styles** group.
Click the **Conditional Formatting** command. A menu will appear with your formatting options.

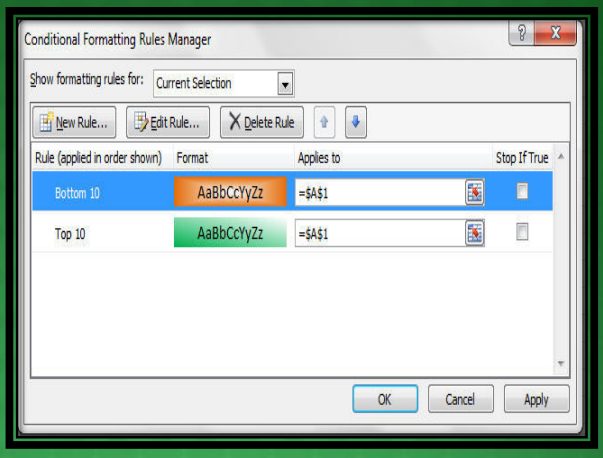


TO REMOVE CONDITIONAL FORMATTING:

Click the **Conditional Formatting** command.
Select **Clear Rules**.
Choose to clear rules from the **entire worksheet** or the **selected cells**.

CONDITIONAL FORMATTING

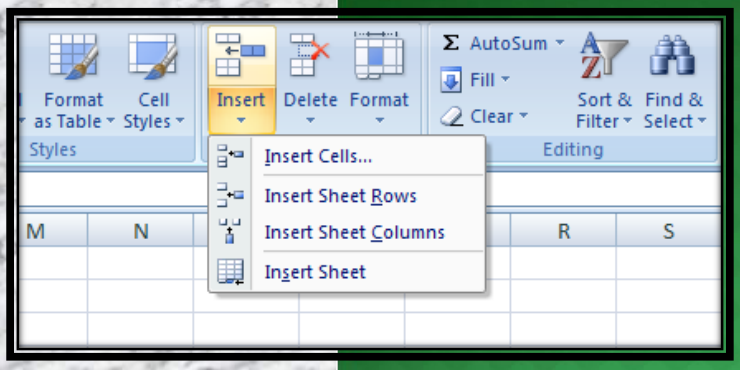




TO APPLY NEW FORMATTING:
Click the **Conditional Formatting** command. Select **New Rules** from the menu. There are different rules, you can apply these rules to differentiate particular cell.

TO MANAGE CONDITIONAL FORMATTING:
Click the **Conditional Formatting** command. Select **Manage Rules** from the menu. The Conditional Formatting Rules Manager dialog box will appear. From here you can edit a rule, delete a rule, or change the order of rules.

TO INSERT ROWS & COLOUMS



NOTE:

1. The new row always appears above the selected row.
2. The new column always appears to the left of the selected column.

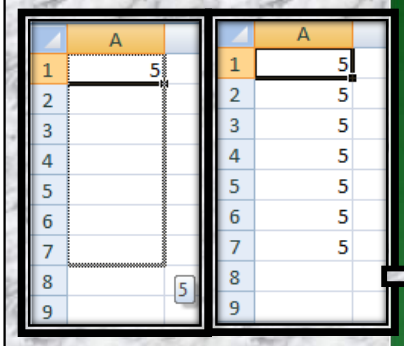
TO INSERT ROWS:

Select the row **below** where you want the new row to appear.
Click the **Insert** command in the Cells group on the Home tab. The row will appear.

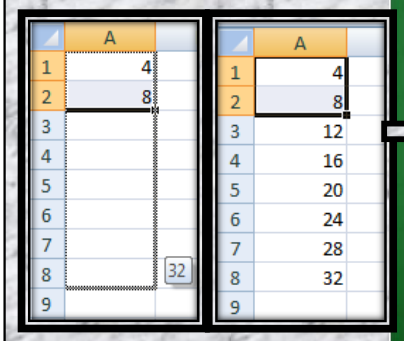
To Insert Columns:

Select the column to the right of where you want the column to appear.
Click the **Insert** command in the Cells group on the Home tab. The column will appear.

EDITING- FILL



- IN THE LOWER RIGHT HAND CORNER OF THE ACTIVE CELL IS EXCEL'S "FILL HANDLE".WHEN YOU HOLD YOUR MOUSE OVER THE TOP OF IT, YOUR CURSOR WILL TURN TO A CROSSHAIR.

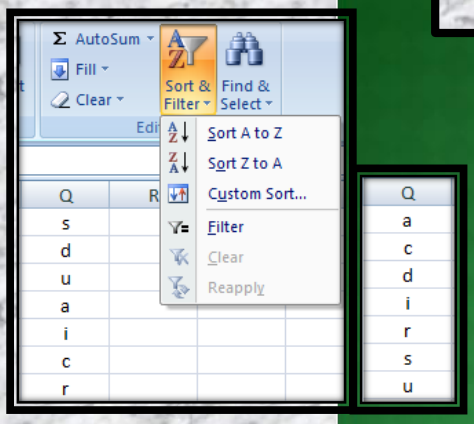


- IF YOU HAVE JUST ONE CELL SELECTED, IF YOU CLICK AND DRAG TO FILL DOWN A COLUMN OR ACROSS A ROW, IT WILL COPY THAT NUMBER OR TEXT TO EACH OF THE OTHER CELLS.

- IF YOU HAVE TWO CELLS SELECTED, EXCEL WILL FILL IN A SERIES. IT WILL COMPLETE THE PATTERN.FOR EXAMPLE,IF YOU PUT 4 AND 8 IN TWO CELLS SELECT THEM,CLICK AND DRAG THE FILL HANDLE ,EXCEL WILL CONTINUE THE PATTERN WITH 12,16,20.ETC.

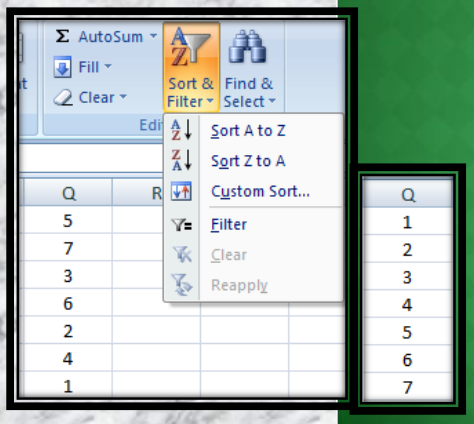
- EXCEL CAN ALSO AUTO- FILL SERIES OF DATES, TIMES, DAYS OF THE WEEK, MONTHS.

SORTING



TO SORT IN ALPHABETICAL ORDER:

Select a cell in the column you want to sort (In this example, we choose a cell in column Q).
Click the **Sort & Filter** command in the **Editing** group on the Home tab.
Select **Sort A to Z**. Now the information in the Category column is organized in alphabetical order.



TO SORT FROM SMALLEST TO LARGEST:

Select a cell in the column you want to sort (In this example, we choose a cell in column Q).
Click the **Sort & Filter** command in the **Editing** group on the Home tab.
Select **From Smallest to Largest**. Now the information is organized from the smallest to largest amount.

CELL REFERENCING

RELATIVE REFERENCE

	A	B	C
1	2	3	=A1+B1
2			
3			

	A	B	C
1	2	3	5
2			
3			

IN CELL (C1) SUM FUNCTION IS USED.
THEN FUNCTION FROM CELL (C1) IS COPY TO CELL (D3).
WHEN THE POSITION OF THE CELL IS CHANGED FROM (C1) TO (D3),THEN THE REFERENCE IS ALSO CHANGED FROM (A1,B1) TO (B3,C3).

	A	B	C	D
1	2	3	=A1+B1	
2				
3				=B3+C3
4				

	A	B	C	D
1	2	3	5	
2				
3		4	6	10
4				

A **RELATIVE CELL REFERENCE** AS (A1) IS BASED ON THE RELATIVE POSITION OF THE CELL. IF THE POSITION OF THE CELL THAT CONTAINS THE REFERENCE CHANGES, THE REFERENCE ITSELF IS CHANGED.

CELL REFERENCING

ABSOLUTE REFERENCE

	A	B	C
1	2	3	=$A$1+$B$1
2			
3			

	A	B	C
1	2	3	5
2			
3			

IN CELL (C1) SUM FUNCTION IS USED.
THEN FUNCTION FROM CELL (C1) IS COPY TO CELL (D3).
WHEN THE POSITION OF THE CELL IS CHANGED FROM (C1) TO (D3),THEN THE ABSOLUTE REFERENCE REMAINS THE SAME(A1,B1).**$ IS USED FOR CONSTANT ROW OR COLUMN.**

	A	B	C	D
1	2	3	=$A$1+$B$1	
2				
3				=$A$1+$B$1
4				

	A	B	C	D
1	2	3	5	
2				
3		4	6	5
4				

AN **ABSOLUTE CELL REFERENCE** AS ($A$1) ALWAYS REFERS TO A CELL IN A SPECIFIC LOCATION. IF THE POSITION OF THE CELL THAT CONTAINS THE FORMULA CHANGES, THE ABSOLUTE REFERENCE REMAINS THE SAME.

CELL REFERENCING

MIXED REFERENCE

	A	B	C
1	2	3	=$A1+$B1
2			
3			

	A	B	C
1	2	3	5
2			
3			

IN CELL (C1) SUM FUNCTION IS USED.
THEN FUNCTION FROM CELL (C1) IS COPY TO CELL (D3).
WHEN THE POSITION OF THE CELL IS CHANGED FROM (C1) TO (D3),THEN ROW REFERENCE IS CHANGED(FROM 1 TO 3) BUT COLUMN REFERENCE REMAINS SAME(A,B).

	A	B	C	D
1	2	3	=$A1+$B1	
2				
3				=$A3+$B3
4				

	A	B	C	D
1	2	3	5	
2				
3	4	6		10
4				

A MIXED REFERENCE HAS EITHER AN ABSOLUTE COLUMN AND RELATIVE ROW OR ABSOLUTE ROW AND RELATIVE COLUMN. AN ABSOLUTE COLUMN REFERENCE TAKES THE FORM $A1, $B1.AN ABSOLUTE ROW REFERENCE TAKES THE FORM A$1, B$1.

FUNCTIONS

DATEDIF FUNCTION

SYNTAX OF DATEDIF

=DATEDIF(START_DATE,END_DATE,"INTERVAL")

A	B	C
MY DATE OF BIRTH	23/06/1993	
TODAY`S DATE	10/01/2013	
	FUNCTIONS	**RESULTS**
NO. OF DAYS	= DATEDIF(B1,B2,"D")	7141
NO. OF MONTHS	= DATEDIF(B1,B2,"M")	234
NO. OF YEARS	= DATEDIF(B1,B2,"Y")	19
NO. OF YEARS	= DATEDIF(B1,B2,"Y")	19
MONTHS OF YEAR	=DATEDIF(B1,B2,"YM")	6
DAYS OVER MONTH	=DATEDIF(B1,B2,"MD")	18

START DATE-
Date from which u want to calculate difference.

END DATE-
Date up to which u want to calculate difference.

INTERVAL-
Form in which u want to calculate difference.

"D"- DAYS
"M"- MONTHS
"Y"- YEARS
"YM"- MONTHS OVER YEAR
"MD"- DAYS OVER MONTH

This says that I am 19 years 6 months & 18 days old

FUNCTIONS

SUMIF FUNCTION

A	B
5	3
1	7
7	4
3	1
9	8
4	6
2	2
FUNCTION	**RESULT**
=SUMIF(A1:A7,"<5")	**10**
=SUMIF(A1:A7,"<5",B1:B7)	**16**

WITHOUT SUM_RANGE

WITH SUM_RANGE

SYNTAX OF SUMIF

=SUMIF(RANGE,CRITERIA,SUM_RANGE)

RANGE-
Range of cells on which conditions are applied.

CRITERIA-
Condition that defines which cell or cells will be added.

SUM RANGE-
Actual cells to sum.

NOTE:-
If sum range is not used then range is used for sum.

FUNCTIONS

IF FUNCTION

A	B	C
	FUNCTION	**RESULT**
5	=IF(A2<5,"TRUE","FALSE")	FALSE
	=IF(A2>5,"TRUE","FALSE")	FALSE
	=IF(A2=5,"TRUE","FALSE")	TRUE
	=IF(A2<5,20,10)	10
	=IF(A2>=5,20,10)	20
	=IF(A2<=5,"A","B")	A
	=IF(A2>5,"A","B")	B

SYNTAX OF IF

=IF(LOGICAL TEXT, VALUE IF TRUE, VALUE IF FALSE)

LOGICAL TEXT-
Any value or expression that can be evaluated to TRUE or FALSE.

VALUE IF TRUE-
Value that is returned if logical text is TRUE.

VALUE IF FALSE-
Value that is returned if logical text is FALSE.

IN COLUMN B DIFFERENT CONDITIONS ARE USED AND BASED ON THIS, IN COLUMN C DIFFERENT RESULTS ARE SHOWN.

COUNT FUNCTIONS

	A	B	C
1	3	FUNCTIONS	RESULT
2	5	= COUNT(A1;A10)	4
3		=COUNTA(A1:A10)	8
4	-	= COUNTBLANK(A1:A10)	2
5	=	=COUNTIF(A1:A10,"<=5")	3
6	.		
7	,		
8			
9	8		
10	0		

SYNTAX OF FUNCTIONS

1. **COUNT**
   =COUNT(VALUE1,VALUE2,...)

2. **COUNTA**
   =COUNTA(VALUE1,VALUE2,...)

3. **COUNTBLANK**
   =COUNTBLANK(RANGE)

4. **COUNTIF**
   =COUNTIF(RANGE,CRITERIA)

1. COUNT ONLY CELLS THAT CONTAINS NUMBER.

2. COUNT CELLS THAT ARE NOT EMPTY.

3. COUNT CELLS THAT ARE BLANK.

4. COUNT NO. OF CELLS THAT MEET GIVEN CONDITION.

TEXT FUNCTIONS

	A	B	C	D
1		LOWER FUNCTION	UPPER FUNCTION	PROPER FUNCTION
2	SmaRt	smart	SMART	Smart
3	BeautiFul	beautiful	BEAUTIFUL	Beautiful
4	DashIng	dashing	DASHING	Dashing
5	GorgeOus	gorgeous	GORGEOUS	Gorgeous
6	PerfEct	perfect	PERFECT	Perfect
7	ExcellEnt	excellent	EXCELLENT	Excellent
8	AwesOme	awesome	AWESOME	Awesome

SYNTAX OF FUNCTIONS

1. **LOWER FUNCTION**
   =LOWER(TEXT)

2. **UPPER FUNCTION**
   =UPPER(TEXT)

3. **PROPER FUNCTION**
   =PROPER(TEXT)

1. TO CONVERT TEXT FROM CAPITAL TO SMALL.

2. TO CONVERT TEXT FROM SMALL TO CAPITAL.

3. TO CAPITALISED EACH WORD OF TEXT.

TEXT FUNCTIONS

	A	B	C	D
1		LEFT FUNCTION	RIGHT FUNCTION	MID FUNCTION
2		=LEFT(A_n,3)	=RIGHT(A_n,3)	=MID(A_n,2,3)
3	smart	sma	art	mar
4	beautiful	bea	ful	eau
5	dashing	das	ing	ash
6	gorgeous	gor	ous	org
7	perfect	per	ect	erf
8	excellent	exc	ent	xce
9	awesome	awe	ome	wes

SYNTAX OF FUNCTIONS

1. **LEFT FUNCTION**
   =LEFT(TEXT,NUM_CHARS)
2. **RIGHT FUNCTION**
   =RIGHT(TEXT,NUM_CHARS)
3. **MID FUNCTION**
   =MID(TEXT,STARTNUM,NUM_CHAR)

1. RETURN SPECIFIED NO. OF CHARACTER FROM START OF TEXT.

2. RETURN SPECIFIED NO. OF CHRACTER FROM END OF TEXT.

3. RETURN CHARACTER FROM MIDDLE OF TEXT,GIVEN A STARTING POSITION.

OTHER FUNCTIONS

	A	B
1	FUNCTIONS	RESULTS
2		
3	=NOW()	14/01/2013 01:55
4		
5		
6	=TODAY()	14/01/2013
7		
8		
9	=MOD(7,3)	1
10		
11		
12	=LEN(A1)	9
13		
14		
15	=SUM(2,3)	5
16		

USES OF FUNCTIONS

⇒ **NOW** RETURNS CURRENT DATE AND TIME.

⇒ **TODAY** RETURNS CURRENT DATE ONLY.

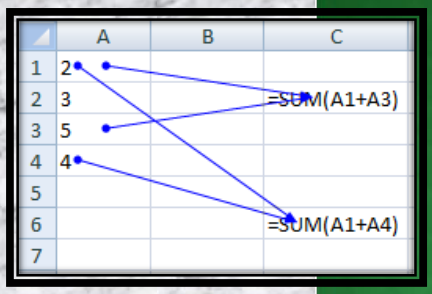
⇒ **MOD** RETURNS THE REMAINDER AFTER A NO. IS DIVIDED BY A DIVISOR.

⇒ **LEN** RETURNS THE NO. OF CHARACTERS IN A TEXT STRING.

⇒ **SUM** ADD ALL THE NUMBERS.

FUNCTION AUDITING

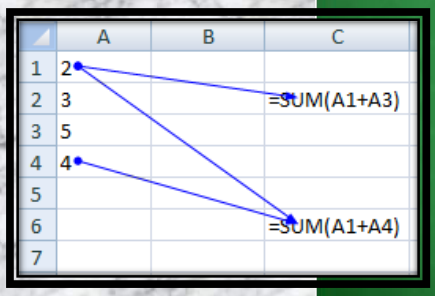
TRACE PRECEDENTS



SHOW ARROW THAT INDICATE WHAT CELLS AFFECT THE VALUE OF THE CURRENTLY SELECTED CELL.

IN THIS EXAMPLE CELLS **A1** & **A3** AFFECT THE VALUE OF CELL **C2** & CELLS **A1** & **A4** AFFECT THE VALUE OF CELL **C6**.

TRACE DEPENDENTS



SHOW ARROW THAT INDICATE WHAT CELLS ARE AFFECTED BY THE VALUE OF THE CURRENTLY SELECTED CELL.

IN THIS EXAMPLE CELL **C2** & **C6** ARE AFFECTED BY THE VALUE OF CELL **A2** & CELL **C6** IS ALSO AFFECTED BY THE CELL **A4**.

SHORTCUT KEYS

PARTICULARS	KEYS
EDIT THE ACTIVE CELL	F_2
CREATE A CHART	F_{11}
INSERT CELL COMMENT	SHIFT + F_2
FUNCTION DIALOGUE BOX	SHIFT + F_3
INSERT A NEW WORKSHEET	SHIFT + F_{11}
NAME MANAGER DIALOGUE BOX	CTRL + F_3
VISUAL BASIC EDITOR	ALT + F_{11}
MACRO DIALOGUE BOX	ALT + F_8
HIDE THE SELECTED COLUMNS	CTRL + 0
UNHIDE THE COLUMNS	CTRL + SHIFT + 0
HIDE THE SELECTED ROWS	CTRL + 9
UNHIDE THE ROWS	CTRL + SHIFT + 9
SELECT ALL CELLS WITH COMMENT	CTRL + SHIFT + O

SHORTCUT KEYS

PARTICULARS	KEYS
DOWN FILL	CTRL + D
RIGHT FILL	CTRL + R
ENTER SUM FUNCTION IN CELL	ALT + =
EURO SYMBOL	ALT + 0128
CENT SYMBOL	ALT + 0162
POUND SYMBOL	ALT + 0163
YEN SYMBOL	ALT + 0165
ENTER NEW LINE IN ACTIVE CELL	ALT + ENTER
CURRENT DATE	CTRL + ;
CURRENT TIME	CTRL + SHIFT + ;
SHOW FORMULA	CTRL + `
SELECT ENTIRE COLUMN	CTRL + SPACEBAR
SELECT ENTIRE ROW	SHIFT + SPACEBAR

SHORTCUT KEYS

PARTICULARS	KEYS
APPLIES NUMBER FORMAT	CTRL + SHIFT + !
APPLIES CURRENCY FORMAT	CTRL + SHIFT + $
APPLIES PERCENTAGE FORMAT	CTRL + SHIFT + %
APPLIES EXPONENTIAL FORMAT	CTRL + SHIFT + ^
APPLIES GENERAL NO. FORMAT	CTRL + SHIFT + ~
APPLIES TIME FORMAT	CTRL + SHIFT + @
APPLIES DATE FORMAT	CTRL + SHIFT + #
APPLIES OUTLINE BORDER	CTRL + SHIFT + &
REMOVE OUTLINE BORDER	CTRL + SHIFT + _

Frequency Distributions

Frequency Distribution

- A table that organizes data values into classes or intervals along with number of values that fall in each class (frequency, f).

1. Ungrouped Frequency Distribution – for data sets with few different values. Each value is in its own class.

2. Grouped Frequency Distribution: for data sets with many different values, which are grouped together in the classes.

Grouped and Ungrouped Frequency Distributions

Ungrouped

Courses Taken	Frequency, f
1	25
2	38
3	217
4	1462
5	932
6	15

Grouped

Age of Voters	Frequency, f
18-30	202
31-42	508
43-54	620
55-66	413
67-78	158
78-90	32

Ungrouped Frequency Distributions

Number of Peas in a Pea Pod Sample Size: 50				
5	5	4	6	4
3	7	6	3	5
6	5	4	5	5
6	2	3	5	5
5	5	7	4	3
4	5	4	5	6
5	1	6	2	6
6	6	6	6	4
4	5	4	5	3
5	5	7	6	5

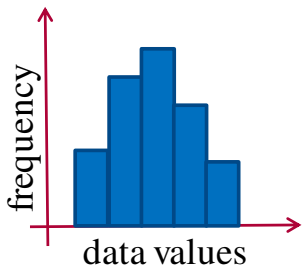
Peas per pod	Freq, f

Peas per pod	Freq, f
1	1
2	2
3	5
4	9
5	18
6	12
7	3

Graphs of Frequency Distributions: Frequency Histograms

Frequency Histogram

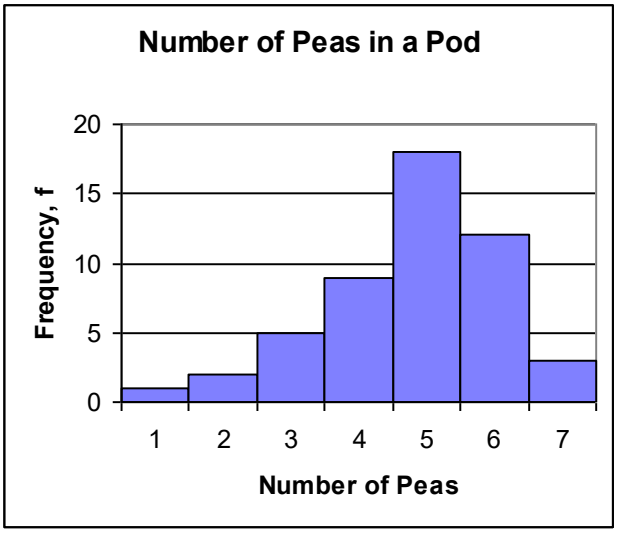
- A bar graph that represents the frequency distribution.
- The horizontal scale is quantitative and measures the data values.
- The vertical scale measures the frequencies of the classes.
- Consecutive bars must touch.



Frequency Histogram

Ex. Peas per Pod

Peas per pod	Freq, f
1	1
2	2
3	5
4	9
5	18
6	12
7	3



Relative Frequency Distributions and Relative Frequency Histograms

Relative Frequency Distribution

- Shows the portion or percentage of the data that falls in a particular class.
- $$\text{relative frequency} = \frac{\text{class frequency}}{\text{Sample size}} = \frac{f}{n}$$

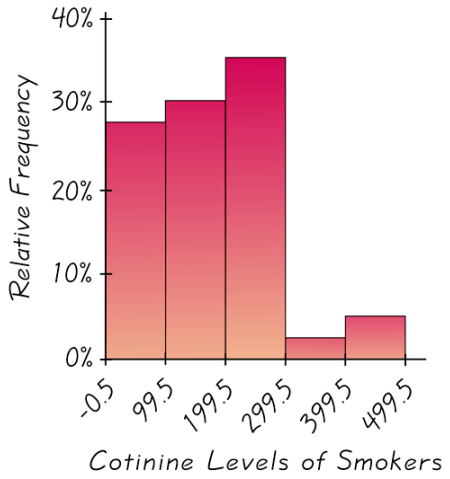
Relative Frequency Histogram

- Has the same shape and the same horizontal scale as the corresponding frequency histogram.
- The vertical scale measures the **relative frequencies**, not frequencies.

Relative Frequency Histogram

Has the same shape and horizontal scale as a histogram, but the vertical scale is marked with relative frequencies.

Cotinine	Relative Frequency
0–99	28%
100–199	30%
200–299	35%
300–399	3%
400–499	5%



Grouped Frequency Distributions

Grouped Frequency Distribution

- For data sets with many different values.
- Groups data into 5-20 classes of equal width.

Exam Scores	Freq, f

Exam Scores	Freq, f
30-39	
40-49	
50-59	
60-69	
70-79	
80-89	
90-99	

Exam Scores	Freq, f
30-39	**1**
40-49	**0**
50-59	**4**
60-69	**9**
70-79	**13**
80-89	**10**
90-99	**3**

Grouped Frequency Distribution Terms

- **Lower class limits:** are the smallest numbers that can actually belong to different classes
- **Upper class limits:** are the largest numbers that can actually belong to different classes
- **Class width:** is the difference between two consecutive lower class limits

Labeling Grouped Frequency Distributions

- **Class midpoints:** the value halfway between LCL and UCL:

$$\frac{(\text{Lower class limit}) + (\text{Upper class limit})}{2}$$

- **Class boundaries**: the value halfway between an UCL and the next LCL

$$\frac{(\text{Upper class limit}) + (\text{next Lower class limit})}{2}$$

Constructing a Grouped Frequency Distribution

1. Determine the range of the data:
   - Range = highest data value – lowest data value
   - May round up to the next convenient number
2. Decide on the number of classes.
   - Usually between 5 and 20; otherwise, it may be difficult to detect any patterns.
3. Find the class width:
   - $\text{class width} = \dfrac{\text{range}}{\text{number of classes}}$
   - *Round up to the next convenient number.*

Constructing a Frequency Distribution

4. Find the class limits.
   - Choose the first LCL: use the minimum data entry or something smaller that is convenient.
   - Find the remaining LCLs: add the class width to the lower limit of the preceding class.
   - Find the UCLs: Remember that classes must cover all data values and cannot overlap.
5. Find the frequencies for each class. (You may add a tally column first and make a tally mark for each data value in the class).

"Shape" of Distributions

Symmetric

- Data is symmetric if the left half of its histogram is roughly a mirror image of its right half.

Skewed

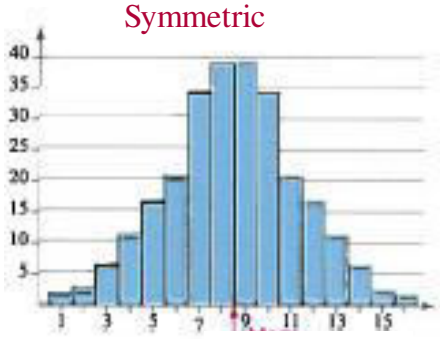
- Data is skewed if it is not symmetric and if it extends more to one side than the other.

Uniform

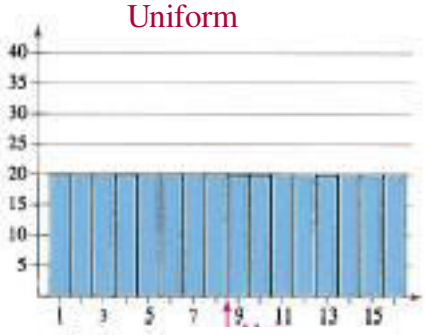
- Data is uniform if it is equally distributed (on a histogram, all the bars are the same height or approximately the same height).

The Shape of Distributions

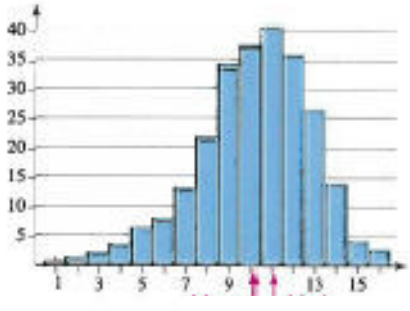
Symmetric



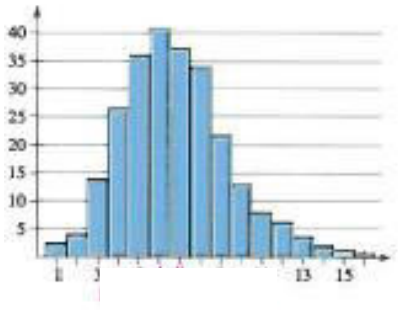
Uniform



Skewed left



Skewed Right



Outliers

Outliers

- Unusual data values as compared to the rest of the set. They may be distinguished by gaps in a histogram.

Section 2.2

More Graphs and Displays

Other Graphs

Besides Histograms, there are other methods of graphing quantitative data:

- **Stem and Leaf Plots**
- **Dot Plots**
- **Time Series**

Stem and Leaf Plots

Represents data by separating each data value into two parts: the stem (such as the leftmost digit) and the leaf (such as the rightmost digit)

Stem-and-Leaf Plot

Stem (tens)	Leaves (units)	
6	449	← Values are 64, 64, 69.
7	01112334444555555666778899	
8	0011122233346899	
9	0024	
10		
11		
12	0	← Value is 120.

Constructing Stem and Leaf Plots

- Split each data value at the same place value to form the **stem** and a **leaf**. (Want 5-20 stems).
- Arrange all possible stems vertically so there are no missing stems.
- Write each leaf to the right of its stem, in order.
- Create a key to recreate the data.
- Variations of stem plots:
  1. Split stems
  2. Back to back stem plots.

Constructing a Stem-and-Leaf Plot

Number of Text Messages Sent

7	8	Key: 15|5 = 155
8		
9		
10	5 8 9 9 9	
11	6 4 2 2 8 8 9 3 7 8 9 9 2	
12	9 6 2 6 2 1 6 2 6 3 1 4 4 9 6	
13	0 9 9 3 4 2 3	
14	4 5 2 0 5 8 7	
15	5 9	

Unordered Stem-and-Leaf Plot

Number of Text Messages Sent

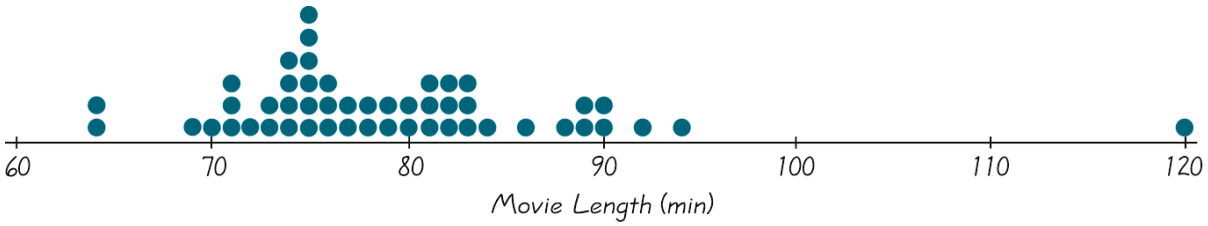
7	8	Key: 15|5 = 155
8		*Include a key to identify*
9		*the values of the data.*
10	5 8 9 9 9	
11	2 2 2 3 4 6 7 8 8 8 9 9 9	
12	1 1 2 2 2 3 4 4 6 6 6 6 6 9 9	
13	0 2 3 3 4 9 9	
14	0 2 4 5 5 7 8	
15	5 9	

Ordered Stem-and-Leaf Plot

Dot Plots

Dot plot

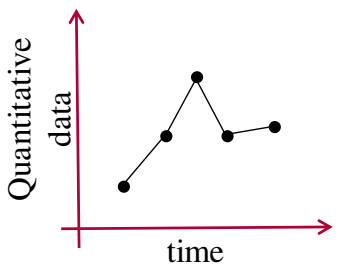
- Consists of a graph in which each data value is plotted as a point along a scale of values



Time Series (Paired data)

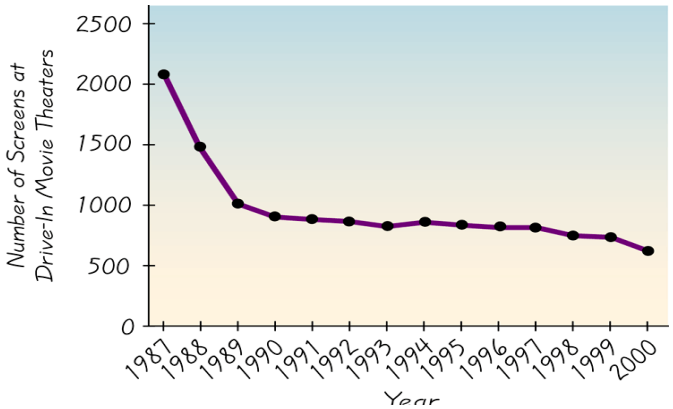
Time Series

- Data set is composed of quantitative entries taken at regular intervals over a period of time.
  - e.g., The amount of precipitation measured each day for one month.
- Use a **time series chart** to graph.



Time-Series Graph

Number of Screens at Drive-In Movies Theaters

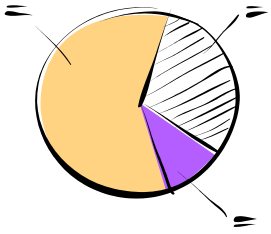


Ex. www.eia.doe.gov/oil_gas/petroleum/

Graphing Qualitative Data Sets

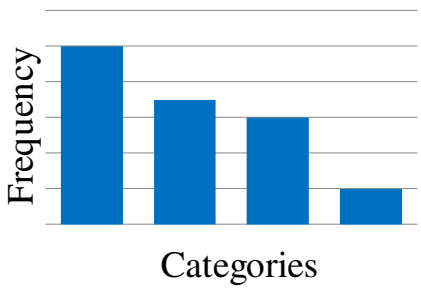
Pie Chart

- A circle is divided into sectors that represent categories.



Pareto Chart

- A vertical bar graph in which the height of each bar represents frequency or relative frequency.



Constructing a Pie Chart

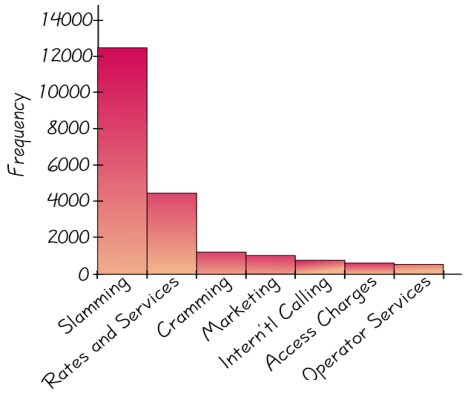
- Find the total sample size.
- Convert the frequencies to relative frequencies (percent).

Marital Status	Frequency, f (in millions)	Relative frequency (%)
Never Married	55.3	$\frac{55.3}{219.7} \approx 0.25$ or 25%
Married	127.7	$\frac{127.7}{219.7} \approx$
Widowed	13.9	$\frac{13.9}{219.7} \approx$
Divorced	22.8	$\frac{22.8}{219.7} \approx$

Total: 219.7

Constructing Pareto Charts

- Create a bar for each category, where the height of the bar can represent frequency or relative frequency.
- The bars are often positioned in order of decreasing height, with the tallest bar positioned at the left.



Boxplots

- Another way to graph the distribution of a numerical variable is through a boxplot (aka box-and-whisker plot).
- A boxplot is a visual representation of the five-number summary of the distribution of a numerical variable. This consists of:
  - The minimum value of the distribution
  - The first quartile
  - The median
  - The third quartile
  - The maximum value of the distribution

Steps to Make a Boxplot

1) Draw a central box (rectangle) from the first quartile to the third quartile
2) Draw a vertical line to mark the median
3) Draw horizontal lines (called *whiskers*) that extend from the box out to the smallest and largest observations that are <u>not</u> outliers
4) If there are any outliers, mark them separately

- Let's go back to our Chris Johnson example.
- Let's reexamine his rushing attempts, along with other key data.

-3 -3 -2 0 0 2 4 4 4 4 6 6 11 16 57 91

$Q_1 = 0 \quad M = 4 \quad Q_3 = 8.5 \quad IQR = 8.5$

Outliers of 57 and 91

Now let's look at the boxplot of this data.

Let's now go back to our Tom Brady example.

Here were his passer ratings, along with other key data we calculated.

62.9	79.6	58.7	93.4	148.3	57.1	124.4	78.9
70.8	143.9	93.3	61.3	63.6	91.6	62.1	

$$Q_1 = 62.1 \quad M = 78.9 \quad Q_3 = 93.4 \quad IQR = 31.3$$

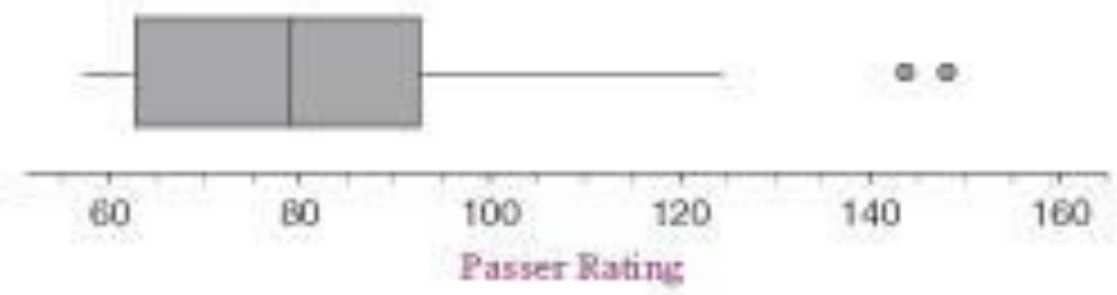
Are there any outliers?

$Q_1 - 1.5IQR$ $\qquad$ $Q_3 + 1.5IQR$

$62.1 - 1.5(31.3) = 15.15$ $\qquad$ $93.4 + 1.5(31.3) = 140.35$

The observations 143.9 and 148.3 are outliers.

…and the boxplot



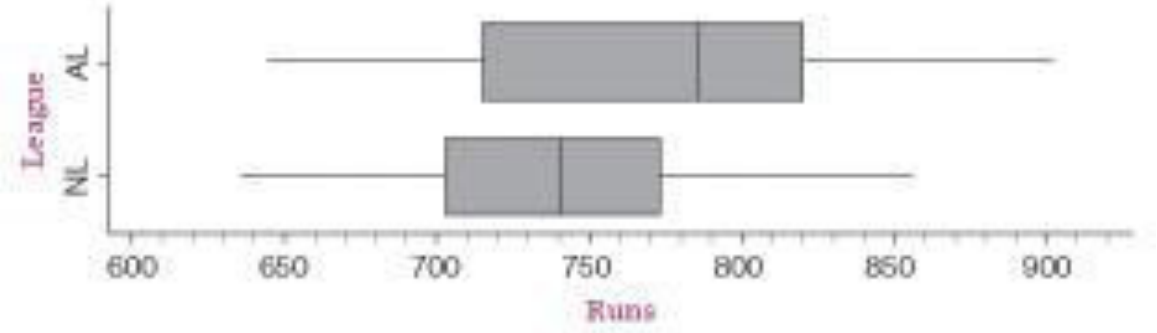
Comparing Distributions

- When asked to compare two distributions, you must address four points:
  - The shape
  - The outliers
  - The center
  - The spread

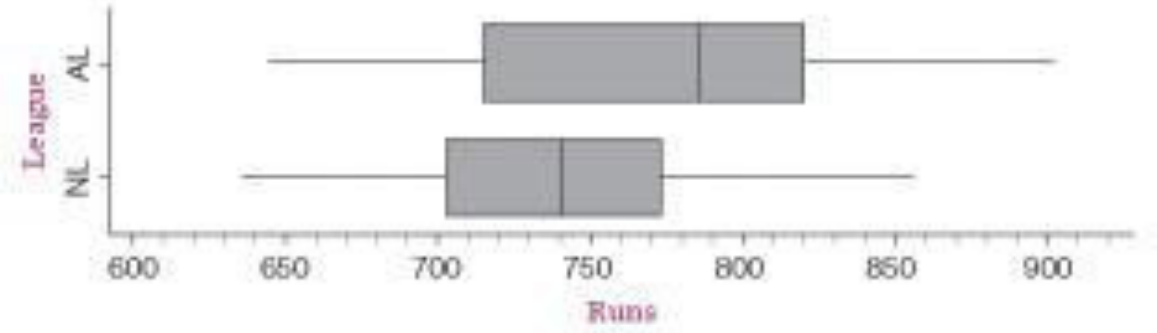
- Think of the acronym SOCS to help you remember what to address.

- The shape of a distribution may be difficult to determine from a boxplot.
- Try comparing the distance from the median to the minimum and maximum values to determine if a distribution is skewed or roughly symmetric.
- You will <u>not</u> be able to tell if a distribution is unimodal from looking at a boxplot.

- Here are boxplots for the number of runs scored in the AL and in the NL during 2008. (Note: the plots are on the same scale for comparison purposes.)



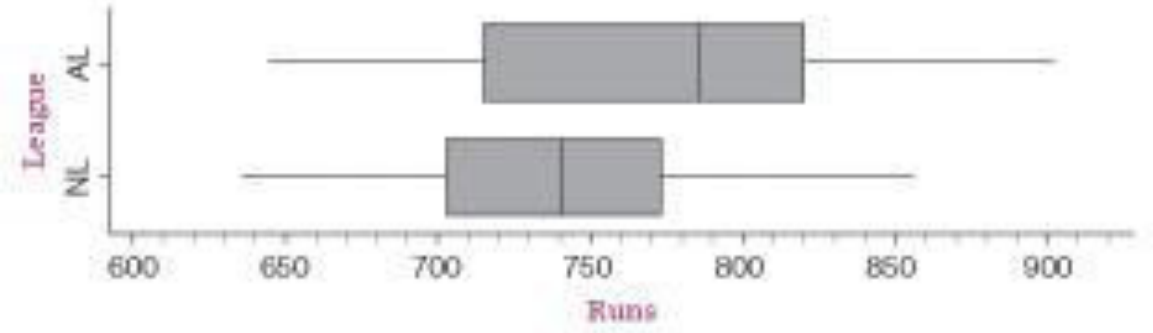
- Let's compare using our four points.



<u>Shape</u>

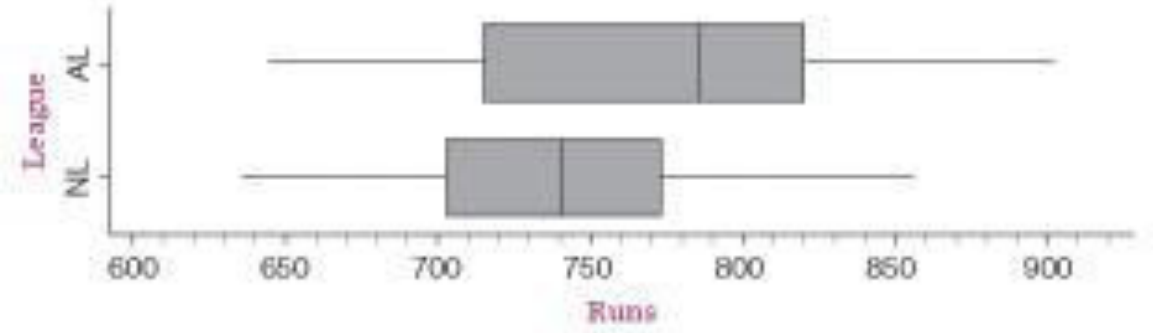
The AL distribution is skewed slightly left (the left half of the distribution appears more spread out).

The NL distribution is approximately symmetric.



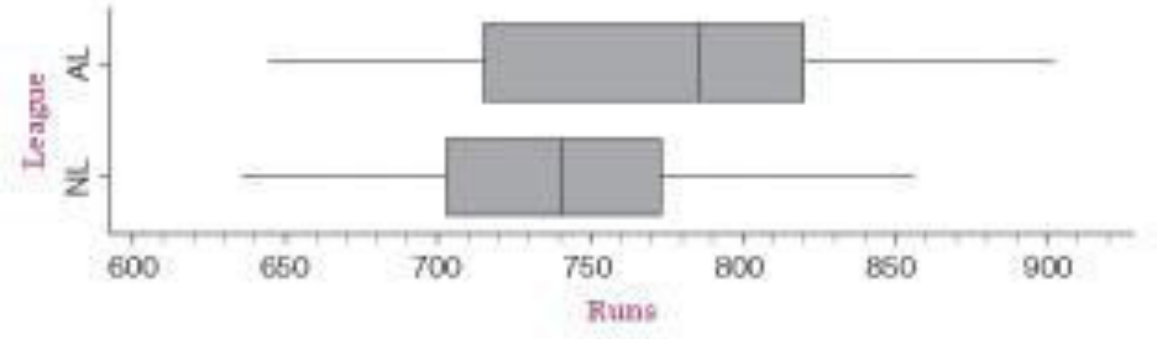
Outliers

Neither distribution contains an outlier.



Center

Typically, teams score more runs in the AL because the median for the AL distributions is higher than the median for the NL distributions.



Spread

- The AL distribution is slightly more spread out because it has both a larger range and larger *IQR*.
- This indicates there is more variability among AL teams and more consistency among NL teams.

Using the TI-84 to Make Graphs and Calculate Summary Statistics

- As fun as it is to calculate everything by hand, the TI-84 calculator can do many of our calculations for us.
- The calculator can create boxplots, histograms, and calculate summary statistics.

Boxplot

- Let's use our 2008 run data.
- Here are the numbers:

AL runs scored:

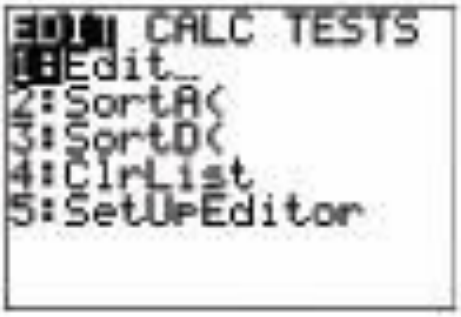
782 845 811 805 821 691 765 829 789
646 671 774 901 714

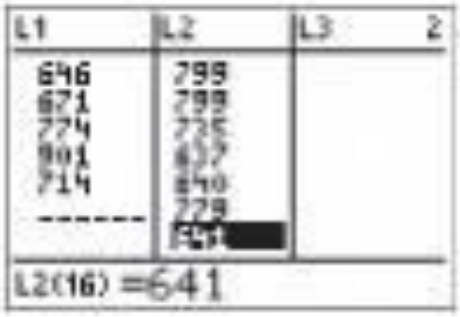
NL runs scored:

720 753 855 704 747 770 712 700 750
799 799 735 637 640 779 641

Write these numbers down or open to pg 120!

- The first thing we have to do is store this data as a list.
- Press STAT and choose the first option EDIT
- Enter the 14 AL data values in L1 and the 16 NL values in L2





Now we are going to set up the boxplot. Exit back into the home screen.

Then press STAT PLOT (2nd and y=).

Choose Plot1. Then, turn Plot1 on.

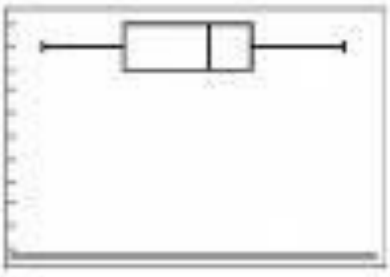
Scroll to Type and choose the boxplot icon (with outliers). It is the first option in the second row.

Enter L1 for Xlist.

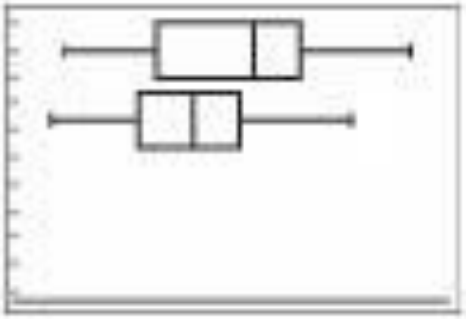
Enter 1 for Freq. Choose a mark for outliers.

Now we will display the graph. Press ZOOM. Then select option 9: ZOOMSTAT. Press enter.

Press TRACE and scroll around to see different statistics for the distribution.



- To see the boxplot for the NL distribution at the same time:
- Go back into STAT PLOT and turn on Plot2. Repeat the steps, but enter L2 for Xlist. To do this, scroll down to Xlist. Then press 2^{nd}-2 (you will see the L2 button on top of the number 2).



PPT is not Available in LMS for some lectures is missing

Applied Biostatistics

Quantiles

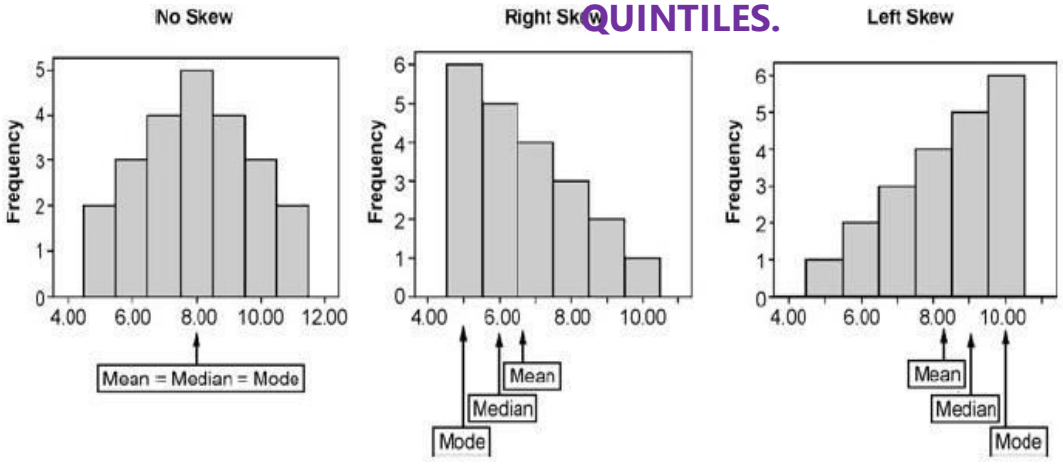
Lecture no 44

Quantiles

The Mean and Median are Special cases of a family of parameters known as location parameters.

