Workshop on Statistics Using Excel and SPSS

3- 5 January 2017

Professor Dr. K. MARIMUTHU

Deputy Vice Chancellor, Academic and International Affairs,

Department of Biotechnology, Faculty of Applied Sciences

AIMST University, Bedong-Semeling, 08100 Semeling, Kedah Darul Aman, Malaysia.

T: +604 - 429 1054 | F: +604 - 429 8102 | HP: +6016 - 4723672

Email: marimuthu@aimst.edu.my | aquamuthu2k@gmail.com

Workshop on Statistics using Excel and SPSS

Workshop Schedule

Day 1

- Introduction to statistics in Research - Lecture – 9.00 am to 10.30 am
- Hands on session - Excel 11 .00 am – 1.00 pm
- Lunch Break 1.00 – 2.00 pm
- Introduction to SPSS – 2.00 pm – 3.00 pm
- Hands on session SPSS 3.00 – 5.00 pm

Day 2

- Hands on session SPSS parametric and Non parametric analysis 9.00 – 4.00 pm

Statistics in Research

Workshop Goals

- Provide knowledge of basic statistical terms and notation
- To understand research process.
- Ability to summarize data and conduct basic statistical analyses using **Excel and SPSS**
- Ability to understand basic statistical analysis in published research papers

4

Lecture outline

- What is research?
- Research Process.
- Motivation in conducting Research.
- Why do we need to have statistical knowledge?
- Biostatistics and its types.
- Important terms related to biostatistics.
- What is sampling and types of sampling methods.
- How to present and describe a set of data **(tables and graphs).**
- Measures of central tendency **(Center),** measures of dispersion **(Spread).**
- **When and how** to use some of the basic analysis, like t tests, chi square, correlation, regression and ANOVA.

5

WHAT IS RESEARCH ?

- **"Research is a systematized effort to gain new knowledge".**
- **Research is the "Systematic process of collecting and analyzing information (data) in