Other location parameters which help us to determine the position of a particular observations in a data are called Measures of

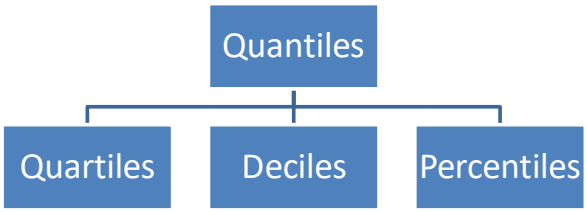
These Measures of Position when help us to divide the sample data into n equal parts, or divide the probability distribution into contiguous interval with equal probabilities are called QUINTILES.



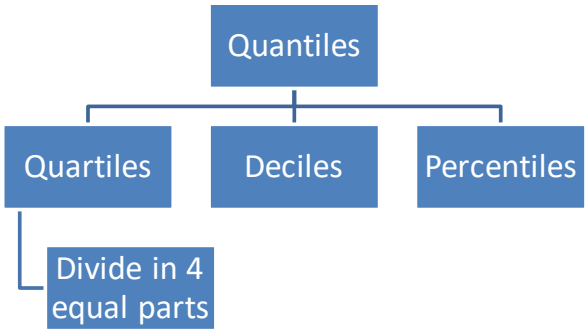
Quantiles

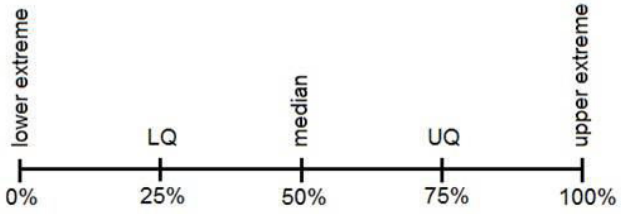
q – quantiles are such q – 1 values that partition a finite set of data into q subsets of equal sizes.

These Quantiles are also called Cut-points. The most simplest and well known cut-point is Median.



Quantiles





https://www.mathplanet.com/education/pre-algebra/probability-and-statistic/stem-and-leaf-plots-and-box-and-whiskers-plot

In an array the Quartiles are 3 cut-points that divides the values into 4 equal parts.

Denoted by Q_k

k = the k-th value

$$Q_k = \left(\frac{k(n+1)}{4}\right)^{th} Value$$

$$Q_1 = \left(\frac{1(n+1)}{4}\right)^{th} Value$$

$$Q_2 = \left(\frac{2(n+1)}{4}\right)^{th} Value$$

$$Q_3 = \left(\frac{3(n+1)}{4}\right)^{th} Value$$

Quantiles

Method for Determining Quartiles.

Step 1: Arrange the data in ascending order.

Step 2: Assign an index to each data point.

Step 3: Find the position of the respective Quartile using following formula.

$$Q_k = \left(\frac{k(n+1)}{4}\right)^{th} Value \qquad k = 1, 2, 3$$

Step 4: Identify the value of the quartile.

a) If a quartile lies on an observation (i.e. if its position is a whole number), the value of the quartiles is the value of that observation. For example, if the position of a quartile is 20, its value is the value of the 20th observation.

b) If a quartile lies between observations, the values of the quartile is the value of the lower observation plus the specified fraction of the difference between the observations. For example, if the position of a quartile is 20¼, it lies between the 20th and 21st observations, and its value is the value of the 20th observation, plus ¼ the difference between the value of the 20th and 21st observations.

https://www.stat.berkeley.edu/users/rabbee/s/dsepd/ss1978/lesson2/section7.html#TXT27

Quantiles

$$Q_k = \left(\frac{k(n+1)}{4}\right)^{th} Value$$

$$Q_1 = \left(\frac{1(n+1)}{4}\right)^{th} Value$$

$$Q_1 = \left(\frac{1(9+1)}{4}\right)^{th} Value = (2.5)^{th}\, value$$

$$Q_2 = \left(\frac{2(n+1)}{4}\right)^{th} Value$$

$$Q_2 = \left(\frac{2(9+1)}{4}\right)^{th} Value = (5)^{th}\, value$$

$$Q_3 = \left(\frac{3(n+1)}{4}\right)^{th} Value$$

$$Q_3 = \left(\frac{4(9+1)}{4}\right)^{th} Value = (7.5)^{th}\, value$$

Example

35, 35, 37, 40, 43, 56, 58, 58, 60

n = 9

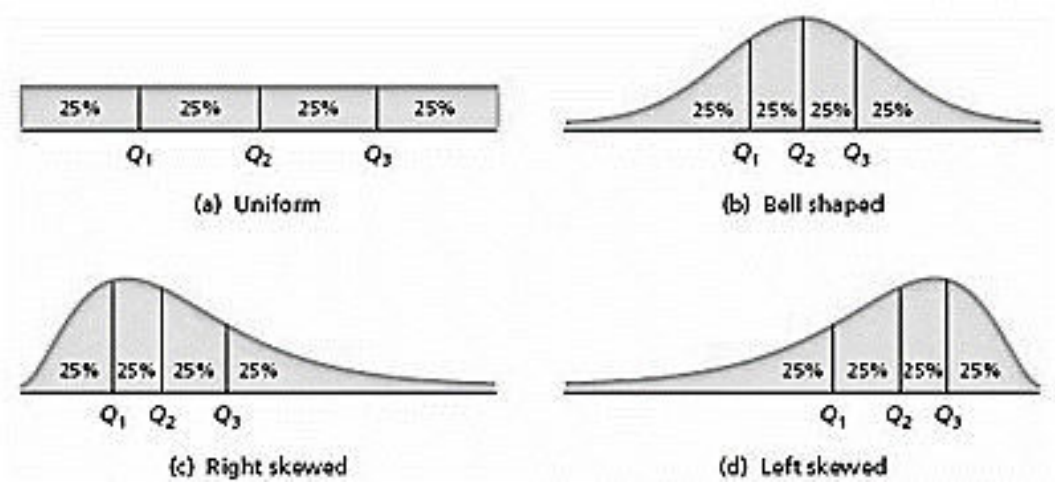
$LQ = \frac{35+37}{2} = 36$ $\qquad$ $UQ = \frac{58+58}{2} = 58$

(35, 35, 37, 40,) 43, (56, 58, 58, 60)

median

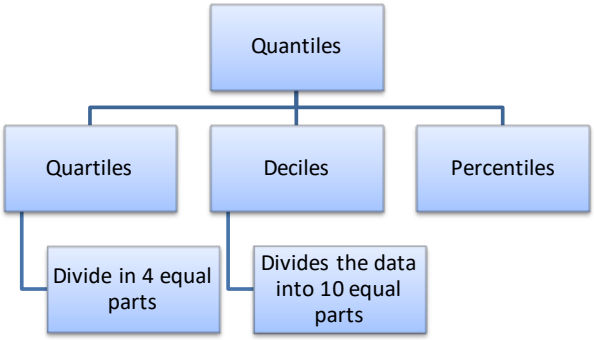
https://www.mathplanet.com/education/pre-algebra/probability-and-statistic/stem-and-leaf-plots-and-box-and-whiskers-plot

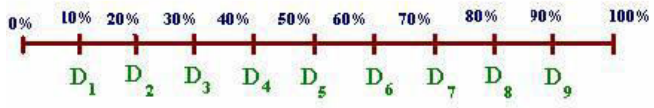
Quantiles



http://www.brainfuse.com/curriculumupload//1363810898934.html

Quantiles





http://dieumsnh.qfb.umich.mx/estadistica/deciles.htm

In an array the Deciles are 9 values/cut-points that divides the values into 10 equal parts.

Denoted by D_k

k = the k-th value

$$D_k = \left(\frac{k(n+1)}{10}\right)^{th} Value$$

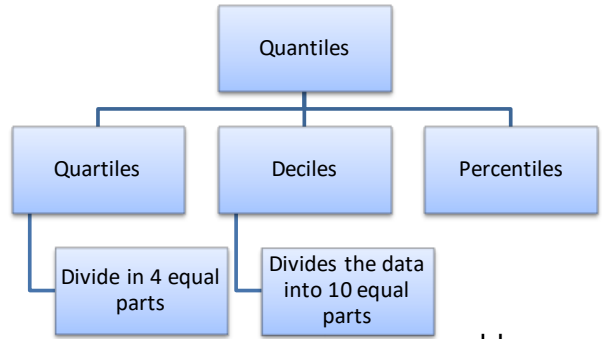
$$D_1 = \left(\frac{(n+1)}{10}\right)^{th} Value$$

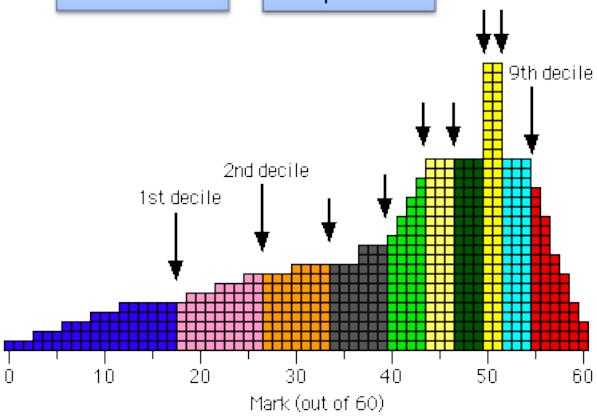
$$D_5 = \left(\frac{5(n+1)}{10}\right)^{th} Value$$

$$\vdots$$

$$D_9 = \left(\frac{9(n+1)}{10}\right)^{th} Value$$

Quantiles





http://www-ist.massey.ac.nz/dstirlin/cast/cast/scentre/centre5.html

In an array the Deciles are 9 values/cut-points that divides the values into 10 equal parts.

Denoted by D_k

k = the k-th value

$$D_k = \left(\frac{k(n+1)}{10}\right)^{th} Value$$

$$D_1 = \left(\frac{(n+1)}{10}\right)^{th} Value$$

$$D_5 = \left(\frac{5(n+1)}{10}\right)^{th} Value$$

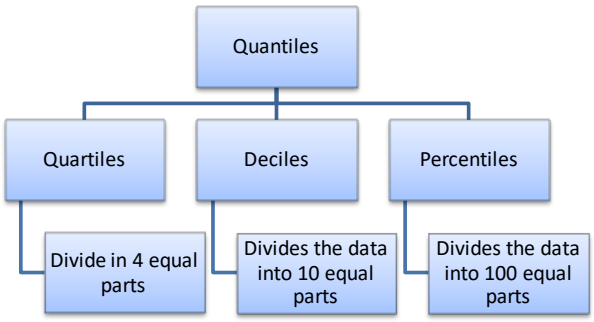
$$\vdots$$

$$D_9 = \left(\frac{9(n+1)}{10}\right)^{th} Value$$

Applied Biostatistics

Quantiles - II

Quantiles - II



In an array the Percentiles are 99 values/cut-points that divides the values into 100 equal parts.

Denoted by P_k

k = the k-th value

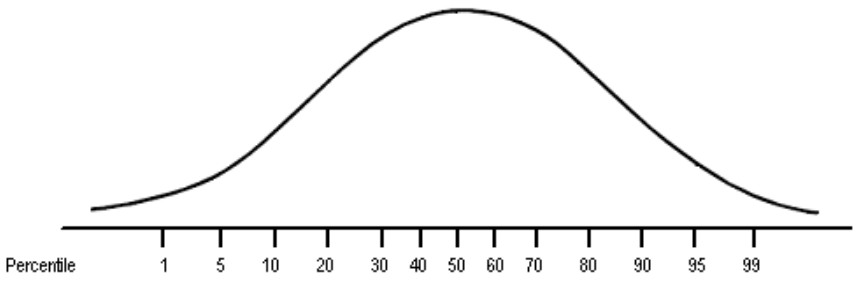
Given a set of n observations $x_1, x_2, \ldots x_n$, the pth percentile P is the value of X such that p percent or less of the observations are less than P and $(100 - p)$ percent or less of the observations are greater than P.

$$P_k = \left(\frac{k(n+1)}{100}\right)^{th} Value \qquad k = 1, 2, 3, \ldots, 99$$

Quantiles - II

Given a set of n observations $x_1, x_2, \ldots x_n$, the pth percentile P is the value of X such that p percent or less of the observations are less than P and $(100 - p)$ percent or less of the observations are greater than P.

$$P_k = \left(\frac{k(n+1)}{100}\right)^{th} Value \qquad k = 1, 2, 3, \ldots, 99$$



http://grants.hhp.uh.edu/MLaughlin/PEP6305/Topic%20005%20Normal%20Distribution.htm

Quantiles - II

Quantiles are more robust and less susceptible than mean if there are outlying observations in the data, or our frequency distribution is Right or Left Tailed.

Quantiles of a random variable are preserved under monotonic transformation.

Standardized test results use percentiles.

Quantiles - II

END

$$Q_1 = P_{25}$$

$$Q_2 = D_5 = P_{50} = Median$$

$$Q_3 = P_{75}$$

$$D_1 = P_{10}$$

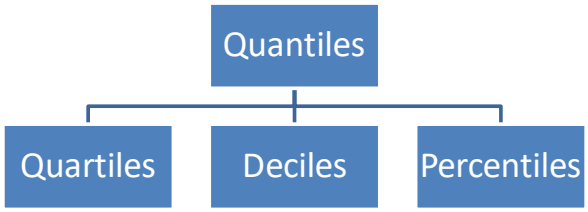
$$D_2 = P_{20}$$

$$D_9 = P_{90}$$

Applied Biostatistics

Examples of Quantiles

Examples of Quantiles



Steps for the Calculation of Percentiles:

Step 1: Order the data into Ascending order.

Step 2: Assign an index to data point

Step 3: Calculate $P_k = \left(\frac{k(n+1)}{100}\right)^{th} Value \quad k = 1, 2, 3, \ldots, 99$

Step 4: A data point that corresponds to the index calculated at Step 3 will be a required percentile.

Examples of Quantiles

GRF Measurements When Trotting of 20 Dogs with a Lame Ligament

14.6	24.3	24.9	27.0	27.2	27.4	28.2	28.8	29.9	30.7
31.5	31.6	32.3	32.8	33.3	33.6	34.3	36.9	38.3	44.0

The effect of velocity on ground reaction forces (GRF) in dogs with lameness from a torn cranial cruciate ligament. The dogs were walked and trotted over a force platform, and the GRF was recorded during a certain phase of their performance. Given 20 observations show the mean of five force measurements per dog when trotting.

Examples of Quantiles

GRF Measurements When Trotting of 20 Dogs with a Lame Ligament

14.6	24.3	24.9	27.0	27.2	27.4	28.2	28.8	29.9	30.7
31.5	31.6	32.3	32.8	33.3	33.6	34.3	36.9	38.3	44.0

$$P_{25} = \left(\frac{25(20+1)}{100}\right)^{th} Value = (5.25)^{th}\ Value$$

$$P_{25} = (5.25)^{th}\, value = \ 5^{th} value + 0.25\left(6^{th} - 5^{th}\right)$$

$$P_{25} = 27.2 + 0.25\,(27.4 - 27.2) = 27.25$$

$$P_{50} = \left(\frac{50(20+1)}{100}\right)^{th} Value = (10.5)^{th}\ Value$$

$$P_{50} = (10.5)^{th}\, value = \ 10^{th} value + 0.5\left(11^{th} - 10^{th}\right)$$

$$P_{50} = 30.7 + 0.50\,(31.5 - 30.7) = 31.1$$

$$P_{75} = \left(\frac{75(20+1)}{100}\right)^{th} Value = (15.75)^{th}\ Value$$

$$P_{75} = (15.75)^{th}\, value = \ 15^{th} value + 0.75\left(16^{th} - 15^{th}\right)$$

$$P_{75} = 33.3 + 0.75\,(33.6 - 33.3) = 33.525$$

Examples of Quantiles

GRF Measurements When Trotting of 20 Dogs with a Lame Ligament									
14.6	24.3	24.9	27.0	27.2	27.4	28.2	28.8	29.9	30.7
31.5	31.6	32.3	32.8	33.3	33.6	34.3	36.9	38.3	44.0

$$P_{25} = Q_1 = 25.25$$

This value shows that in our data 25% i.e. 5 dogs have their GRF measurements less than 25.25 Newtons, and 75% i.e. 15 have their GRF measurements above 25.25 Newtons

$$P_{50} = Q_2 = 31.1$$

This value shows that in our data 50% i.e. 10 dogs have their GRF measurements less than 31.1 Newtons and the rest of the 50% have their GRF measurements above 31.1 Newtons

$$P_{75} = Q_3 = 33.525$$

This value shows that in our data 75% i.e. 15 dogs have their GRF measurements less than 33.525 Newtons and the rest of the 25% have their GRF measurements above 33.525 Newtons

Examples of Quantiles

END

In a similar way one can calculate deciles, and other Quantile / cut points.

Quantiles also help us to determine the location and spread of the distribution.

MIDTERM EXAM Syllabus

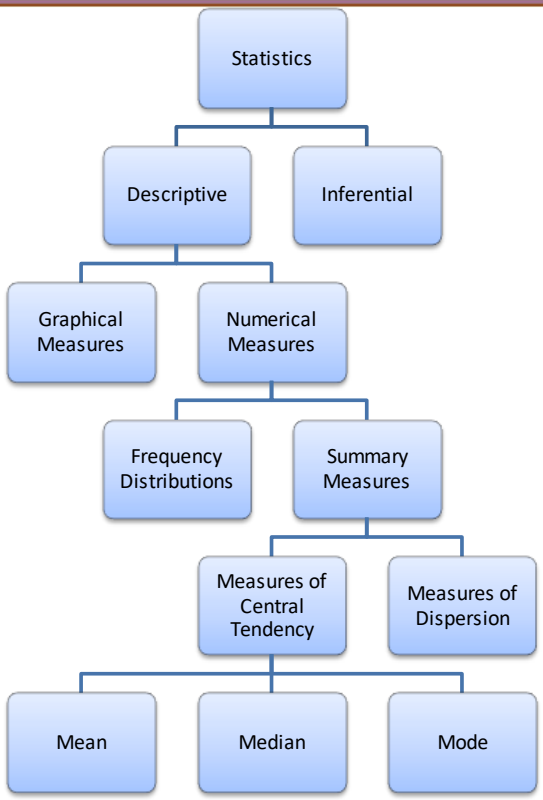
Applied Biostatistics

Measures of Dispersion

Lecture no 47

Final Term start

Measures of Dispersion



Measures of Dispersion

Dispersion is defined as the extent to which the observations in the dataset are spread away from the central value.

Measures of Dispersion are the methods that convey information regarding the amount of variability present in the set of data.

If all the values in the data are the same then there is no dispersion.

Other terms used synonymously with dispersion includes variation, spread and scatter.

Measures of Dispersion

Fasting Plasma Glucose levels of 7 individuals in two groups are given.

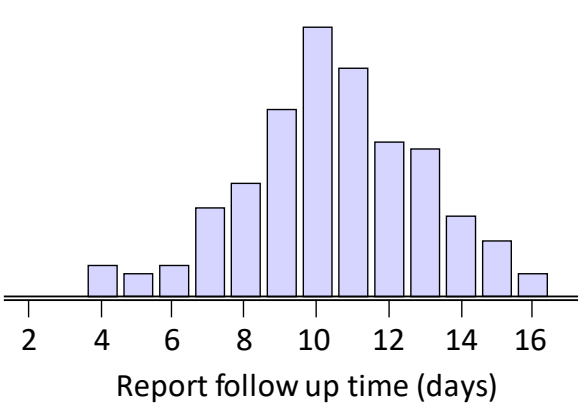
Group 1	Group 2
x	y
100	120
110	105
90	80
105	100
95	95
101	103
99	97

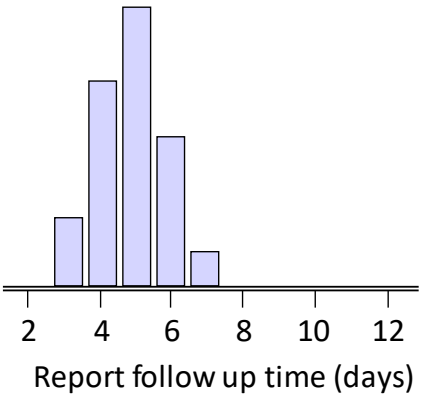
$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{700}{7} = 100$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{700}{7} = 100$$

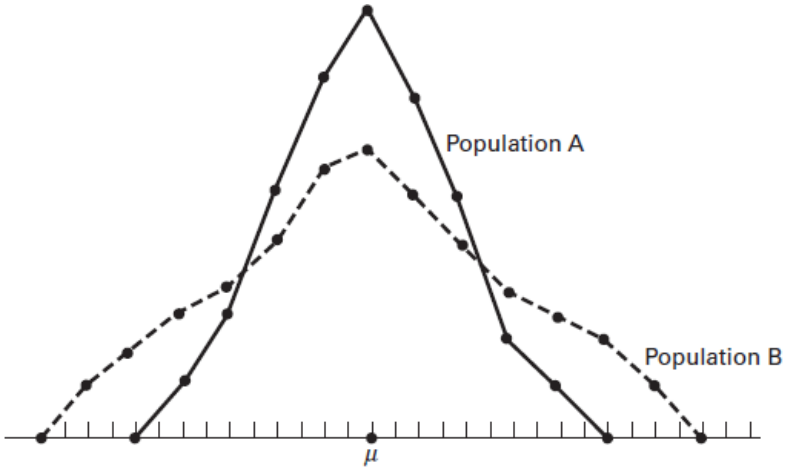
Measures of Dispersion

- **The dispersion in a set of data is the variation among the set of data values.**
- **It measures whether they are all close together, or more scattered.**





Measures of Dispersion

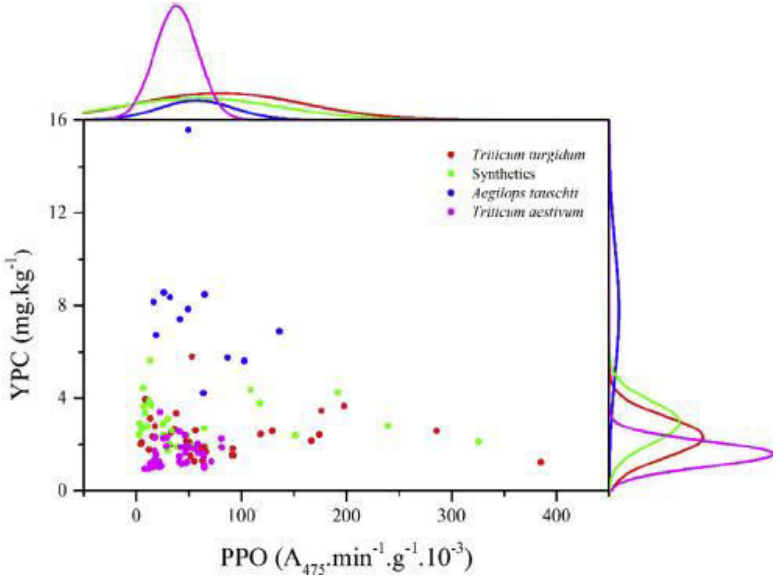


Two frequency distributions with equal means but different amounts of dispersion.

Measures of Dispersion

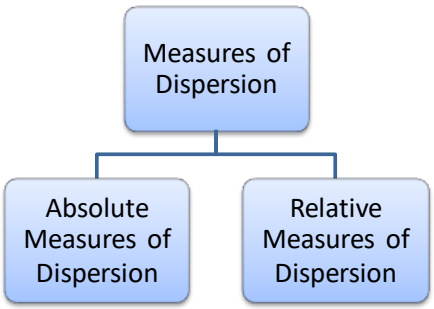
PPO = Polyphenol Oxidase activity

YPC = Yellow Pigment Content.



Li F. Y *et al* (2015), "Polyphenol oxidase activity and yellow pigment content in Aegilops tauschii, Triticum Turgidum, Triticum aestivum, syntehtic hexaploid wheat and its parents." Journal of cereal sciences (65) pp 192 – 201.

Measures of Dispersion



Measures which are expressed in terms of the same units as the original data are termed as Absolute Measures of Dispersion.

Measures which are expressed in terms of the ratios and percentages these measures are independent of the units of measurement and termed as Relative Measures of Dispersion.

Measures of Dispersion

END

Each absolute measure of dispersion can be converted into its relative measure.

Relative measure of dispersion are used to make comparisons between different datasets, irrespective of their units of measurements.

Applied Biostatistics

Absolute Measures of Dispersion – I

Lecture no 48

Absolute Measures of Dispersion - I

Absolute Measures of Dispersion are of different types.

- **Range**
- **Quartile Deviation**
- **Mean Absolute Deviation**
- **Standard Deviation and Variance**

Absolute Measures of Dispersion - I

Range

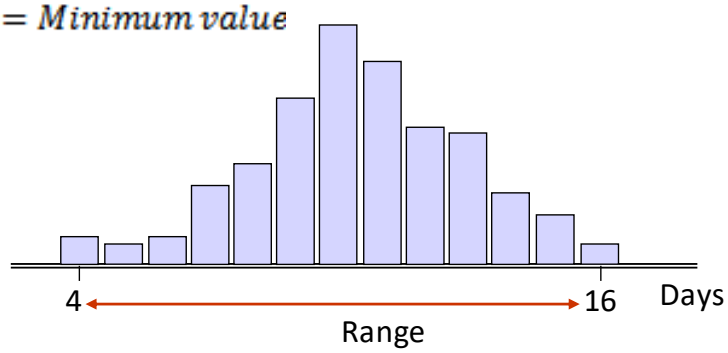
Range is defined as the difference between the maximum and the minimum value in the data.

It is sensitive to only the most extreme values in the list. The range of a list is 0 if and only if all the data-points in the list are equal.

$$Range = x_m - x_o$$

$$x_m = Maximum\ value$$

$$x_o = Minimum\ value$$



Absolute Measures of Dispersion - I

Example:

GRF Measurements When Trotting of 20 Dogs with a Lame Ligament

14.6	24.3	24.9	27.0	27.2	27.4	28.2	28.8	29.9	30.7
31.5	31.6	32.3	32.8	33.3	33.6	34.3	36.9	38.3	44.0

$$x_o = Minimum\ value = 14.6$$

$$x_m = Maximum\ value = 44.0$$

$$Range = x_m - x_o = 44.0 - 14.6 = 29.4\ Newtons$$

Absolute Measures of Dispersion - I

Pros and Cons of Range

- **Pros**
  - best for symmetric data with no outliers
  - easy to compute and understand
  - good option for ordinal data
- **Cons**
  - doesn't use all of the data, only the extremes
  - very much affected if the extremes are outliers
  - only shows maximum spread, does not show shape

Absolute Measures of Dispersion - I

Absolute Measures of Dispersion are of different types.

- **Range**
- **Quartile Deviation**
- **Mean Absolute Deviation**
- **Standard Deviation and Variance**

Absolute Measures of Dispersion - I

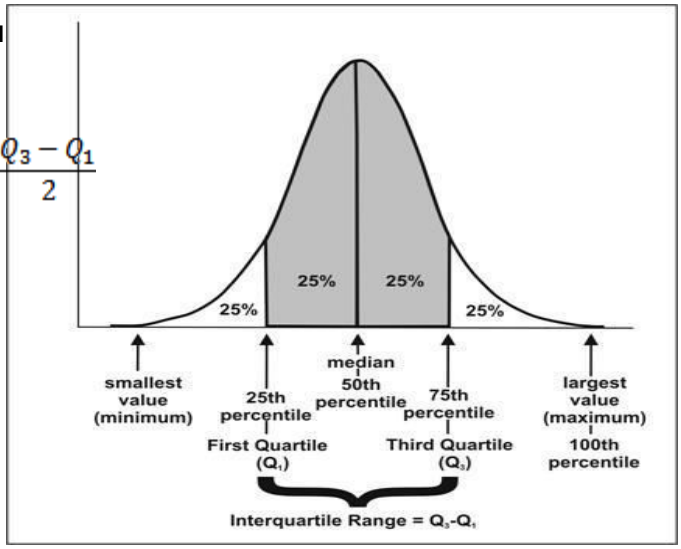
Quartile Deviation (Q.D.)

- **Quartile Deviation is the half distance between the Upper Quartile Q_3 (i.e. 75th percentile) and Lower Quartile Q_1 (i.e. 25th Percentile).**
- **Essentially describes that how much the middle 50% of the data varies.**
- **Also known as semi – interqu**

$$Q.D. = \frac{Inter\ Quartile\ Range\ (IQR)}{2} = \frac{Q_3 - Q_1}{2}$$

$$Q_1 = \left(\frac{1(n+1)}{4}\right)^{th} Value$$

$$Q_3 = \left(\frac{3(n+1)}{4}\right)^{th} Value$$



https://www.cdc.gov/ophss/csels/dsepd/ss1978/lesson2/section7.html#TXT27

Absolute Measures of Dispersion - I

Inter – Quartile Range (IQR)

Method for Determining the Inter – Quartile Range.

Step 1: Arrange the data in and ascending order.

Step 2: Find the position of the lower Quartile i.e. Q_1 and upper Quartile i.e. Q_3 using following formula.

$$Q_1 = \left(\frac{1(n+1)}{4}\right)^{th} Value$$

$$Q_3 = \left(\frac{3(n+1)}{4}\right)^{th} Value$$

Absolute Measures of Dispersion - I

Inter – Quartile Range (IQR)

Method for Determining the Inter – Quartile Range.

Step 3: Identify the value of the 1st and 3rd quartiles.

1. If a quartile lies on an observation (i.e. if its position is a whole number), the value of the quartiles is the value of that observation.

2. If a quartile lies between observations, the values of the quartile is the value of the lower observation plus the specified fraction of the difference between the observations. For example, if the position of a quartile is 20¼, it lies between the 20th and 21st observations, and its value is the value of the 20th observation, plus ¼ the difference between the value of the 20th and 21st observations.

Step 4: Calculate Inter Quartile Range (IQR) using the following formula.

$$Inter\ Quartile\ Range\ (IQR) = Q_3 - Q_1$$

https://www.cdc.gov/ophss/csels/dsepd/ss1978/lesson2/section7.html#TXT27

Absolute Measures of Dispersion - I

GRF Measurements When Trotting of 20 Dogs with a Lame Ligament

14.6	24.3	24.9	27.0	27.2	27.4	28.2	28.8	29.9	30.7
31.5	31.6	32.3	32.8	33.3	33.6	34.3	36.9	38.3	44.0

$$Q_3 = \left(\frac{3(20+1)}{4}\right)^{th} Value = (15.75)^{th}\ Value$$

$$Q_3 = (15.75)^{th}\ value = 15^{th} value + 0.75\left(16^{th} - 15^{th}\right)$$

$$Q_3 = 33.3 + 0.75\,(33.6 - 33.3) = 33.525\ \mathbf{N}$$

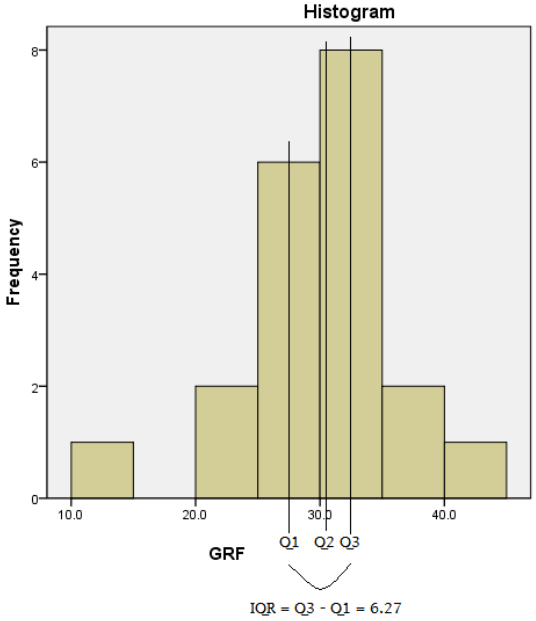
$$Q_1 = \left(\frac{1(20+1)}{4}\right)^{th} Value = (5.25)^{th}\ Value$$

$$Q_1 = (5.25)^{th}\ value = 5^{th} value + 0.25\left(6^{th} - 5^{th}\right)$$

$$Q_1 = 27.2 + 0.25\,(27.4 - 27.2) = 27.25\ \mathbf{N}$$

$$IQR = Q_3 - Q_1 = 33.525 - 27.25 = 6.275$$

$$Q.D = \frac{IQR}{2} = \frac{6.275}{2} = 3.138$$ **N**



Absolute Measures of Dispersion - I

Properties and uses of the Inter-Quartile Range

- The IQR is generally used in conjunction with the median. Together, they are useful for characterizing the central location and spread of any frequency distribution, but particularly those that are skewed.
- For a more complete characterization of a frequency distribution, the 1st and 3rd quartiles are sometimes used with the minimum value, the median, and the maximum value to produce a five number summary of the distribution.
- Together these five numbers provide a good description of the centre, spread and shape of the distribution.

Absolute Measures of Dispersion - I

Pros and Cons of Quartile Deviation

- **Pros**
  - Good for ordinal data.
  - Good for Skewed data.
  - Ignores extreme values
  - More stable than the range because it ignores outliers
- **Cons**
  - Harder to calculate and understand
  - Doesn't use all the information (ignores half of the data-points, not just the outliers)
    - Tails almost always matter in data and these aren't included
    - Outliers can also sometimes matter and again these aren't included.

Absolute Measures of Dispersion - I

END

Each absolute measure of dispersion can be converted into its relative measure.

Relative measure of dispersion are used to make comparisons between different datasets, irrespective of their units of measurements.

Applied Biostatistics

Absolute Measures of Dispersion - II

Absolute Measures of Dispersion - II

Absolute Measures of Dispersion are of different types.

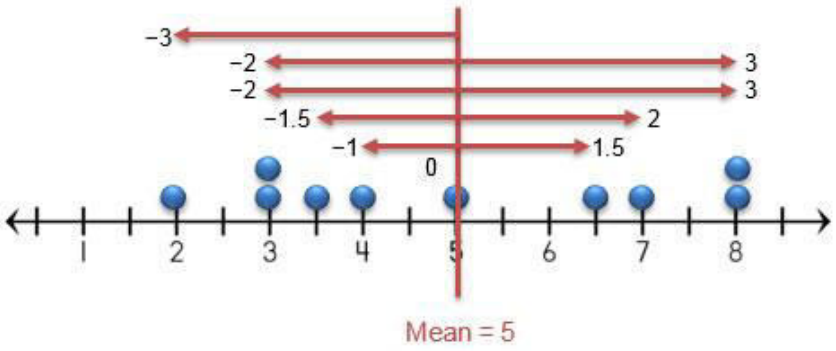
- **Range**
- **Quartile Deviation**
- **Mean Absolute Deviation**
- **Standard Deviation and Variance**

Absolute Measures of Dispersion - II

Mean Absolute Deviation

Arithmetic Mean of the absolute deviation measured either from the mean or median.

$$Mean\ Absolute\ Deviation = \frac{\sum_{i=1}^{n}|x_i - \bar{x}|}{n}$$



https://www.texasgateway.org/resource/mean-absolute-deviation-0

Absolute Measures of Dispersion - II

Calculating Mean Absolute Deviation

Step 1: Calculate mean of all the values

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

Step 2: Find the absolute distance (difference) of each value from that mean by subtracting the mean from each value in the data, ignoring minus signs.

$$absolute\ deviations = |x_i - \bar{x}|$$

Step 3: Find the mean of those distances

$$Mean\ Absolute\ Deviation = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

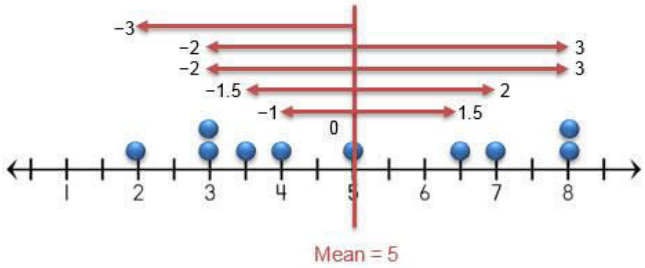
Absolute Measures of Dispersion - II

Mean Absolute Deviation

Example: In a clinical study 10 breast cancer patients were followed up for the period of the time until all of them died. Their survival times (in months) are given:

2, 3, 3, 3.5, 4, 5, 6.5, 7, 8, 8

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{50}{10} = 5$$



$$\sum_{i=1}^{n} |x_i - \bar{x}| = |-3| + |-2| + |-2| + |-1.5| + |-1| + |0| + |1.5| + |2| + |3| + |3| = 19$$

$$Mean\ Absolute\ Deviation = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n} = \frac{19}{10} = 1.9$$

https://www.texasgateway.org/resource/mean-absolute-deviation-0

Absolute Measures of Dispersion - II

Pros and Cons of Mean Absolute Deviation

- **Pros**
  - Easy to calculate and easy to understand.
  - Based on all the observation in the data.
  - Shows the scatter of observations from its central value.
  - It facilitated comparison between different items in a series.
- **Cons**
  - It violates the algebraic principle by ignoring the signs of the values.
  - Affected by fluctuations in the sampling.
  - Depending on the type of central value being used it produces different results.

Absolute Measures of Dispersion - II

Absolute Measures of Dispersion are of different types.

- Range
- Quartile Deviation
- Mean Absolute Deviation
- Standard Deviation and Variance

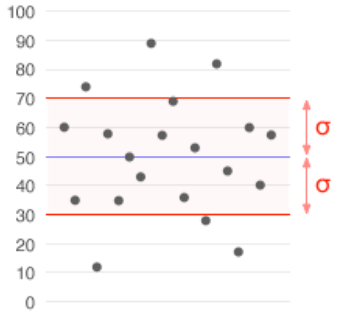
Absolute Measures of Dispersion - II

Variance

It is defined as the mean of the squared deviations of the observations from their mean.

If the variance is calculated for the population data then it is denoted by σ^2

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



If the variance is calculated for the sample data obtained from a population then it is denoted by:

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

Absolute Measures of Dispersion - II

Various formula to calculate sample variance

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} \quad \text{.....} \quad \textbf{(1)}$$

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \quad \text{.....} \quad \textbf{(2)}$$

$$s^2 = \frac{\sum_{i=1}^{n}x_i^2}{n} - \left(\frac{\sum_{i=1}^{n}x_i}{n}\right)^2 \quad \text{.....} \quad \textbf{(3)}$$

The formula in Eq. (1) a formula derived using Maximum likelihood estimation principles.

The formula in Eq. (2) is more suitable for interpreting the sample variance. Moreover variance expression in Eq. (2) results in an unbiased estimate of the variance.

The formula in Eq. (3) is considered to be the most suitable for the sake of calculations.

Absolute Measures of Dispersion - II

Method for calculating the sample variance

Step 1: Calculate the Arithmetic Mean

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

Step 2: Subtract the mean from each observation

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

Step 3: Square the difference

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

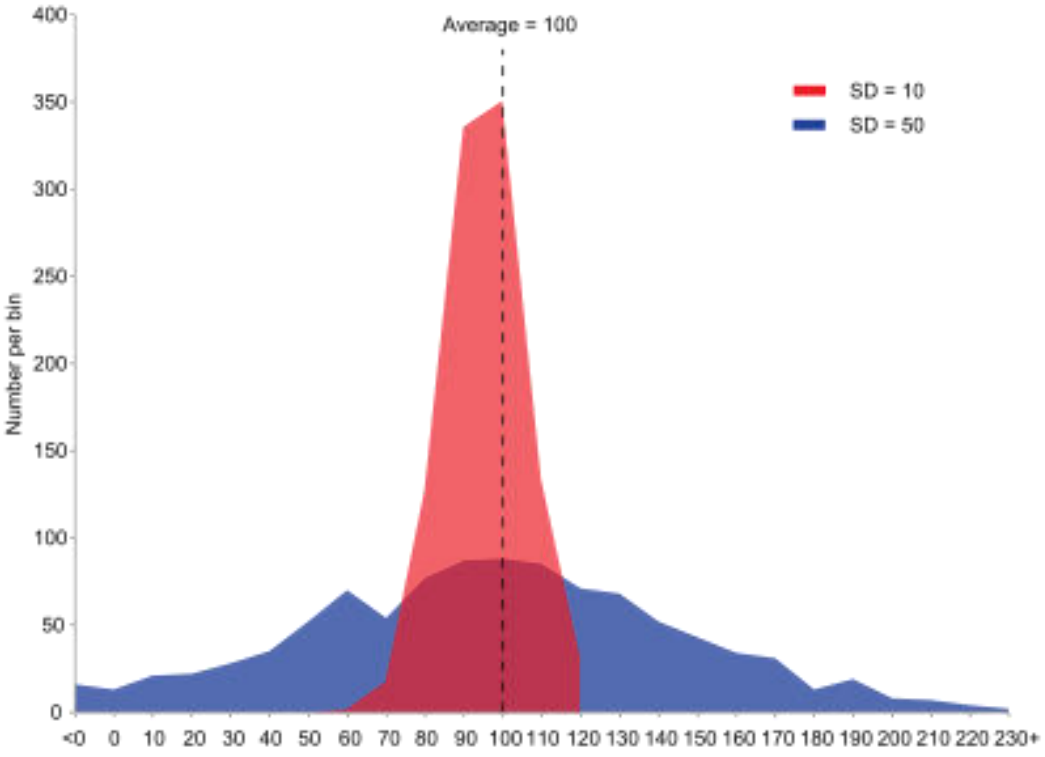
Step 4: Sum the Squared difference

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

Step 5: Divide the Sum of the squared difference by

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

Absolute Measures of Dispersion - II



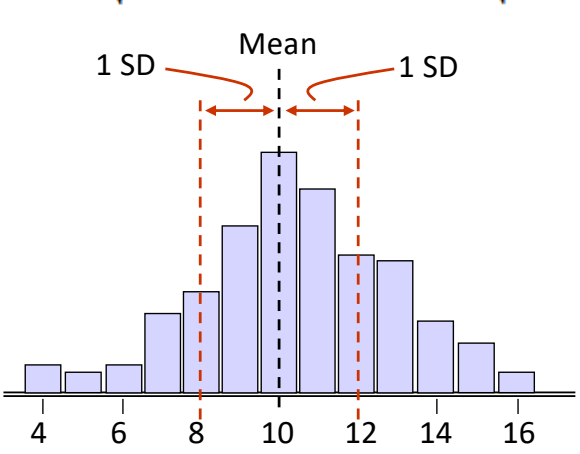
https://en.wikipedia.org/wiki/Standard_deviation

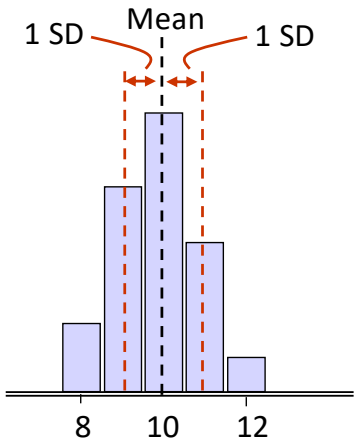
Absolute Measures of Dispersion - II

Standard Deviation

- The standard deviation (SD) is the square root of the variance.
  - small SD = values cluster closely around the mean
  - large SD = values are scattered

$$\sigma^2 = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}} \qquad S^2 = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} \qquad s^2 = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$





Absolute Measures of Dispersion - II

Method for calculating the sample Standard Deviation

Step 1: Calculate the Arithmetic Mean $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

Step 2: Subtract the mean from each observation $(x_i - \bar{x})$

Step 3: Square the difference $(x_i - \bar{x})^2$

Step 4: Sum the Squared difference $\sum_{i=1}^{n}(x_i - \bar{x})^2$

Step 5: Divide the Sum of the squared difference by $(n-1)$

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

Step 6: Take the square root of the value obtained in Step 5. The result is standard deviation.

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

Absolute Measures of Dispersion - II

Example

The incubation period of Hepatitis A is given:

27, 32, 25, 30, 22

Calculate amount of Spread?

Solution:

Step 1: Calculate the Arithmetic Mean $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

$$\bar{x} = \frac{27 + 31 + 15 + 30 + 22}{5} = \frac{125}{5} = 25.0\ days$$

https://www.cdc.gov/ophss/csels/dsepd/ss1978/lesson2/section7.html

Absolute Measures of Dispersion - II

Step 2 & 3: Subtract the mean from each observation & Square the difference

Value x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
27	27 – 25 = + 2	4
31	31 – 25 = + 6	36
15	15 – 25 = - 10	100
30	30 – 25 = + 5	25
22	22 – 25 = - 3	9

Step 4: Sum the Squared Differences.

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = 4 + 36 + 100 + 25 + 9 = 174$$

Absolute Measures of Dispersion - II

Step 5: Divide the sum of squared differences by $(n-1)$, such that $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$

$$Variance = s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{174}{4} = 43.5\ (days)^2$$

Step 6: Take the square root of the value obtained in Step 5 i.e. variance. The result is standard deviation.

$$Standard\ Deviation = S.D. = s = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}} = \sqrt{\frac{174}{4}} = \sqrt{43.5}\ (days) = 6.6\ (days)$$

Applied Biostatistics

Absolute Measures of Dispersion - III

Absolute Measures of Dispersion - III

Properties of Variance

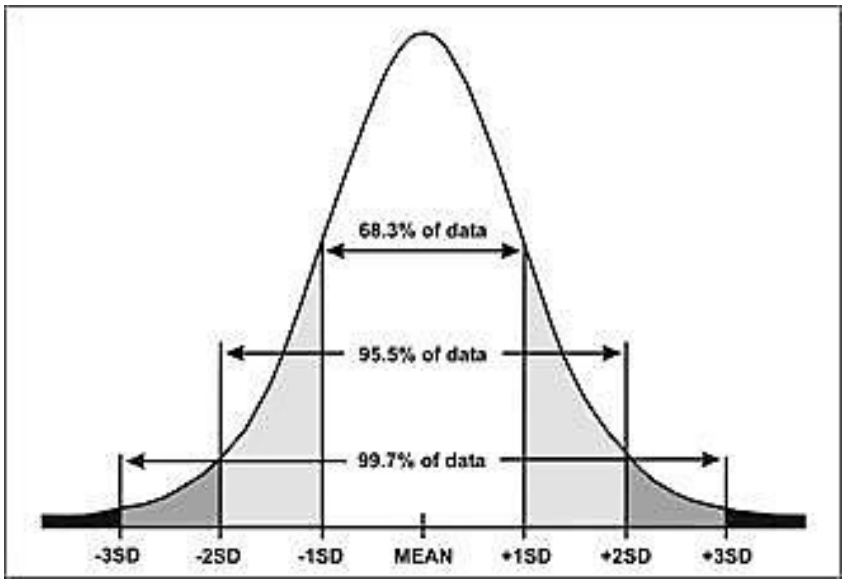
1. **The variance of a constant is equal to zero.**

2. **The variance is independent of Origin, i.e. it remains unchanged when a constant is added to or subtracted from the each and every observation of the data.**

3. **The variance is multiplied or divided by the square of the constant if each observation is multiplied or divided by the same constant.**

4. **The variance of the sum of difference of two independent variable is equal to the sum of their respective variances.**

Absolute Measures of Dispersion - III



https://www.cdc.gov/ophss/csels/dsepd/ss1978/lesson2/section7.html

Absolute Measures of Dispersion - III

Pros and Cons of Variance / S.D.

- **Pros**
  - Both measures summarizes the deviation of a large distribution from mean in one figure.
  - These measures indicates that if the variation of difference of individual observations from is mean is real or by chance.
  - These measures are preferred when the data is symmetric.
- **Cons**
  - The numeric value of the S.D. Does not have easy non-statistical interpretation.
  - Variance gives the answer in squared units.
  - The value of both the measures is affected by having extreme observations in our data.

Absolute Measures of Dispersion - III

END

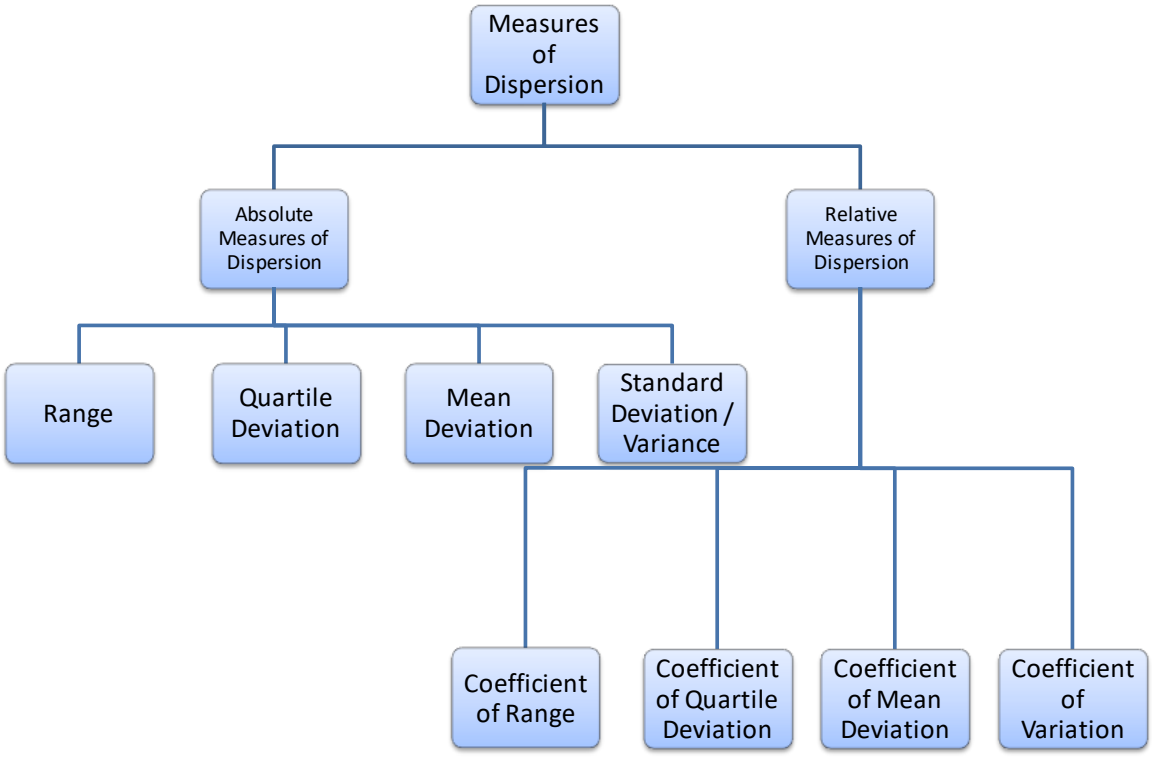
All the absolute measures of dispersion plays a very vital role in estimating other statistical measures.

Applied Biostatistics

Relative Measures Of Dispersion

Lecture no 50

Relative Measures of Dispersion



Relative Measures of Dispersion

Coefficient of Range

$$Coefficient\ of\ Range = \frac{x_m - x_o}{x_m + x_o} \times 100$$

GRF Measurements When Trotting of 20 Dogs with a Lame Ligament

14.6	24.3	24.9	27.0	27.2	27.4	28.2	28.8	29.9	30.7
31.5	31.6	32.3	32.8	33.3	33.6	34.3	36.9	38.3	44.0

$x_o = Minimum\ value = 14.6$

$x_m = Maximum\ value = 44.0$

$Range = x_m - x_o = 44.0 - 14.6 = 29.4\ Newtons$

$$Coefficient\ of\ Range = \frac{44.0 - 14.6}{44.0 + 14.6} \times 100 = \frac{29.4}{58.6} \times 100 = 50.17$$

Relative Measures of Dispersion

Coefficient of Quartile Deviation

$$Coefficient\ of\ Q.D. = \frac{Q_3 - Q_1}{Q_3 + Q_1} \times 100$$

$$Q_1 = \left(\frac{1(n+1)}{4}\right)^{th} Value$$

$$Q_3 = \left(\frac{3(n+1)}{4}\right)^{th} Value$$

Relative Measures of Dispersion

GRF Measurements When Trotting of 20 Dogs with a Lame Ligament

14.6	24.3	24.9	27.0	27.2	27.4	28.2	28.8	29.9	30.7
31.5	31.6	32.3	32.8	33.3	33.6	34.3	36.9	38.3	44.0

$$Q_3 = \left(\frac{3(20+1)}{4}\right)^{th} Value = (15.75)^{th}\ Value$$

$$Q_3 = 33.3 + 0.75\,(33.6 - 33.3) = 33.525$$

$$Q_1 = \left(\frac{1(20+1)}{4}\right)^{th} Value = (5.25)^{th}\ Value$$

$$Q_1 = 27.2 + 0.25\,(27.4 - 27.2) = 27.25$$

$$Q.D = \frac{IQR}{2} = \frac{6.275}{2} = 3.138$$

$$Coefficient\ of\ Q.D. = \frac{Q_3 - Q_1}{Q_3 + Q_1} \times 100 = \frac{33.525 - 27.25}{33.525 + 27.25} \times 100 = 10.32\%$$

Relative Measures of Dispersion

Coefficient of Mean Absolute Deviation from Mean

$$Coefficient\ of\ Mean\ Absolute\ Deviation = \frac{M.D_{\bar{x}}}{\bar{x}} \times 100 = \frac{\sum_{i=1}^{n}|x_i - \bar{x}|/n}{\bar{x}} \times 100$$

Example: In a clinical study 10 breast cancer patients were followed up for the period of the time until all of them died. Their survival times (in months) are given:

2, 3, 3, 3.5, 4, 5, 6.5, 7, 8, 8

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{50}{10} = 5$$

$$\sum_{i=1}^{n}|x_i - \bar{x}| = |-3| + |-2| + |-2| + |-1.5| + |-1| + |0| + |1.5| + |2| + |3| + |3| = 19$$

$$Mean\ Absolute\ Deviation = \frac{\sum_{i=1}^{n}|x_i - \bar{x}|}{n} = \frac{19}{10} = 1.9$$

$$Coefficient\ of\ Mean\ Absolute\ Deviation = \frac{\sum_{i=1}^{n}|x_i - \bar{x}|/n}{\bar{x}} \times 100 = \frac{1.9}{5} \times 100 = 38\%$$

Relative Measures of Dispersion

Coefficient of Variation

- **The Standard Deviation is useful as a measure of variation within a given set of data.**
- **However, comparing dispersion in two datasets on the basis of standard deviation may lead to fallacious results.**
- **Two means may be quite different.**
- **when one desires to compare the dispersion between two or more than two data sets then we prefer using coefficient of variation.**

Relative Measures of Dispersion

Coefficient of Variation (C.V)

$$C.V.(x) = \frac{s}{\bar{x}} \times 100$$

Suppose two samples of human males yield the following results:

	Sample 1	Sample 2
Age	25 years	11 years
Mean weight	145 pounds	80 pounds
Standard deviation	10 pounds	10 pounds

We wish to know which is more variable, the weights of the 25-year-olds or the weights of the 11-year-olds.

Relative Measures of Dispersion

Coefficient of Variation : Example

	Sample 1	Sample 2
Age	25 years	11 years
Mean weight	145 pounds	80 pounds
Standard deviation	10 pounds	10 pounds

A comparison of the standard deviations might lead one to conclude that the two samples possess equal variability. If we compute the coefficients of variation, however, we have for the 25-year-olds

$$\text{C.V.} = \frac{10}{145}(100) = 6.9\%$$

and for the 11-year-olds

$$\text{C.V.} = \frac{10}{80}(100) = 12.5\%$$

If we compare these results, we get quite a different impression. It is clear from this example that variation is much higher in the sample of 11-year-olds than in the sample of 25-year-olds.

Relative Measures of Dispersion

Properties of Coefficient of Variation

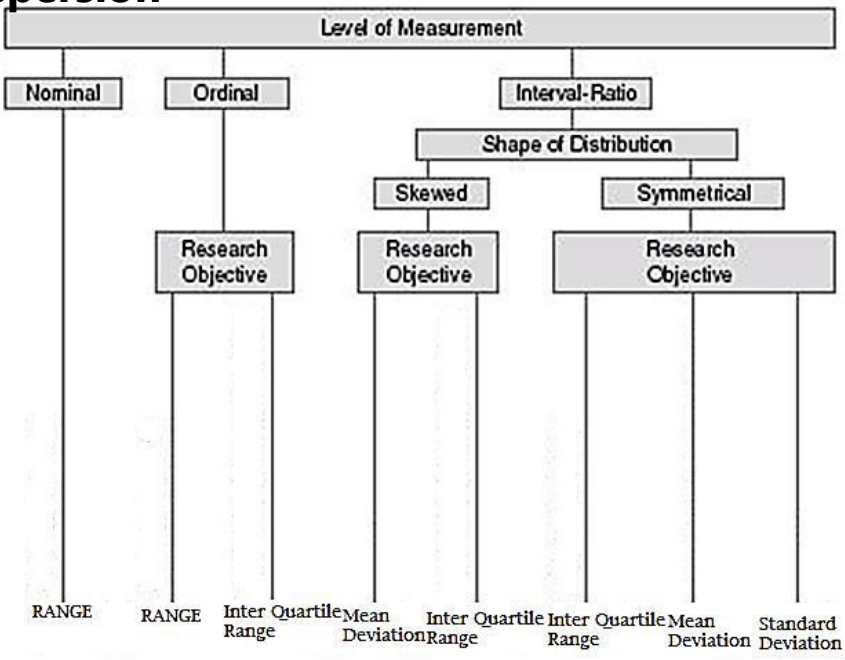
- **The Coefficient of Variation is a useful measure, when comparing the results obtained by different persons, who are conducting investigations involving same variable.**

- **Since the coefficient of Variation in independent of the scale of measurement, it is useful statistic for comparing the variability of two or more variables measured on different scales.**

for example, using the coefficient of variation to compare the variability in weights of one sample of subjects whose weights are expressed in pounds with the variability in weights of another sample of subjects whose weights are expressed in kilograms.

Relative Measures of Dispersion

Appropriate usage of the measures of Dispersion



Relative Measures of Dispersion

END

While reporting measures of Dispersion, we tend to report them along with appropriate measures of Central Tendency.

For symmetric data we prefer reporting Mean along with Standard Deviation

For Skewed data we prefer reporting Median with Inter quartile Range.

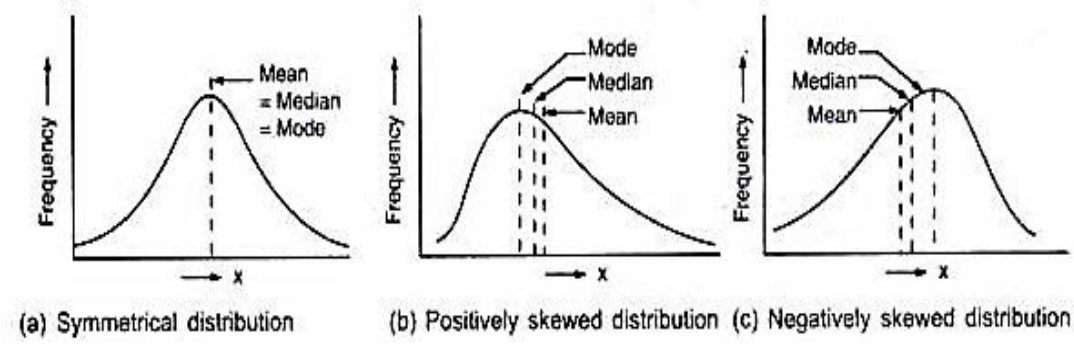
Applied Biostatistics

Skewness
And
Kurtosis

Relative Measures of Dispersion

Skewness

"It is defined as the departure from Symmetry"



(a) Symmetrical distribution (b) Positively skewed distribution (c) Negatively skewed distribution

Relative Measures of Dispersion

Measures of Skewness

Coefficient of Skewness help us to measure amount of the departure from symmetry.

It is denoted by Sk

Karl Pearson (1857 – 1936)

$$Sk = \frac{Mean - Mode}{Standard\ Deviation}$$

Relative Measures of Dispersion

Measures of Skewness

Since we know that mode is sometimes ill defined to located by simple methods.

It is replaced by its equivalent for moderately skewed distributions

$$Sk = \frac{3(Mean - Median)}{Standard\ Deviation}$$

Arthur Lyon Bowley (1869 – 1957)

$$Sk = \frac{Q_1 + Q_3 - 2Q_2}{Q_3 - Q_1}$$

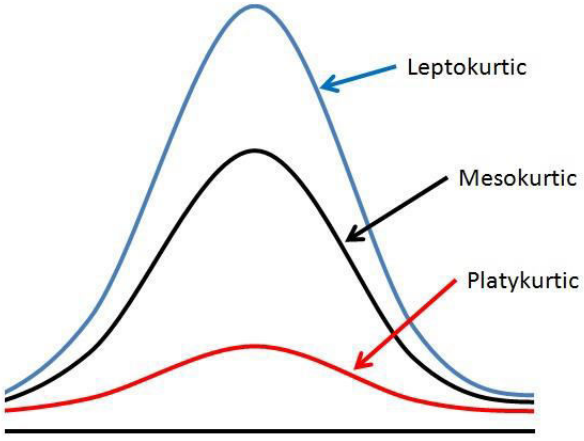
Relative Measures of Dispersion

Measures of Skewness

Relative Measures of Dispersion

Kurtosis

Peakedness of the data



https://www.bogleheads.org/wiki/File:Kurtosis1.jpg

Relative Measures of Dispersion

Measures of Kurtosis

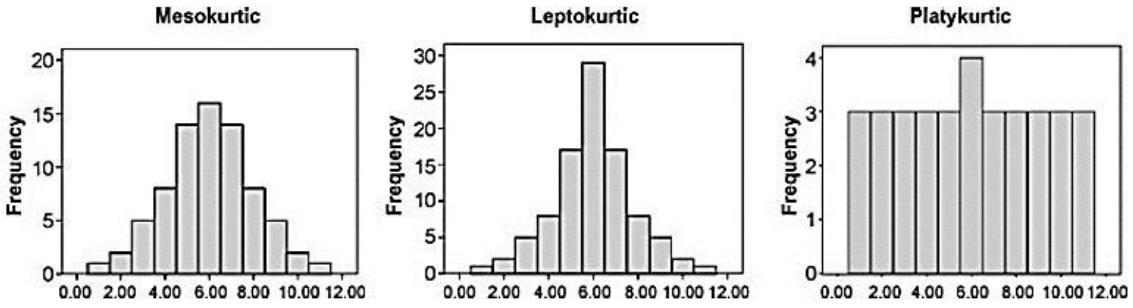
$$Kurtosis = \frac{n\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} - 3 = \frac{n\sum_{i=1}^{n}(x_i - \bar{x})^4}{(n-1)^2 s^4} - 3$$

$$k = \frac{Q.D.}{P_{90} - P_{10}}$$

Relative Measures of Dispersion

Example Kurtosis

	Mesokurtic	Leptokurtic	Platykurtic
Mean	6.0000	6.0000	6.0000
Median	6.0000	6.0000	6.0000
Mode	6.00	6.00	6.00
Skewness	.000	.608	−1.158



Relative Measures of Dispersion

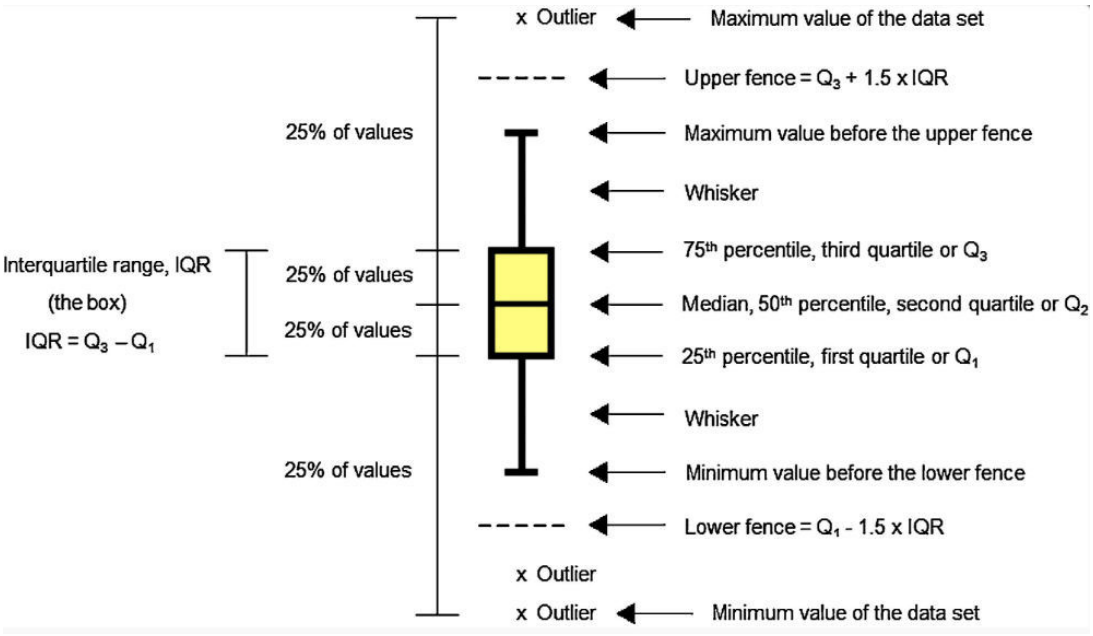
END

While conducting the data analysis we do not rely on only one measure of skewness, we do use

- Graphical
- Numerical Measures
  - We first observe them using mean, median and mode.
  - We also use various measures of Skewness.
- Manual calculations using formula for kurtosis is usually not necessary. packages provide this information. As a part of descriptive statistics.

Relative Measures of Dispersion

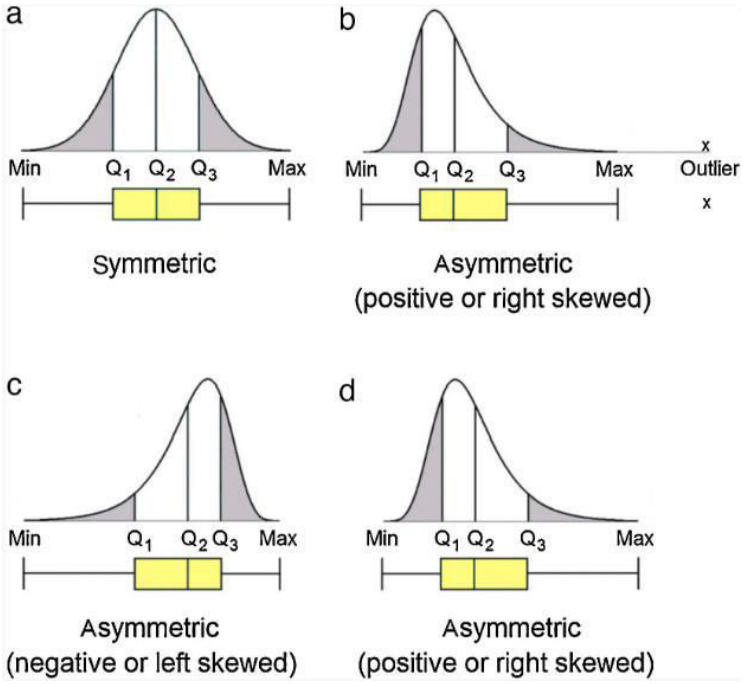
Box and Whisker Plot



http://www.elsevier.es/pt-revista-educacion-quimica-78-articulo-graphical-representation-chemical-periodicity-main-S0187893X16300106

Relative Measures of Dispersion

Shapes from Box and Whisker Plot



http://www.elsevier.es/pt-revista-educacion-quimica-78-articulo-graphical-representation-chemical-periodicity-main-S0187893X16300106

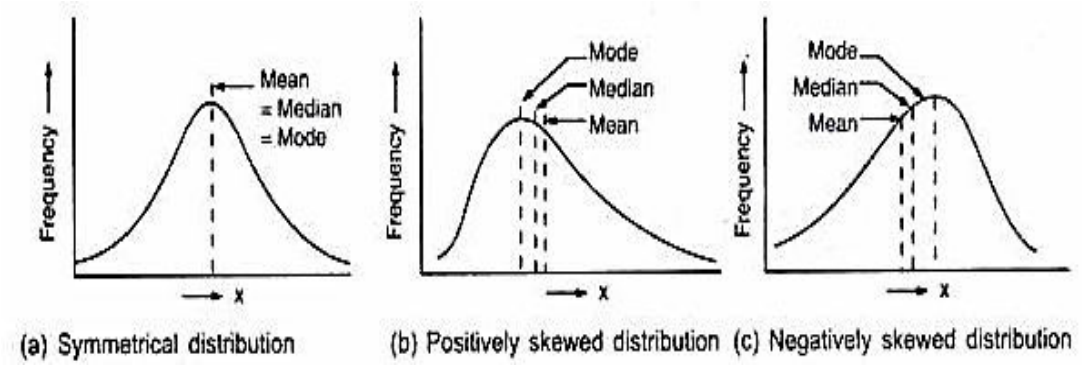
Applied Biostatistics

Skewness And Kurtosis

Skewness and Kurtosis

Skewness

"It is defined as the departure from Symmetry"



(a) Symmetrical distribution (b) Positively skewed distribution (c) Negatively skewed distribution

Skewness and Kurtosis

Measures of Skewness

Coefficient of Skewness help us to measure amount of the departure from symmetry.

It is denoted by Sk

Karl Pearson (1857 – 1936)

$$Sk = \frac{Mean - Mode}{Standard\ Deviation}$$

Skewness and Kurtosis

Measures of Skewness

Since we know that mode is sometimes ill defined to be located by simple methods.

It is replaced by its equivalent for moderately skewed distributions

$$Sk = \frac{3(Mean - Median)}{Standard\ Deviation}$$

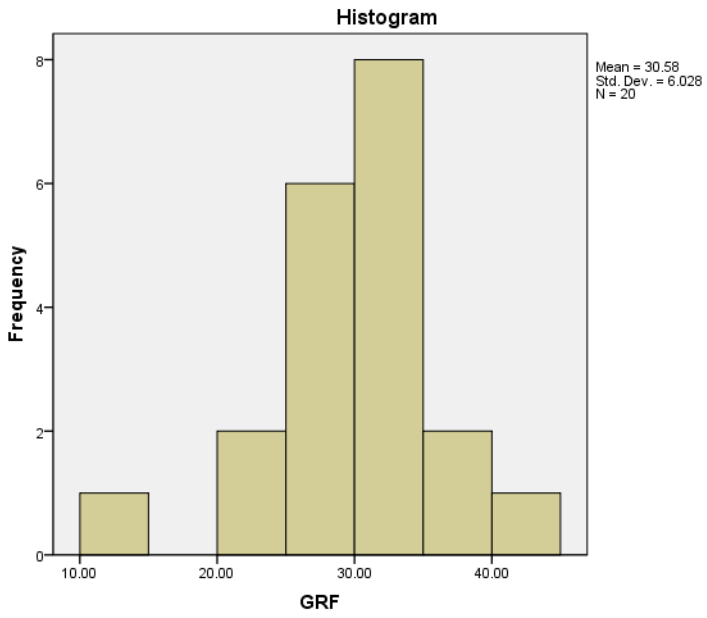
Arthur Lyon Bowley (1869 – 1957)

$$Sk = \frac{Q_1 + Q_3 - 2Q_2}{Q_3 - Q_1}$$

Skewness and Kurtosis

GRF Measurements When Trotting of 20 Dogs with a Lame Ligament

14.6	24.3	24.9	27.0	27.2	27.4	28.2	28.8	29.9	30.7
31.5	31.6	32.3	32.8	33.3	33.6	34.3	36.9	38.3	44.0



Mean = 30.58 N
Median = 31.10 N

Since Median > Mean

There is no most frequent value hence there is "No Mode".

Therefore we can not use the formula for skewness that contains Mode.

$$Sk = \frac{Mean - Mode}{Standard\ Deviation}$$

Skewness and Kurtosis

GRF Measurements When Trotting of 20 Dogs with a Lame Ligament

14.6	24.3	24.9	27.0	27.2	27.4	28.2	28.8	29.9	30.7
31.5	31.6	32.3	32.8	33.3	33.6	34.3	36.9	38.3	44.0

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{611.60}{20} = 30.58\ N$$

$$Median = \tilde{x} = \left(\frac{20+1}{2}\right)^{th} Value = (10.5)^{th}\ value$$

$$Median\ = (10.5)^{th}\ value = \ 10^{th} value + 0.50\left(11^{th} - 10^{th}\right)$$

$$Median = 30.7 + 0.50\,(31.5 - 30.7)\ = 31.10\ N$$

$$s = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2} = \sqrt{\frac{19393.22}{20} - \left(\frac{611.6}{20}\right)^2} = 6.03\ N$$

$$Sk = \frac{3(Mean - Median)}{Standard\ Deviation} = \frac{3\,(30.58 - 31.10)}{6.03} = -0.259$$

Skewness and Kurtosis

$$Q_1 = \left(\frac{1(20+1)}{4}\right)^{th} Value = (5.25)^{th}\ Value$$

$$Q_1 = (5.25)^{th}\ value = \ 5^{th} value + 0.25\left(6^{th} - 5^{th}\right)$$

$$Q_1 = 27.2 + 0.25\,(27.4 - 27.2) = 27.25$$

$$Q_2 = \left(\frac{2(20+1)}{4}\right)^{th} Value = (10.50)^{th}\ Value$$

$$Q_2\ = (10.5)^{th}\ value = \ 10^{th} value + 0.50\left(11^{th} - 10^{th}\right)$$

$$Q_2 = 30.7 + 0.50\,(31.5 - 30.7) = 31.10\ N$$

$$Q_3 = \left(\frac{3(20+1)}{4}\right)^{th} Value = (15.75)^{th}\ Value$$

$$Q_3 = (15.75)^{th}\ value = \ 15^{th} value + 0.75\left(16^{th} - 15^{th}\right)$$

$$Q_3 = 33.3 + 0.75\,(33.6 - 33.3) = 33.525$$

$$Sk = \frac{Q_1 + Q_3 - 2Q_2}{Q_3 - Q_1} = \frac{27.25 + 33.525 - 2\,(31.10)}{33.525 - 27.25} = -0.23$$

Skewness and Kurtosis

Measures of Skewness

Qualitative Variables:

There is not need to measure coefficient of skewness

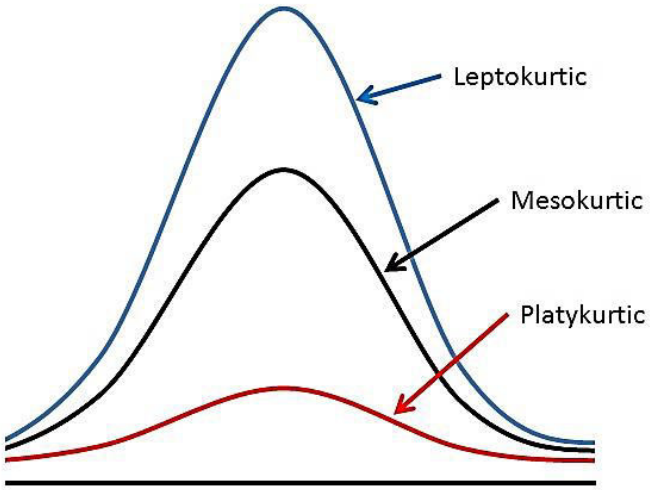
Quantitative Variables

- Frequency Histogram
- Frequency Curve
- Comparison of Mean, Median and Mode.
- Coefficients of Skewness

Skewness and Kurtosis

Kurtosis

Peakedness of the data



https://www.bogleheads.org/wiki/File:Kurtosis1.jpg

Skewness and Kurtosis

Measures of Kurtosis

$$Kurtosis = ku = \frac{n\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} - 3$$

$ku = 0 \quad for\ Mesokurtic$

$ku > 0 \quad for\ Leptokurtic$

$ku < 0 \quad for\ Platykurtic$

$$k = \frac{Q.D.}{P_{90} - P_{10}}$$

$0 < k < 0.5$

$k = 0.263\ for\ Mesokurtic$

$k > 0.263\ for\ leptokurtic$

$k < 0.263\ for\ platykurtic$

Skewness and Kurtosis

Measures of Kurtosis

$$Kurtosis = \frac{n\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} - 3 = \frac{n\sum_{i=1}^{n}(x_i - \bar{x})^4}{(n-1)^2 s^4} - 3$$

$$Kurtosis = ku = \frac{n\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} - 3 = \frac{20\ (106190.9)}{(690.492)^2} - 3 = \frac{2123817.938}{476779.2} - 3 = 4.45 - 3$$

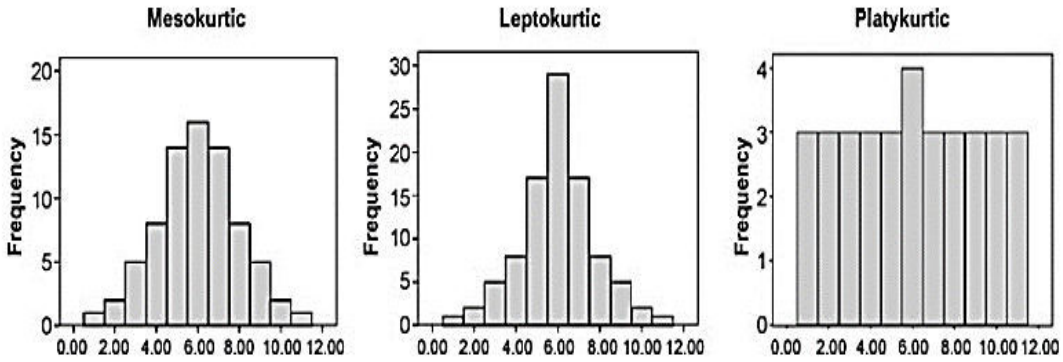
$ku = 1.45$

$$k = \frac{Q.D.}{P_{90} - P_{10}}$$

Skewness and Kurtosis

Example Kurtosis

	Mesokurtic	Leptokurtic	Platykurtic
Mean	6.0000	6.0000	6.0000
Median	6.0000	6.0000	6.0000
Mode	6.00	6.00	6.00
Skewness	.000	.608	−1.158



Skewness and Kurtosis

END

While conducting the data analysis we do not rely on only one measure of skewness, we do use

- **Graphical**
- **Numerical Measures**
  - **We first observe them using mean, median and mode.**
  - **We also use various measures of Skewness.**
- **Manual calculations using formula for kurtosis is usually not necessary. packages provide this information. As a part of descriptive statistics.**

Applied Biostatistics

Box
&
Whisker Plot

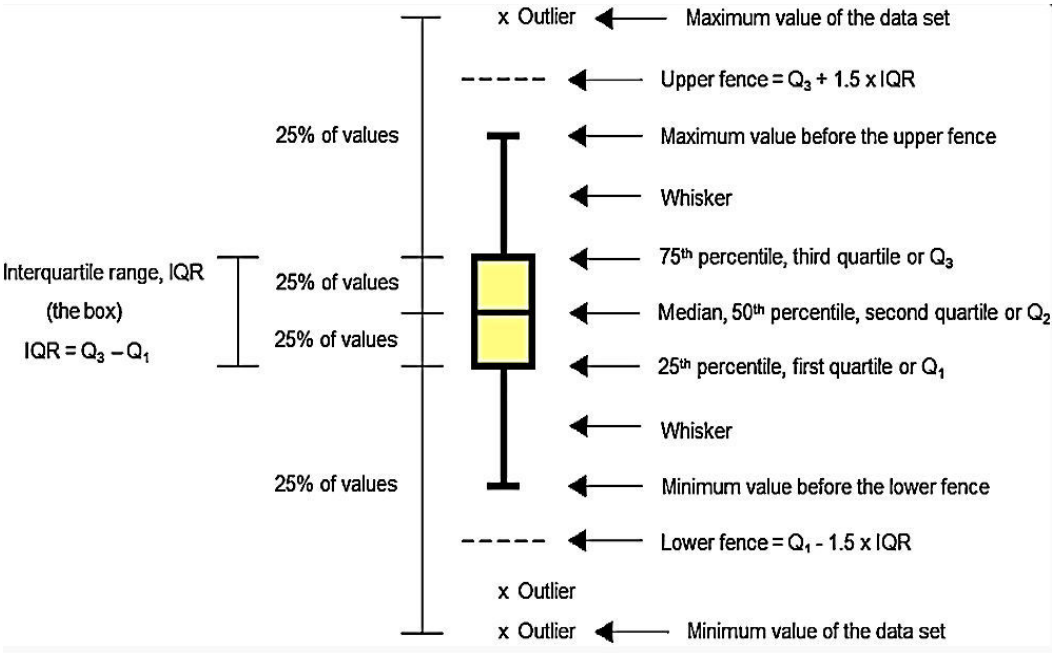
Box and Whisker Plot

Box and Whisker Plot

- **A useful visual device for communicating the information contained in a dataset.**
- **Sometimes simply called "Box Plot".**
- **Makes use of the quartiles.**

Box and Whisker Plot

Box and Whisker Plot



http://www.elsevier.es/pt-revista-educacion-quimica-78-articulo-graphical-representation-chemical-periodicity-main-S0187893X16300106

Box and Whisker Plot

Box and Whisker Plot

Step 1: Represent the variable of interest on the horizontal axis

Step 2: Draw a box in the space above the horizontal axis in such a way that the left end of the box aligns with the lower Quartile i.e. Q_1 and the right end of the box aligns with the upper quartile i.e. Q_3.

Step 3: Divide the box into two parts by a vertical line that aligns with the median i.e. Q_2.

Step 4: Draw a horizontal line called whisker from the left end of the box to a point that aligns with the smallest measurement in the data set.

Step 5: Draw another horizontal line or whisker from the right end of the box to a point that aligns with the largest measurement in the data set.

Box and Whisker Plot

GRF Measurements When Trotting of 20 Dogs with a Lame Ligament

14.6	24.3	24.9	27.0	27.2	27.4	28.2	28.8	29.9	30.7
31.5	31.6	32.3	32.8	33.3	33.6	34.3	36.9	38.3	44.0

$$Q_1 = \left(\frac{1(20+1)}{4}\right)^{th} Value = (5.25)^{th}\ Value = 27.25\ N$$

$$Minimum\ value = 14.6$$

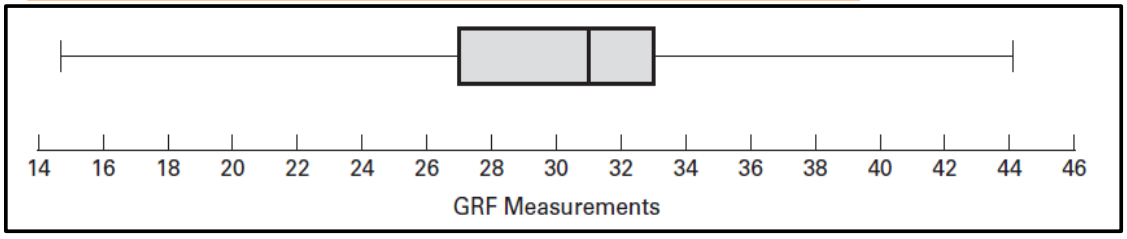
$$Median = \tilde{x} = \left(\frac{20+1}{2}\right)^{th} Value = (10.5)^{th}\ value = 31.10\ N$$

$$Maximum\ value = 44.0$$

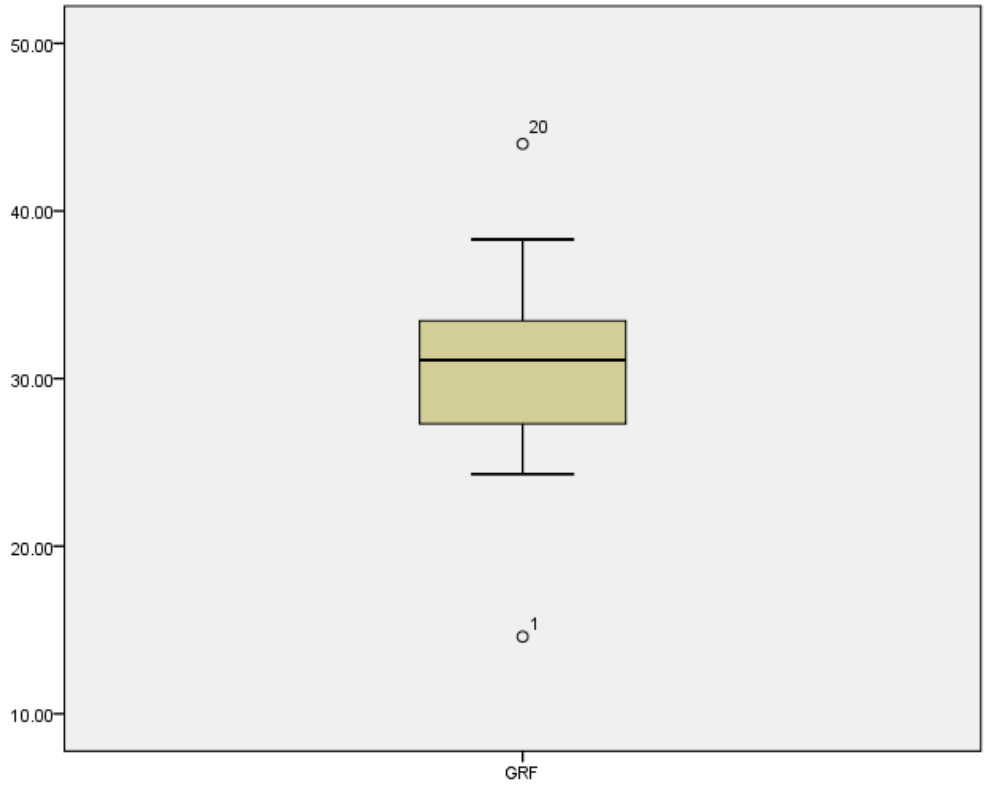
$$Q_3 = \left(\frac{3(20+1)}{4}\right)^{th} Value = (15.75)^{th}\ Value = 33.525\ N$$

$$Upper\ fence = Q_3 + 1.5 \times IQR = 33.525 + 1.5\,(33.525 - 27.25) = 42.938$$

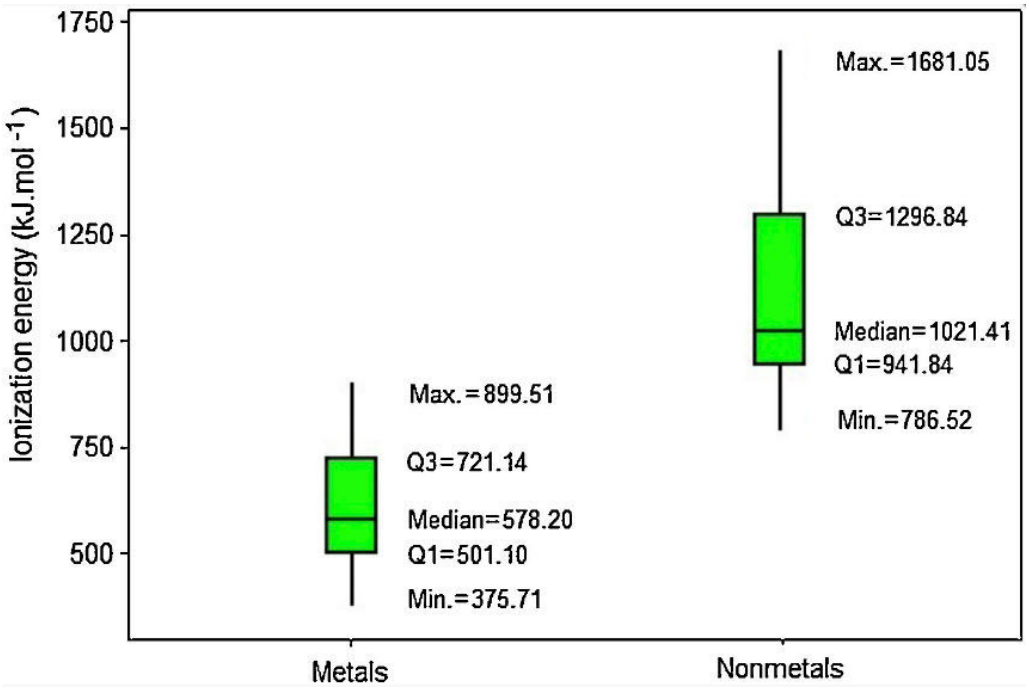
$$Lower\ fence = Q_1 - 1.5 \times IQR = 27.25 - 1.5\,(33.525 - 27.25) = 17.838$$



Box and Whisker Plot



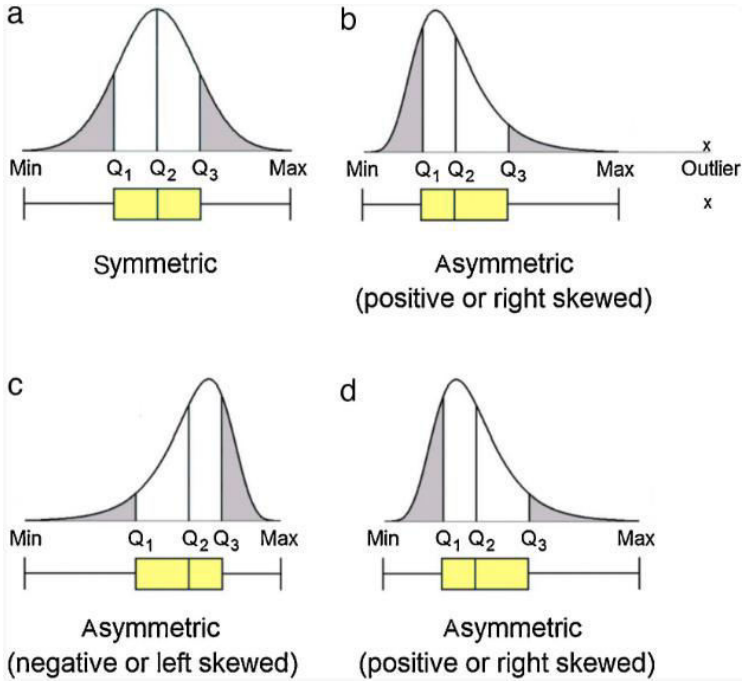
Box and Whisker Plot



http://www.elsevier.es/pt-revista-educacion-quimica-78-articulo-graphical-representation-chemical-periodicity-main-S0187893X16300106

Box and Whisker Plot

Shapes from Box and Whisker Plot



http://www.elsevier.es/pt-revista-educacion-quimica-78-articulo-graphical-representation-chemical-periodicity-main-S0187893X16300106

Box and Whisker Plot

END

Box and Whisker plot provides a very quick summary of our quantitative variable irrespective of it to be Symmetric or Skewed.

Box and Whisker plot is drawn for quantitative variables only.

If we want to see the distribution of a quantitative variable across different levels of a qualitative variable then Box plots provides a good and quick summary.

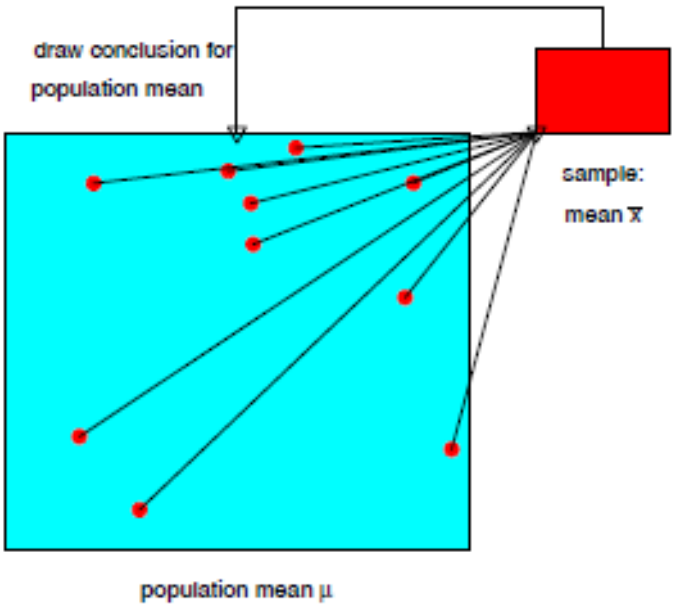
Applied Biostatistics

Introduction
To
Probability - I

Introduction to Probability - I

Link between sample and Population:

- **Generalize results from sample to the population.**
- **The population is a theoretical and usually undefined quantity.**
- **Needed for statistical tests and confidence intervals.**



Introduction to Probability - I

Have you at a certain time asked yourself the following questions?

How do you deal with these questions? Were you able to answer them with certainty?

https://www.slideshare.net/jmpalero/mathematics-8-basic-concepts-of-probability

Introduction to Probability - I

- **A patient has a 50 – 50 chance of surviving a certain surgery.**
- **A person has 95% certainty of curing after the treatment.**
- **A nurse may say that nine times out of ten person will miss a dose.**
- **Quantifying such terms like**
  - **Certain**
  - **probable**
  - **likely**
  - **Chance**

Introduction to Probability - I

Intuitive:

$$\text{Probability} = \text{relative frequency in the population}$$

Formal:

Random experiment

$\downarrow$

Events

$\downarrow$

Probabilities

Introduction to Probability - I

When we talk about probabilities we talk about the probability of the events that might occur in some random experiments.

An experiment is some activity with an observable outcome.

An experiment which produces different outcomes in multiple repetitions of the experiment, performed under similar conditions, is called a Random Experiment.

Introduction to Probability - I

Examples of a Random Experiments:

- Tossing of a fair coin
- Diagnosing a person for specific disease
- Measure the body height of an individual
- Rolling of a dice

Introduction to Probability - I

When the blood type of a patient is determined, the sample space my be represented by:

$$S = \{A, AB, B, O\}$$

Here in the above example:

Random Experiment: **Blood Test**

Sample Space is $S = \{A, AB, B, O\}$

Outcome = **Any possible value of a sample space. i.e. A , AB, B or O.**

Introduction to Probability - I

Intuitive:

Probability = relative frequency in the population

Formal:

Random experiment

↓

Events

↓

Probabilities

Introduction to Probability - I

Sample space Ω = set of all possible results of a random experiment

Examples:

Diagnosis $\longrightarrow \Omega = \{$ "sick", "healthy" $\}$
Roll the dice $\longrightarrow \Omega = \{1, 2, 3, 4, 5, 6\}$
Body height $\longrightarrow \Omega = \{x | x > 0\}$

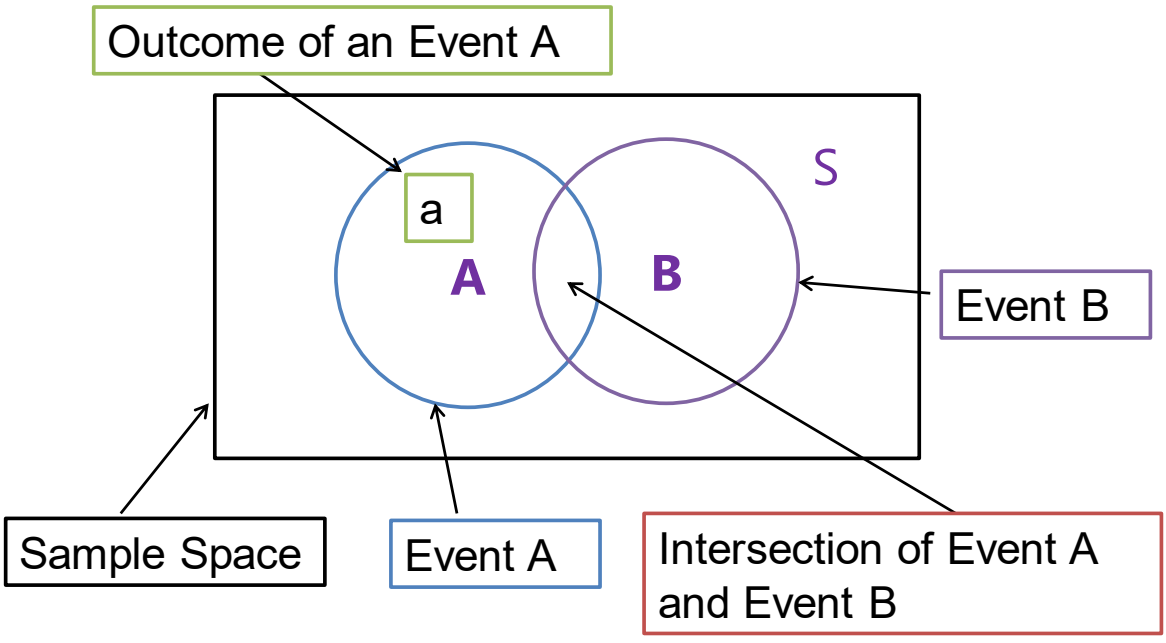
Event A = subset of Ω

Examples:

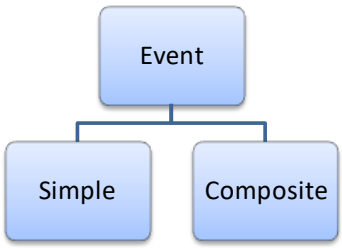
$A = \{2, 4, 6\}$ even number on the dice
$A = \{1\}$
$A = \{$Body height > 180 cm$\}$
$A = \{170$ cm $\leq$ Body height ≤ 180 cm$\}$
$A = \Omega =$ sure event
$A = \emptyset =$ impossible event

Introduction to Probability - I

VENN DIAGRAM

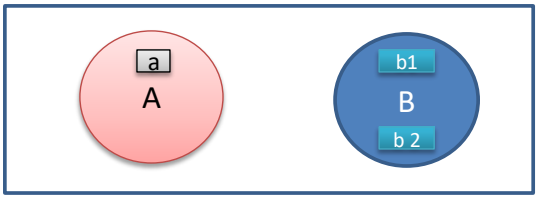


Introduction to Probability - I



If an Event consists of exactly one outcome, it is called Simple Event.

If an event consists of more than one outcomes, it is called compound event.



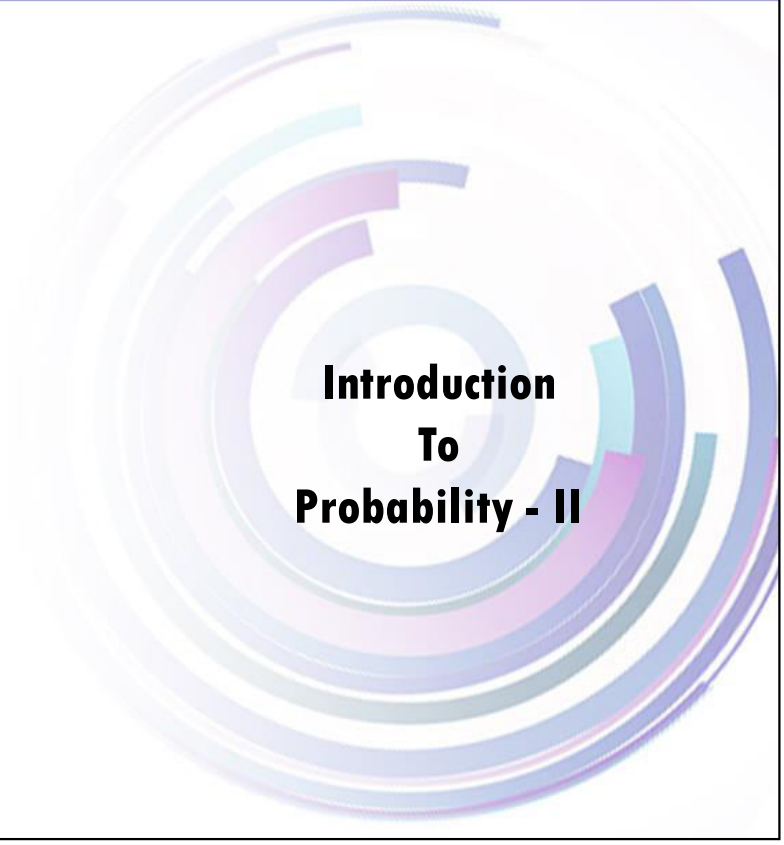
Introduction to Probability - I

END

There are various other types of events like

- **Mutually Exclusive Events**
- **Collectively Exhaustive Events**
- **Equally Likely Events**
- **Independent and Dependent Events**

Applied Biostatistics



Introduction to Probability - II

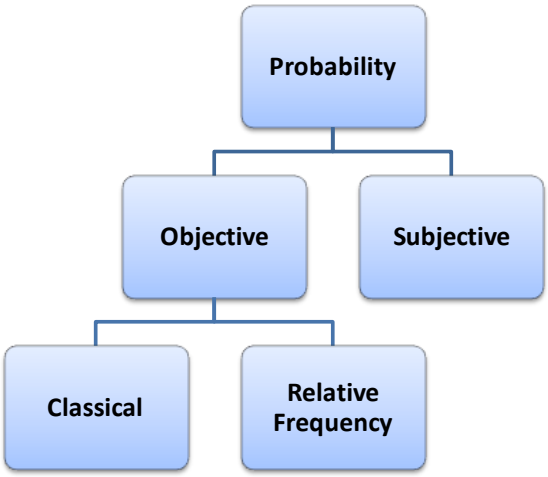
Intuitive:

Probability = relative frequency in the population

Formal:

Random experiment
↓
Events
↓
Probabilities

Introduction to Probability - II



In the early 1950s, L.J. Savage gave considerable impetus to what is called the "Personalistic" or subjective concept of Probability.

The Subjective definition of probability utilizes intuition, experience, and collective wisdom to assign a degree of belief than an event will occur.

Introduction to Probability - II

Subjective Probability

Example: The probability that a cure for cancer will be discovered with the next 10 years

Example: A medical doctor tells a patient with a newly diagnosed cancer that the probability of successfully treating the cancer is 90%.

- **The doctor is assigning a subjective probability of 0.90 to the event that the cancer can be successfully treated.**
- **Such a probability can not be determined by objective definitions of probability.**

Introduction to Probability - II

Classical Probability (*Priori*)

- This concept dates back to 17th century and the work of two mathematicians Pascal and Fermat.
- Much of this theory was developed out of the attempts to solve problems related to the games of chance, such as those involving the rolling of dice.

DEFINITION

If an event can occur in N mutually exclusive and equally likely ways, and if m of these possess a trait E, the probability of the occurrence of E is equal to m/N.

If we read $P(E)$ as "the probability of E," we may express this definition as

$$P(E) = \frac{m}{N}$$

Introduction to Probability - II

Classical Probability (*Priori*)

- For an experiment consisting of n outcomes, The Classical definition of the probability assigns probability 1/n to each outcome or simple event.
- For an event A consisting of k outcomes, the probability of event A is given as

$$P(A) = \frac{k}{n}$$

Introduction to Probability - II

Classical Probability (*Priori*)

- **The following table gives the information concerning 50 organ transplants in the state of Nebraska during a recent year. Each patient represented below had only one transplant.**

	Waiting Time for Transplant	
Type of transplant	Less than one year	One year or more
Heart	10	5
Kidney	7	3
Liver	5	5
Pancreas	3	2
Eyes	5	5

- **If one of the 50 patient records is randomly selected, the probability that a patient had a heart transplant is 15/50 = 0.30**

Introduction to Probability - II

Relative Frequency (*Posteriori*)

- **The Relative Frequency approach to the probability depends upon the repeatability of some process and the ability to count the number of repetitions, as well as the number of times, some event of interest occurs.**

DEFINITION

If some process is repeated a large number of times, n, and if some resulting event with the characteristic E occurs m times, the relative frequency of occurrence of E, m/n, will be approximately equal to the probability of E.

To express this definition in compact form, we write

$$P(E) = \frac{m}{n}$$

We must keep in mind, however, that, strictly speaking, m/n is only an estimate of $P(E)$.

Introduction to Probability - II

Relative Frequency (*Posteriori*)

- **A study by the Lahore Traffic Police found that when 750 drivers were randomly stopped 471 were found to be wearing seta belts.**
- **The relative frequency probability that driver wears a seat belt in Lahore while driving is**

$$P\,(seat\ belt) = \lim_{n\to\infty} \frac{m}{n} = \frac{471}{750} = 0.63$$

Introduction to Probability - II

END

Classical Definition of the probability can not be applied if the assumption of equally likely does not hold.

This definition become vague when the number of possible outcomes become infinite.

Applied Biostatistics

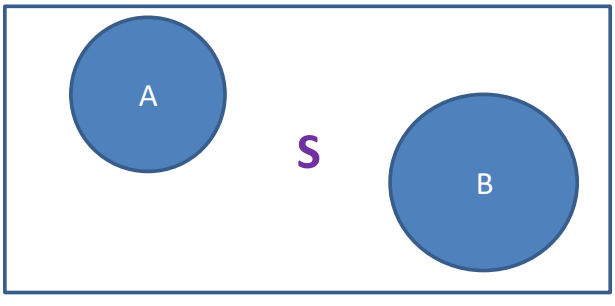
Probability: Basic Terminology

Probability: Basic Terminology

Mutually Exclusive Events

Two or more events are said to be mutually exclusive if the events do not have any outcomes in common. They are events that can not occur together.

If A and B are Mutually Exclusive events then the joint probability of A and B equals zero, that is, P(A and B) = 0



Probability: Basic Terminology

Mutually Exclusive Events

A random experiment consists in observing the gender of two randomly selected individuals.

The Event , A, that both individuals are male
The Event , B, that both individuals are female.

Here events A and B are mutually exclusive, because if both are male, then both cannot be female

i.e. P(A and B) = 0

Probability: Basic Terminology

Equally Likely Events

Two or more events are said to be equally Likely if one event is as likely to occur as the other.

When tossing a fair coin:

The chances for the occurrence of head are the same as the chances for the occurrence of tail.

$$P(H) = P(T) = \frac{1}{2}$$

Probability: Basic Terminology

Collectively Exhaustive Events

Two or more events are said to be Collectively Exhaustive events when union of the mutually exclusive events makes the entire sample space.

When tossing a fair coin: $S = \{H, T\}$

Both the events for the occurrence of Head or tail may not occur together, hence they are mutually exclusive, and the union of both makes the entire sample space.

$$P(H) = P(T) = \frac{1}{2}$$

Probability: Basic Terminology

Dependent and Independent Events

Dependent Events: Two events A and B are dependent events when the occurrence of event A has an influence on the occurrence of an event B.

$$P(A \text{ and } B) = P(A) \times P\left({}^{B}/_{A}\right)$$

The event of being a diabetic and having a family history of diabetes are dependent events.

Probability: Basic Terminology

Dependent and Independent Events

Independent Events: **Two events A and B are said to be independent events when the occurrence of an event A has no effect on the occurrence of an event B.**

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

The events of having 10 letters in your last name and being a biological sciences major are independent events.

Many times its not obvious whether two events are independent or not. In such cases we use following formula.

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

Probability: Basic Terminology

Complementary Events

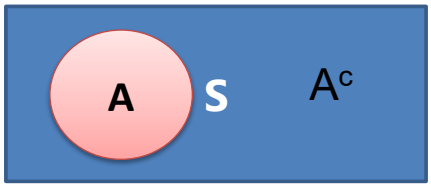
To every event A, there corresponds another event A^c, called the complement of A that consists of all other outcomes in the sample space not in event A.

The word NOT is used to describe the complement.

Since the event and its complement must account for all the outcomes of an experiment, their probabilities must add up to one.

$$P(A) + P\ (A^c) = 1$$

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



Probability: Basic Terminology

Complementary Events

Example: **Approximately 2% of the Pakistani population is diabetic.**

The probability that a randomly chosen Pakistani is non-diabetic is 0.98.

$$let\ A\ \Rightarrow an\ event\ that\ a\ Pakistani\ is\ Diabetic$$

$$P(A) = 0.02$$

$$P(A^c) = 1 - P(A) = 1 - 0.02 = 0.98$$

Probability: Basic Terminology

END

Complementary events are always mutually exclusive events but mutually exclusive events are not always complementary event.

Applied Biostatistics

Probability: Axioms Of Probability

Axioms of Probability

https://en.wikipedia.org/wiki/Andrey_Kolmogorov

Elementary Properties of Probability

In 1933 the axiomatic approach to probability was formalized by the Russian mathematician A.N. Kolmogorov.

The basis of this approach is embodied in three properties from which a whole system of probability theory is constructed through the use of mathematical logic.

Axioms of Probability

Elementary Properties of Probability

There are three axiomatic properties of Probability

- **Axiom of Non – Negativity**
- **Axiom of Exhaustiveness**
- **Axiom of Additive - ness**

Axioms of Probability

Axiom of Non - Negativity

The axiom of non – negativity states that, Given some process (or experiment) with *n* mutually exclusive outcomes (called events), , the probability of any event is assigned a non – negative number i.e.

$$P(E_i) \geq 0$$

Therefore, we can say that all the events must have probability greater than or equal to zero.

A Key concept in the statement of this property is the concept of mutually exclusive outcomes.

Axioms of Probability

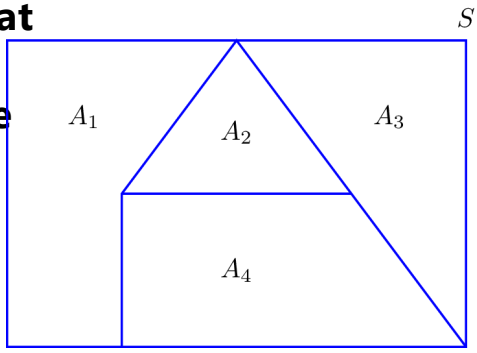
Axiom of Exhaustiveness

The axiom of exhaustiveness states that, the sum of the mutually exclusive events is equal to 1.

$$P(E_1) + P(E_2) + \cdots + P(E_n) = 1$$

This property refers to the fact that Observer of the probabilistic Process must allow for all possible Event.

Again this property requires the Events to by mutually exclusive.



https://www.probabilitycourse.com/chapter1/1_2_2_set_operations.php

Axioms of Probability

Axiom of Additive - ness

The axiom of Additiveness states that, for any two mutually exclusive events . The probability of the occurrence of either is equal to the sum of their individual probabilities.

$$P(E_i \cup E_j) = P(E_i) + P(E_j)$$

Suppose the two events are not mutually exclusive; that is, suppose they could occur at the same time. In attempting to compute the probability for either would be very complicated.

Axioms of Probability

END

Three Axioms of the probability help us understand that probability of any event can never be negative and it can never be more than 1,

Hence we can say that

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

Applied Biostatistics

Calculating Probability: Simple And Conditional

Lecture no 55

Calculating Probability - Simple

To investigate the effect of age at onset of bipolar disorder on the course of the illness. One of the variables investigated was family history of mood disorders. Following table shows the frequency of a family history of mood disorders in the two groups of interest.

Group 1: Early Age onset (i.e. 18 years or younger)
Group 2: Later Age onset (i.e. later than 18 years)

Suppose we pick a person at random from this sample.
What is the probability that this person will be 18 years old or younger?

Frequency of Family History of Mood Disorder by Age Group Among Bipolar Subjects

Family History of Mood Disorders	Early = 18 (*E*)	Later > 18 (*L*)	Total
Negative (*A*)	28	35	63
Bipolar disorder (*B*)	19	38	57
Unipolar (*C*)	41	44	85
Unipolar and bipolar (*D*)	53	60	113
Total	141	177	318

Source: Tasha D. Carter, Emanuela Mundo, Sagar V. Parkh, and James L. Kennedy, "Early Age at Onset as a Risk Factor for Poor Outcome of Bipolar Disorder," *Journal of Psychiatric Research, 37* (2003), 297–303.

Calculating Probability - Simple

Frequency of Family History of Mood Disorder by Age Group Among Bipolar Subjects

Family History of Mood Disorders	Early = 18 (*E*)	Later > 18 (*L*)	Total
Negative (*A*)	28	35	63
Bipolar disorder (*B*)	19	38	57
Unipolar (*C*)	41	44	85
Unipolar and bipolar (*D*)	53	60	113
Total	141	177	318

Source: Tasha D. Carter, Emanuela Mundo, Sagar V. Parkh, and James L. Kennedy, "Early Age at Onset as a Risk Factor for Poor Outcome of Bipolar Disorder," *Journal of Psychiatric Research, 37* (2003), 297–303.

What is the probability that this person will be 18 years old or younger?

let E ⇒ an event of interest that the age of a person selected is 18 years or younger

$$P(E) = \text{number of Early subjects /total number of subjects}$$
$$= 141/318 = .4434$$

Calculating Probability - Conditional

Joint Probability Sometimes we want to find the probability that a subject picked at random from a group of subjects possesses two characteristics at the same time. Such a probability is referred to as a *joint probability.* We illustrate the calculation of a joint probability with the following example.

What is the probability that a person picked at random from 318 subjects will be early (E) and will be a person who has no family history of mood disorder (A) ?

The probability we are seeking may be written in symbolic notation as $P(E \cap A)$ in which the symbol $\cap$ is read either as "intersection" or "and." The statement $E \cap A$ indicates the joint occurrence of conditions E and A.

Calculating Probability - Conditional

Frequency of Family History of Mood Disorder by Age Group Among Bipolar Subjects

Family History of Mood Disorders	Early = 18 (*E*)	Later > 18 (*L*)	Total
Negative (*A*)	28	35	63
Bipolar disorder (*B*)	19	38	57
Unipolar (*C*)	41	44	85
Unipolar and bipolar (*D*)	53	60	113
Total	141	177	318

Source: Tasha D. Carter, Emanuela Mundo, Sagar V. Parkh, and James L. Kennedy, "Early Age at Onset as a Risk Factor for Poor Outcome of Bipolar Disorder," *Journal of Psychiatric Research, 37* (2003), 297–303.

What is the probability that a person picked at random from 318 subjects will be early (E) and will be a person who has no family history of mood

The number of subjects satisfying both of the desired conditions is found in Table at the intersection of the column labeled E and the row labeled A and is seen to be 28. Since the selection will be made from the total set of subjects, the denominator is 318. Thus, we may write the joint probability as

$$P(E \cap A) = \frac{n(E \cap A)}{n(S)} = \frac{28}{318} = 0.0881$$

Calculating Probability - Conditional

Marginal Probability refers to the probability in which the numerator of the probability is a marginal total from a table and the denominator is the grand total from the table.

Frequency of Family History of Mood Disorder by Age Group Among Bipolar Subjects

Family History of Mood Disorders	Early = 18 (*E*)	Later > 18 (*L*)	Total
Negative (*A*)	28	35	63
Bipolar disorder (*B*)	19	38	57
Unipolar (*C*)	41	44	85
Unipolar and bipolar (*D*)	53	60	113
Total	141	177	318

Source: Tasha D. Carter, Emanuela Mundo, Sagar V. Parkh, and James L. Kennedy, "Early Age at Onset as a Risk Factor for Poor Outcome of Bipolar Disorder," *Journal of Psychiatric Research, 37* (2003), 297–303.

Calculating Probability - Conditional

Frequency of Family History of Mood Disorder by Age Group Among Bipolar Subjects

Family History of Mood Disorders	Early = 18 (*E*)	Later > 18 (*L*)	Total
Negative (*A*)	28	35	63
Bipolar disorder (*B*)	19	38	57
Unipolar (*C*)	41	44	85
Unipolar and bipolar (*D*)	53	60	113
Total	141	177	318

Source: Tasha D. Carter, Emanuela Mundo, Sagar V. Parkh, and James L. Kennedy, "Early Age at Onset as a Risk Factor for Poor Outcome of Bipolar Disorder," *Journal of Psychiatric Research, 37* (2003), 297–303.

When we compute the probability that a person picked at random from the 318 person is an early age of onset subject?

let E⇒ an event of interest that the age of a person selected is 18 years or younger

$$P(E) = \frac{n(E)}{n(S)} = \frac{141}{318} = 0.4434$$

Calculating Probability - Conditional

DEFINITION

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

$$P(A_i) = \sum P(A_i \cap B_j), \qquad \text{for all values of } j$$

Frequency of Family History of Mood Disorder by Age Group Among Bipolar Subjects

Family History of Mood Disorders	Early = 18 (E)	Later > 18 (L)	Total
Negative (A)	28	35	63
Bipolar disorder (B)	19	38	57
Unipolar (C)	41	44	85
Unipolar and bipolar (D)	53	60	113
Total	141	177	318

Source: Tasha D. Carter, Emanuela Mundo, Sagar V. Parkh, and James L. Kennedy, "Early Age at Onset as a Risk Factor for Poor Outcome of Bipolar Disorder," *Journal of Psychiatric Research, 37* (2003), 297–303.

Calculating Probability - Conditional

Frequency of Family History of Mood Disorder by Age Group Among Bipolar Subjects

Family History of Mood Disorders	Early = 18 (E)	Later > 18 (L)	Total
Negative (A)	28	35	63
Bipolar disorder (B)	19	38	57
Unipolar (C)	41	44	85
Unipolar and bipolar (D)	53	60	113
Total	141	177	318

Source: Tasha D. Carter, Emanuela Mundo, Sagar V. Parkh, and James L. Kennedy, "Early Age at Onset as a Risk Factor for Poor Outcome of Bipolar Disorder," *Journal of Psychiatric Research, 37* (2003), 297–303.

$$P(E \cap A) = 28/318 = .0881$$
$$P(E \cap B) = 19/318 = .0597$$
$$P(E \cap C) = 41/318 = .1289$$
$$P(E \cap D) = 53/318 = .1667$$

Calculating Probability - Conditional

$$P(E \cap A) = 28/318 = .0881$$
$$P(E \cap B) = 19/318 = .0597$$
$$P(E \cap C) = 41/318 = .1289$$
$$P(E \cap D) = 53/318 = .1667$$

We obtain the marginal probability $P(E)$ by adding these four joint probabilities as follows:

$$P(E) = P(E \cap A) + P(E \cap B) + P(E \cap C) + P(E \cap D)$$
$$= .0881 + .0597 + .1289 + .1667$$
$$P(E) = .4434$$

Calculating Probability - Conditional

DEFINITION

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

That is,

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \qquad P(B) \neq 0$$

Calculating Probability - Conditional

Frequency of Family History of Mood Disorder by Age Group Among Bipolar Subjects

Family History of Mood Disorders	Early = 18 (E)	Later > 18 (L)	Total
Negative (A)	28	35	63
Bipolar disorder (B)	19	38	57
Unipolar (C)	41	44	85
Unipolar and bipolar (D)	53	60	113
Total	141	177	318

Source: Tasha D. Carter, Emanuela Mundo, Sagar V. Parkh, and James L. Kennedy, "Early Age at Onset as a Risk Factor for Poor Outcome of Bipolar Disorder," *Journal of Psychiatric Research, 37* (2003), 297–303.

What is the probability that a subject has no family history of mood disorders (A), given that the selected subject is Early (E)?

This is a conditional probability and is written as $P(A \mid E)$ in which the vertical line is read "given."

$$P(A|E) = \frac{P(A \cap E)}{P(E)} = \frac{0.0881}{0.4434} = 0.1986$$

Calculating Probability - Conditional

Frequency of Family History of Mood Disorder by Age Group Among Bipolar Subjects

Family History of Mood Disorders	Early = 18 (E)	Later > 18 (L)	Total
Negative (A)	28	35	63
Bipolar disorder (B)	19	38	57
Unipolar (C)	41	44	85
Unipolar and bipolar (D)	53	60	113
Total	141	177	318

Source: Tasha D. Carter, Emanuela Mundo, Sagar V. Parkh, and James L. Kennedy, "Early Age at Onset as a Risk Factor for Poor Outcome of Bipolar Disorder," *Journal of Psychiatric Research, 37* (2003), 297–303.

$$P(A|E) = \frac{28}{141} = 0.1986$$

Calculating Probability - Conditional

END

The set of "all possible outcomes" may constitute a subset of the total group.

The size of the group of interest may be reduced by conditions not applicable to the total group.

When probabilities are calculated with the subset of the total group as the denominator, the result is a conditional probability

Applied Biostatistics

Probability: Multiplicative Rule

Probability: Multiplicative Rule

Independent Events: Two events A and B are said to be independent events when the occurrence of an event A has no effect on the occurrence of an event B.

Dependent Events: Two events A and B are dependent events when the occurrence of event A has an influence on the occurrence of an event B.

Probability: Multiplicative Rule

The Multiplicative Rule is a way to find the probability of two events occurring at the same time.

Lets say there are two events A and E. The multiplicative rule is a way to calculate the probability that events A and E occurs at the same time.

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

General Rule: (For dependent Events)

$$P(A \text{ and } E) = P(A \cap E) = P(A) \times P(E|A)$$
$$P(A \text{ and } E) = P(A \cap E) = P(E) \times P(A|E)$$

Specific Rule: (For Independent Events)

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

Probability: Multiplicative Rule

Frequency of Family History of Mood Disorder by Age Group Among Bipolar Subjects

Family History of Mood Disorders	Early = 18 (E)	Later > 18 (L)	Total
Negative (A)	28	35	63
Bipolar disorder (B)	19	38	57
Unipolar (C)	41	44	85
Unipolar and bipolar (D)	53	60	113
Total	141	177	318

Source: Tasha D. Carter, Emanuela Mundo, Sagar V. Parkh, and James L. Kennedy, "Early Age at Onset as a Risk Factor for Poor Outcome of Bipolar Disorder," *Journal of Psychiatric Research, 37* (2003), 297–303.

$$P(A|E) = \frac{28}{141} = 0.1986$$

$$P(E) = \frac{n(E)}{n(S)} = \frac{141}{318} = 0.4434$$

We wish to compute the joint probability of Early age at onset (E) and a negative family history of mood disorders (A) from knowledge of an appropriate marginal probability and an appropriate conditional probability.

Probability: Multiplicative Rule

We wish to compute the joint probability of Early age at onset (E) and a negative family history of mood disorders (A) from knowledge of an appropriate marginal probability and an appropriate conditional probability.

$$P(A|E) = \frac{28}{141} = 0.1986$$

$$P(E) = \frac{n(E)}{n(S)} = \frac{141}{318} = 0.4434$$

$$P(A \text{ and } E) = P(A \cap E) = P(E) \times P(A|E)$$

$$P(A \cap E) = P(E) \times P(A|E) = 0.4434 \times 0.1986 = 0.0881$$

Probability: Multiplicative Rule

Specific Multiplicative Rule of Probability states that " for any two independent events A and E, the probability for the joint occurrence of both A and E is the product of their individual probabilities".

In general one can think of the Multiplicative Rule as "and" rule.

If both event A and event E must happen in order for a certain outcome to occur, and if A and E are independent of each other then one can use Specific multiplicative rule to calculate the probability of the outcome.

Probability: Multiplicative Rule

Consider a cross between two heterozygous (Aa) individuals. What is the probability for an (aa) individual in the next generation.

The only way to get an (aa) off spring is that mother contributes (a) gamete and a father contributes (a) gamete as well.

Each parent has 1/2 chance of making an (a) gamete. Thus the chance of (aa) offspring is:

$$Let\ A \Rightarrow an\ event\ that\ mother\ contributes\ "a"\ gamete$$

$$Let\ B \Rightarrow an\ event\ that\ father\ contributes\ "a"\ gamete$$

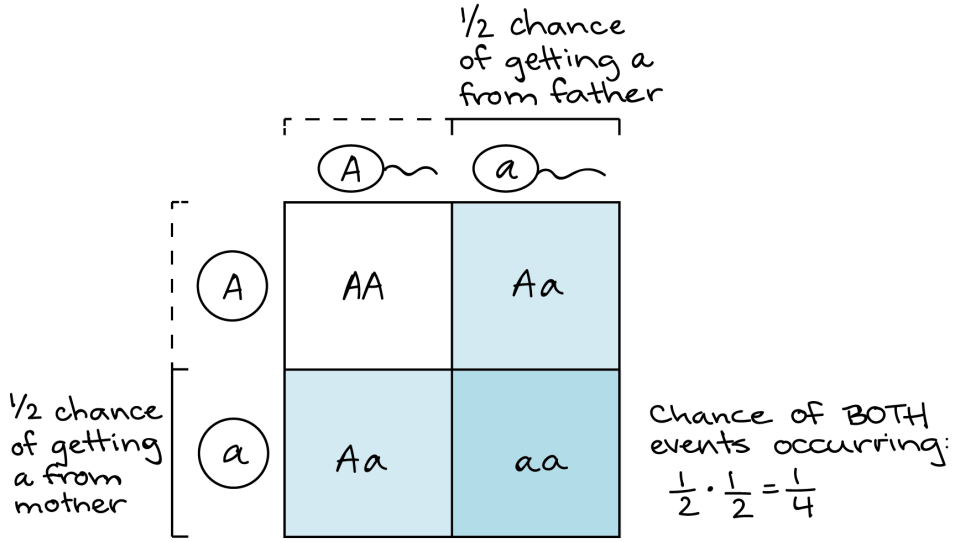
$$P(mother\ contributes\ a\ and\ father\ contributes\ a) = P(A \cap B)$$

$$P(A \cap B) = P(A) \times P(B) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

https://www.khanacademy.org/science/biology/classical-

Probability: Multiplicative Rule

Lets look at the Punnett Square.



https://www.khanacademy.org/science/biology/classical-genetics/mendelian--genetics/a/probabilities-in-genetics

Probability: Multiplicative Rule

END

We see through the algebraic manipulation of the multiplicative rule stated earlier, that it may be used to find any one of the three probabilities in its statement if the other two are known .

Applied Biostatistics

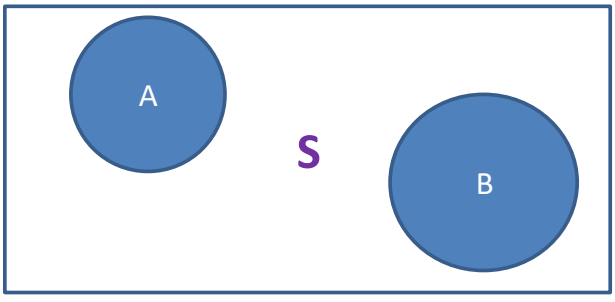
Probability: Additive Rule

Probability: Additive Rule

Mutually Exclusive Events

Two or more events are said to be mutually exclusive if the events do not have any outcomes in common. They are events that can not occur together.

If A and B are Mutually Exclusive events then the joint probability of A and B equals zero, that is, P(A and B) = 0

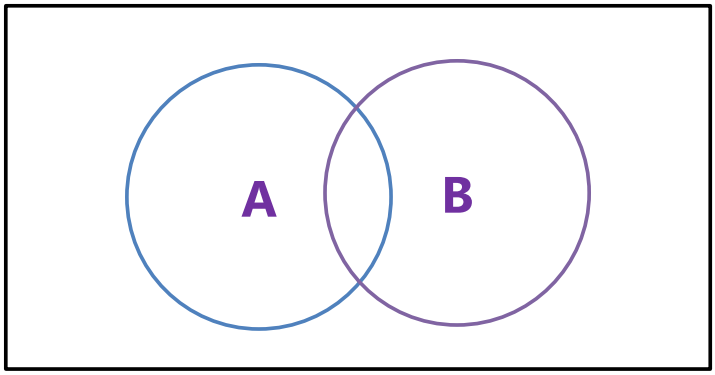


Probability: Additive Rule

Not - Mutually Exclusive Events

Two or more events are said to be not mutually exclusive if the events have some outcomes in common. They are events that can occur together.

If A and B are Mutually Exclusive events then the joint probability of A and B exists, that is, P(A and B) ≠ 0



Probability: Additive Rule

Additive Rule for Mutually Exclusive Events

Frequency of Family History of Mood Disorder by Age Group Among Bipolar Subjects

Family History of Mood Disorders	Early = 18 (*E*)	Later > 18 (*L*)	Total
Negative (*A*)	28	35	63
Bipolar disorder (*B*)	19	38	57
Unipolar (*C*)	41	44	85
Unipolar and bipolar (*D*)	53	60	113
Total	141	177	318

Source: Tasha D. Carter, Emanuela Mundo, Sagar V. Parkh, and James L. Kennedy, "Early Age at Onset as a Risk Factor for Poor Outcome of Bipolar Disorder," *Journal of Psychiatric Research, 37* (2003), 297–303.

What is the probability that a person selected will be an Early age at onset (E) or Later age at onset (L).

$$P(Early\ Age\ at\ onset\ OR\ Later\ age\ at\ onset) = P(E \cup L) = P(E) + P(L)$$

$$P(E \cup L) = \frac{141}{318} + \frac{177}{318} = 1$$

Probability: Additive Rule

Additive Rule for Mutually Exclusive Events

For an example lets use the Additive rule to predict the fraction of off spring from an Aa x Aa cross that will have the dominant phenotype. (AA or Aa genotype).

In this cross there are three events that can lead to a dominant phenotype.

- **Two A gametes meet (giving AA genotype) OR**
- **A gamete from Mom meets with a gamete from Dad (giving Aa genotype) OR**
- **a gamete from Mom meets with A gamete from Dad (giving Aa genotype).**

Probability: Additive Rule

Additive Rule for Mutually Exclusive Events

In this cross there are three events that can lead to a dominant phenotype.

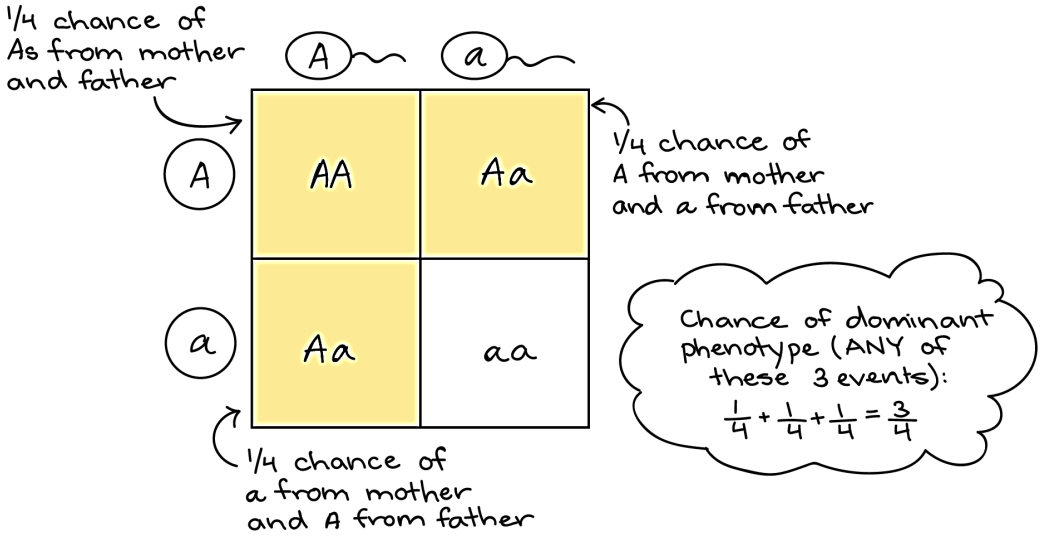
- **Two A gametes meet (giving AA genotype) OR**
- **A gamete from Mom meets with a gamete from Dad (giving Aa genotype) OR**
- **a gamete from Mom meets with A gamete from Dad (giving Aa genotype).**

Since each individual event has probability of occurrence = 1/4

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) = \frac{1}{4} + \frac{1}{4} + \frac{1}{4} = \frac{3}{4}$$

Probability: Additive Rule

Additive Rule for Mutually Exclusive Events



https://www.khanacademy.org/science/biology/classical-genetics/mendelian--genetics/a/probabilities-in-genetics

Probability: Additive Rule

DEFINITION

Given two events A and B, the probability that event A, or event B, or both occur is equal to the probability that event A occurs, plus the probability that event B occurs, minus the probability that the events occur simultaneously.

The addition rule may be written

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

When events A and B cannot occur simultaneously, $P(A \cap B)$ is sometimes called "exclusive or," and $P(A \cap B) = 0$. When events A and B can occur simultaneously, $P(A \cap B)$ is sometimes called "inclusive or," and we use the addition rule to calculate $P(A \cup B)$. Let us illustrate the use of the addition rule by means of an example.

Probability: Additive Rule

Frequency of Family History of Mood Disorder by Age Group Among Bipolar Subjects

Family History of Mood Disorders	Early = 18 (E)	Later > 18 (L)	Total
Negative (A)	28	35	63
Bipolar disorder (B)	19	38	57
Unipolar (C)	41	44	85
Unipolar and bipolar (D)	53	60	113
Total	141	177	318

Source: Tasha D. Carter, Emanuela Mundo, Sagar V. Parkh, and James L. Kennedy, "Early Age at Onset as a Risk Factor for Poor Outcome of Bipolar Disorder," *Journal of Psychiatric Research, 37* (2003), 297–303.

If we select a person at random from the 318 subjects represented in Table what is the probability that this person will be an Early age of onset subject (E) or will have no family history of mood disorders (A) or both?

The probability we seek is $P(E \cup A)$.

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

$$P(E \cup A) = \frac{141}{318} + \frac{63}{318} - \frac{28}{318} = 0.5534$$

Probability: Additive Rule

END

There are various other rules which are used to calculate probability of an event.

- **Rule for Complementary events**
- **Bayes Rule**

Applied Biostatistics

Diagnostic Testing – I

Lecture no 57

Diagnostic Testing - I

- **Tests are used in Clinical Diagnosis and Screening.**
- **In the health sciences, a widely used application of probability laws and concepts is found in the evaluation of screening tests and diagnostic criteria.**
- **Clinicians look for enhanced ability to correctly predict the presence or absence of a particular disease from the knowledge of test results.**
- **Biostatistician look for the Information regarding the likelihood of positive and negative test results and likelihood of the presence or absence of a particular symptom in patients with and without a particular disease.**

Diagnostic Testing - I

- How well is a subject classified into disease or non disease category?
  - Ideally all subjects having the disease should be classified as "having the disease" and vice versa.
  - Practically, we must be aware of the fact that "Tests are not always infallible".
  - The ability to classify individuals into the correct disease status depends on the accuracy of the tests, among other things

Diagnostic Testing - I

- A diagnostic test is used to determine the presence or absence of a disease when a subject shows signs or symptoms of the disease.
- A screening test identifies asymptotic individuals who may have the disease.
- The diagnostic test is performed after a positive screening test to establish definitive diagnosis.

http://ocw.jhsph.edu/courses/FundEpi/PDFs/Lecture11.pdf

Diagnostic Testing - I

Some Common Screening Tests

- Fasting blood cholesterol for heart disease.
- Fasting blood sugar for diabetes.
- Blood Pressure for hypertension.
- Mammography for breast cancer.
- PSA test for prostate cancer.
- Fecal Occult blood test for colon cancer.

http://ocw .jhsph.edu/courses/FundEpi/PDFs/Lecture11.pdf

Diagnostic Testing - I

Variation in Biological Values

- Many test results have a continuous scale (are continuous variables)
  - Blood Glucose level (70 – 100 mg/dL)
  - Cholesterol level (less than 100 mg/dL)
  - Blood Pressure
    - Systolic Blood Pressure 120
    - Diastolic Blood Pressure 80
- Distribution of Biological measurements in humans may or may not permit easy separation of diseased from non-diseased individuals, based upon the value of the measurements.

http://ocw .jhsph.edu/courses/FundEpi/PDFs/Lecture11.pdf

Diagnostic Testing - I

In summary, the following questions must be answered in order to evaluate the usefulness of test results and symptom status in determining whether or not a subject has some disease:

1. Given that a subject has the disease, what is the probability of a positive test result (or the presence of a symptom)?
2. Given that a subject does not have the disease, what is the probability of a negative test result (or the absence of a symptom)?
3. Given a positive screening test (or the presence of a symptom), what is the probability that the subject has the disease?
4. Given a negative screening test result (or the absence of a symptom), what is the probability that the subject does not have the disease?

Diagnostic Testing - I

- **One must know the correct disease status prior to calculation.**
- **Gold Standard test is the best test available**
  - **It is often invasive or expensive**
- **A new test is, for example, as new screening test or a less expensive diagnostic test.**
- **Use a 2 x 2 table to compare the performance of the new test to the gold standard test.**

http://ocw .jhsph.edu/courses/FundEpi/PDFs/Lecture11.pdf

Diagnostic Testing - I

Test	Disease +	Disease −
+	a (True positives)	b (False positives)
−	c (False negatives)	d (True negatives)

http://ocw .jhsph.edu/courses/FundEpi/PDFs/Lecture11.pdf

Diagnostic Testing - I

A testing procedure may yield:

- **False Positive**
  - **When a test indicates a positive status (as having disease) when the true status is negative (not having disease).**
- **False Negative**
  - **When a test indicates a negative status (not having disease) when the true status is positive (having disease).**

Diagnostic Testing - I

Sensitivity of a test (or symptom) is the probability of a positive test result (or the presence of the symptom) given the presence of the disease.

Sensitivity is the ability of the test to identify correctly those who have the disease (a) from all individuals with the disease (a+c)

Sensitivity is a fixed characteristic of the test

New test	Disease +	Disease −
+	a (True positives)	b
−	c	d (True negatives)

$$Sensitivity = \Pr(^{T+}/_{D+})$$

$$Sensitivity = \frac{a}{a+c} = \frac{True\ Positive}{Disease+}$$

http://ocw .jhsph.edu/courses/FundEpi/PDFs/Lecture11.pdf

Diagnostic Testing - I

Specificity of a test (or symptom) is the probability of a negative test result (or the absence of the symptom) given the absence of the disease.

Specificity is the ability of the test to identify correctly those who do not have the disease (d) from all individuals free from the disease (b + d)

eristic of the test

New test	Disease +	Disease −
+	a (True positives)	b
−	c	d (True negatives)

$$Sepecificity = Pr\ (^{T-}/_{D-})$$

$$Specificity = \frac{d}{b+d} = \frac{True\ Negatives}{Disease-}$$

http://ocw .jhsph.edu/courses/FundEpi/PDFs/Lecture11.pdf

Diagnostic Testing - I

If a person tests positive, what is the probability that he or she has the disease?

And if the person tests negative, what is the probability that he or she does not have the disease?

http://ocw .jhsph.edu/courses/FundEpi/PDFs/Lecture11.pdf

Diagnostic Testing - I

The Positive Predictive Value (PPV) of a screening test (or symptom) is the probability that a subject has the disease given that the symptom has positive screening test result (or has the symptom)

Positive Predictive Value (PPV) is the proportion of patients who [illegible] the disease

$$PPV = Pr\left(D+ / T+\right)$$

Test \ Disease	+	−
+	a (True positives)	b (False positives)
−	c (False negatives)	d (True negatives)

$$PPV = \frac{a}{a+b} = \frac{True\ Positive}{Test+}$$

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

http://ocw .jhsph.edu/courses/FundEpi/PDFs/Lecture11.pdf

Diagnostic Testing - I

The Negative Predictive Value (NPV) of a screening test (or symptom) is the probability that a subject does not have the disease given that the symptom has negative screening test result (or does not have the symptom)

Negative Predictive Value (NPV) is the proportion of patients who test Negative who actually do not have the disease.

Test \ Disease	+	−
+	a (True positives)	b (False positives)
−	c (False negatives)	d (True negatives)

$$NPV = Pr(D^{-}/_{T}{}_{-})$$

$$NPV = \frac{d}{c+d} = \frac{True\ Negatives}{Test-}$$

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

http://ocw .jhsph.edu/courses/FundEpi/PDFs/Lecture11.pdf

Diagnostic Testing - I

Estimates of the Positive Predictive Value and Negative Predicative Value of a test (or symptom) may be obtained from knowledge of a test's (or symptom's) sensitivity and specificity and the probability of the relevant disease in the general population.

- For those who are interested

$$PPV = \frac{\text{sensitivity x prevalence}}{\text{(sensitivity x prevalence)} + \text{(1- specificity) x (1- prevalence)}}$$

$$NPV = \frac{\text{specificity x (1– prevalence)}}{\text{[(specificity x (1– prevalence)]} + \text{[(1- sensitivity) x prevalence]}}$$

http://ocw .jhsph.edu/courses/FundEpi/PDFs/Lecture11.pdf

Diagnostic Testing - I

END

Sensitivity, Specificity Positive Predictive Value and Negative Predictive Value

All these measures help to discuss the effectiveness of a newly introduced diagnostic test.

Having higher sensitivity doesn't necessarily means that specificity will also be higher and Vice Versa.

One need to find a good tradeoff among all these values to get the best test.

Applied Biostatistics

Diagnostic Testing - II

Diagnostic Testing - II

- Assume a population of 1,000 people
- 100 have a disease
- 900 do not have the disease
- A screening test is used to identify the 100 people with the disease
- The results of the screening appears in this table

Screening Results	True Characteristics in Population		Total
	Disease	No Disease	
Positive	80	100	180
Negative	20	800	820
Total	100	900	1,000

http://ocw .jhsph.edu/courses/FundEpi/PDFs/Lecture11.pdf

Diagnostic Testing - II

Screening Results	True Characteristics in Population		Total
	Disease	No Disease	
Positive	80	100	180
Negative	20	800	820
Total	100	900	1,000

Sensitivity = 80/100 = 80%

Specificity = 800/900 = 89%

http://ocw .jhsph.edu/courses/FundEpi/PDFs/Lecture11.pdf

Diagnostic Testing - II

Positive predictive value = 80/180 = 44%

Screening Results	True Characteristics in Population		Total
	Disease	No Disease	
Positive	80	100	180
Negative	20	800	820
Total	100	900	1,000

Negative predictive value = 800/820 = 98%

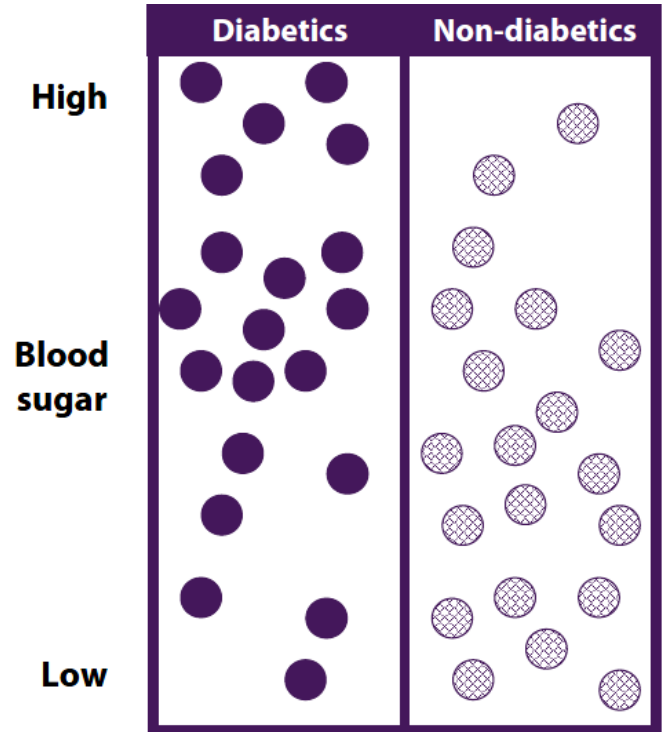
http://ocw .jhsph.edu/courses/FundEpi/PDFs/Lecture11.pdf

Diagnostic Testing - II

- Example: type II diabetes mellitus
  - Highly prevalent in the older, especially obese, U.S. population
  - Diagnosis requires oral glucose tolerance test
  - Subjects drink a glucose solution, and blood is drawn at intervals for measurement of glucose
  - Screening test is fasting plasma glucose
    - Easier, faster, more convenient, and less expensive

http://ocw .jhsph.edu/courses/FundEpi/PDFs/Lecture11.pdf

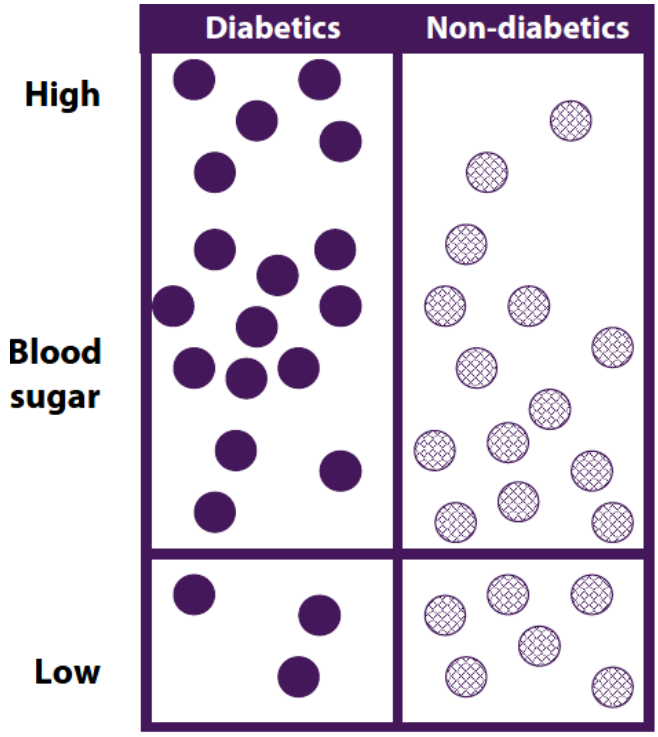
Diagnostic Testing - II



20 diabetics
20 non-diabetics

http://ocw .jhsph.edu/courses/FundEpi/PDFs/Lecture11.pdf

Diagnostic Testing - II

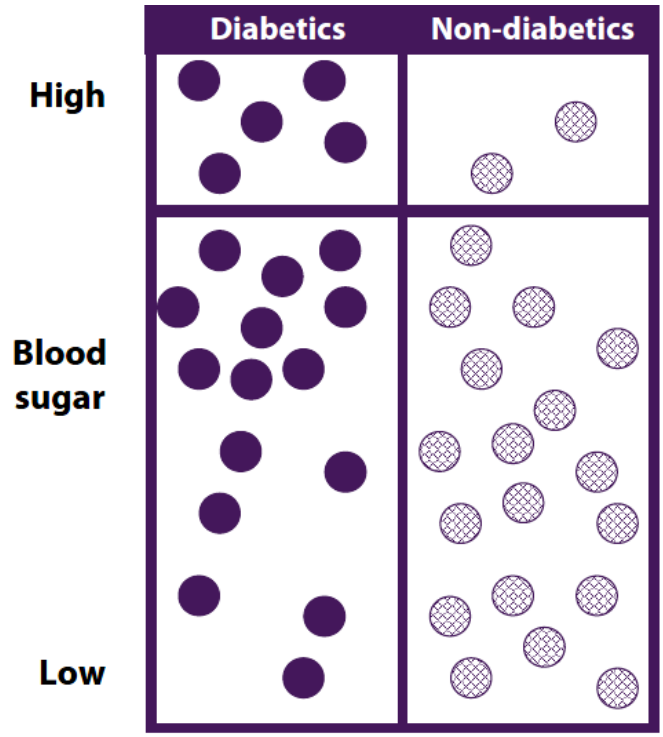


Subjects are screened using fasting plasma glucose with a low (blood sugar) cut-point

	Diabetics	Non-Diabetics
+	17	14
–	3	6
	20	20
	Sens=85%	Spec=30%

http://ocw .jhsph.edu/courses/FundEpi/PDFs/Lecture11.pdf

Diagnostic Testing - II

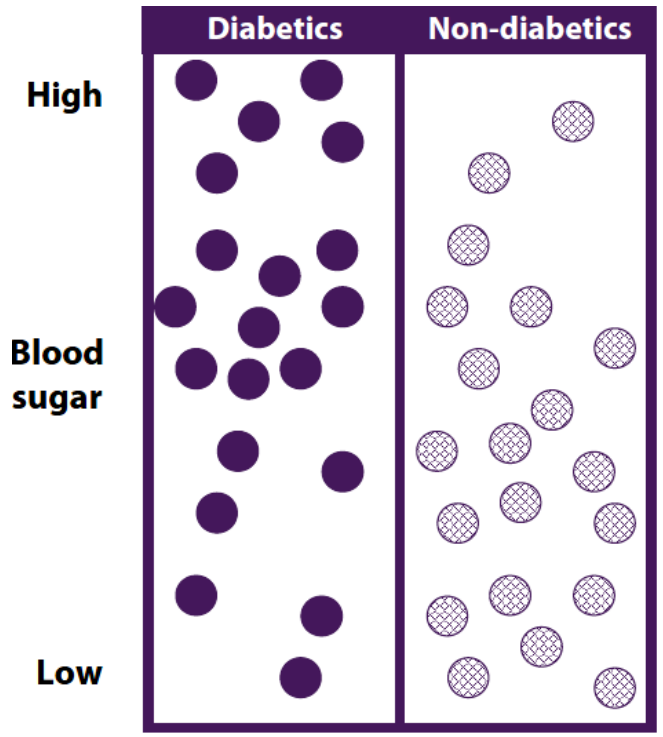


Subjects are screened using fasting plasma glucose with a high cut-point

	Diabetics	Non-Diabetics
+	5	2
–	15	18
	20	20
	Sens=25%	Spec=90%

http://ocw .jhsph.edu/courses/FundEpi/PDFs/Lecture11.pdf

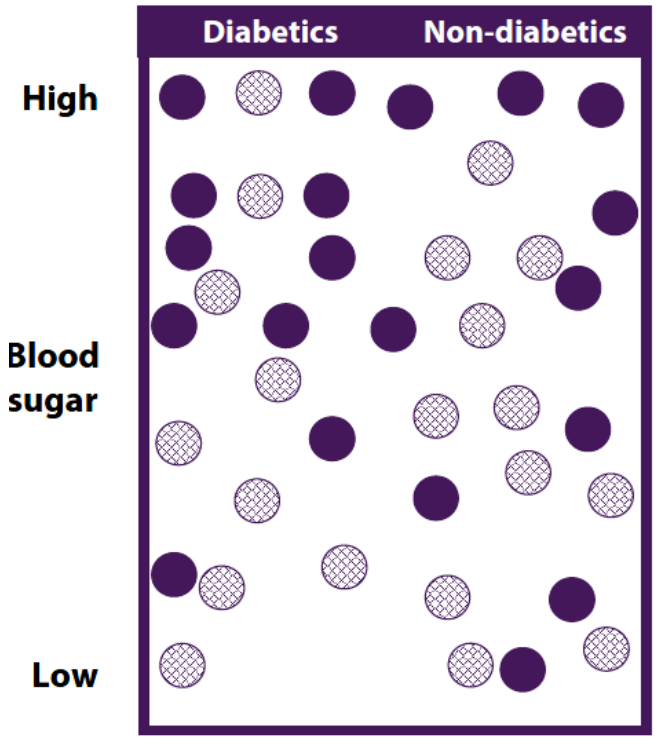
Diagnostic Testing - II



In a typical population, there is no line separating the two groups, and the subjects are mixed

http://ocw .jhsph.edu/courses/FundEpi/PDFs/Lecture11.pdf

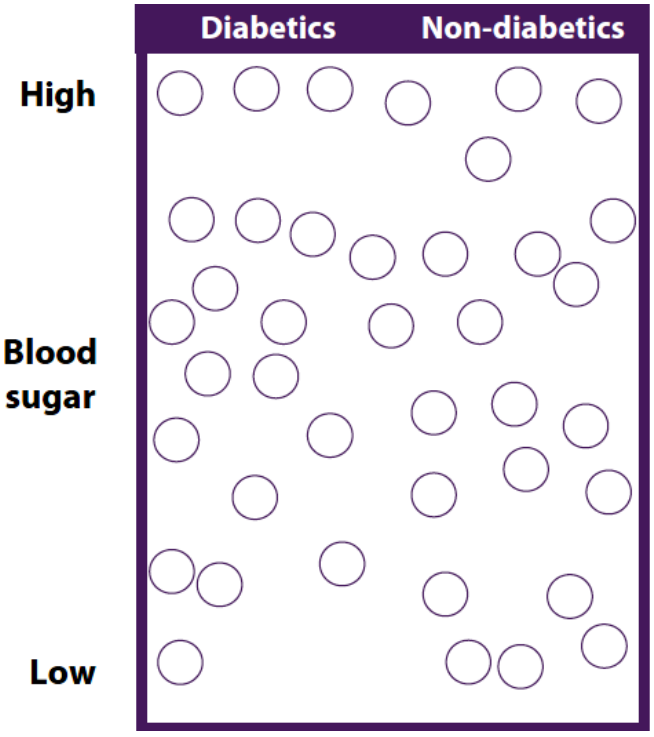
Diagnostic Testing - II



In a typical population, there is no line separating the two groups, and the subjects are mixed

http://ocw .jhsph.edu/courses/FundEpi/PDFs/Lecture11.pdf

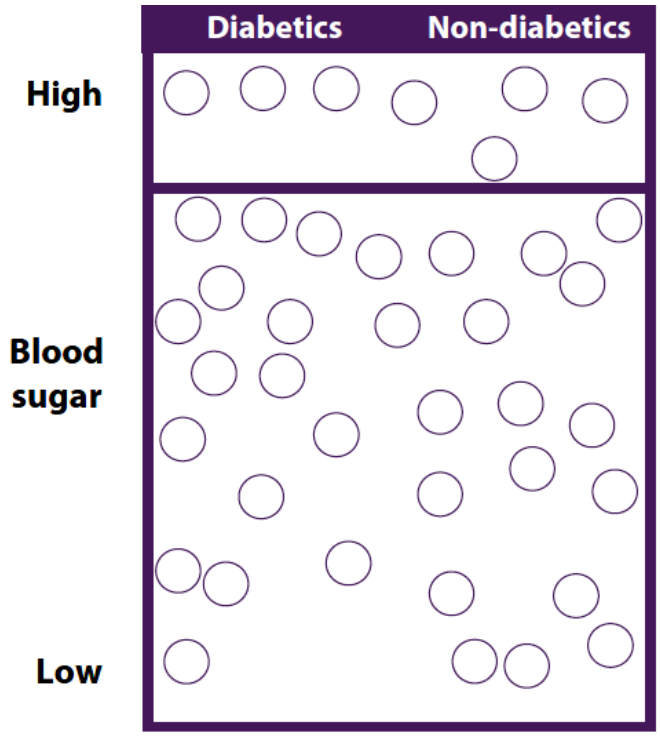
Diagnostic Testing - II



In fact, there is no color or label

http://ocw .jhsph.edu/courses/FundEpi/PDFs/Lecture11.pdf

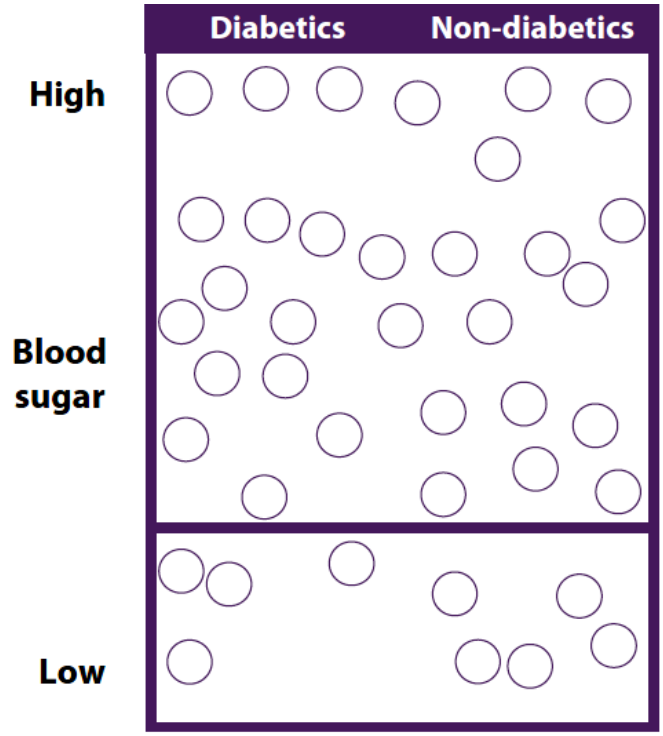
Diagnostic Testing - II



A screening test using a high cut-point will treat the bottom box as normal and will identify the 7 subjects above the line as having diabetes

http://ocw .jhsph.edu/courses/FundEpi/PDFs/Lecture11.pdf

Diagnostic Testing - II



A screening test using a high cut-point will treat the bottom box as normal and will identify the 7 subjects above the line as having diabetes;
But a low cut-point will result in identifying 31 subjects as having diabetes

http://ocw .jhsph.edu/courses/FundEpi/PDFs/Lecture11.pdf

Diagnostic Testing - II

Lessons Learned

- Different cut-points yield different sensitivities and specificities
- The cut-point determines how many subjects will be considered as having the disease
- The cut-point that identifies more true negatives will also identify more false negatives
- The cut-point that identifies more true positives will also identify more false positives

Diagnostic Testing - II

Where to Draw the Cut-Point

- If the diagnostic (confirmatory) test is expensive or invasive:
  - Minimize false positives
    or
  - Use a cut-point with high specificity
- If the penalty for missing a case is high (e.g., the disease is fatal and treatment exists, or disease easily spreads):
  - Maximize true positives
    - That is, use a cut-point with high sensitivity
- Balance severity of false positives against false negatives

Diagnostic Testing - II

END

Sensitivity and **Specificity** values alone may be highly misleading

The "Worst-case" sensitivity and specificity must be calculated in order to avoid reliance on experiments with few results.

1 – Spec = False Positive Rate
1- Sens = False Negative Rate
Power = Sensitivity

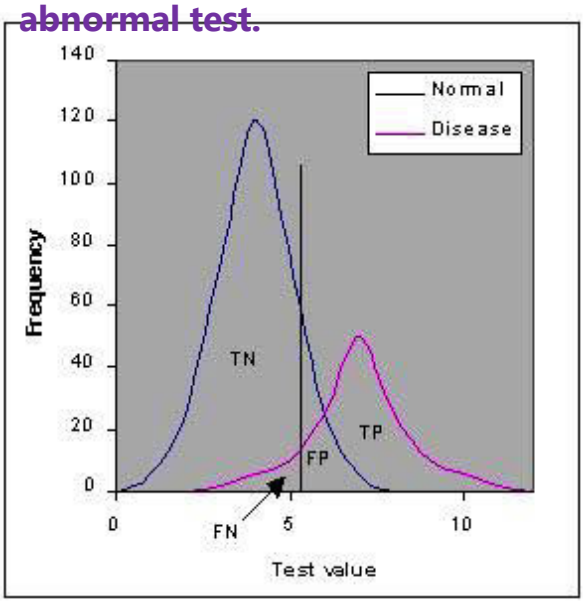
Applied Biostatistics

ROC CURVES

ROC CURVES

The Sensitivity and Specificity of a diagnostic test depends on more than just the "quality " of the test.

They also depends on the definition of what constitutes an abnormal test.



Look at the idealized graph showing the number of patients with and without a disease arranged according to the value of a diagnostic test.

The area of overlap indicates where the test cannot distinguish normal from disease.

In practice, we choose a cutpoint above which we consider the test to be abnormal and below which we consider the test to be normal.

http://gim.unmc.edu/dxtests/ROC1.htm

ROC CURVES

Consider the following data on patients with suspected hypothyroidism reported by Goldstein and Mushlin (J Gen Intern Med 1987;2:20-24.). They measured T4 and TSH values in ambulatory patients with suspected hypothyroidism and used the TSH values as a gold standard for determining which patients were truly hypothyroid.

T4 value	Hypothyroid	Euthyroid
5 or less	18	1
5.1 - 7	7	17
7.1 - 9	4	36
9 or more	3	39
Totals:	32	93

The lower the T4 value, the more likely the patients are to be hypothyroid.

http://gim.unmc.edu/dxtests/Default.htm

ROC CURVES

To illustrate how sensitivity and specificity change depending on the choice of T4 level that defines hypothyroidism.

T4 value	Hypothyroid	Euthyroid
5 or less	18	1
5.1 - 7	7	17
7.1 - 9	4	36
9 or more	3	39
Totals:	32	93

T4 values of 5 or less are considered to by hypothyroid

T4 value	Hypothyroid	Euthyroid
5 or less	18	1
> 5	14	92
Totals:	32	93

the sensitivity is 0.56 and the specificity is 0.99

http://gim.unmc.edu/dxtests/Default.htm

ROC CURVES

To illustrate how sensitivity and specificity change depending on the choice of T4 level that defines hypothyroidism.

T4 value	Hypothyroid	Euthyroid
5 or less	18	1
5.1 - 7	7	17
7.1 - 9	4	36
9 or more	3	39
Totals:	32	93

T4 values of 7 or less are considered to by hypothyroid

T4 value	Hypothyroid	Euthyroid
7 or less	25	18
> 7	7	75
Totals:	32	93

the sensitivity is 0.78 and the specificity is 0.81

http://gim.unmc.edu/dxtests/Default.htm

ROC CURVES

To illustrate how sensitivity and specificity change depending on the choice of T4 level that defines hypothyroidism.

T4 value	Hypothyroid	Euthyroid
5 or less	18	1
5.1 - 7	7	17
7.1 - 9	4	36
9 or more	3	39
Totals:	32	93

T4 values of 9 or less are considered to by hypothyroid

T4 value	Hypothyroid	Euthyroid
9 or less	29	54
> 9	3	39
Totals:	32	93

the sensitivity is 0.91 and the specificity is 0.42

http://gim.unmc.edu/dxtests/Default.htm

ROC CURVES

T4 value	Hypothyroid	Euthyroid
5 or less	18	1
5.1 - 7	7	17
7.1 - 9	4	36
9 or more	3	39
Totals:	32	93

Hence the table of Sensitivity and Specificity at various cut points is given as

Cutpoint	Sensitivity	Specificity
5	0.56	0.99
7	0.78	0.81
9	0.91	0.42

- **Improve the sensitivity by moving to cut-point to a *higher* T4 value**
- **You can make the criterion for a positive test *less* strict.**
- **improve the specificity by moving the cut-point to a lower T4 value**
- **you can make the criterion for a positive test more strict**
- **there is a tradeoff between sensitivity and specificity**
- **change the definition of a positive test to improve one but the other will decline**

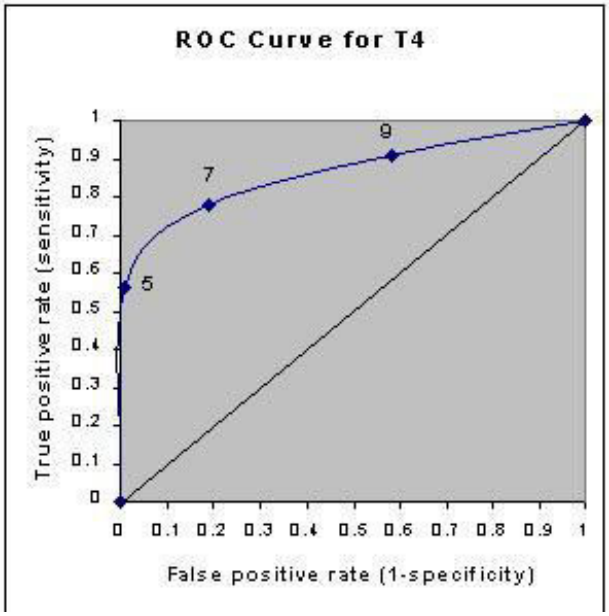
http://gim.unmc.edu/dxtests/Default.htm

ROC CURVES

Hence the table of Sensitivity and Specificity at various cut points is given as

Cutpoint	Sensitivity	Specificity
5	0.56	0.99
7	0.78	0.81
9	0.91	0.42

Cutpoint	True Positives	False Positives
5	0.56	0.01
7	0.78	0.19
9	0.91	0.58



http://gim.unmc.edu/dxtests/Default.htm

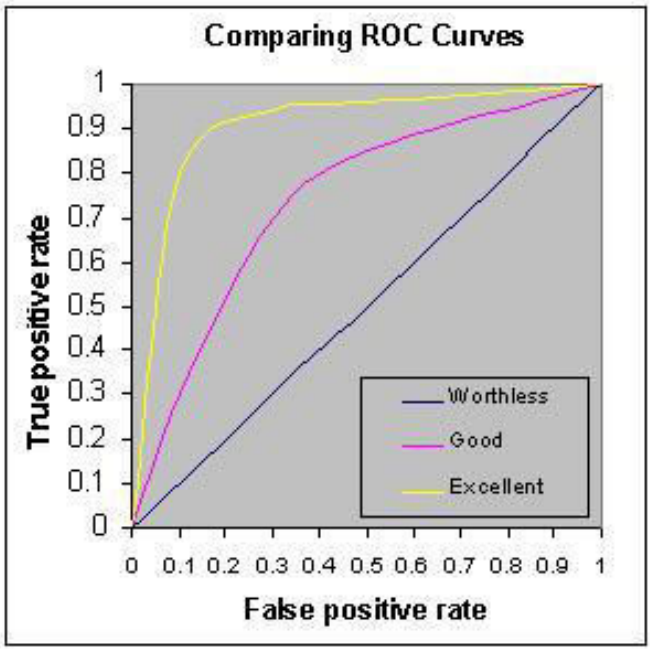
ROC CURVES

Receiver Operating Characteristic (ROC) Curve, is a plot of the true positive rate against the false positive rate for the different possible cut points of the diagnostic test.

1. **It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).**
2. **The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.**
3. **The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.**
4. **The slope of the tangent line at a cut point gives the likelihood ratio (LR) for that value of the test.**

http://gim.unmc.edu/dxtests/Default.htm

ROC CURVES



0.90-1 = excellent (A)
0.80-0.90 = good (B)
0.70-0.80 = fair (C)
0.60-0.70 = poor (D)
0.50-0.60 = fail (F)

http://gim.unmc.edu/dxtests/Default.htm

ROC CURVES

END

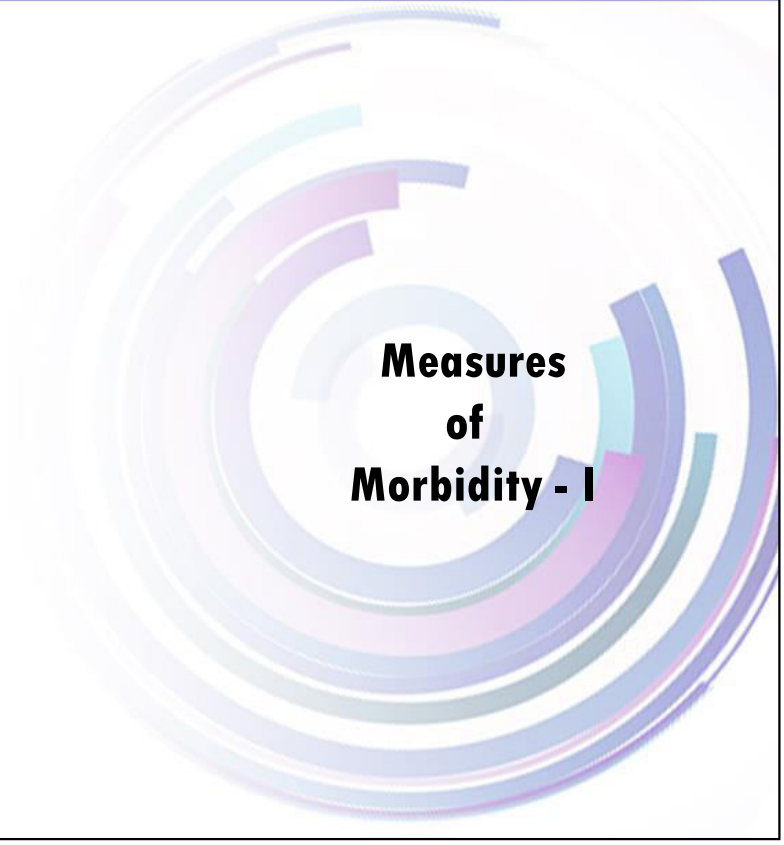
ROC analysis is a part of field called " Signal Detection Theory"

Which was developed during World War II for the analysis of radar images.

It was not until 1970's that signal detection theory was recognized as useful for interpreting medical test results.

http://gim.unmc.edu/dxtests/Default.htm

Applied Biostatistics



Measures of Morbidity - I

Morbidity has been defined as any departure , subjective, objective, from a state of physiological, psychological well – being.

In Practice, Morbidity encompasses disease, injury, and disability, and number of persons who are ill.

Measures of morbidity frequency characterize the number of persons in a population who become ill (incidence) or are ill at a given time (prevalence).

https://www.cdc.gov/ophss/csels/dsepd/ss1978/lesson3/section2.html

Measures of Morbidity - I

Measure	Numerator	Denominator
Incidence Proportion (or Attack Rate or Risk)	Number of New Cases of Disease during Specified Time Interval	Population at start of Time interval
Incidence Rate (Or Person Time Rate)	Number of New Cases of Disease during Specified Time Interval	Summed person-years of observation or average population during time interval
Point prevalence	Number of current cases (new and preexisting) at a specified point in time	Population at the same specified point in time
Period prevalence	Number of current cases (new and preexisting) over a specified period of time	Average or mid-interval population

https://www.cdc.gov/ophss/csels/dsepd/ss1978/lesson3/section2.html

Measures of Morbidity - I

Incidence refers to the occurrence of new cases of disease or injury in a population over a specified period of time.

Although some epidemiologists use incidence to mean the number of new cases in a community,

others use incidence to mean the number of new cases per unit in the population.

- **Incidence Proportion**
- **Incidence Rate**

https://www.cdc.gov/ophss/csels/dsepd/ss1978/lesson3/section2.html

Measures of Morbidity - I

Incidence Proportion

**Incidence proportion is the proportion of an initially disease-free population that develops disease,
becomes injured, or dies during a specified**

Synonyms include attack rate, risk, probability of getting disease, and cumulative incidence.

https://www.cdc.gov/ophss/csels/dsepd/ss1978/lesson3/section2.html

Measures of Morbidity - I

Incidence Proportion

Incidence proportion is a proportion because the persons in the numerator, those who develop disease, are all included in the denominator (the entire population).

$$Incidence\ Proportion\ (Risk) = \frac{Number\ of\ new\ cases\ of\ disesase\ or\ injury\ during\ specified\ period}{Size\ of\ Population\ at\ the\ start\ of\ period}$$

https://www.cdc.gov/ophss/csels/dsepd/ss1978/lesson3/section2.html

Measures of Morbidity - I

Example A: In the study of diabetics, 100 of the 189 diabetic men died during the 13-year follow-up period. Calculate the risk of death for these men.

Numerator = 100 deaths among the diabetic men
Denominator = 189 diabetic men
10^n = 10^2 = 100

Risk = (100 / 189) × 100 = 52.9%

https://www.cdc.gov/ophss/csels/dsepd/ss1978/lesson3/section2.html

Measures of Morbidity - I

Example B: In an outbreak of gastroenteritis among attendees of a corporate picnic, 99 persons ate potato salad, 30 of whom developed gastroenteritis. Calculate the risk of illness among persons who ate potato salad.

Numerator = 30 persons who ate potato salad and developed gastroenteritis
Denominator = 99 persons who ate potato salad
10^n = 10^2 = 100

Risk = "Food-specific attack rate" = (30 / 99) × 100
= 0.303 × 100
= 30.3%

https://www.cdc.gov/ophss/csels/dsepd/ss1978/lesson3/section2.html

Measures of Morbidity - I

Properties and uses of Incidence

The probability of developing the disease during the specified period.

It includes only new cases of disease

The denominator is the number of persons in the population at the start of the observation period.

A risk is also a proportion.

https://www.cdc.gov/ophss/csels/dsepd/ss1978/lesson3/section2.html

Measures of Morbidity - I

Incidence Rate

Incidence rate or person-time rate is a measure of incidence that incorporates time directly into the denominator.

A person-time rate is generally calculated from a long-term cohort follow-up study

https://www.cdc.gov/ophss/csels/dsepd/ss1978/lesson3/section2.html

Measures of Morbidity - I

Incidence Rate

Typically, each person is observed from an established starting time until one of four "end points" is reached:

- **Onset of disease**
- **Death**
- **Migration out of the study ("lost to follow-up")**

Similar to the incidence proportion, the numerator of the incidence rate is the number of new cases identified during the period of observation. However, the denominator differs.

https://www.cdc.gov/ophss/csels/dsepd/ss1978/lesson3/section2.html

Measures of Morbidity - I

Example

In 2003, 44,232 new cases of acquired immunodeficiency syndrome (AIDS) were reported in the United States. The estimated mid-year population of the U.S. in 2003 was approximately 290,809,777. Calculate the incidence rate of AIDS in 2003.

Numerator = 44,232 new cases of AIDS
Denominator = 290,809,777 estimated mid-year population
10^n = 100,000

Incidence rate = (44,232 / 290,809,777) × 100,000
= 15.21 new cases of AIDS per 100,000 population

https://www.cdc.gov/ophss/csels/dsepd/ss1978/lesson3/section2.html

Measures of Morbidity - I

Properties and Uses of Incidence Rate

An incidence rate describes how quickly disease occurs in a population

Because person-time is calculated for each subject, it can accommodate persons coming into and leaving the study.

The denominator accounts for study participants who are lost to follow-up or who die during the study period.

https://www.cdc.gov/ophss/csels/dsepd/ss1978/lesson3/section2.html

Measures of Morbidity - I

Properties and Uses of Incidence Rate

Person-time has one important drawback. Person-time assumes that the probability of disease during the study period is constant,

Long-term cohort studies of the type described here are not very common

https://www.cdc.gov/ophss/csels/dsepd/ss1978/lesson3/section2.html

Measures of Morbidity - I

END

Finally, if you report the incidence rate of, say, the heart disease study as 2.5 per 1,000 person-years,

epidemiologists might understand, but most others will not.

simply replace "person-years" with "persons per year."

Applied Biostatistics

Measures
of
Morbidity - II

Measures of Morbidity - II

Prevalence: Sometimes referred to as **Prevalence Rate**.

It is the **Proportion of persons** in a population who have a **particular disease or attribute** at a **specified point in time** or over a **specified period of time**.

Point prevalence refers to the prevalence measured at a particular point in time. It is the proportion of persons with a particular disease or attribute on a particular date.

Period prevalence refers to prevalence measured over an interval of time. It is the proportion of persons with a particular disease or attribute at any time during the interval.

https://www.cdc.gov/ophss/csels/dsepd/ss1978/lesson3/section2.html

Measures of Morbidity - II

Method for calculating prevalence of disease

$$\frac{\text{All new and preexisting cases durign a given time period}}{\text{Population during the same time period}} \times 10^n$$

Method for calculating prevalence of an attribute

$$\frac{\text{Persons having a particular attribute durign a given time period}}{\text{Population during the same time period}} \times 10^n$$

The value of 10^n is usually 1 or 100 for common attributes. The value of 10^n might be 1,000, 100,000, or even 1,000,000 for rare attributes and for most diseases.

https://www.cdc.gov/ophss/csels/dsepd/ss1978/lesson3/section2.html

Measures of Morbidity - II

EXAMPLE: Calculating Prevalence

In a survey of 1,150 women who gave birth in Maine in 2000, a total of 468 reported taking a multivitamin at least 4 times a week during the month before becoming pregnant. Calculate the prevalence of frequent multivitamin use in this group.

Numerator = 468 multivitamin users

Denominator = 1,150 women

Prevalence = (468 / 1,150) × 100 = 0.407 × 100 = 40.7%

https://www.cdc.gov/ophss/csels/dsepd/ss1978/lesson3/section2.html

Measures of Morbidity - II

Properties and uses of Prevalence

Prevalence and incidence are usually confused. The key difference is in their numerator

Numerator of incidence = new cases that occurred during a given time period

Numerator of prevalence = all cases present during a given time period

Prevalence is based on both incidence and duration of illness.

https://www.cdc.gov/ophss/csels/dsepd/ss1978/lesson3/section2.html

Measures of Morbidity - II

Properties and uses of Prevalence

High prevalence of a disease within a population might reflect high incidence or prolonged survival without cure or both.

Conversely, low prevalence might indicate low incidence, a rapidly fatal process, or rapid recovery.

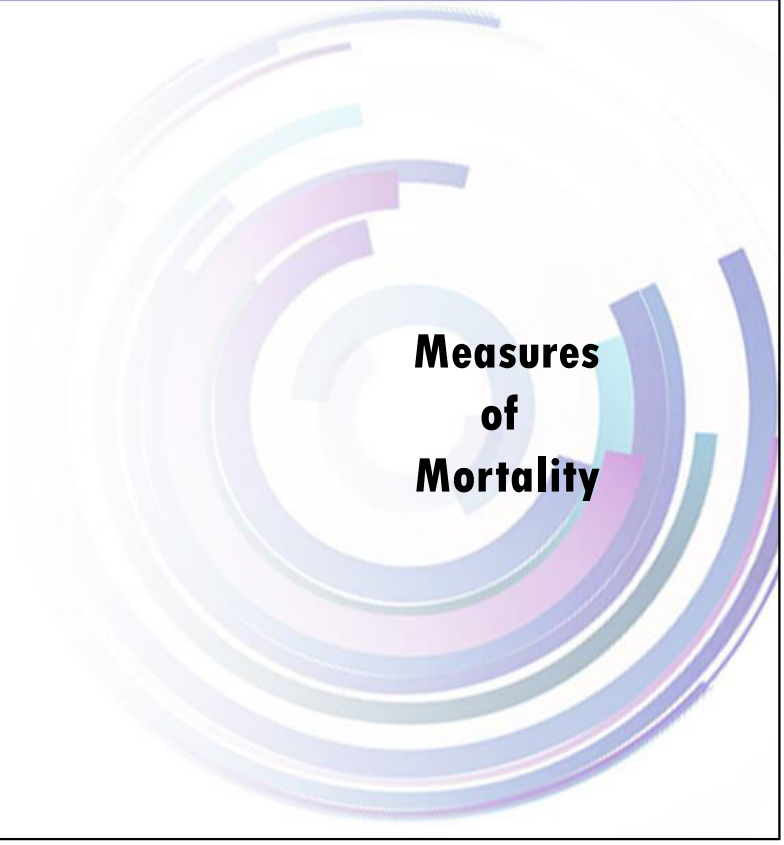
https://www.cdc.gov/ophss/csels/dsepd/ss1978/lesson3/section2.html

Measures of Morbidity - II

END

Prevalence rather than incidence is often measured for chronic diseases such as diabetes or osteoarthritis which have long duration and dates of onset that are difficult to pinpoint.

Applied Biostatistics



Measures of Mortality

Mortality Rate

A **Mortality rate** is a measure of the **frequency of occurrence of death** in a defined population during **a specified interval.**

Mortality and **Morbidity** measures are often **the same mathematically**; its just a matter of what you choose to measure, **illness** or **death**

$$Mortality\ Rate = \frac{Deaths\ occuring\ during\ a\ given\ time\ period}{Size\ of\ Population\ among\ which\ the\ deaths\ occured} \times 10^n$$

https://www.cdc.gov/ophss/csels/dsepd/ss1978/lesson3/section2.html

Measures of Mortality

Mortality Rate

$$Mortality\ Rate = \frac{\begin{array}{c}Deaths\ occuring\ during \\ a\ given\ time\ period\end{array}}{\begin{array}{c}Size\ of\ Population \\ among\ which\ the\ deaths\ occured\end{array}} \times 10^n$$

When mortality rates are based on vital statistics (e.g., counts of death certificates), the denominator most commonly used is the size of the population at the middle of the time period.

https://www.cdc.gov/ophss/csels/dsepd/ss1978/lesson3/section2.html

Measures of Mortality

Crude Mortality (Death) Rate

$$Crude\ Mortality\ Rate = \frac{\begin{array}{c}Total\ number\ of\ deaths\ during \\ a\ given\ time\ interval\end{array}}{Mid\ Year\ Population} \times 100{,}000$$

The crude mortality rate is the mortality rate from all causes of death for a population.

In Pakistan the Crude Mortality Rate has been non-increasing for past many years.

https://www.cdc.gov/ophss/csels/dsepd/ss1978/lesson3/section2.html

Measures of Mortality

Crude Mortality (Death) Rate

The population (in thousands) of Pakistan in 2016 was estimated to be 199,710 and the total number of death (in thousands) for the same year were 1318. Then the Crude Mortality Rate can be measured as:

$$Crude\ Mortality\ Rate = \frac{1{,}318{,}000}{199{,}710{,}000} \times 1{,}000 = 6.6$$

This values means that, in the year 2016, there were 6.6 deaths occurred per 1,000 people living in a country

Measures of Mortality

Crude Mortality (Death) Rate

To compare the crude death rates of two communities is hazardous.

Unless it is known that the communities are comparable with respect to the many characteristics

These comparable characteristics should be other than health conditions, that influence the death rate.

Measures of Mortality

Cause Specific Mortality Rate

$$Cause\ Specific\ Mortality\ Rate = \frac{Number\ of\ deaths\ assigned\ to\ a\ specific\ cause\ during\ a\ given\ time\ interval}{Mid-Interval\ Population} \times 10^n$$

The cause-specific mortality rate is the mortality rate from a specified cause for a population.

The numerator is the number of deaths attributed to a specific cause.

The denominator remains the size of the population at the midpoint of the time period

https://www.cdc.gov/ophss/csels/dsepd/ss1978/lesson3/section2.html

Measures of Mortality

Age Specific Mortality Rate

$$Age\ Specific\ Mortality\ Rate = \frac{Number\ of\ deaths\ in\ specific\ age\ group}{Number\ of\ Person\ in\ that\ age\ group\ in\ the\ population} \times 10^n$$

An age-specific mortality rate is a mortality rate limited to a particular age group.

The numerator is the number of deaths in that age group

The denominator is the number of persons in that age group in the population.

https://www.cdc.gov/ophss/csels/dsepd/ss1978/lesson3/section2.html

Measures of Mortality

Infant Mortality Rate

$$Infant\ Mortality\ Rate = \frac{Number\ of\ deaths\ among\ children < 1\ year\ of\ age\ reported\ during\ a\ given\ time\ period}{Number\ of\ Live\ births\ reported\ during\ the\ same\ time\ period} \times 10^n$$

The infant mortality rate is generally calculated on an annual basis.

It is a widely used measure of health status because it reflects the health of the mother and infant during pregnancy and the year thereafter.

Is the infant mortality rate a ratio? Yes

https://www.cdc.gov/ophss/csels/dsepd/ss1978/lesson3/section2.html

Measures of Mortality

END

There are various other Measures of Mortality:

- **Neonatal Mortality Rate**
- **Fetal Death Rate**
- **Child Mortality Rate**
- **Adjusted or Standardized Death Rates**
- **Maternal Mortality Rate**

https://www.cdc.gov/ophss/csels/dsepd/ss1978/lesson3/section2.html

Applied Biostatistics

Bernoulli Distribution

Lecture no 58

Bernoulli Distribution

Bernoulli Trials

http://www.turkiyeningercekleri.com/1w2o3r4d5p6r7e8s9s0/venus-ve-mars-sembolleri/

https://stpeteurology.com/treatment-success-overactive-bladder/

https://www.terminix.com/blog/education/why-mosquitoes-bite-me-so-much

Bernoulli Distribution

Bernoulli Trial

An experiment or a trial, whose outcome can be classified as either a success or a failure, is called a Bernoulli Trial.

- Jacob Bernoulli (1965 – 1705)
- Swiss Mathematician

https://en.wikipedia.org/wiki/Jacob_Bernoulli#/media/File:Jakob_Bernoulli.jpg

Bernoulli Distribution

Bernoulli Distribution

The Bernoulli Distribution is a discrete probability Distribution having two possible outcomes.

Let X be a Bernoulli random variable, that occurs as a result of a Bernoulli trial.

X = 1 ; When the outcome is success.
X = 0 ; When the outcome is failure.

Bernoulli Distribution

Bernoulli Distribution

If p denotes the probability of Success
1 – p denotes the probability of Failure

Then the Probability Mass Function of X, which is a Bernoulli Random Variable can be given as:

$$f(x) = p^x(1-p)^{1-x} \qquad x = 0, 1$$

Bernoulli Distribution

Bernoulli Distribution (Example)

Random Experiment: The birth of a child

No. of Possible Outcomes = 2
1. Baby Boy
2. Baby Girl

http://www.turkiyeningercekleri.com/1w2o3r4d5p6r7e8s9s0/venus-ve-mars-sembolleri/

Success = Birth of a Baby girl = p
Failure = Not a birth of a baby girl = 1- p

Event of Interest = Birth of a Baby girl

Let the random variable "X" denotes the birth of a baby girl
X = 0, 1

$$p(X = x) = \begin{cases} p^x(1-p)^{1-x} & for\ x \in \{0,1\} \\ 0 & Otherwise \end{cases}$$

Bernoulli Distribution

Bernoulli Distribution (Example)

No. of Possible Outcomes = 2
1. Baby Boy
2. Baby Girl

Event of Interest = Birth of a Baby girl

Let the random variable "X" denotes the birth of a baby girl
X = 0, 1

Success = p = 0.60
Failure = 1- p = 1 – 0.60 = 0.40

$$p(X = x) = \begin{cases} p^x(1-p)^{1-x} & for\ x \in \{0,1\} \\ 0 & Otherwise \end{cases}$$

$$P(X = 0) = (0.60)^0(1-0.60)^{1-0} = 0.40$$

$$P(X = 1) = (0.60)^1(1-0.60)^{1-1} = 0.60$$

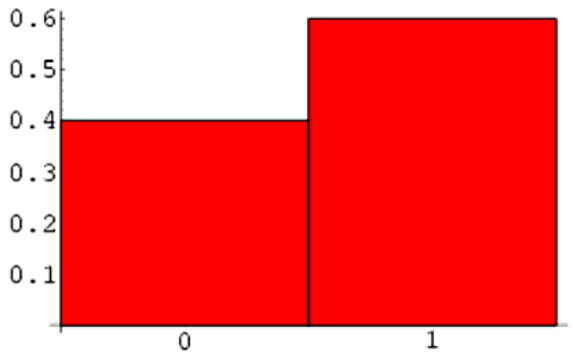
Bernoulli Distribution

Bernoulli Distribution (Example)

$$p(X = x) = \begin{cases} p^x(1-p)^{1-x} & for\ x \in \{0,1\} \\ 0 & Otherwise \end{cases}$$

$$P(X = 0) = (0.60)^0(1-0.60)^{1-0} = 0.40$$

$$P(X = 1) = (0.60)^1(1-0.60)^{1-1} = 0.60$$



Bernoulli Distribution

Bernoulli Distribution (Example)

$$p(X = x) = \begin{cases} p^x(1-p)^{1-x} & for\ x \in \{0,1\} \\ 0 & Otherwise \end{cases}$$

$$P(X = 0) = (0.60)^0(1-0.60)^{1-0} = 0.40$$

$$P(X = 1) = (0.60)^1(1-0.60)^{1-1} = 0.60$$

$$Mean = E(X) = \sum_{X=0,1} X\,P(X) = \big(0 \times (1-p)\big) + (1 \times p) = p$$

$$Mean = E(X) = \sum_{X=0,1} X\,P(X) = (0 \times 0.40) + (1 \times 0.60) = 0.60$$

$$Variance = E(X^2) - \big(E(X)\big)^2 = [0^2 \times (1-p) + 1^2 \times p] - (p)^2 = p(1-p)$$

$$Variance = p(1-p) = 0.60\,(1-0.60) = 0.24$$

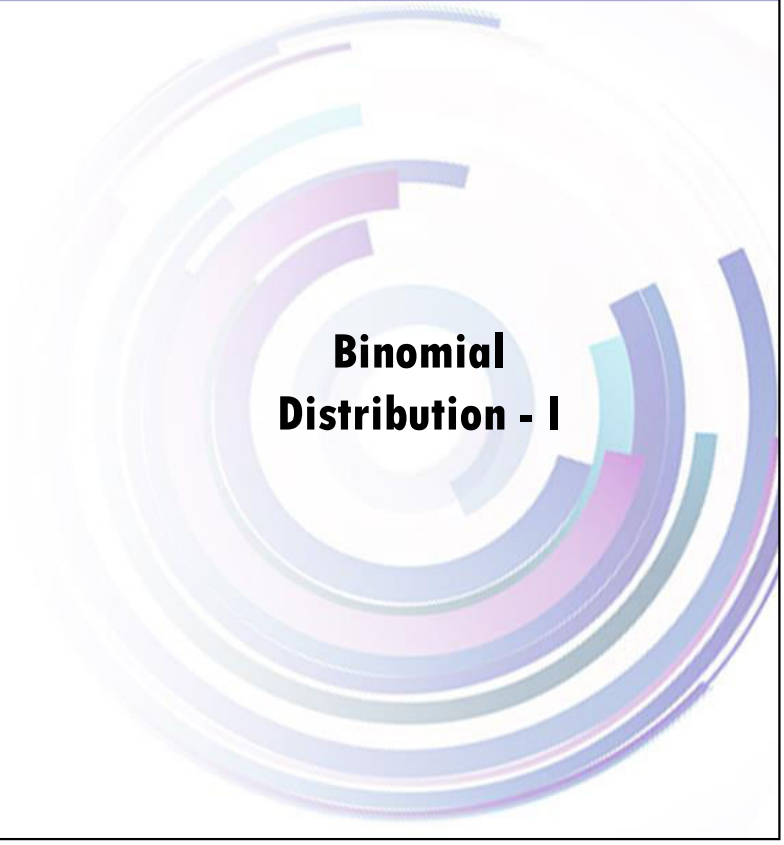
Bernoulli Distribution

END

- **The Bernoulli distribution is the simplest discrete Probability distribution.**
- **It is the building block for other more complicated discrete distributions.**

1. **Binomial Distribution**
2. **Geometric Distribution**
3. **Negative Binomial Distribution**

Applied Biostatistics



Binomial Distribution - I

Bernoulli Process

A Sequence of Bernoulli Trials forms a Bernoulli Process, under the following conditions

1. **Each Trial results in one of the two possible, Mutually exclusive, outcomes.**
    1. **Success**
    2. **Failure**

2. **The Probability of Success, denoted by "p" remains constant from trial to trial.**

3. **The trials are independent.**

Binomial Distribution - I

Binomial Probability Distribution

A Binomial Probability Distribution results from a Bernoulli Process that meets all the following requirements.

1. **The Procedure has a fixed number of Trials.**
2. **Each trial must have all outcomes classified into two categories (Success, Failure)**
3. **The Probability of success remains constant from trial to trial.**
4. **Successive trials must be independent.**

Binomial Distribution - I

Binomial Probability Mass Function

$$P(X = x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & x = 0, 1, 2, \cdots, n \\ 0 & Otherwise \end{cases}$$

S and F (Success and Failure) will denote two possible outcomes

p and q will denote the probabilities of S and F, respectively, so.

P(S) = p (p = probability of success)
P(F) = 1 – p = q (q = probability of failure)

Binomial Distribution - I

Binomial Probability Distribution: Notation

n = Number of fixed Trials

x = Specific number of successes in "n" trials.

p = The probability of success in one of the "n" trials

q = The probability of failure in one of the "n" trials

P(x) = The probability of getting exactly "x" success among the "n" trials.

Binomial Distribution - I

Binomial Probability Distribution: Rationale

$$b(x;n,p) = P(X = x) = \frac{n!}{x!\,(n-x)!}\,p^x(1-p)^{n-x} \qquad x = 0, 1, 2, \cdots, n$$

No. of outcome with exactly "x" successes aomng "n" trials

The Probability of "x" successesamong "n" trials for any one particular order

Binomial Distribution - I

Methods for Finding Probability

We will discuss three methods for finding the probabilities corresponding to the random variable "x" in a binomial distribution.

1. **Using the Binomial Probability Formula**
2. **Using the Table of Probabilities**
3. **Using technology**

Binomial Distribution - I

Method 1: Using the Binomial Probability Formula

$$P(X=x)=\begin{cases}\binom{n}{x}p^x(1-p)^{n-x} & x=0,1,2,\cdots,n\\ 0 & Otherwise\end{cases}$$

$$P(X=x)=\frac{n!}{x!\,(n-x)!}p^x(1-p)^{n-x} \qquad x=0,1,2,\cdots,n$$

Binomial Distribution - I

Method 1: Example

Using Multiplication Rule:

Suppose we have *n = 5* and *p = 0.60*. The probability that binomial trials yield four successes will be given as (one way to do only):

$$P(SSSSF) = pppqq = (0.60)^4 \times 0.40 = 0.05184$$

Binomial Distribution - I

Method 1: Steps

Step 1: Identify a success

Step 2: Determine , p , the success probability

Step 3: Determine, n , the number of trials

Step 4: The binomial probability formula for the number of successes, x , is

$$P(X = x) = \frac{n!}{x!\,(n-x)!}\, p^x (1-p)^{n-x} \qquad x = 0, 1, 2, \cdots, n$$

Binomial Distribution - I

Method 1: Example

Using Binomial Probability Distribution:

Suppose we have *n = 5* and *p = 0.60*. The probability that binomial trials yield four successes will be given as :

n = 5
p = 0.60
1 – p = q = 0.40
x = 4

$$P(X = x) = \frac{n!}{x!\,(n-x)!}\,p^x(1-p)^{n-x} \qquad x = 0, 1, 2, \cdots, n$$

Binomial Distribution - I

Method 1: Example

Using Binomial Probability Distribution:

p = 0.60
1 – p = q = 0.40
x = 4

$$P(X = x) = \frac{n!}{x!\,(n-x)!}\,p^x(1-p)^{n-x} \qquad x = 0, 1, 2, \cdots, n$$

$$P(X = 4) = \frac{5!}{4!\,(5-4)!}\,(0.60)^4(1-0.60)^{5-4} = 5\,(0.1296)(0.4) = 0.2592$$

Binomial Distribution - I

Method 1: Example

Using Binomial Probability Distribution:

n = 5
p = 0.60
1 – p = q = 0.40

$$P(X=0)=\frac{5!}{0!5!}(0.60)^0(0.40)^5=0.0102.$$

$$P(X=1)=\frac{5!}{1!4!}(0.60)^1(0.40)^4=0.0768.$$

$$P(X=2)=\frac{5!}{2!3!}(0.60)^2(0.40)^3=0.2304.$$

$$P(X=3)=\frac{5!}{3!2!}(0.60)^3(0.40)^2=0.3456.$$

$$P(X=4)=\frac{5!}{4!1!}(0.60)^4(0.40)^1=0.2592.$$

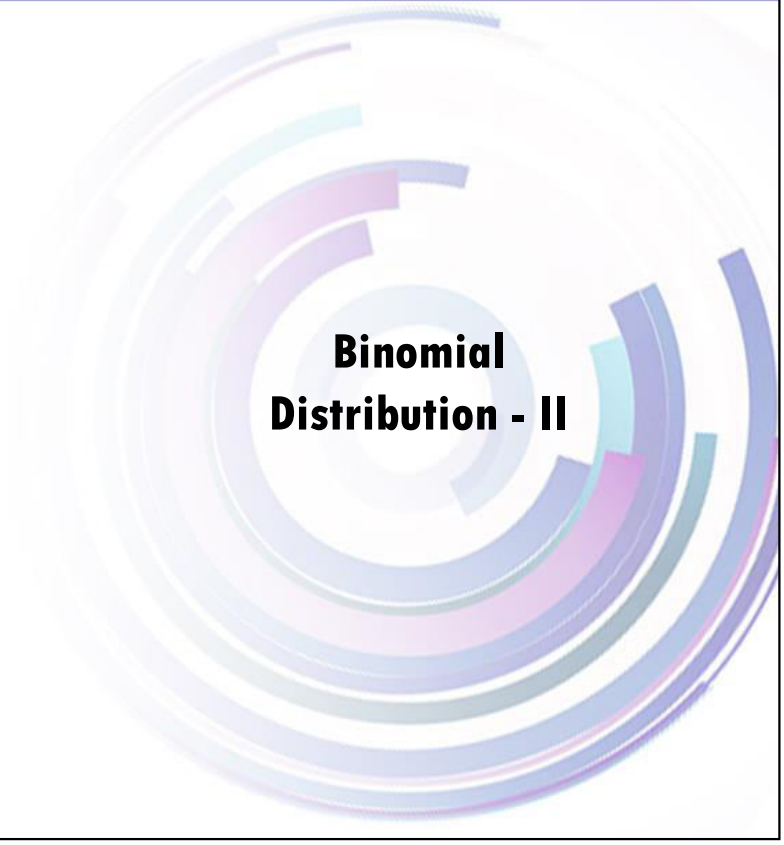
$$P(X=5)=\frac{5!}{5!0!}(0.60)^5(0.40)^0=0.0778.$$

Binomial Distribution - I

END

- **One has to be very certain that "x" and "p" both refers to the same category, being called a success.**
- **When the sampling is conducted without replacement, then consider events to be independent if n<0.05 N**

Applied Biostatistics



Binomial Distribution - II

Method 2: Using the Table of Probabilities

		p											
n	x	.01	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50	.55	.60
2	0	.980	.902	.810	.723	.640	.563	.490	.423	.360	.303	.250	.203	.160
	1	.020	.095	.180	.255	.320	.375	.420	.455	.480	.495	.500	.495	.480
	2	.000	.002	.010	.023	.040	.063	.090	.123	.160	.203	.250	.303	.360
3	0	.970	.857	.729	.614	.512	.422	.343	.275	.216	.166	.125	.091	.064
	1	.029	.135	.243	.325	.384	.422	.441	.444	.432	.408	.375	.334	.288
	2	.000	.007	.027	.057	.096	.141	.189	.239	.288	.334	.375	.408	.432
	3	.000	.000	.001	.003	.008	.016	.027	.043	.064	.091	.125	.166	.216
4	0	.961	.815	.656	.522	.410	.316	.240	.179	.130	.092	.062	.041	.026
	1	.039	.171	.292	.368	.410	.422	.412	.384	.346	.300	.250	.200	.154
	2	.001	.014	.049	.098	.154	.211	.265	.311	.346	.368	.375	.368	.346
	3	.000	.000	.004	.011	.026	.047	.076	.112	.154	.200	.250	.300	.346
	4	.000	.000	.000	.001	.002	.004	.008	.015	.026	.041	.062	.092	.130
5	0	.951	.774	.590	.444	.328	.237	.168	.116	.078	.050	.031	.019	.010
	1	.048	.204	.328	.392	.410	.396	.360	.312	.259	.206	.156	.113	.077
	2	.001	.021	.073	.138	.205	.264	.309	.336	.346	.337	.312	.276	.230
	3	.000	.001	.008	.024	.051	.088	.132	.181	.230	.276	.312	.337	.346
	4	.000	.000	.000	.002	.006	.015	.028	.049	.077	.113	.156	.206	.259
	5	.000	.000	.000	.000	.000	.001	.002	.005	.010	.019	.031	.050	.078

Binomial Distribution - II

Method 2: Using the Table of Probabilities

n = 5

p = 0.60

1 – p = q = 0.40

$$P(X=0)=\frac{5!}{0!5!}(0.60)^0(0.40)^5=0.0102.$$

$$P(X=1)=\frac{5!}{1!4!}(0.60)^1(0.40)^4=0.0768.$$

$$P(X=2)=\frac{5!}{2!3!}(0.60)^2(0.40)^3=0.2304.$$

$$P(X=3)=\frac{5!}{3!2!}(0.60)^3(0.40)^2=0.3456.$$

$$P(X=4)=\frac{5!}{4!1!}(0.60)^4(0.40)^1=0.2592.$$

$$P(X=5)=\frac{5!}{5!0!}(0.60)^5(0.40)^0=0.0778.$$

		p				
n	x	.40	.45	.50	.55	.60
2	0	.360	.303	.250	.203	.160
	1	.480	.495	.500	.495	.480
	2	.160	.203	.250	.303	.360
3	0	.216	.166	.125	.091	.064
	1	.432	.408	.375	.334	.288
	2	.288	.334	.375	.408	.432
	3	.064	.091	.125	.166	.216
4	0	.130	.092	.062	.041	.026
	1	.346	.300	.250	.200	.154
	2	.346	.368	.375	.368	.346
	3	.154	.200	.250	.300	.346
	4	.026	.041	.062	.092	.130
5	0	.078	.050	.031	.019	.010
	1	.259	.206	.156	.113	.077
	2	.346	.337	.312	.276	.230
	3	.230	.276	.312	.337	.346
	4	.077	.113	.156	.206	.259
	5	.010	.019	.031	.050	.078

Binomial Distribution - II

Method 2: Cumulative Probability Function

Distribution Function

$$B(x;n,p)=\sum_{k=0}^{n}\binom{n}{k}p^k(1-p)^{n-k} \qquad x=0,1,2,\cdots,n$$

$$B(x;n,p)=\sum_{k=0}^{n}b(k;n,p) \qquad x=0,1,2,\cdots,n$$

$$b(x;n,p)=B(x;n,p)-B(x-1;n,p)$$

Binomial Distribution - II

Method 2: Cumulative Probability Function

$$P[X \le c] = \sum_{x=0}^{c} \binom{n}{x} p^x (1-p)^{n-x}$$

		p						
	c	0.05	0.10	0.20	0.30	0.40	0.50	0.60
n = 1	0	0.950	0.900	0.800	0.700	0.600	0.500	0.400
	1	1.000	1.000	1.000	1.000	1.000	1.000	1.000
n = 2	0	0.903	0.810	0.640	0.490	0.360	0.250	0.160
	1	0.998	0.990	0.960	0.910	0.840	0.750	0.640
	2	1.000	1.000	1.000	1.000	1.000	1.000	1.000
n = 3	0	0.857	0.729	0.512	0.343	0.216	0.125	0.064
	1	0.993	0.972	0.896	0.784	0.648	0.500	0.352
	2	1.000	0.999	0.992	0.973	0.936	0.875	0.784
	3	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Binomial Distribution - II

Method 2: Cumulative Probability Distribution

n = 5
p = 0.60
1 – p = q = 0.40

$$P(X = 4) = b(4; 5, 0.60)$$

$$b(4; 5, 0.60) = B(5; 5, 0.60) - B(4; 5, 0.60)$$

$$b(4; 5, 0.60) = 1 - 0.922 = 0.078$$

$$P[X \le c] = \sum_{x=0}^{c} \binom{n}{x} p^x (1-p)^{n-x}$$

		p			
	c	0.05	0.10	0.50	0.60
n = 1	0	0.950	0.900	0.500	0.400
	1	1.000	1.000	1.000	1.000
n = 2	0	0.903	0.810	0.250	0.160
	1	0.998	0.990	0.750	0.640
	2	1.000	1.000	1.000	1.000
n = 5	0	0.774	0.590	0.031	0.010
	1	0.977	0.919	0.188	0.087
	2	0.999	0.991	0.500	0.317
	3	1.000	1.000	0.813	0.663
	4	1.000	1.000	0.969	0.922
	5	1.000	1.000	1.000	1.000

Binomial Distribution - II

Mean, Variance and Standard Deviation of the Binomial Distribution

$$Mean = E(X) = \sum_{x=0}^{n} XP(X) = np$$

$$Variance = E(X^2) - [E(X)]^2 = npq$$

$$SD = \sqrt{variance} = \sqrt{npq}$$

n = 5
p = 0.60
1 – p = q = 0.40

Mean = 5 x 0.60 = 3
Variance = 5x0.60x0.40 = 1.2
SD = 1.90

Binomial Distribution - II

Interpretation of the results

It is especially important to interpret results. The Range Rule of Thumb suggests that values are unusual if they are outside of these limits.

Maximum Usual Value = $\mu + 2\sigma$

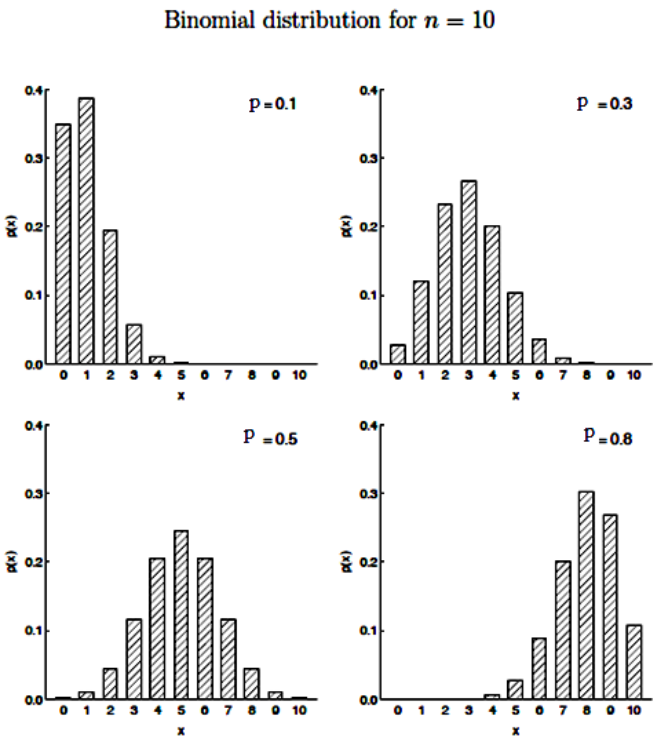
Minimum Usual Value = $\mu - 2\sigma$

Binomial Distribution - II

Effects of "n" and "p" on the shape

- For small "p" and small "n", the binomial distribution will be right skewed.
- For large "p" and small "n", the binomial distribution will be left skewed.
- For p = 0.5 and large or small "n" the binomial distribution will be symmetric.
- For small p but large "n" the binomial distribution approaches symmetry.

Binomial Distribution - II



Binomial Distribution - II

END

Methods 3:

Using Technology

Using MS Excel we will use a function

BINOMDIST(x,n,p,cumm)

Applied Biostatistics

Binomial Distribution Demo Using MS Excel

Applied Biostatistics

Poisson Distribution

Poisson Distribution

Interest:

Number of events occurring

- **In a specific period of time.**
- **In a specific area of volume.**

1. **Number of death claims received per day by an insurance company.**

2. **Number of families with child mortality.**

3. **Number of Aplha particles emitted from a radioactive source during a given period of time.**

Poisson Distribution

The Discrete Distribution used for such instances is called Poisson Probability Distribution.

Named after a French mathematician Simeon Denis Poisson

https://en.wikipedia.org/wiki/Sim%C3%A9on_Denis_Poisson

Poisson Distribution

Characteristics of a Poisson Random Variable

Let "X" be the number of times a certain event occurs during a given unit of time (or in given area)

- **The probability that event occurs in a given unit of time is the same for all the units.**
- **The number of events that occur in one unit of time is independent of the number of events in the other units.**
- **The mean (or expected) rate is λ.**

Poisson Distribution

Characteristics of a Poisson Random Variable

Let "X" be the number of times a certain event occurs during a given unit of time (or in given area) and it satisfies given characteristics

then "X" will be considered as a Poisson Random Variable, with parameter λ.

$$p(x) = \frac{\lambda^x}{x!} e^{-\lambda}, \qquad x = 0, 1, 2, \ldots$$

λ = The average number of events per interval
e = The number **2.71828**... (Euler's number) the base of the natural logarithms

Poisson Distribution

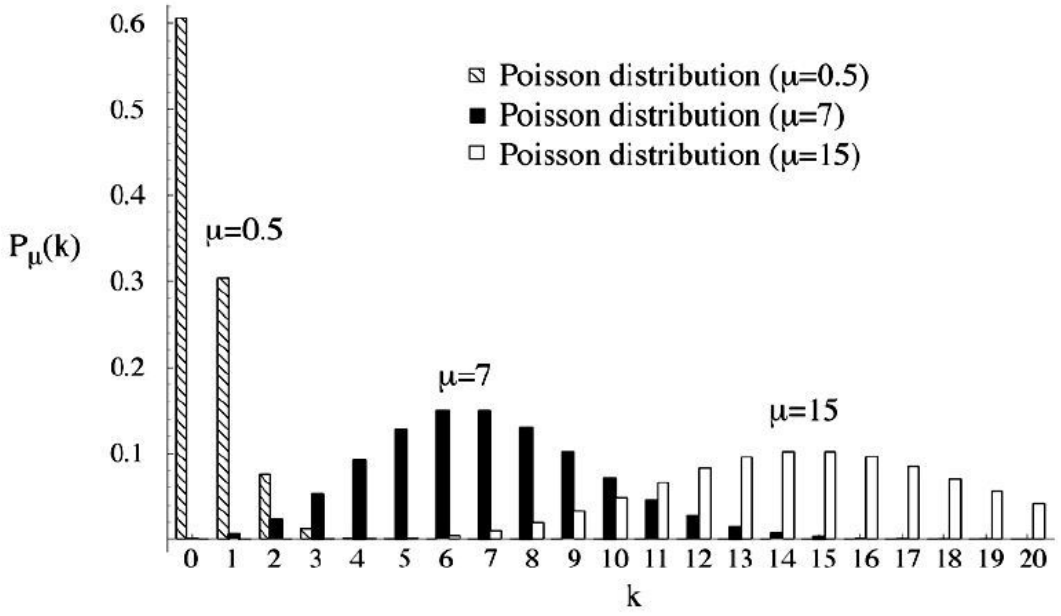
Poisson Process

The Poisson process generates a Poisson distribution.

Which is characterized by following three traits.

1. **Outcomes are discrete**
2. **The number of Success, in any interval are independent of success in any other interval**
3. **The probability of two or more successes over a sufficiently small interval is essentially zero.**

Poisson Distribution



Poisson Distribution

Example

In a study of drug-induced anaphylaxis among patients taking rocuronium bromide as part of their anesthesia, Laake and Røttingen (A-7) found that the occurrence of anaphylaxis followed a Poisson model with $\lambda = 12$ incidents per year in Norway. Find the probability that in the next year, among patients receiving rocuronium, exactly three will experience anaphylaxis.

Solution:

$$P(X = 3) = \frac{e^{-12}12^3}{3!} = .00177$$

Poisson Distribution

Example

What is the probability that at least three patients in the next year will experience anaphylaxis if rocuronium is administered with anesthesia?

Solution: We can use the concept of complementary events in this case. Since $P(X \leq 2)$ is the complement of $P(X \geq 3)$, we have

$$
\begin{aligned}
P(X \geq 3) &= 1 - P(X \leq 2) = 1 - [P(X = 0) + P(X = 1) + P(X = 2)] \\
&= 1 - \left[\frac{e^{-12}12^0}{0!} + \frac{e^{-12}12^1}{1!} + \frac{e^{-12}12^2}{2!}\right] \\
&= 1 - [.00000614 + .00007373 + .00044238] \\
&= 1 - .00052225 \\
&= .99947775
\end{aligned}
$$

Poisson Distribution

Poisson Approximation o the Binomial

Poisson Distribution is often used as an approximation for Binomial probabilities when:

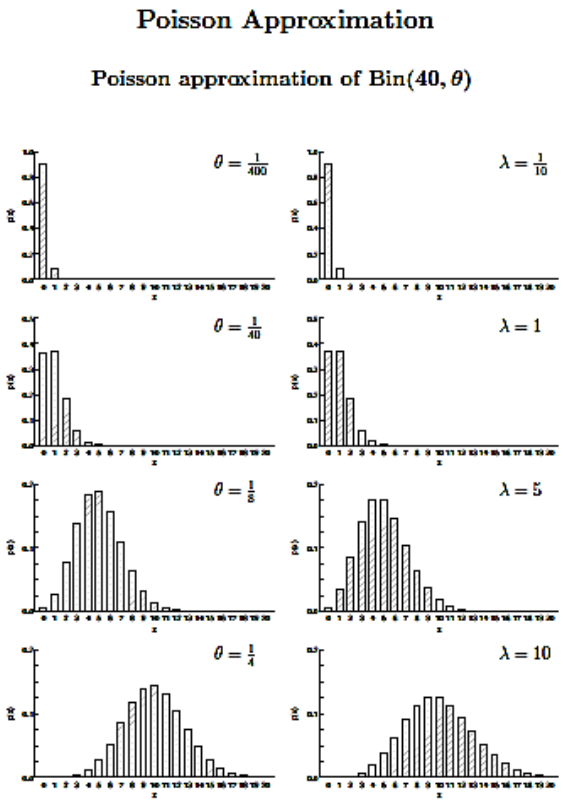
"n" is large and "p" is small

Rule of Thumb

$$n \geq 100$$

$$np \leq 10$$

Poisson Distribution



Poisson Distribution

END

The Poisson Distribution is important because it is often used for describing the behavior of rare events. (with small probabilities)

Lecture No 60

Sample and population (ASW, 15)

- A **population** is the collection of all the elements of interest.
- A **sample** is a subset of the population.
  - Good or bad samples.
  - Representative or non-representative samples. A researcher hopes to obtain a sample that represents the population, at least in the variables of interest for the issue being examined.
  - Probabilistic samples are samples selected using the principles of probability. This may allow a researcher to determine the sampling distribution of a sample statistic. If so, the researcher can determine the probability of any given sampling error and make statistical inferences about population characteristics.

Why sample?

- Time of researcher and those being surveyed.
- Cost to group or agency commissioning the survey.
- Confidentiality, anonymity, and other ethical issues.
- Non-interference with population. Large sample could alter the nature of population, eg. opinion surveys.
- Do not destroy population, eg. crash test only a small sample of automobiles.
- Cooperation of respondents – individuals, firms, administrative agencies.
- Partial data is all that is available, eg. fossils and historical records, climate change.

Methods of sampling – nonprobabilistic

- Friends, family, neighbours, acquaintances.
- Students in a class or co-workers in a workplace.
- Convenience (ASW, 286).
- Volunteers.
- Snowball sample.
- Judgment sample (ASW, 286).
- Quota sample – obtain a cross-section of a population, eg. by age and sex for individuals or by region, firm size, and industry for businesses. This may be reasonably representative.
- Sampling distribution of statistics cannot be obtained using any of the above methods, so statistical inference is not possible.

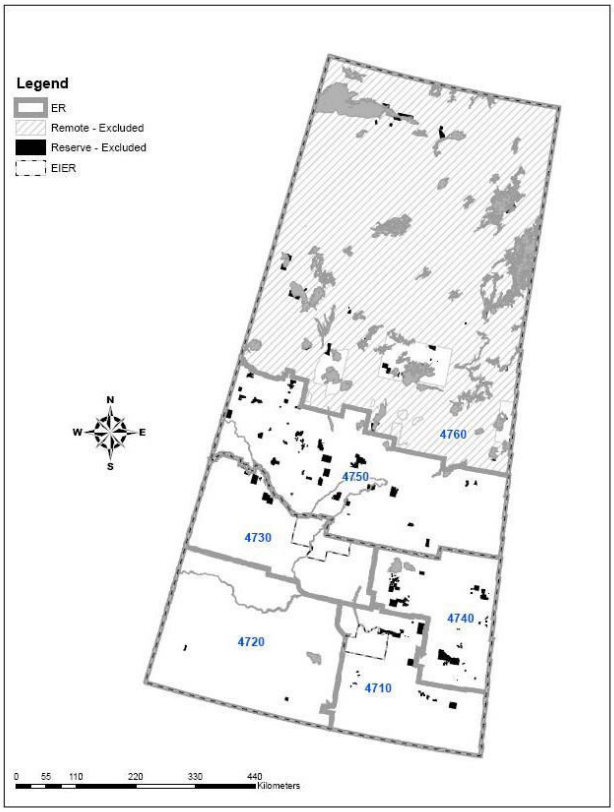
Methods of sampling – probabilistic

- Random sampling methods – each member has an equal probability of being selected.
- Systematic – every k^{th} case. Equivalent to random if patterns in list are unrelated to issues of interest. Eg. telephone book.
- Stratified samples – sample from each stratum or subgroup of a population. Eg. region, size of firm.
- Cluster samples – sample only certain clusters of members of a population. Eg. city blocks, firms.
- Multistage samples – combinations of random, systematic, stratified, and cluster sampling.
- If probability involved at each stage, then distribution of sample statistics can be obtained.

Map of Economic Regions in Saskatchewan for strata used in the monthly Labour Force Survey.

Source: Statistics Canada, catalogue number 71-526-X.

Clusters and individuals are selected from each of the 5 southern economic regions. In addition, the two CMAs of Regina and Saskatoon are strata. Note that the north of the province is treated as a remote region. Remote regions and Indian Reserves are not sampled in the Survey.



Some terms used in sampling

- **Sampled population** – population from which sample drawn (ASW, 258). Researcher should clearly define.
- **Frame** – list of elements that sample selected from (ASW, 258). Eg. telephone book, city business directory. May be able to construct a frame.
- **Parameter** – characteristics of a population (ASW, 259). Eg. total (annual GDP or exports), proportion *p* of population that votes Liberal in federal election. Also, μ or σ of a probability distribution are termed parameters.
- **Statistic** – numerical characteristics of a sample. Eg. monthly unemployment rate, pre-election polls.
- **Sampling distribution** of a statistic is the probability distribution of the statistic.

Selecting a sample (ASW, 259-261)

- *N* is the symbol given for the size of the population or the number of elements in the population.
- *n* is the symbol given for the size of the sample or the number of elements in the sample.
- **Simple random sample** is a sample of size *n* selected in a manner that each possible sample of size *n* has the same probability of being selected.
- In the case of a random sample of size *n* = 1, each element has the same chance of being selected.

Selecting a simple random sample

- **Sample with replacement** – after any element randomly selected, replace it and randomly select another element. But this could lead to the same element being selected more than once.
- More common to **sample without replacement**. Make sure that on each stage, each element remaining in the population has the same probability of being selected.
- Use a random number table or a computer generated random selection process. Or use a coin, die, or bingo ball popper, etc.

Simple random sample of size 2 from a population of 4 elements – without replacement

Population elements are A, B, C, D. N=4, n=2.

1st element selected could be any one of the 4 elements and this leaves 3, so there are 4 x 3 = 12 possible samples, each equally likely: AB, AC, AD, BA, BC, BD, CA, CB, CD, DA, DB, DC.

$$P_n^N = \frac{N!}{(N-n)!} = \frac{4!}{(4-2)!} = 12$$

If the order of selection does not matter (ie. we are interested only in what elements are selected), then this reduces to 6 combination. If {AB} is AB or BA, etc., then the equally likely random samples are {AB}, {AC}, {AD}, {BC}, {BD}, {CD}. This is the number of combinations (ASW, 261, note 1).

$$C_n^N = \frac{N!}{n!(N-n)!} = \frac{4!}{2!(4-2)!} = 6$$

First N = 18 companies on US 200 list

1. 3M
2. Abbott
3. Adobe
4. Aetna
5. Aflac
6. Air products
7. Alcoa
8. Allergan
9. Allstate
10. Alfria
11. Amazon
12. American Electric
13. American Express
14. American Tower
15. Amgen
16. Andarko
17. Anheuser Busch
18. Apache

Using random number table

Part of Table 7.1:

71744 51102 15141

95436 79115 08303

Suppose you were asked to select a simple random sample of size n =5.

Since 18 cases, two digits required and, in order, these are: 71 74 45 11 02 15 14 19 54 36 79 11 50 83 03.

Select cases 11, 2, 15, 14, and 3.

Keep track of where you last used the table and begin the next selection at that point.

Using Excel(ASW, 292)

- Suppose the data are in rows 2 through 46 in columns A through H.
- To arrange the rows in random order
  - Enter =RAND() in H2
  - Copy cell H2 to cells H3:H46 and each cell has a random number assigned – these later change
  - Select any cell in H
  - For Excel 2003, click **Data**, then **Sort**, and **Sort by Ascending**.
  - For Excel 2007, on the **Home** tab, in the **Editing** group, click **Sort and Filter** and **Sort Smallest to Largest.**
- The rows are now in random order. For a random sample of size *n*, select the data in the first *n* rows.

Sampling from a process (ASW, 261)

- It my be difficult or impossible or to obtain or construct a frame.
  - Larger or potentially infinite population – fish, trees, manufacturing processes.
  - Continuous processes – production of milk or other liquids, transporting commodities to a warehouse.
- Random sample is one where any element selected in the sample:
  - Is selected independently of any other element.
  - Follows the same probability distribution as the elements in the population.
- Careful design for sample is especially important.
  - Sample production of milk at random times.
  - Forest products – randomly select clusters from maps or previous surveys of tree types, size, etc.

Point Estimation (ASW, 263)

Measure	Parameter	Statistic or point estimator	Sampling error
Mean	μ	$\bar{x}$	$\lvert\bar{x} - \mu\rvert$
Standard deviation	σ	s	$\lvert s - \sigma\rvert$
Proportion	*p*	$\bar{p}$	$\lvert\bar{p} - p\rvert$
No. of elements	*N*	*n*	

The proportion is the frequency of occurrence of a characteristic divided by the total number of elements. The proportion of elements of a population that take on the characteristic is *p* and the proportion of the elements in the sample selected with this same characteristic is $\bar{p}$.

Terms for estimation

- **Parameters** are characteristics of a **population** or, more specifically, a **target population** (ASW, 265). Parameters may also be termed **population values**.
- A **statistic** is also referred to as a **sample statistic** or, when estimating a parameter, a **point estimator** of a parameter. A specific value of a point estimator is referred to as a **point estimate** of a parameter.
- The **sampling error** is the difference between the point estimate (value of the estimator) and the value of the parameter. This is the error caused by sampling only a subset of elements of a population, rather than all elements in a population. A researcher hopes to minimize the sampling error, but all samples have some such error associated with them.

Percentage of respondents, votes, and number of seats by party, November 5, 2003 Saskatchewan provincial election

Political Party	CBC Poll, Oct. 20-26 $\bar{P}$	Cutler Poll, Oct. 29 – Nov. 5 $\bar{P}$	Election Result P	Number of Seats
NDP	42%	47%	44.5%	30
Saskatchewan Party	39%	37%	39.4%	28
Liberal	18%	14%	14.2%	0
Other	1%	2%	1.9%	0
Total	100%	100%	100.0%	58
Undecided	15%	16%		
Sample size *(n)*	800	773		

Sources: CBC Poll results from Western Opinion Research, "Saskatchewan Election Survey for The Canadian Broadcasting Corporation," October 27, 2003. Obtained from web site. http://sask.cbc.ca/regional/servlet/View?filename=poll_one031028, November 7, 2003. Cutler poll results provided by Fred Cutler and from the *Leader-Post*, November 7, 2003, p. A5.

Sampling error in Saskatchewan polls

The actual results from the election are provided in the last two columns, with the second last column giving the parameters for the population. These are percentages, rather than proportions, so I have labelled them as upper case P. The second and third columns provide statistics on point estimators $\bar{P}$ of P from two different polls. For any party, the difference between these two provides a measure of the sampling error.

For example, the Cutler Poll has a sampling error of only 0.2 percentage points for the Liberals, but a sampling error of 2.4 percentage points for the Saskatchewan Party.

Sampling distributions

- A **sampling distribution** is the probability distribution for all possible values of the sample statistic.
- Each sample contains different elements so the value of the sample statistic differs for each sample selected. These statistics provide different estimates of the parameter. The sampling distribution describes how these different values are distributed.
- For the most part, we will work with the sampling distribution of the sample mean. With the sampling distribution of $\bar{x}$, we can "make probability statements about how close the sample mean is to the population mean μ" (ASW, 267). Alternatively, it provides a way of determining the probability of various levels of sampling error.

Sampling distribution of the sample mean

- When a sample is selected, the sampling method may allow the researcher to determine the sampling distribution of the sample mean $\bar{x}$. The researcher hopes that the mean of the sampling distribution will be μ, the mean of the population. If this occurs, then the expected value of the statistic $\bar{x}$ is μ. This characteristic of the sample mean is that of being an unbiased estimator of μ. In this case, $E(\bar{x}) = \mu$

- If the variance of the sampling distribution can be determined, then the researcher is able to determine how variable $\bar{x}$ is when there are repeated samples. The researcher hopes to have a small variability for the sample means, so most estimates of μ are close to μ.

Sampling distribution of the sample mean when random sampling

- If a simple random sample is drawn from a normally distributed population, the sampling distribution of $\bar{x}$ is normally distributed (ASW, 269).
- The mean of the distribution of $\bar{x}$ is μ, the population mean.
- If the sample size *n* is a reasonably small proportion of the population size, then the standard deviation of $\bar{x}$ is the population standard deviation σ divided by the square root of the sample size. That is, samples that contain, say, less than 5% of the population elements, the **finite population correction factor** is not required since it does not alter results much (ASW, 270).

Random sample from a normally distributed population

	Normally distributed population	Sampling distribution of $\bar{x}$ when sample is random
No. of elements	*N*	*n*
Mean	μ	μ
Standard deviation	σ	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

Note: If $n/N > 0.05$, it may be best to use the finite population correction factor (ASW, 270).

Central limit theorem – CLT (ASW, 271)

The sampling distribution of the sample mean, $\bar{x}$, is approximated by a normal distribution when the sample is a simple random sample and the sample size, *n*, is large.

In this case, the mean of the sampling distribution is the population mean, μ, and the standard deviation of the sampling distribution is the population standard deviation, σ, divided by the square root of the sample size. The latter is referred to as the **standard error** of the mean.

A sample size of 100 or more elements is generally considered sufficient to permit using the CLT. If the population from which the sample is drawn is symmetrically distributed, *n* > 30 may be sufficient to use the CLT.

Large random sample from any population

	Any population	Sampling distribution of $\bar{x}$ when sample is random
No. of elements	*N*	*n*
Mean	μ	μ
Standard deviation	σ	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

A sample size n of greater than 100 is generally considered sufficiently large to use.

Simulation example

- 192 random samples from population that is not normally distributed.
- Sample size of *n* = 50 for each of the random samples.

- Handouts in Monday's class provide these results.

Sampling distribution in theory and practice

- Population mean μ = 2352 and standard deviation σ = 1485.
- Random sample of size *n = 50.*
- Sample mean $\bar{x}$ is normally distributed with a mean of μ = 2352 and a standard deviation, or standard error, of

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1485}{\sqrt{50}} = \frac{1485}{7.071} = 210$$

In the simulation, the mean of the 192 random samples is 2337 and the standard deviation is 206.

Correlation & Regression

Lecture 66

Dr. Moataza Mahmoud Abdel Wahab
Lecturer of Biostatistics
High Institute of Public Health
University of Alexandria

Important Terms

- **Sporadic**: disease occurs occasionally, irregularly
- **Endemic**: disease stays in population at low frequency
- **Epidemic**: sudden outbreak in disease above typical level
- **Pandemic**: epidemic over wide area (may be entire world).
- **Morbidity**: all reported cases of disease, illness, and disability
- **Mortality**: reported deaths due to a disease

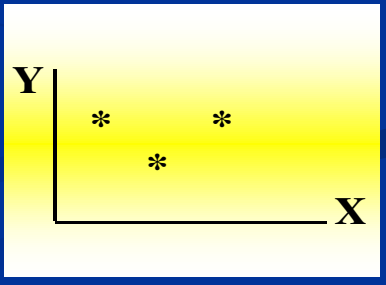
Correlation

Finding the relationship between two quantitative variables without being able to infer causal relationships

Correlation is a statistical technique used to determine the degree to which two variables are related

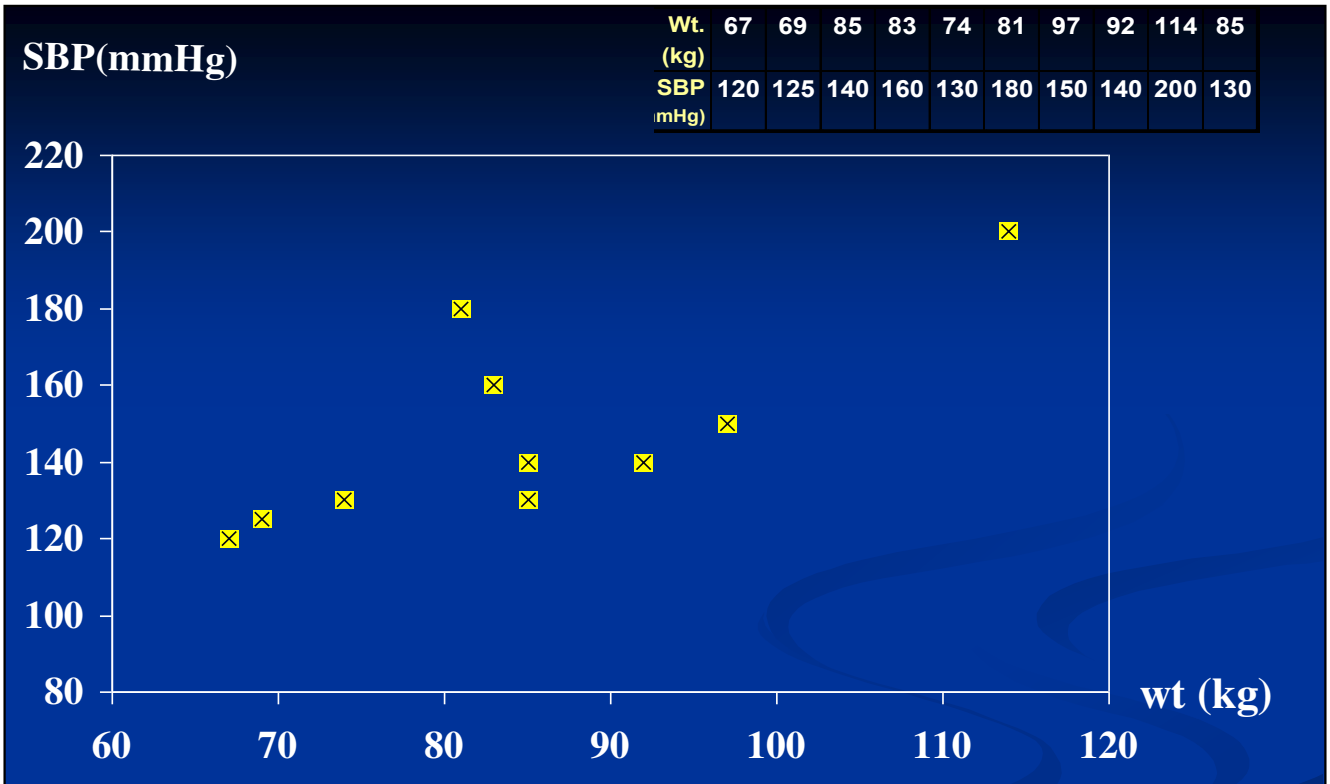
Scatter diagram

- Rectangular coordinate
- Two quantitative variables
- One variable is called independent (X) and the second is called dependent (Y)
- Points are not joined
- No frequency table

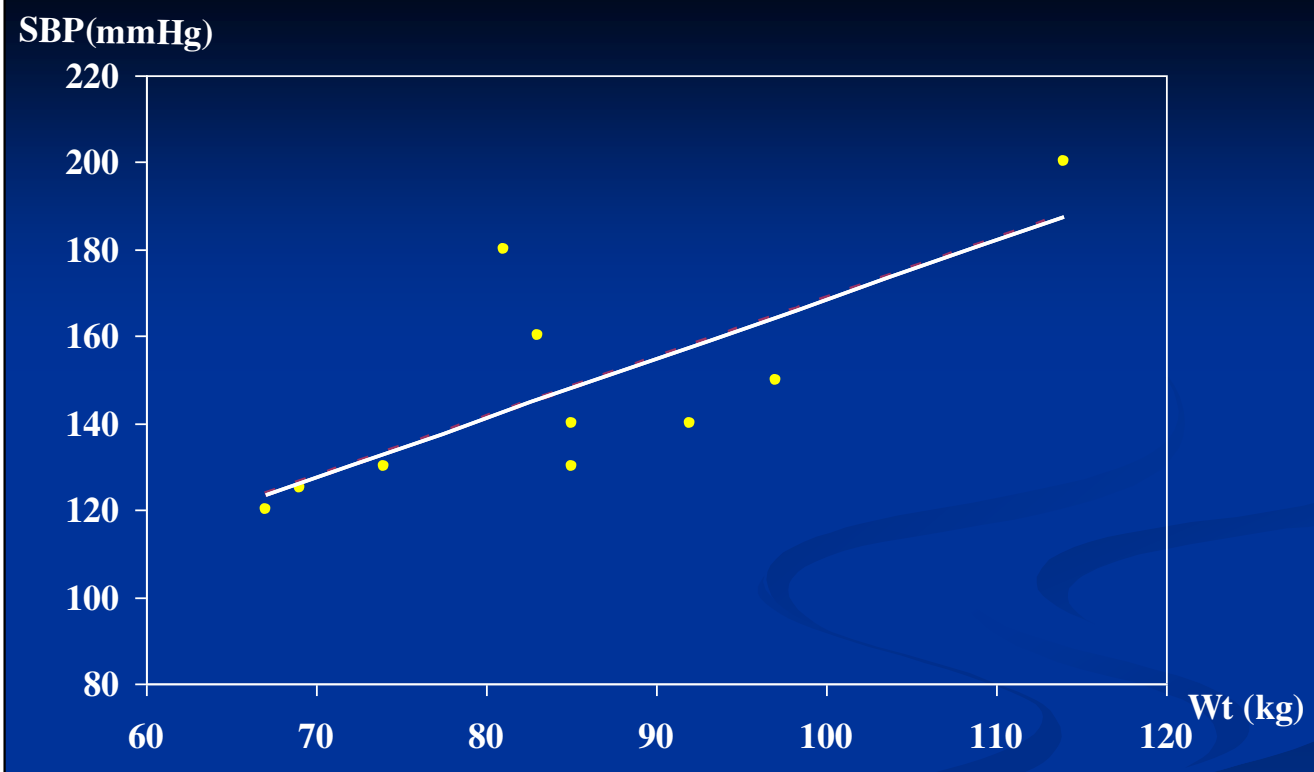


Example

Wt. (kg)	67	69	85	83	74	81	97	92	114	85
SBP mHg)	120	125	140	160	130	180	150	140	200	130



Scatter diagram of weight and systolic blood pressure



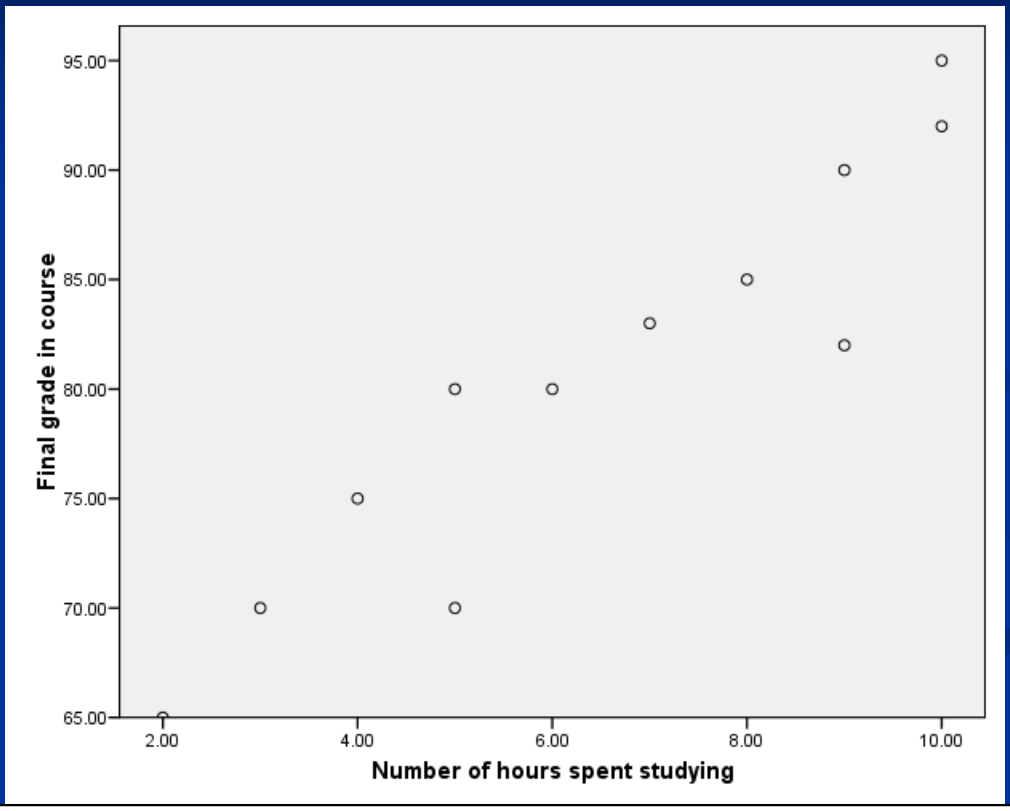
Scatter diagram of weight and systolic blood pressure

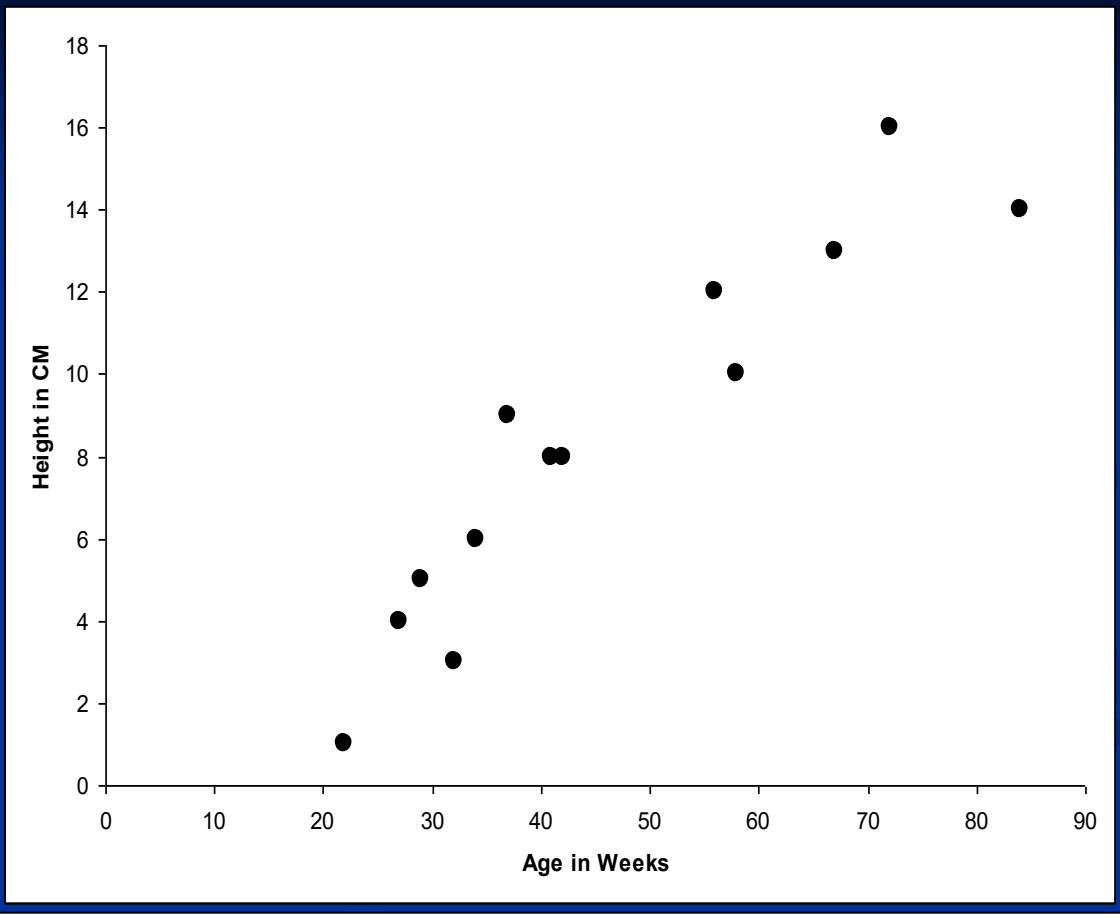
Scatter plots

The pattern of data is indicative of the type of relationship between your two variables:

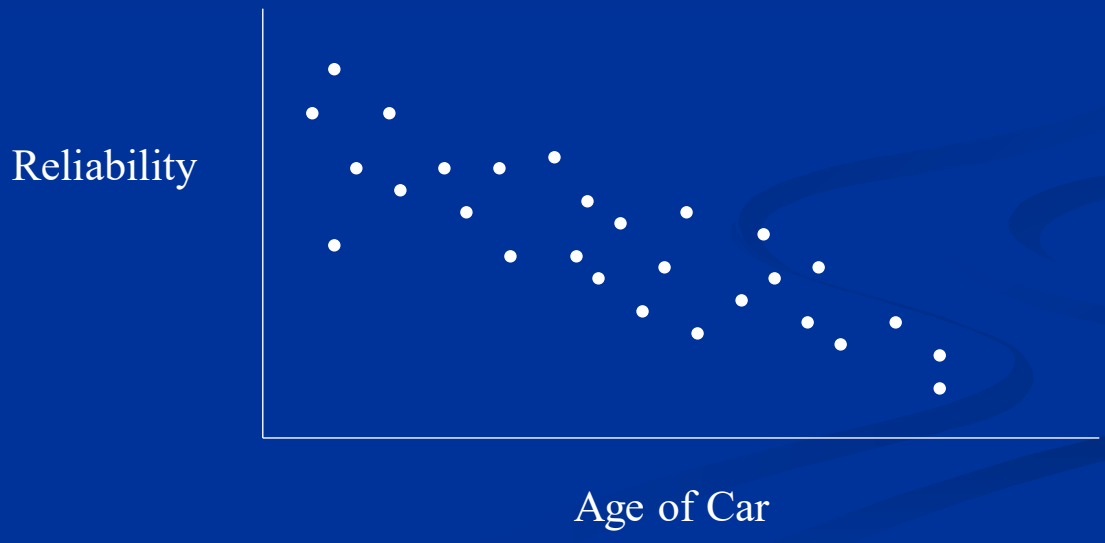
- positive relationship
- negative relationship
- no relationship

Positive relationship

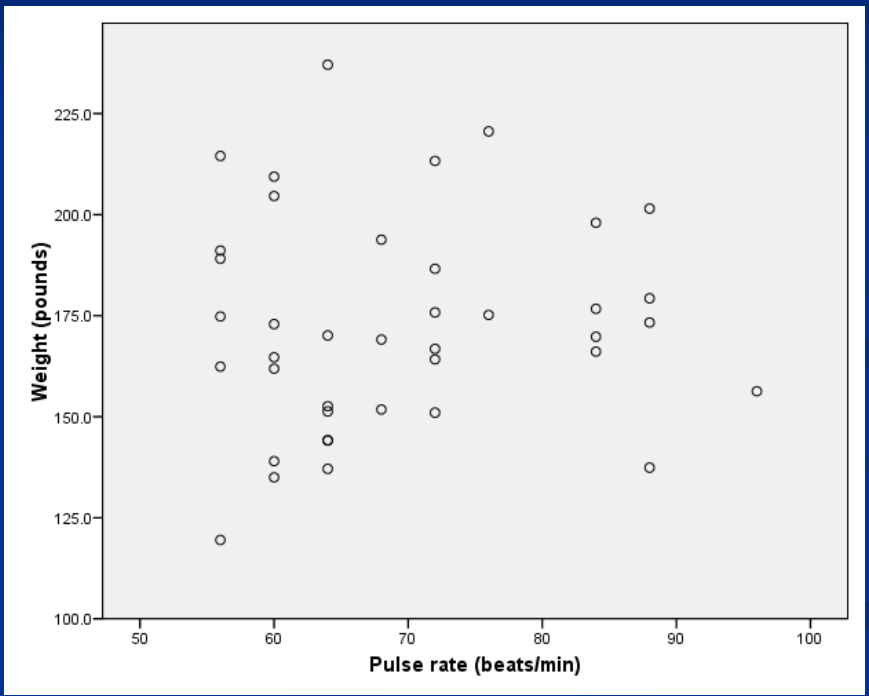




Negative relationship



No relation



Correlation Coefficient

Statistic showing the degree of relation between two variables

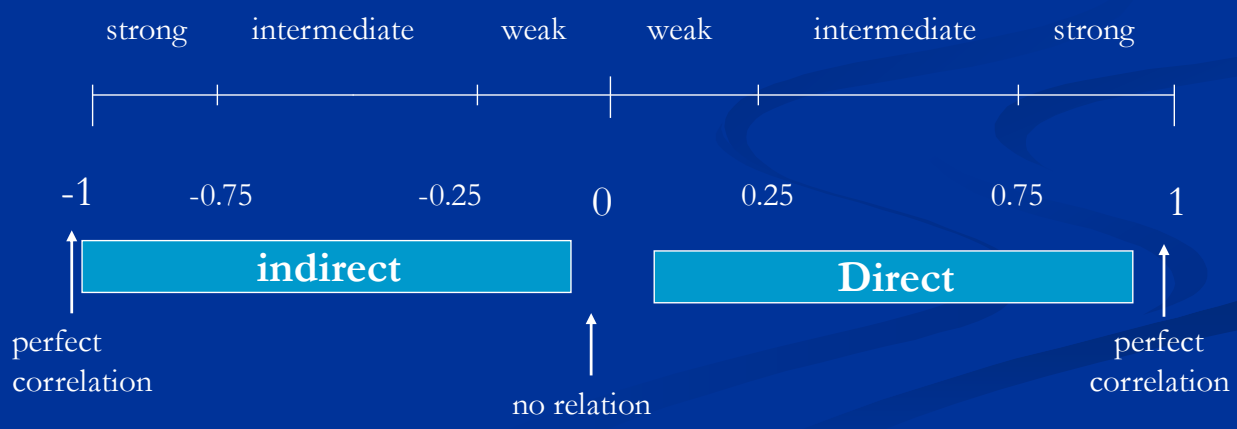
Simple Correlation coefficient (r)

- It is also called Pearson's correlation or product moment correlation coefficient.
- It measures the nature and strength between two variables of the quantitative type.

- The sign of r denotes the nature of association
- while the value of r denotes the strength of association.

- If the sign is +ve this means the relation is direct (an increase in one variable is associated with an increase in the other variable and a decrease in one variable is associated with a decrease in the other variable).
- While if the sign is -ve this means an inverse or indirect relationship (which means an increase in one variable is associated with a decrease in the other).

- The value of r ranges between (-1) and (+1)
- The value of r denotes the strength of the association as illustrated by the following diagram.



- If r = Zero this means no association or correlation between the two variables.
- If 0 < r < 0.25 = weak correlation.
- If 0.25 ≤ r < 0.75 = intermediate correlation.
- If 0.75 ≤ r < 1 = strong correlation.
- If r = I = perfect correlation.

How to compute the simple correlation coefficient (r)

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left(\sum x^2 - \frac{(\sum x)^2}{n}\right)\left(\sum y^2 - \frac{(\sum y)^2}{n}\right)}}$$

Example:

A sample of 6 children was selected, data about their age in years and weight in kilograms was recorded as shown in the following table . It is required to find the correlation between age and weight.

serial No	Age (years)	Weight (Kg)
1	7	12
2	6	8
3	8	12
4	5	10
5	6	11
6	9	13

These 2 variables are of the quantitative type, one variable (Age) is called the independent and denoted as (X) variable and the other (weight) is called the dependent and denoted as (Y) variables to find the relation between age and weight compute the simple correlation coefficient using the following formula:

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left(\sum x^2 - \frac{(\sum x)^2}{n}\right)\left(\sum y^2 - \frac{(\sum y)^2}{n}\right)}}$$

Serial n.	Age (years) (x)	Weight (Kg) (y)	xy	X^2	Y^2
1	7	12	84	49	144
2	6	8	48	36	64
3	8	12	96	64	144
4	5	10	50	25	100
5	6	11	66	36	121
6	9	13	117	81	169
Total	∑x= 41	∑y= 66	∑xy= 461	∑x2= 291	∑y2= 742

$$r = \frac{461 - \frac{41 \times 66}{6}}{\sqrt{\left[291 - \frac{(41)^2}{6}\right].\left[742 - \frac{(66)^2}{6}\right]}}$$

r = 0.759

strong direct correlation

EXAMPLE: Relationship between Anxiety and Test Scores

Anxiety (X)	Test score (Y)	X^2	Y^2	XY
10	2	100	4	20
8	3	64	9	24
2	9	4	81	18
1	7	1	49	7
5	6	25	36	30
6	5	36	25	30
$\sum X = 32$	$\sum Y = 32$	$\sum X^2 = 230$	$\sum Y^2 = 204$	$\sum XY=129$

Calculating Correlation Coefficient

$$r = \frac{(6)(129) - (32)(32)}{\sqrt{\left(6(230) - 32^2\right)\left(6(204) - 32^2\right)}} = \frac{774 - 1024}{\sqrt{(356)(200)}} = -.94$$

r = - 0.94

Indirect strong correlation

Spearman Rank Correlation Coefficient (r_s)

- It is a non-parametric measure of correlation.
- This procedure makes use of the two sets of ranks that may be assigned to the sample values of x and Y.
- Spearman Rank correlation coefficient could be computed in the following cases:
- Both variables are quantitative.
- Both variables are qualitative ordinal.
- One variable is quantitative and the other is qualitative ordinal.

Procedure:

1. Rank the values of X from 1 to n where n is the numbers of pairs of values of X and Y in the sample.
2. Rank the values of Y from 1 to n.
3. Compute the value of di for each pair of observation by subtracting the rank of Yi from the rank of Xi
4. Square each di and compute ∑di2 which is the sum of the squared values.

5. Apply the following formula

$$r_s = 1 - \frac{6\sum(di)^2}{n(n^2 - 1)}$$

- The value of r_s denotes the magnitude and nature of association giving the same interpretation as simple r.

Example

In a study of the relationship between level education and income the following data was obtained. Find the relationship between them and comment.

sample numbers	level education (X)	Income (Y)
A	Preparatory.	25
B	Primary.	10
C	University.	8
D	secondary	10
E	secondary	15
F	illiterate	50
G	University.	60

Answer:

	(X)	(Y)	Rank X	Rank Y	di	di^2
A	Preparatory	25	5	3	2	4
B	Primary.	10	6	5.5	0.5	0.25
C	University.	8	1.5	7	-5.5	30.25
D	secondary	10	3.5	5.5	-2	4
E	secondary	15	3.5	4	-0.5	0.25
F	illiterate	50	7	2	5	25
G	university.	60	1.5	1	0.5	0.25

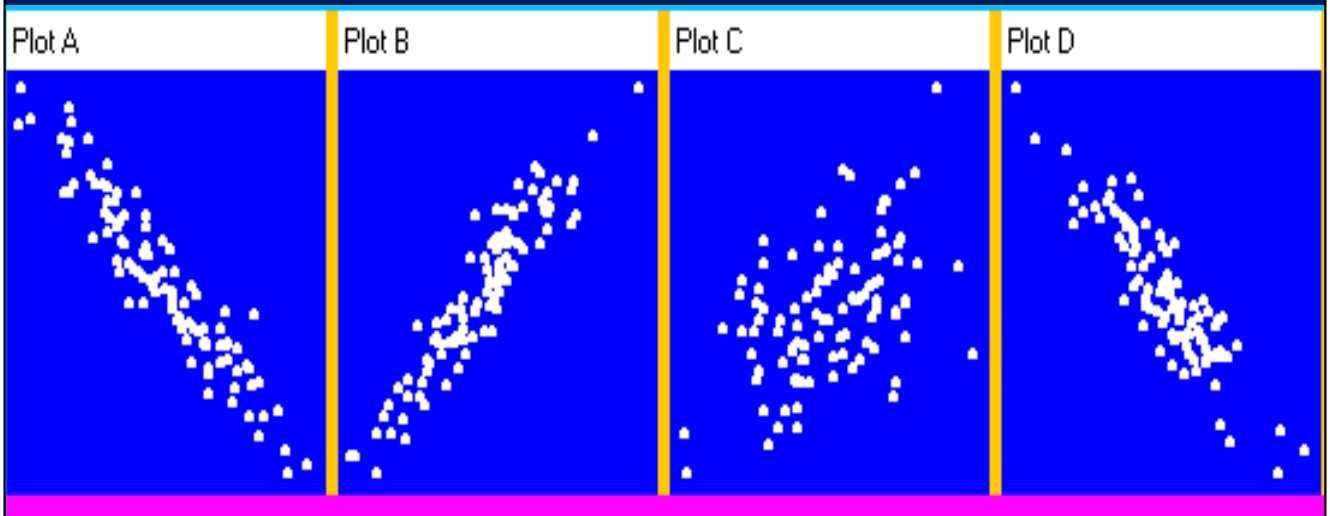
$$\sum di^2 = 64$$

$$r_s = 1 - \frac{6 \times 64}{7(48)} = -0.1$$

Comment:

There is an indirect weak correlation between level of education and income.

exercise



r = -0.95	A	B	C	D
r = -0.9	A	B	C	D
r = 0.49	A	B	C	D
r = 0.93	A	B	C	D

Regression Analyses

- Regression: technique concerned with predicting some variables by knowing others
- The process of predicting variable Y using variable X

Regression

- Uses a variable (x) to predict some outcome variable (y)
- Tells you how values in y change as a function of changes in values of x

Correlation and Regression

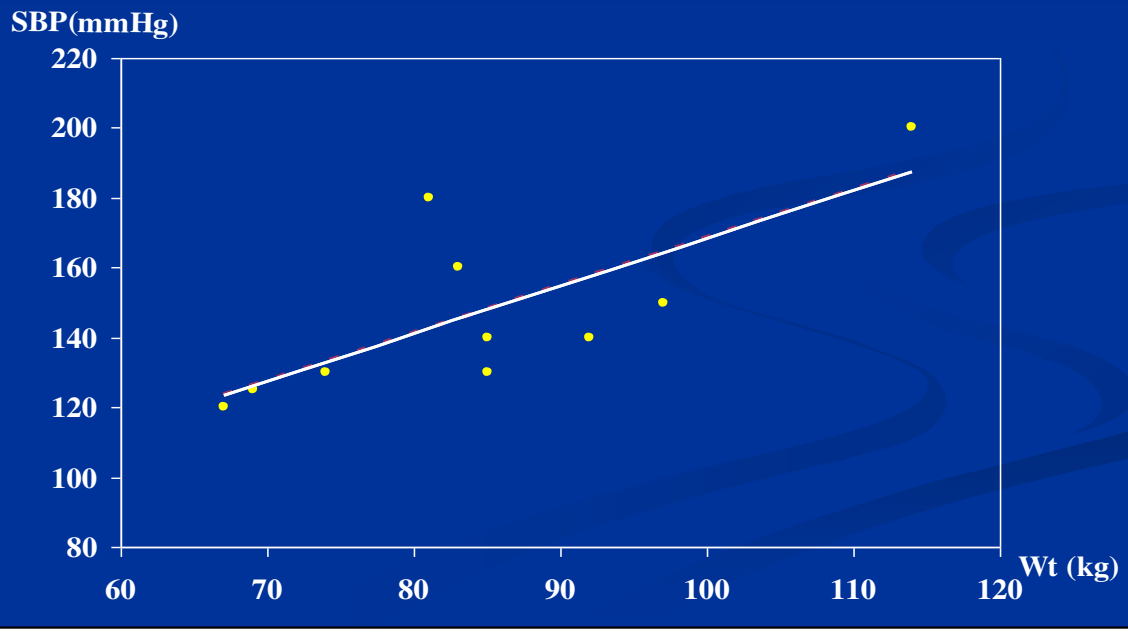
- Correlation describes the strength of a **linear** relationship between two variables
- **Linear** means "**straight line**"
- **Regression** tells us how to draw the straight line described by the correlation

Regression

- Calculates the "best-fit" line for a certain set of data

The regression line makes the sum of the squares of the residuals smaller than for any other line

Regression minimizes residuals



By using the least squares method (a procedure that minimizes the vertical deviations of plotted points surrounding a straight line) we are able to construct a best fitting straight line to the scatter diagram points and then formulate a regression equation in the form of:

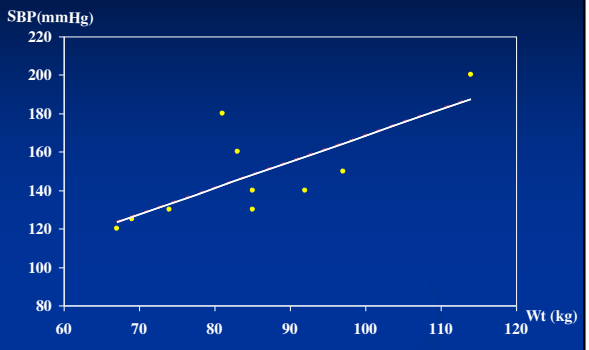
$$\hat{y} = a + bX$$

$$\hat{y} = \bar{y} + b(x - \bar{x})$$

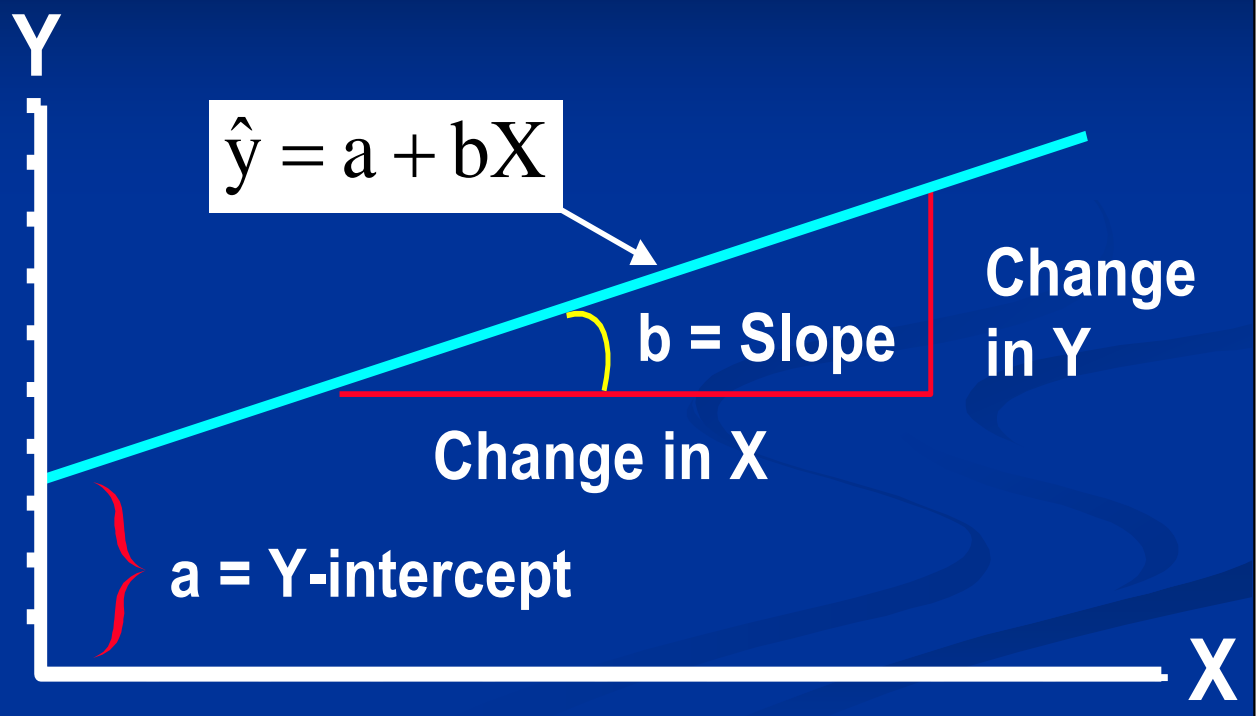
$$b = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

Regression Equation

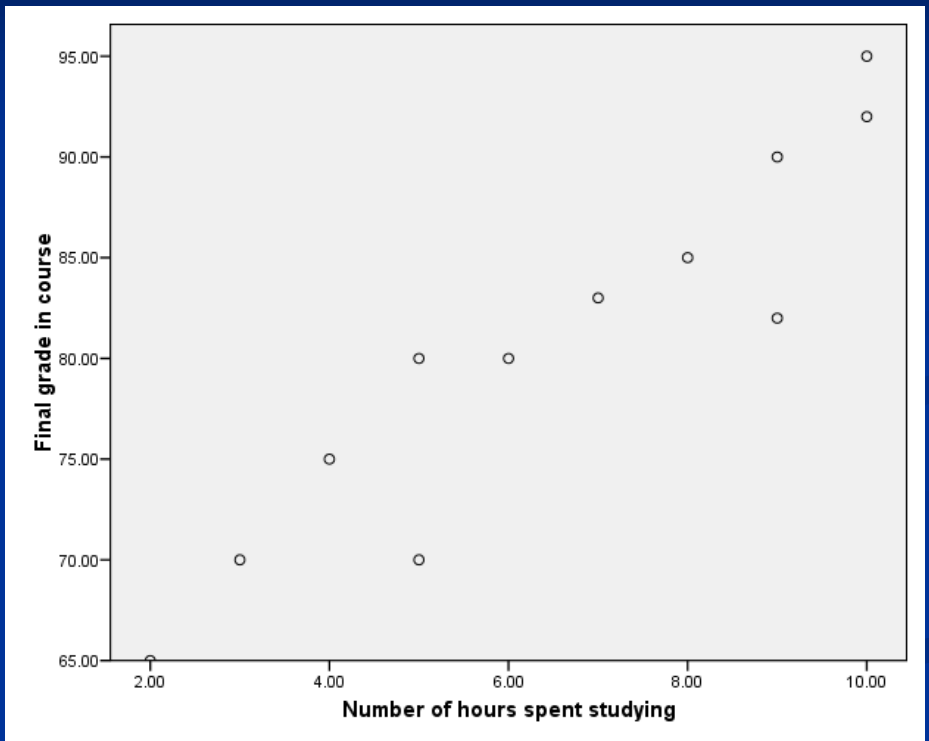
- Regression equation describes the regression line mathematically
  - Intercept
  - Slope



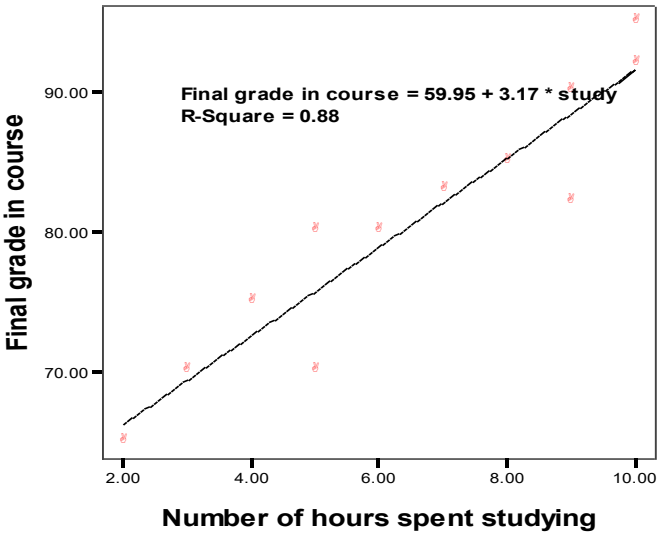
Linear Equations



Hours studying and grades



Regressing grades on hours



Predicted final grade in class =

59.95 + 3.17*(number of hours you study per week)

Predicted final grade in class = 59.95 + 3.17*(hours of study)

Predict the final grade of…

- Someone who studies for 12 hours
- Final grade = 59.95 + (3.17*12)
- Final grade = 97.99

- Someone who studies for 1 hour:
- Final grade = 59.95 + (3.17*1)
- Final grade = 63.12

Exercise

A sample of 6 persons was selected the value of their age (x variable) and their weight is demonstrated in the following table. Find the regression equation and what is the predicted weight when age is 8.5 years.

Serial no.	Age (x)	Weight (y)
1	7	12
2	6	8
3	8	12
4	5	10
5	6	11
6	9	13

Answer

Serial no.	Age (x)	Weight (y)	xy	X^2	Y^2
1	7	12	84	49	144
2	6	8	48	36	64
3	8	12	96	64	144
4	5	10	50	25	100
5	6	11	66	36	121
6	9	13	117	81	169
Total	41	66	461	291	742

$$\bar{x} = \frac{41}{6} = 6.83$$

$$\bar{y} = \frac{66}{6} = 11$$

$$b = \frac{461 - \frac{41 \times 66}{6}}{291 - \frac{(41)^2}{6}} = 0.92$$

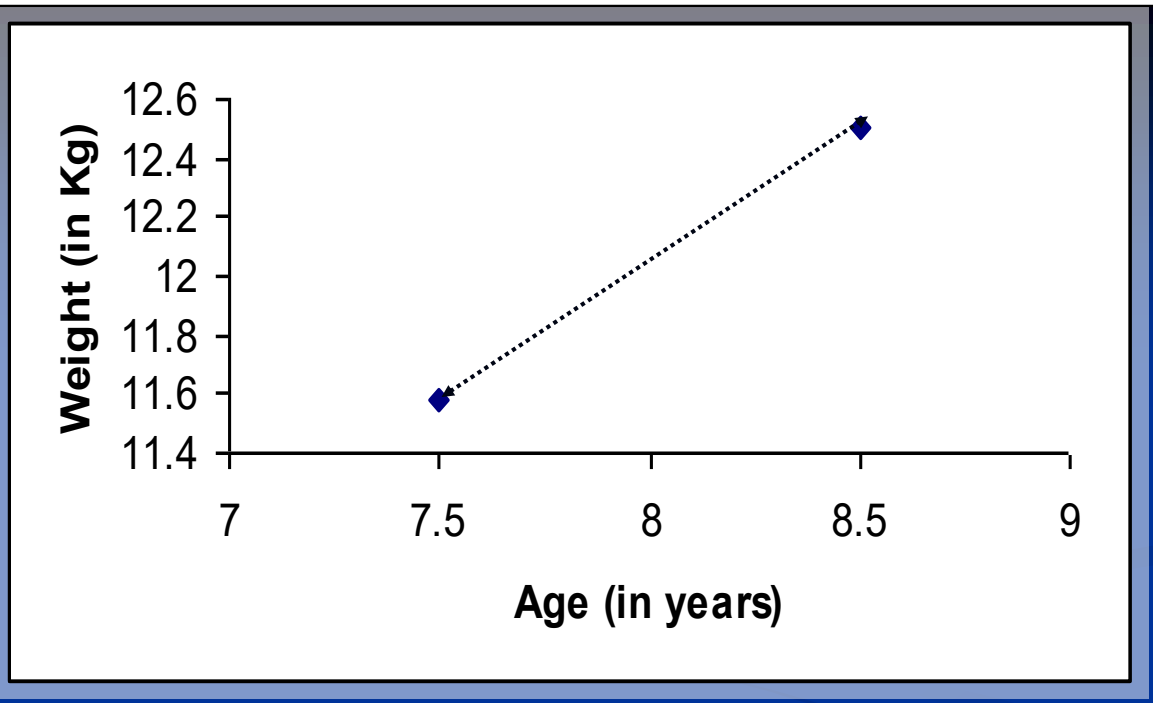
Regression equation

$$\hat{y}_{(x)} = 11 + 0.9(x - 6.83)$$

$$\hat{y}_{(x)} = 4.675 + 0.92x$$

$$\hat{y}_{(8.5)} = 4.675 + 0.92 * 8.5 = 12.50\text{Kg}$$

$$\hat{y}_{(7.5)} = 4.675 + 0.92 * 7.5 = 11.58\text{Kg}$$



we create a regression line by plotting two estimated values for y against their X component, then extending the line right and left.

Exercise 2

The following are the age (in years) and systolic blood pressure of 20 apparently healthy adults.

Age (x)	B.P (y)	Age (x)	B.P (y)
20	120	46	128
43	128	53	136
63	141	60	146
26	126	20	124
53	134	63	143
31	128	43	130
58	136	26	124
46	132	19	121
58	140	31	126
70	144	23	123

- Find the correlation between age and blood pressure using simple and Spearman's correlation coefficients, and comment.
- Find the regression equation?
- What is the predicted blood pressure for a man aging 25 years?

Serial	x	y	xy	x2
1	20	120	2400	400
2	43	128	5504	1849
3	63	141	8883	3969
4	26	126	3276	676
5	53	134	7102	2809
6	31	128	3968	961
7	58	136	7888	3364
8	46	132	6072	2116
9	58	140	8120	3364
10	70	144	10080	4900

Serial	x	y	xy	x2
11	46	128	5888	2116
12	53	136	7208	2809
13	60	146	8760	3600
14	20	124	2480	400
15	63	143	9009	3969
16	43	130	5590	1849
17	26	124	3224	676
18	19	121	2299	361
19	31	126	3906	961
20	23	123	2829	529
Total	852	2630	114486	41678

$$b_1 = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} = \frac{114486 - \frac{852 \times 2630}{20}}{41678 - \frac{852^2}{20}} = 0.4547$$

$\hat{y}$ =112.13 + 0.4547 x

for age 25

B.P = 112.13 + 0.4547 * 25=123.49 = 123.5 mm hg

Multiple Regression

Multiple regression analysis is a straightforward extension of simple regression analysis which allows more than one independent variable.

Introduction: What is SPSS?

Lecture no 76

- Originally it is an acronym of Statistical Package for the Social Science but now it stands for Statistical Product and Service Solutions

- One of the most popular statistical packages which can perform highly complex data manipulation and analysis with simple instructions

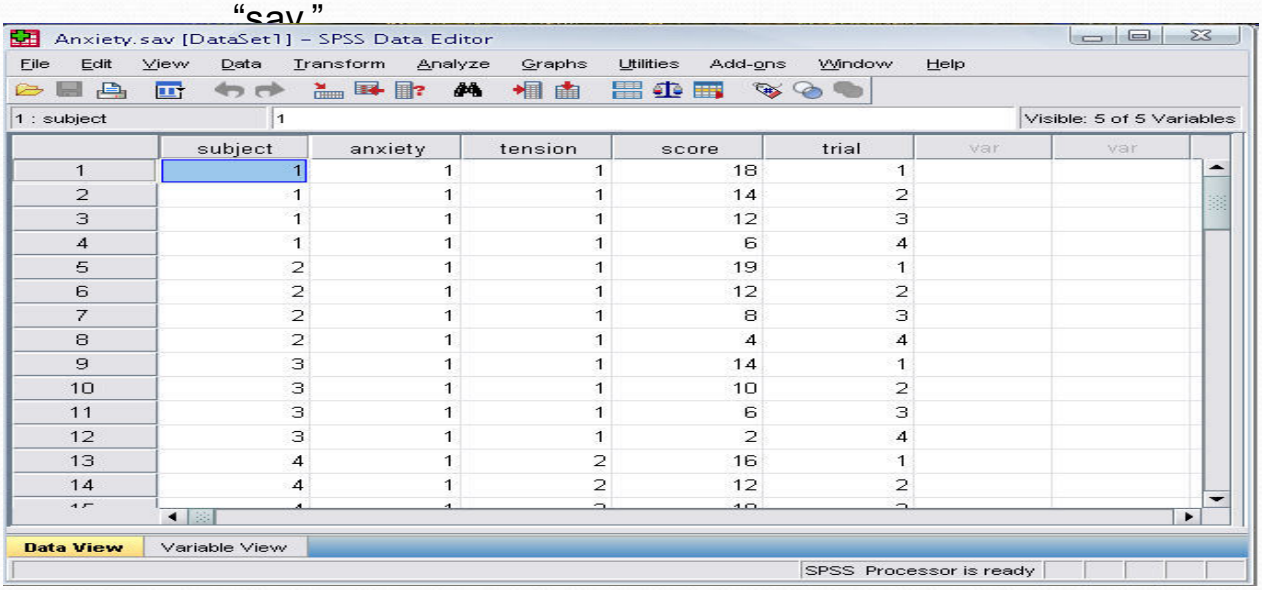
The Four Windows:

Data editor
Output viewer
Syntax editor
Script window

The Four Windows: Data Editor

- Data Editor

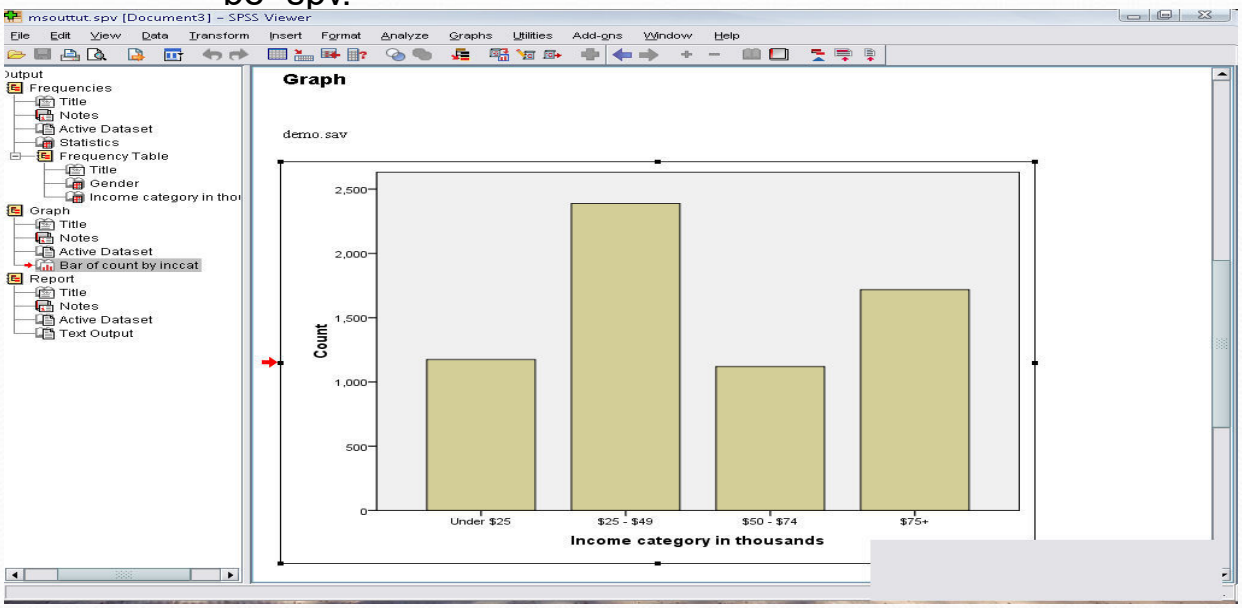
  Spreadsheet-like system for defining, entering, editing, and displaying data. Extension of the saved file will be "sav."



The Four Windows: Output Viewer

- Output Viewer

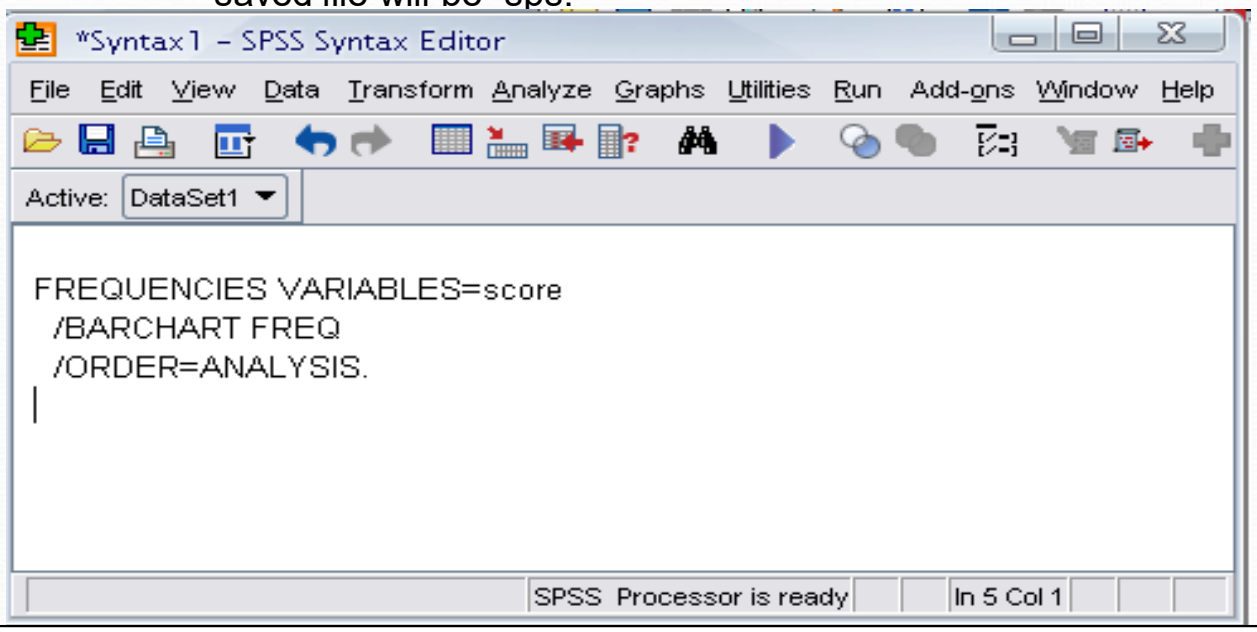
  Displays output and errors. Extension of the saved file will be "spv."



The Four Windows: Syntax editor

- Syntax Editor

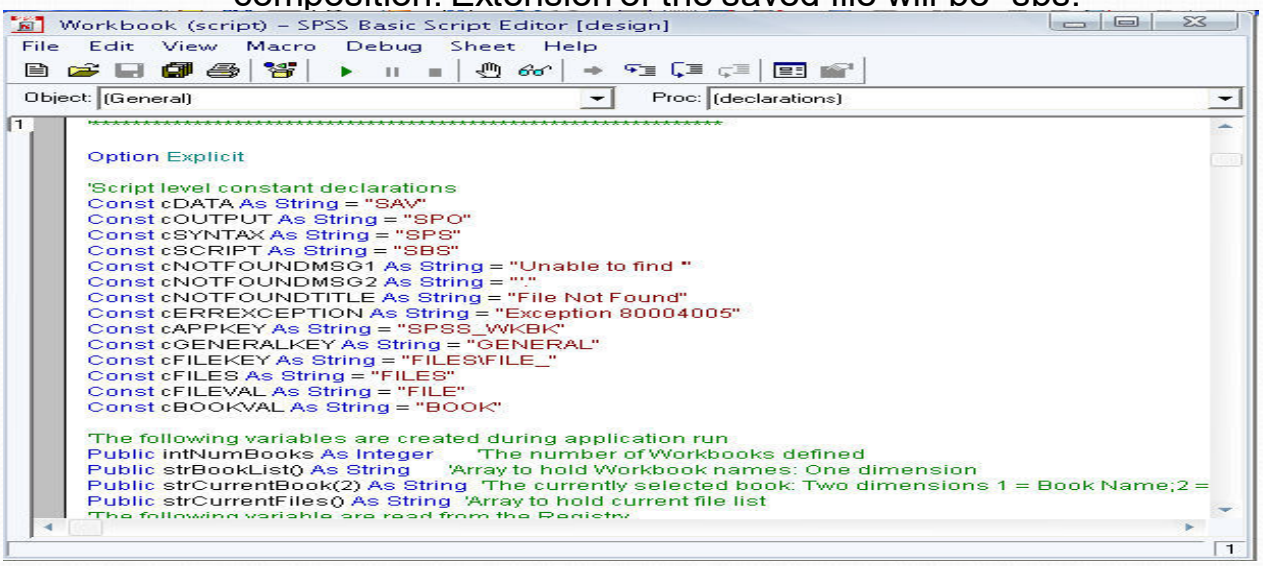
  Text editor for syntax composition. Extension of the saved file will be “sps.”



The Four Windows: Script Window

- Script Window

  Provides the opportunity to write full-blown programs, in a BASIC-like language. Text editor for syntax composition. Extension of the saved file will be “sbs.”



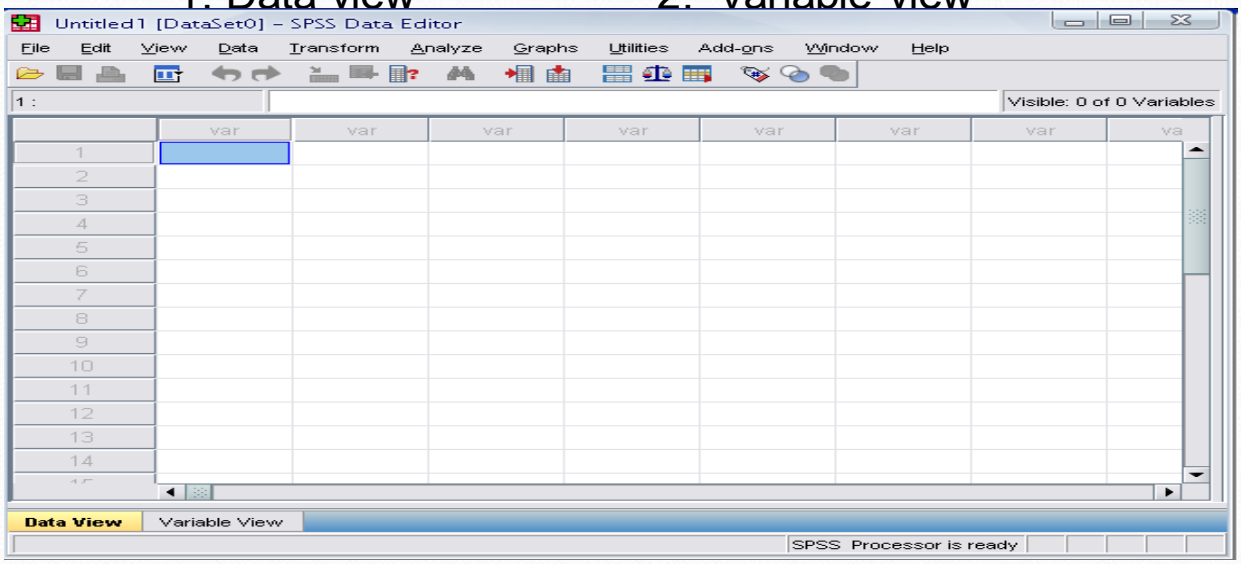
The basics of managing data files

Opening SPSS

- Start → All Programs → SPSS Inc→ SPSS 16.0 → SPSS 16.0

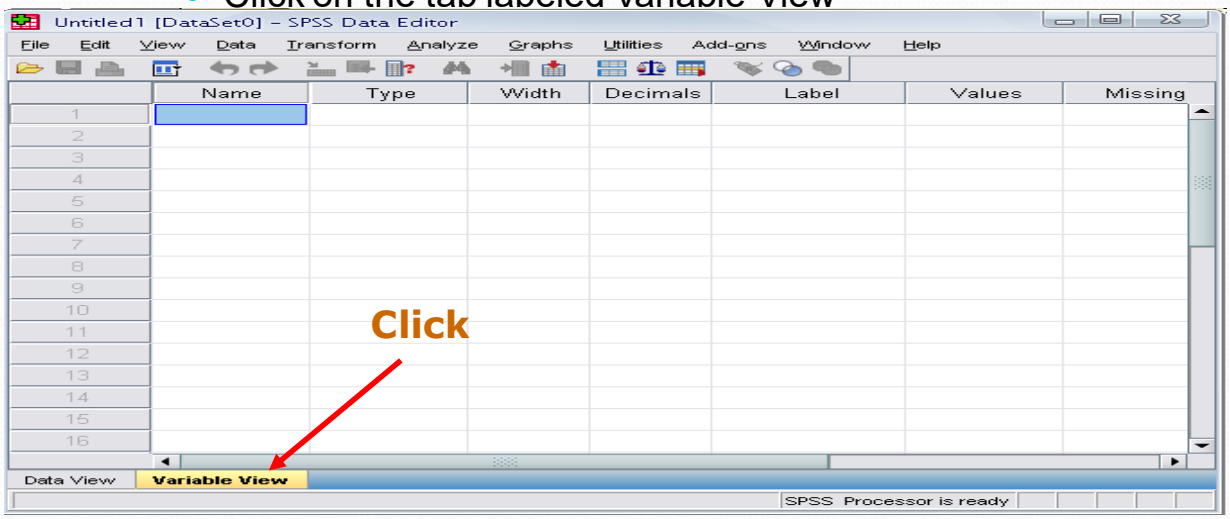
Opening SPSS

- The default window will have the data editor
- There are two sheets in the window:
  1. Data view
  2. Variable view



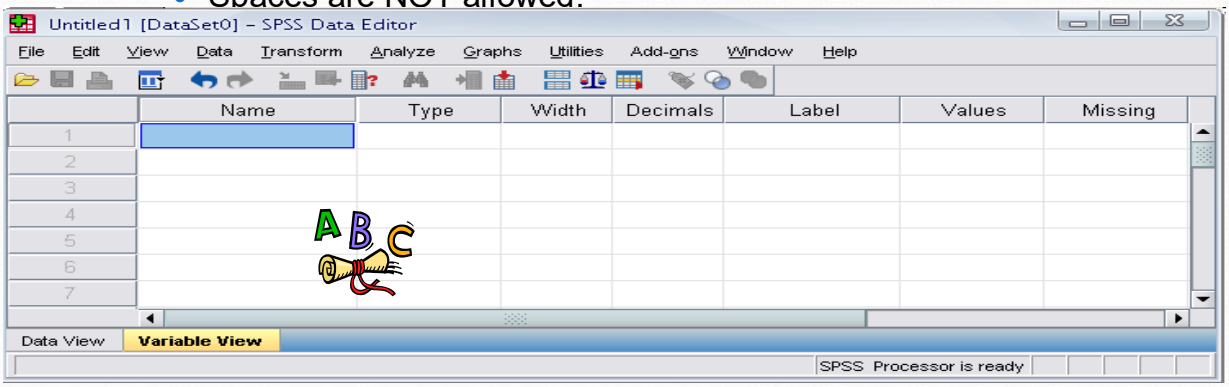
Data View window

- The Data View window
  This sheet is visible when you first open the Data Editor and this sheet contains the data
- Click on the tab labeled Variable View



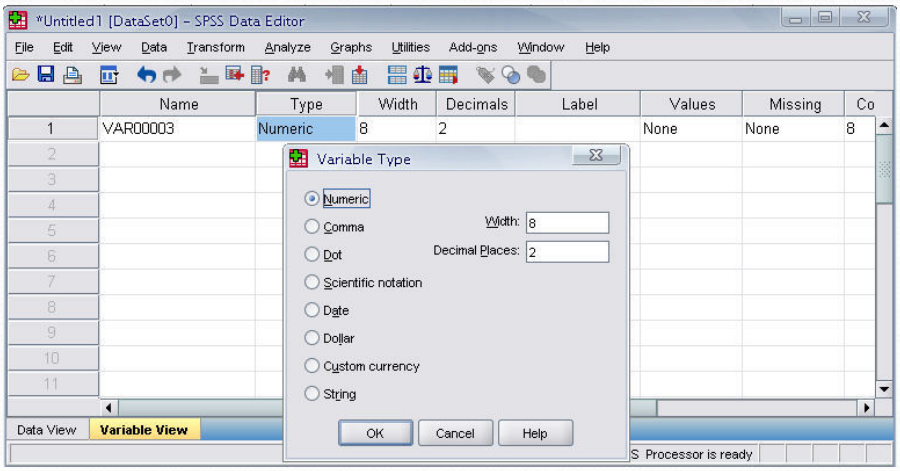
Variable View window

- This sheet contains information about the data set that is stored with the dataset
- Name
  - The first character of the variable name must be alphabetic
  - Variable names must be unique, and have to be less than 64 characters.
  - Spaces are NOT allowed.



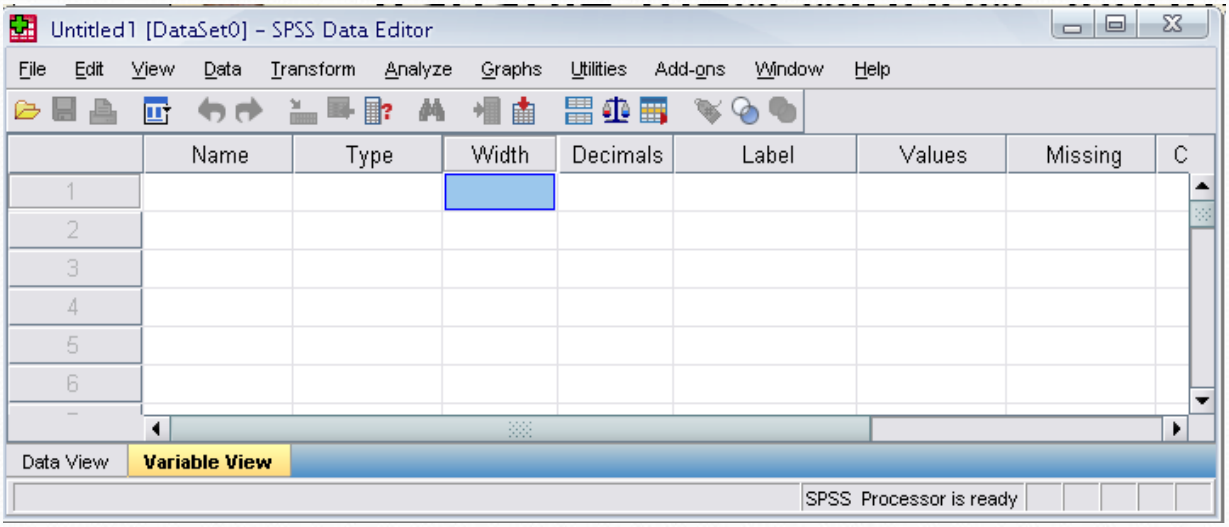
Variable View window: Type

- Type
  - Click on the 'type' box. The two basic types of variables that you will use are numeric and string. This column enables you to specify the type of variable.



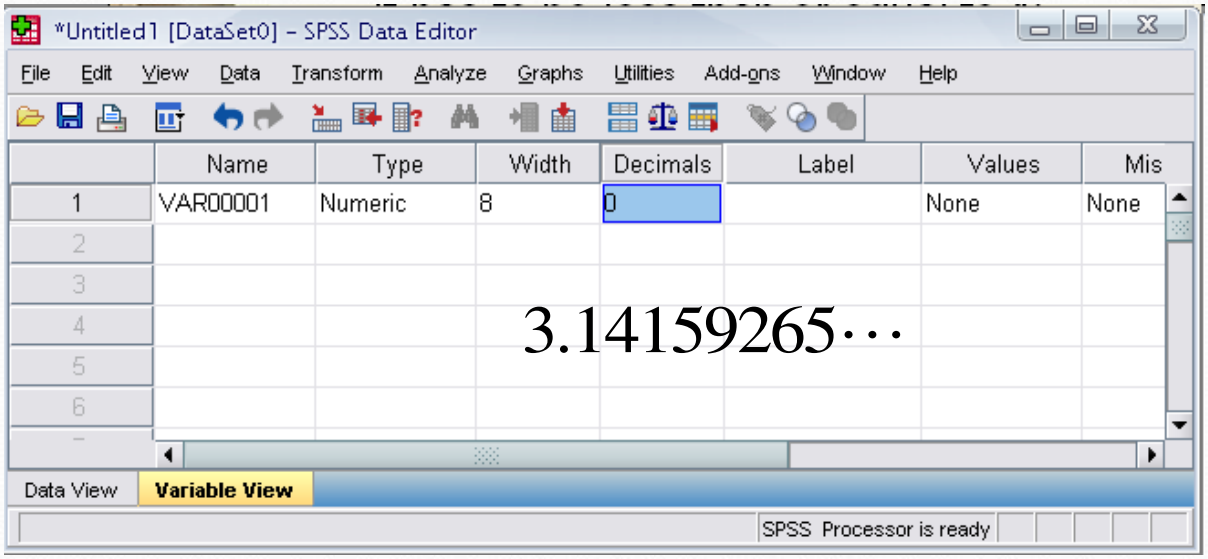
Variable View window: Width

- Width
  - Width allows you to determine the number of characters SPSS will allow to be entered for the variable



Variable View window: Decimals

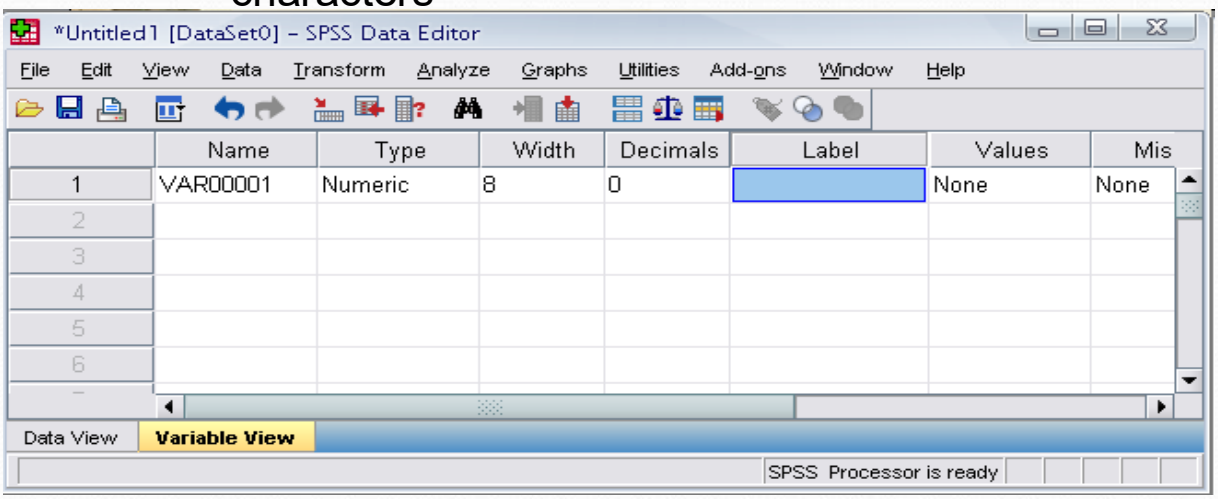
- Decimals
  - Number of decimals
  - It has to be less than or equal to 16



$3.14159265\cdots$

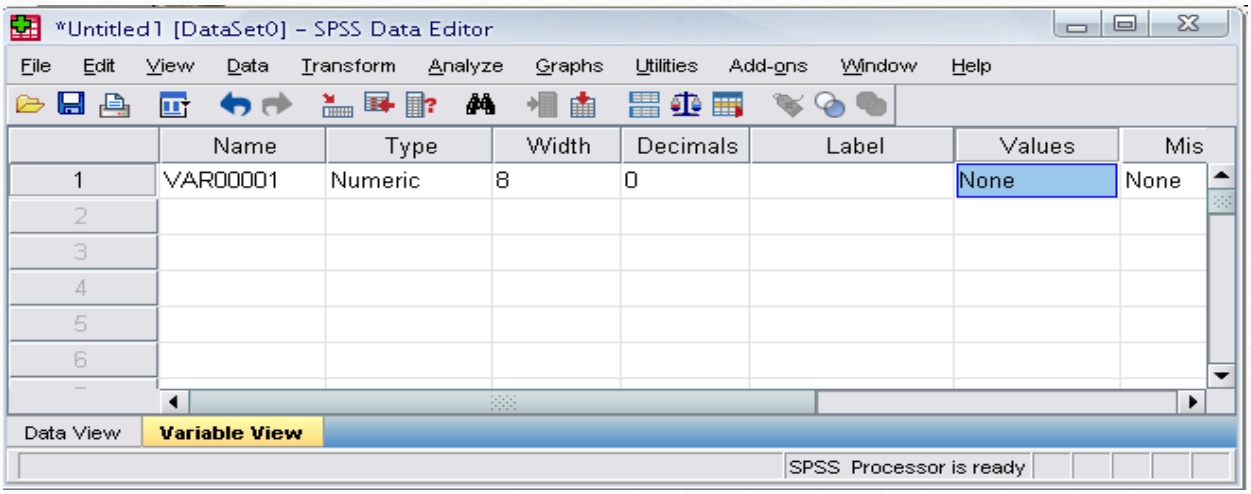
Variable View window: Label

- Label
  - You can specify the details of the variable
  - You can write characters with spaces up to 256 characters



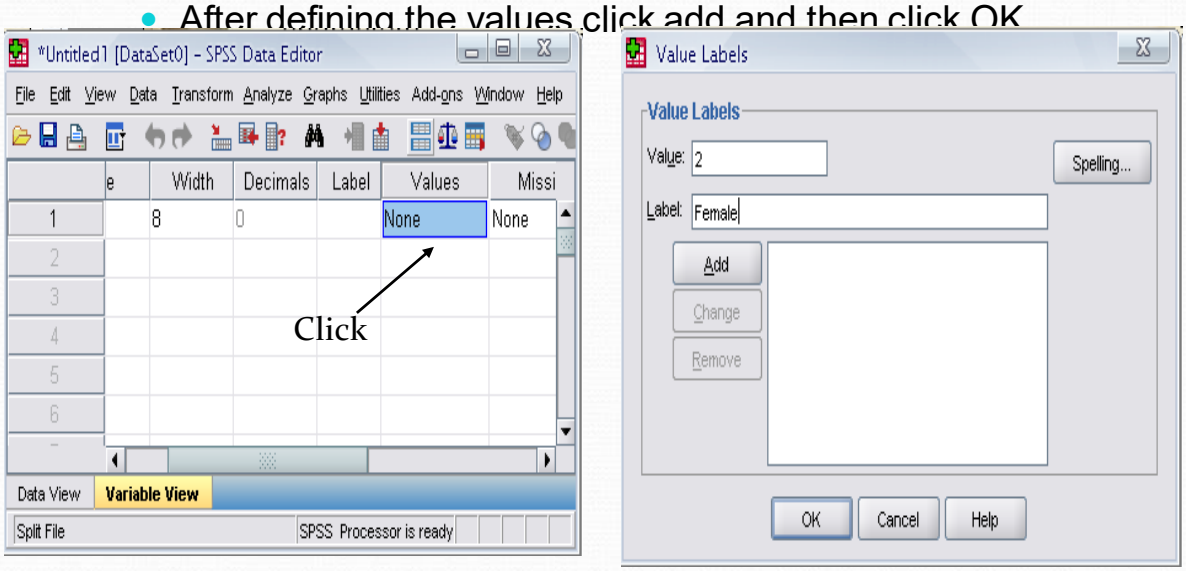
Variable View window: Values

- Values
  - This is used and to suggest which numbers represent which categories when the variable represents a category



Defining the value labels

- Click the cell in the values column as shown below
- For the value, and the label, you can put up to 60 characters.
- After defining the values click add and then click OK



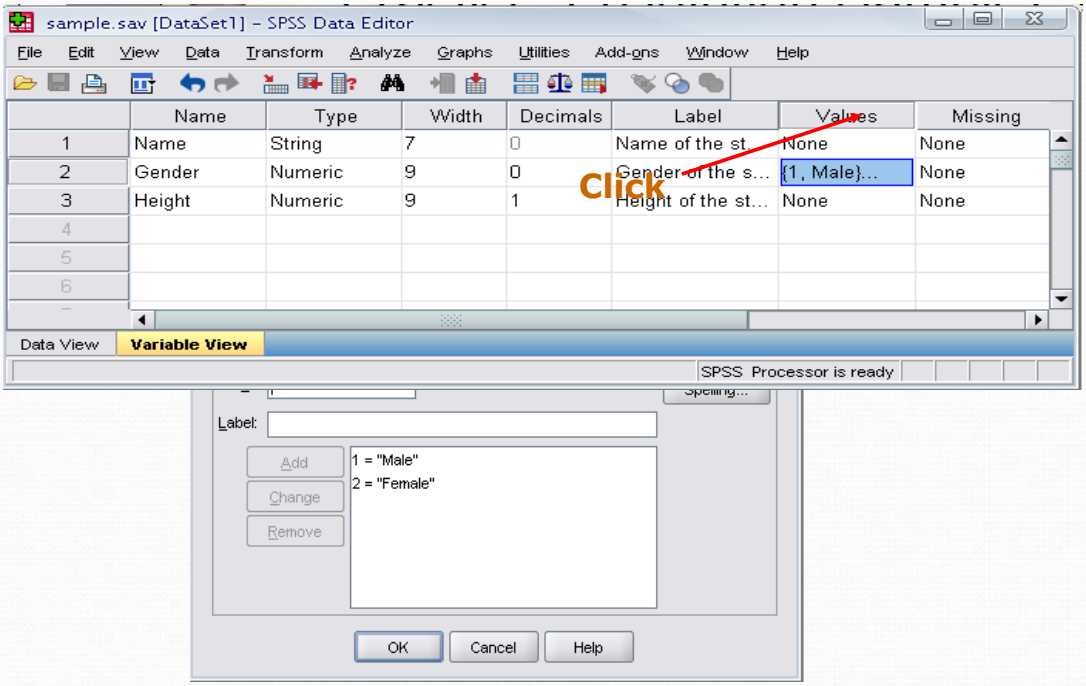
Practice 1

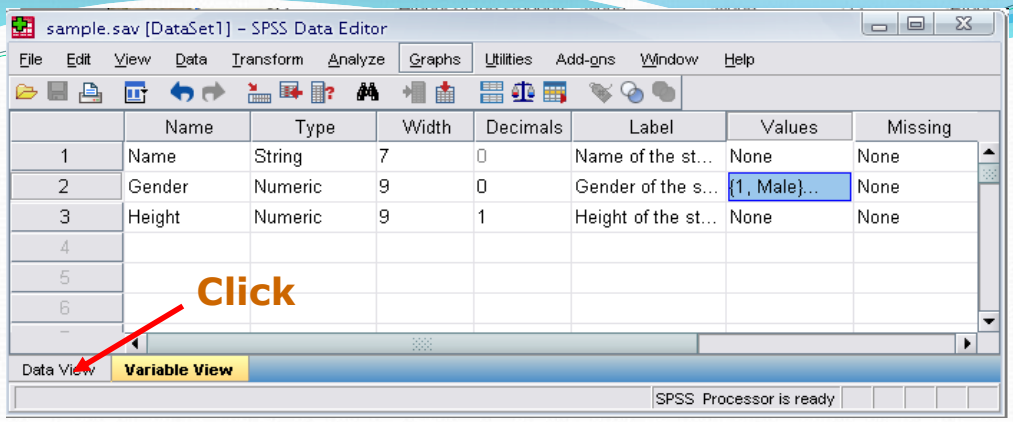
- How would you put the following information into SPSS?

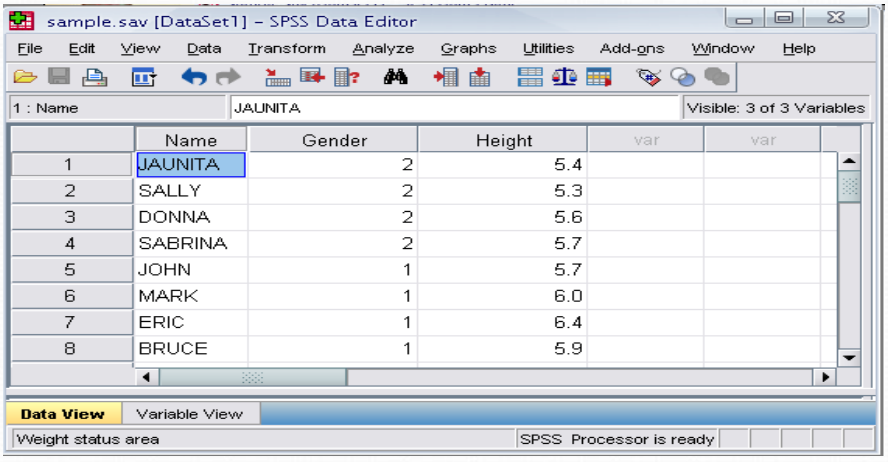
Name	Gender	Height
JAUNITA	2	5.4
SALLY	2	5.3
DONNA	2	5.6
SABRINA	2	5.7
JOHN	1	5.7
MARK	1	6
ERIC	1	6.4
BRUCE	1	5.9

Value = 1 represents Male and Value = 2 represents Female

Practice 1 (Solution Sample)

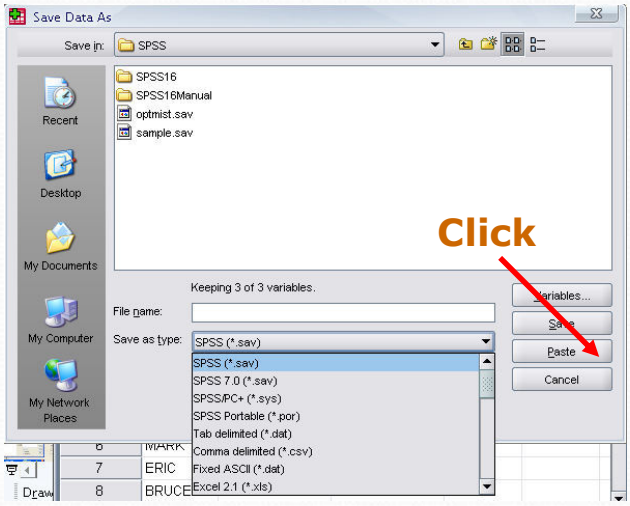






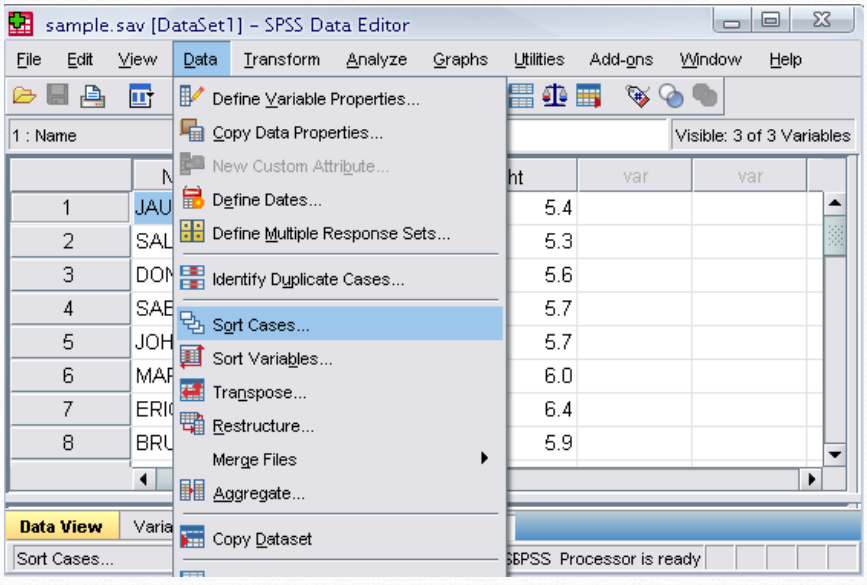
Saving the data

- To save the data file you created simply click ‘file’ and click ‘save as.’ You can save the file in different forms by clicking “Save as type.”



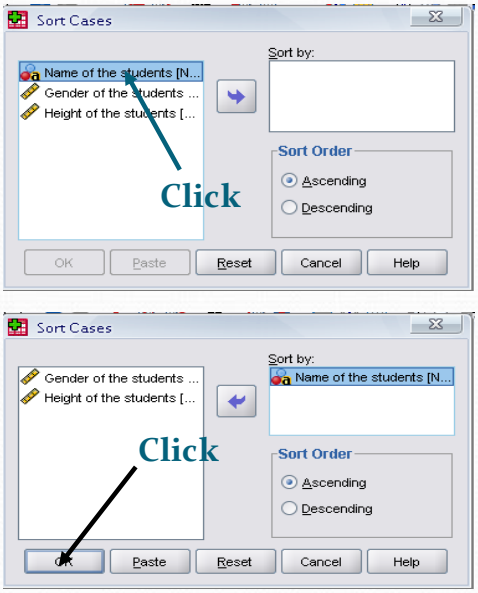
Sorting the data

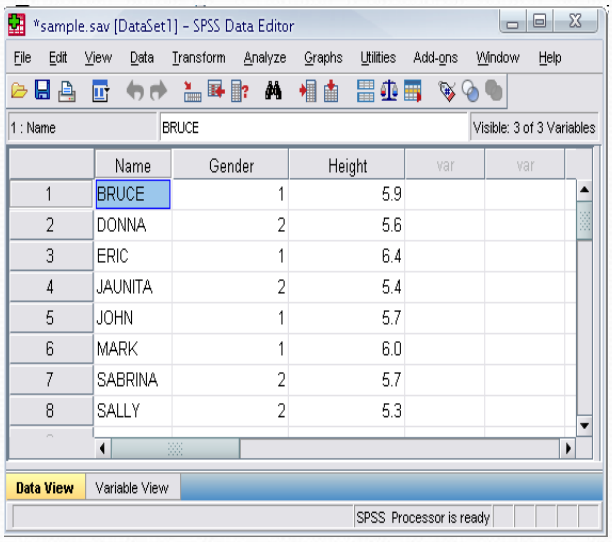
- Click ‘Data’ and then click Sort Cases



Sorting the data (cont'd)

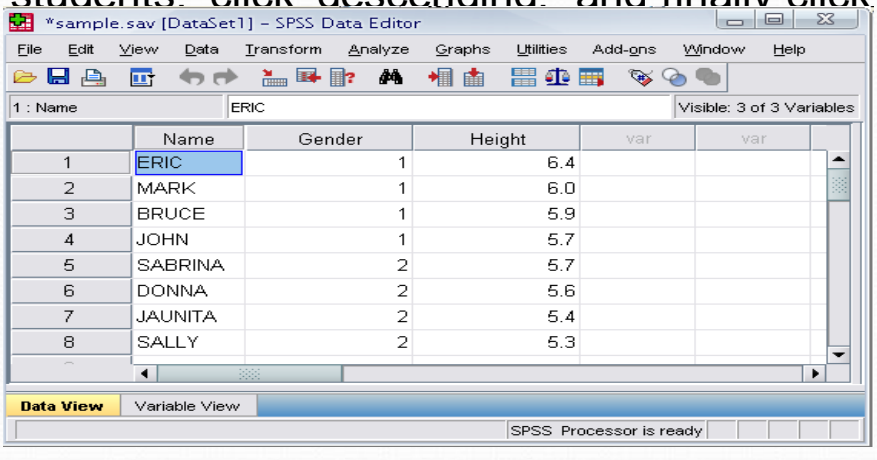
- Double Click 'Name of the students.' Then click ok.





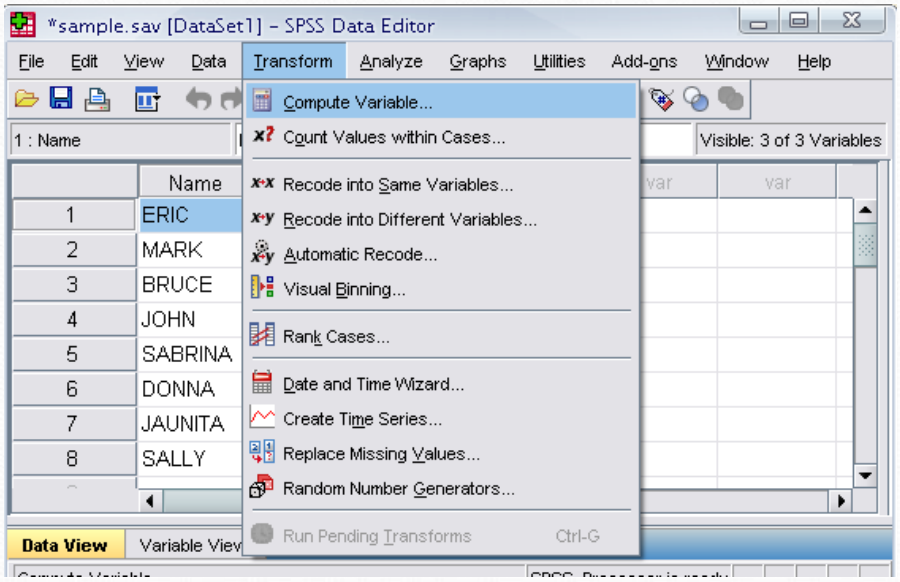
Practice 2

- How would you sort the data by the 'Height' of students in descending order?
- Answer
  - Click data, sort cases, double click 'height of students,' click 'descending,' and finally click ok.



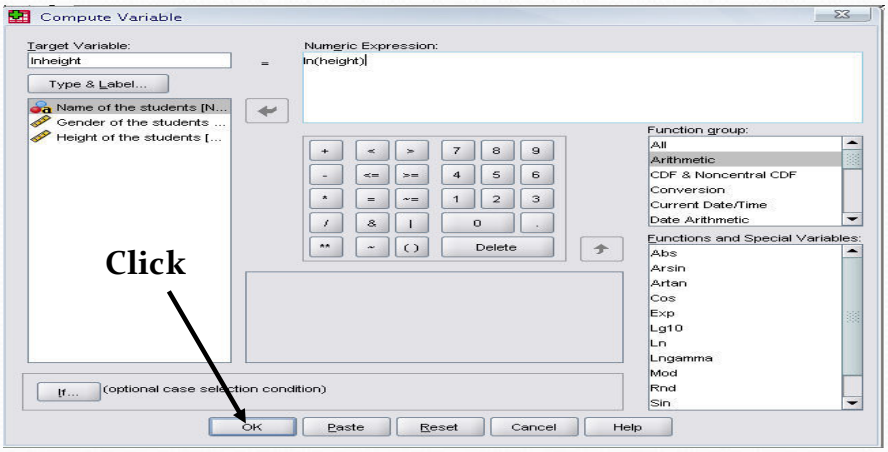
Transforming data

- Click ‘Transform’ and then click ‘Compute Variable…’



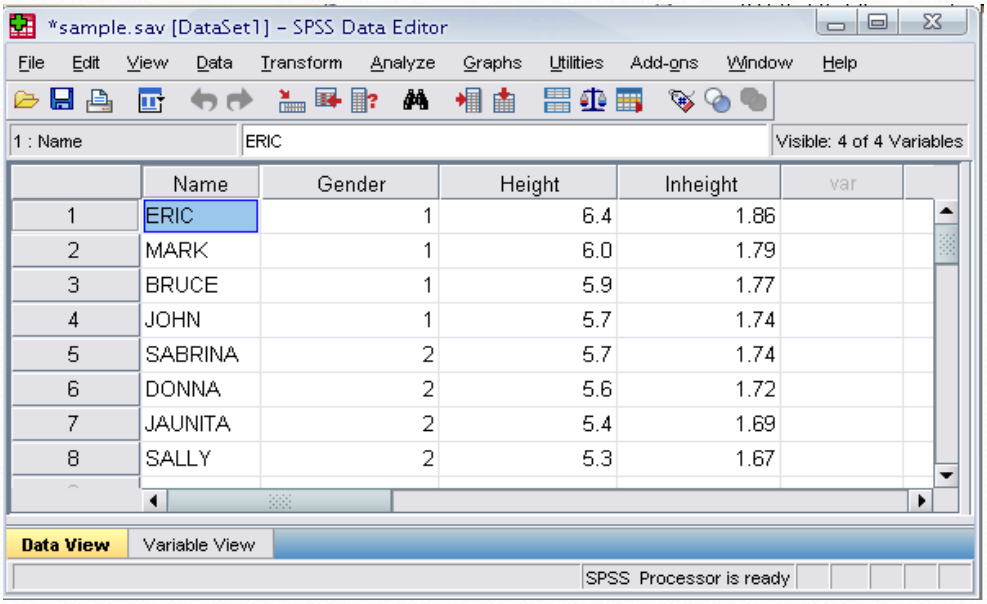
Transforming data (cont’d)

- Example: Adding a new variable named ‘lnheight’ which is the natural log of height
  - Type in lnheight in the ‘Target Variable’ box. Then type in ‘ln(height)’ in the ‘Numeric Expression’ box. Click OK



Transforming data (cont'd)

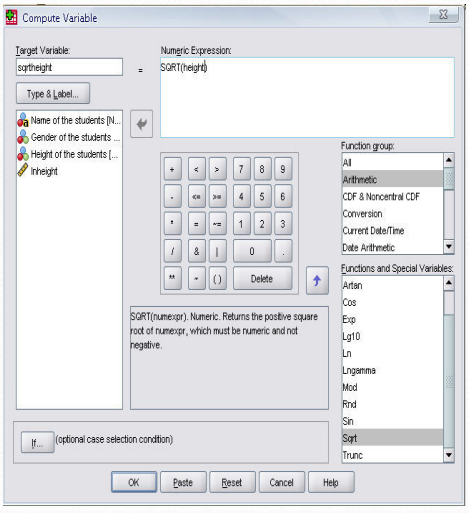
- A new variable 'Inheight' is added to the table

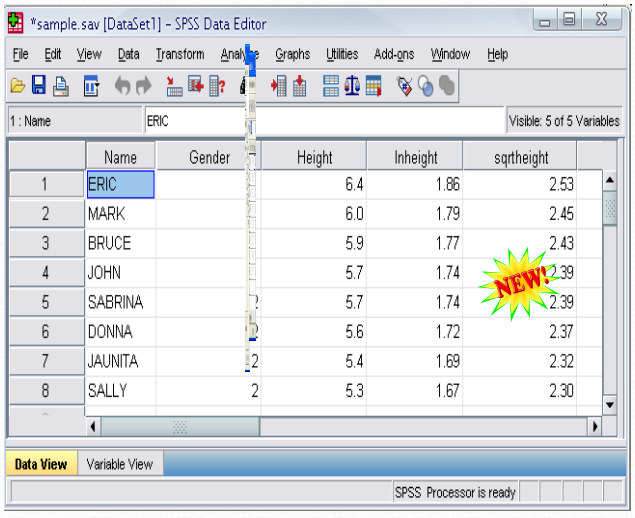


	Name	Gender	Height	Inheight	var
1	ERIC	1	6.4	1.86	
2	MARK	1	6.0	1.79	
3	BRUCE	1	5.9	1.77	
4	JOHN	1	5.7	1.74	
5	SABRINA	2	5.7	1.74	
6	DONNA	2	5.6	1.72	
7	JAUNITA	2	5.4	1.69	
8	SALLY	2	5.3	1.67	

Practice 3

- Create a new variable named "sqrtheight" which is the square root of height.
- Answer





	Name	Gender	Height	Inheight	sqrtheight
1	ERIC		6.4	1.86	2.53
2	MARK		6.0	1.79	2.45
3	BRUCE		5.9	1.77	2.43
4	JOHN		5.7	1.74	2.39
5	SABRINA	2	5.7	1.74	2.39
6	DONNA	2	5.6	1.72	2.37
7	JAUNITA	2	5.4	1.69	2.32
8	SALLY	2	5.3	1.67	2.30

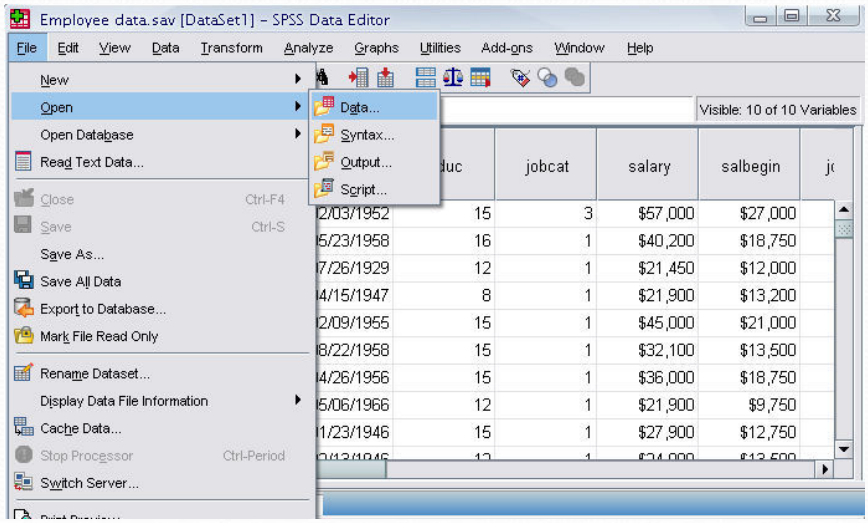
The basic analysis

The basic analysis of SPSS that will be introduced in this class

- Frequencies
  - This analysis produces frequency tables showing frequency counts and percentages of the values of individual variables.
- Descriptives
  - This analysis shows the maximum, minimum, mean, and standard deviation of the variables
- Linear regression analysis
  - Linear Regression estimates the coefficients of the linear equation

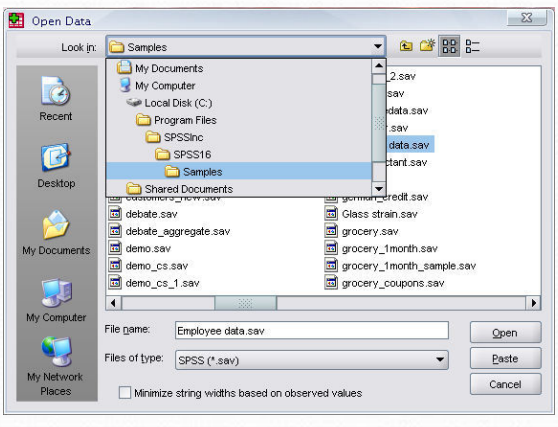
Opening the sample data

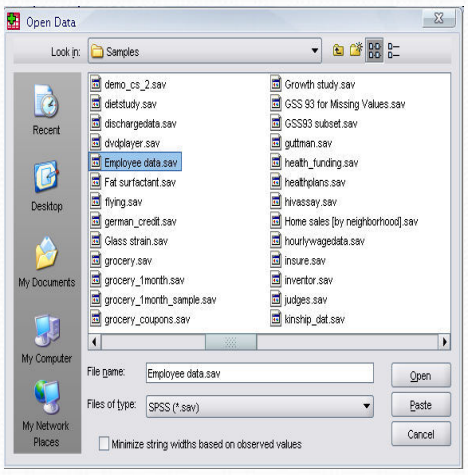
- Open ‘Employee data.sav’ from the SPSS
  - Go to “File,” “Open,” and Click Data



Opening the sample data

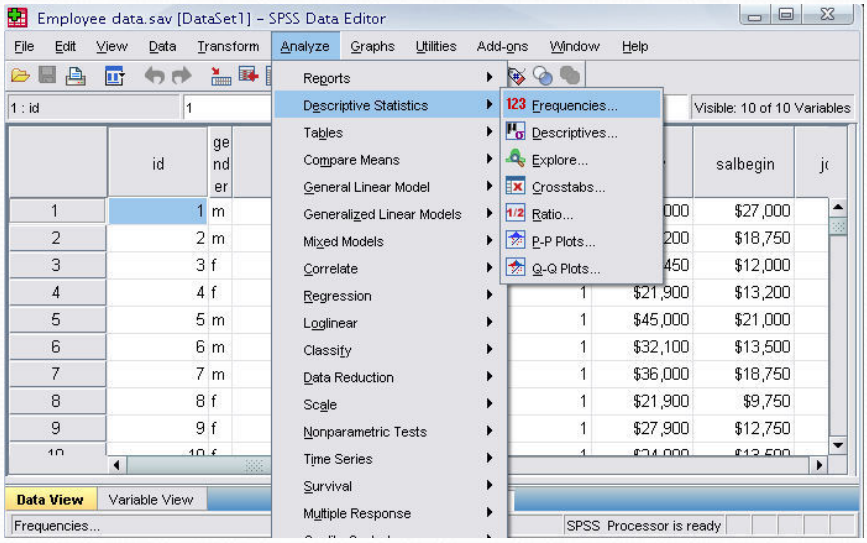
- Go to Program Files,” “SPSSInc,” “SPSS16,” and “Samples” folder.
- Open “Employee Data.sav” file





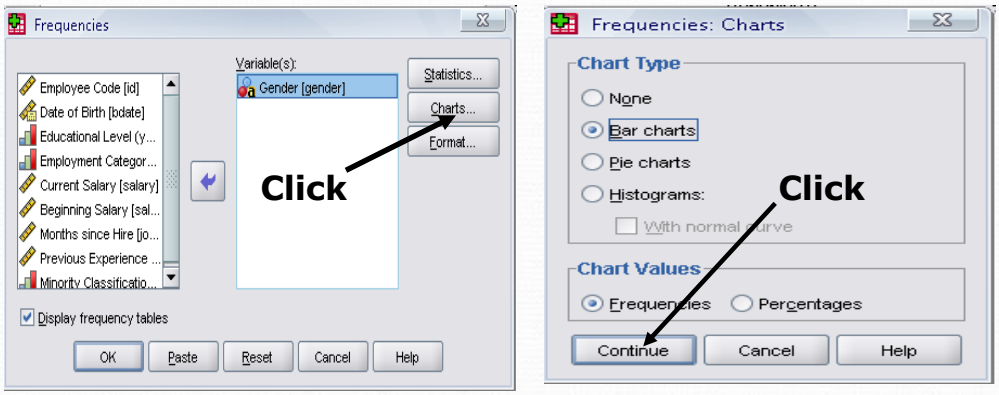
Frequencies

- Click ‘Analyze,’ ‘Descriptive statistics,’ then click ‘Frequencies’



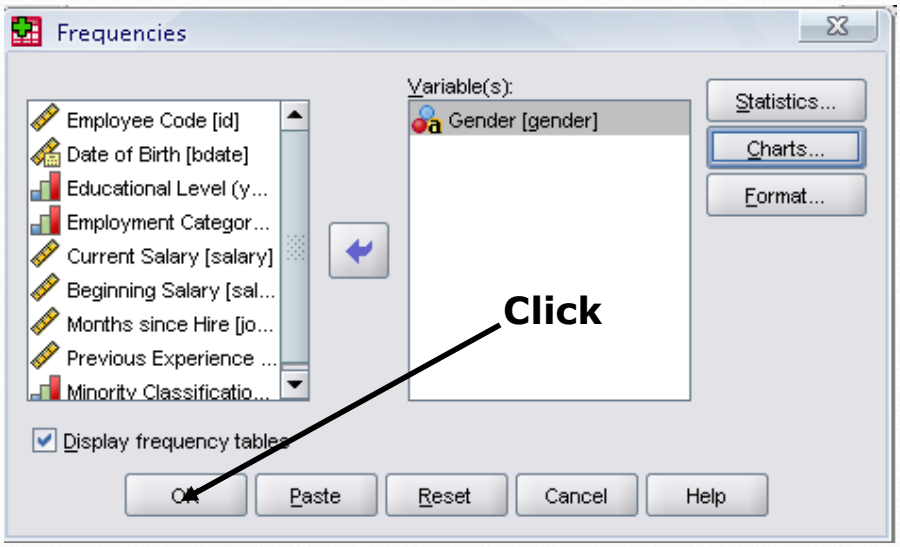
Frequencies

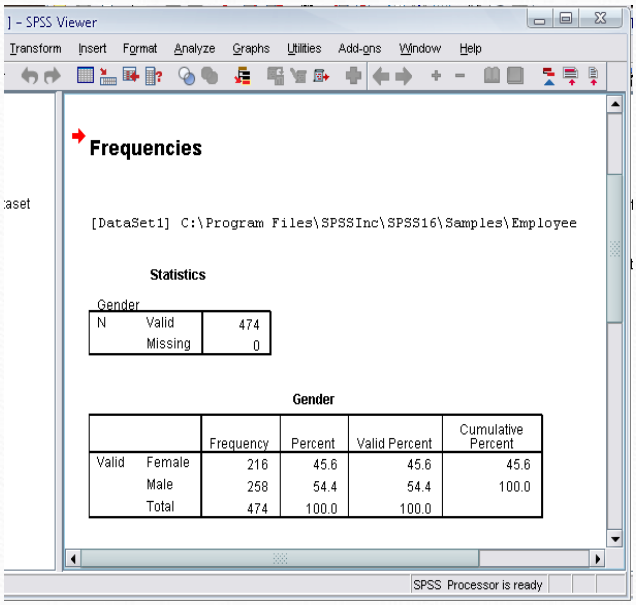
- Click gender and put it into the variable box.
- Click ‘Charts.’
- Then click ‘Bar charts’ and click ‘Continue.’

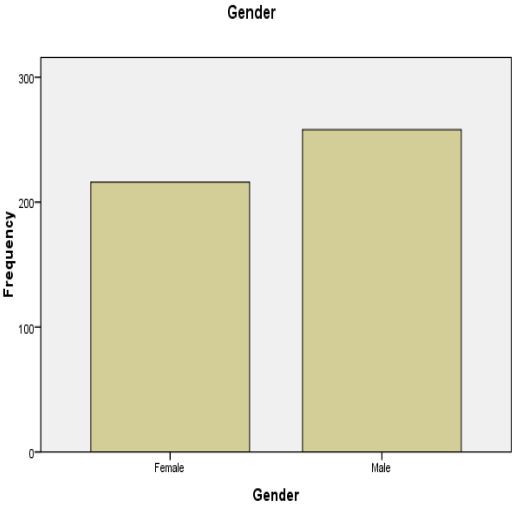


Frequencies

- Finally Click OK in the Frequencies box.

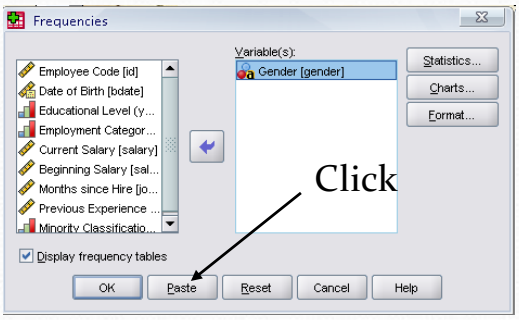


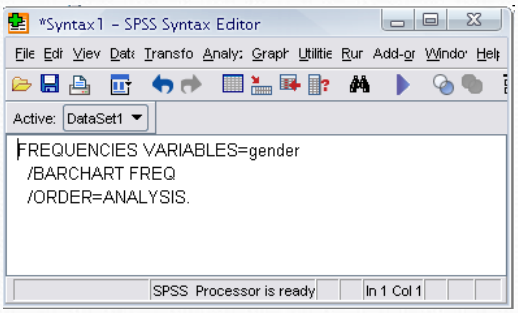




Using the Syntax editor

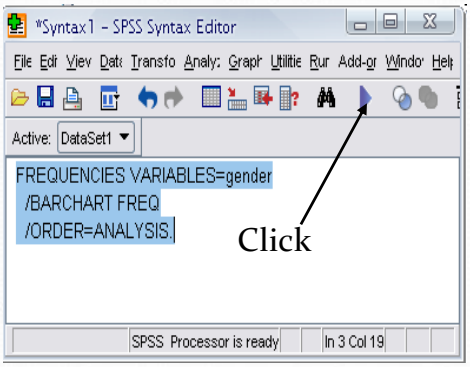
- Click ‘Analyze,’ ‘Descriptive statistics,’ then click ‘Frequencies.’
- Put ‘Gender’ in the Variable(s) box.
- Then click ‘Charts,’ ‘Bar charts,’ and click ‘Continue.’
- Click ‘Paste.’

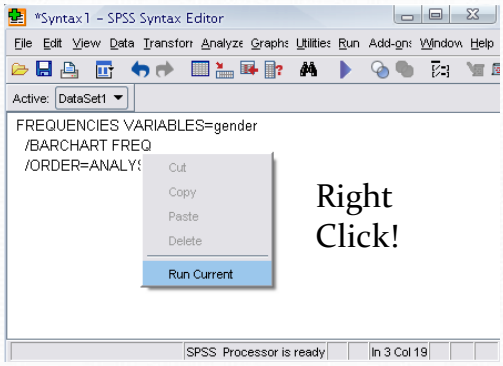




Using the Syntax editor

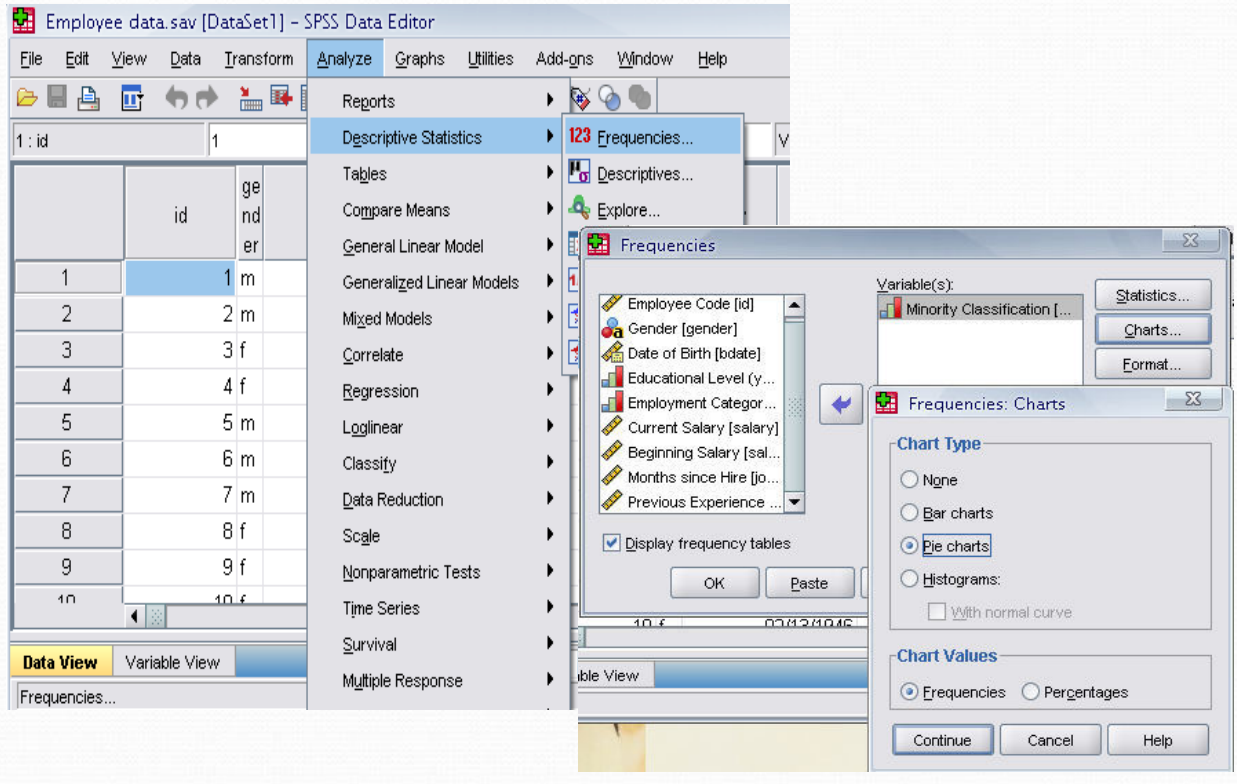
- Highlight the commands in the Syntax editor and then click the run icon.
- You can do the same thing by right clicking the highlighted area and then by clicking ‘Run Current’



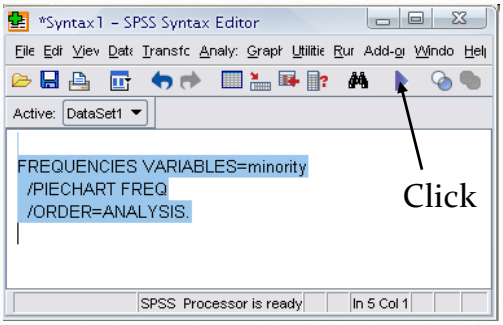


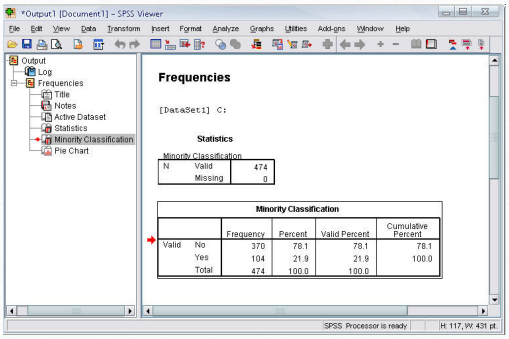
Practice 4

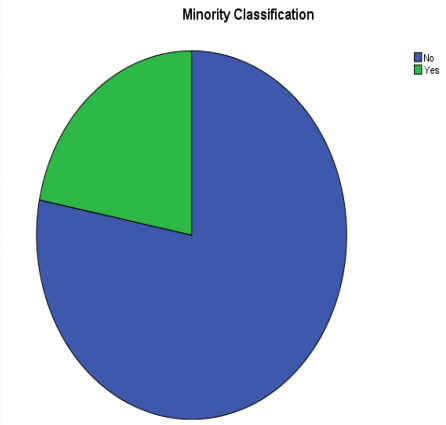
- Do a frequency analysis on the variable “minority”
- Create pie charts for it
- Do the same analysis using the syntax editor



Answer

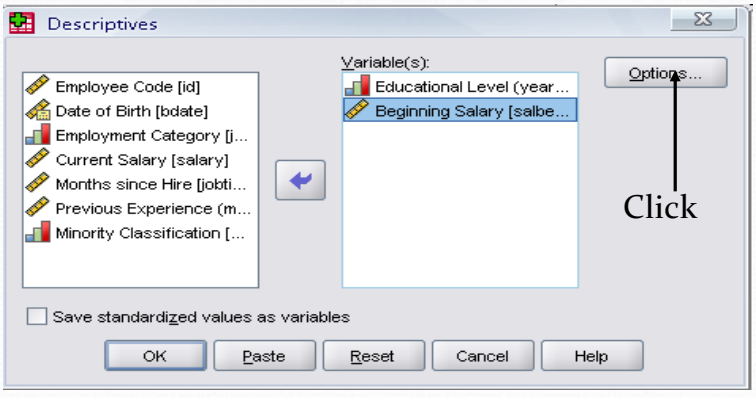






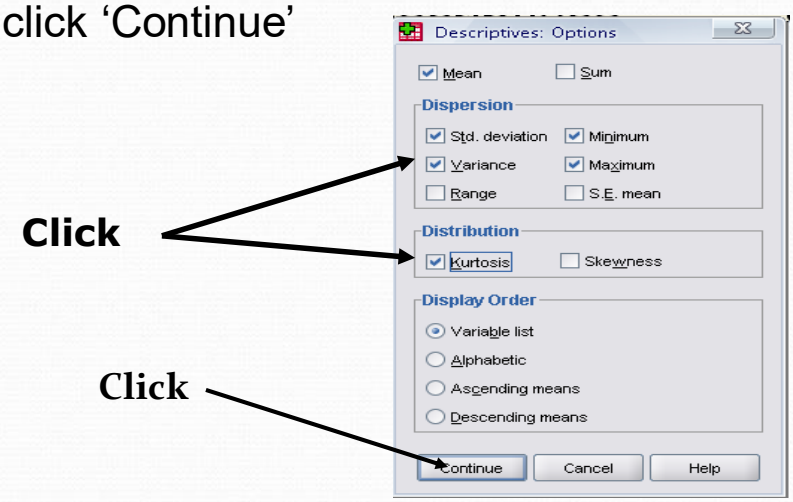
Descriptives

- Click ‘Analyze,’ ‘Descriptive statistics,’ then click ‘Descriptives…’
- Click ‘Educational level’ and ‘Beginning Salary,’ and put it into the variable box.
- Click Options



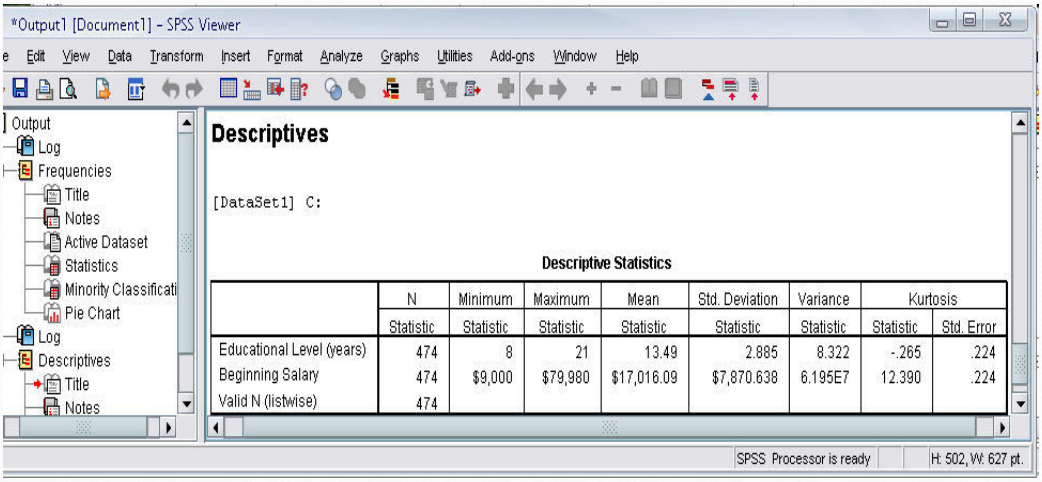
Descriptives

- The options allows you to analyze other descriptive statistics besides the mean and Std.
- Click ‘variance’ and ‘kurtosis’
- Finally click ‘Continue’



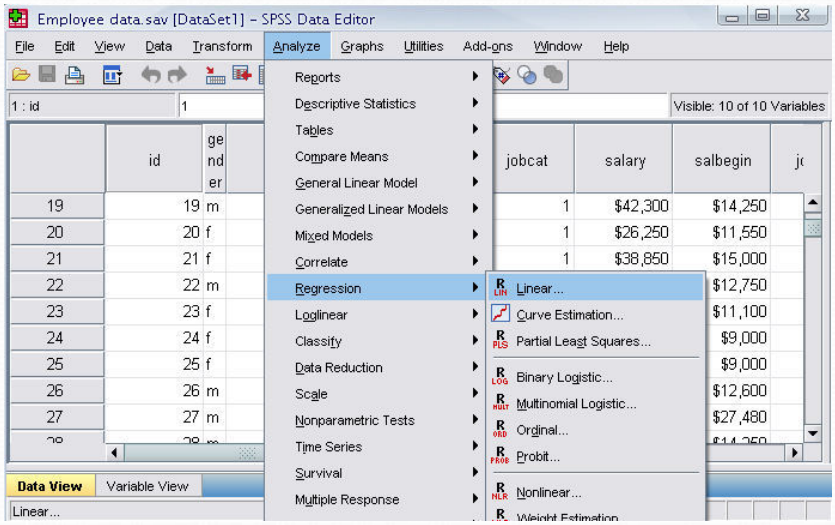
Descriptives

- Finally Click OK in the Descriptives box. You will be able to see the result of the analysis.



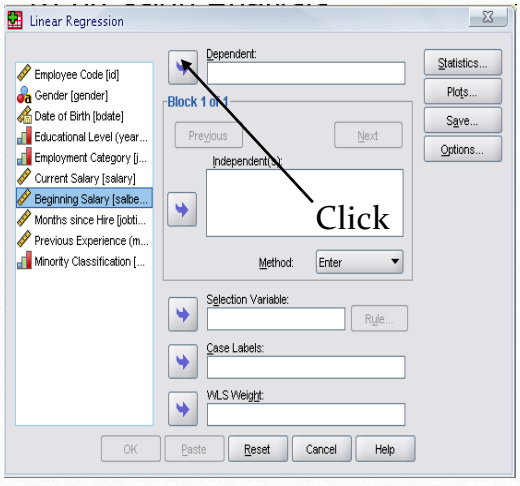
Regression Analysis

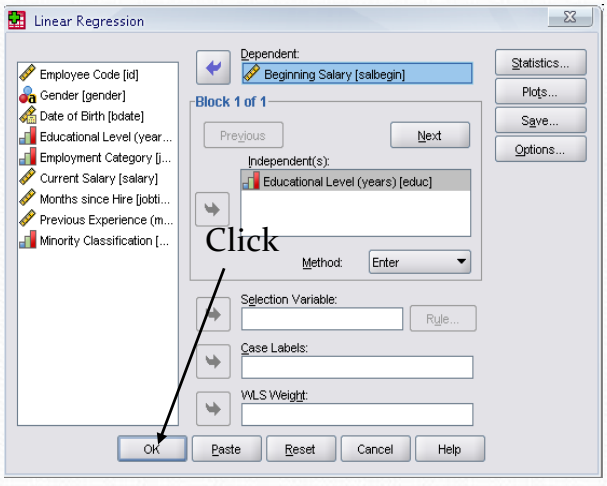
- Click ‘Analyze,’ ‘Regression,’ then click ‘Linear’ from the main menu.



Regression Analysis

- For example let’s analyze the model $\text{salbegin} = \beta_0 + \beta_1 \text{edu} + \varepsilon$
- Put ‘Beginning Salary’ as Dependent and ‘Educational Level’ as Independent.





Regression Analysis

- Clicking OK gives the result

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.633[a]	.401	.400	$6,098.259

a. Predictors: (Constant), Educational Level (years)

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.175E10	1	1.175E10	315.897	.000[a]
	Residual	1.755E10	472	3.719E7		
	Total	2.930E10	473			

a. Predictors: (Constant), Educational Level (years)

b. Dependent Variable: Beginning Salary

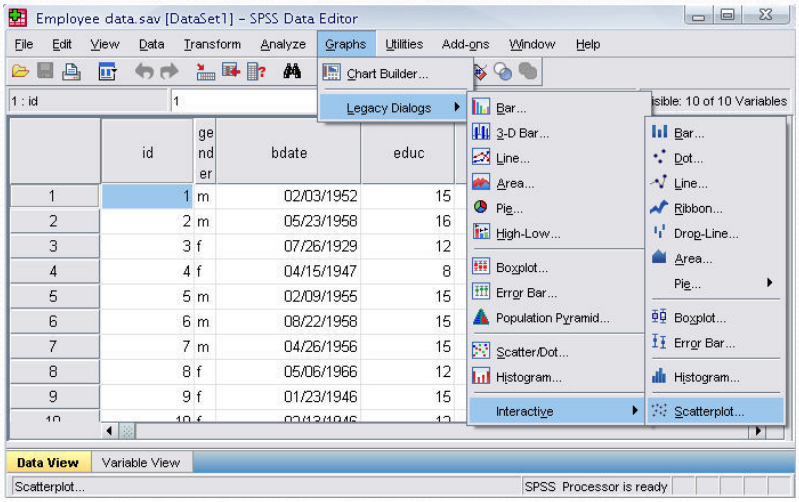
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-6290.967	1340.920		-4.692	.000
	Educational Level (years)	1727.528	97.197	.633	17.773	.000

a. Dependent Variable: Beginning Salary

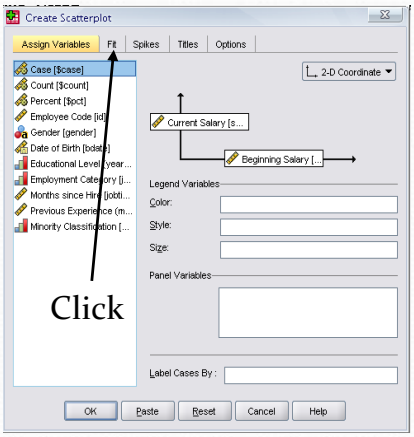
Plotting the regression line

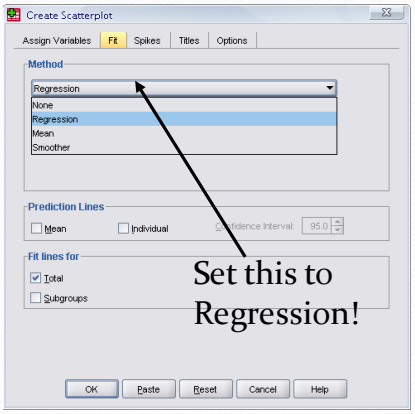
- Click ‘Graphs,’ ‘Legacy Dialogs,’ ‘Interactive,’ and ‘Scatterplot’ from the main menu.

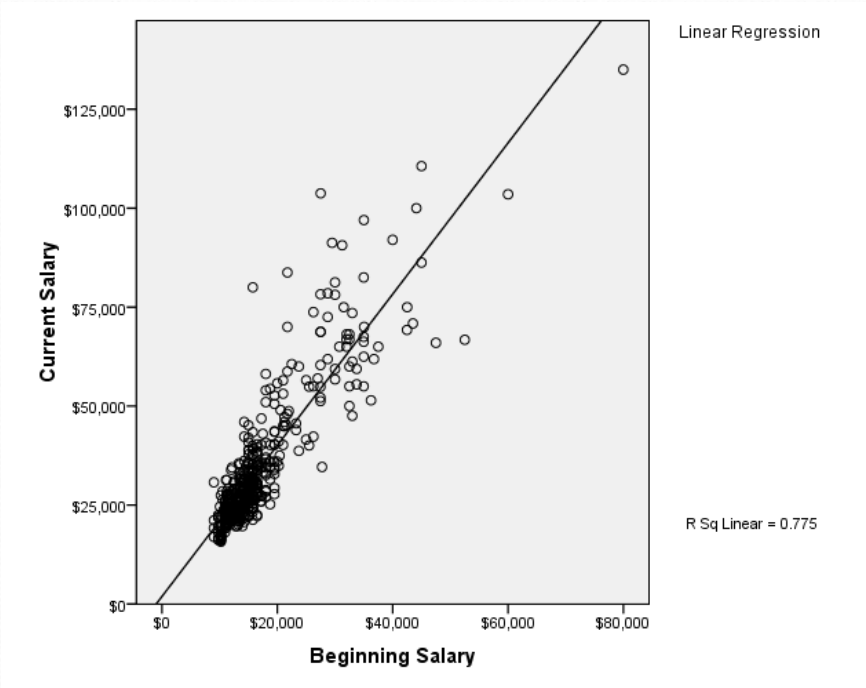


Plotting the regression line

- Drag ‘Current Salary’ into the vertical axis box and ‘Beginning Salary’ in the horizontal axis box.
- Click ‘Fit’ bar. Make sure the Method is regression in the Fit box. Then click ‘OK’.







www.vuCtn.com

Good Luck

Regards : Zarva Chaudhary

Admin:

* Zarva Chaudhary*

Chaudhary Moazzam

* Laiba Mahi *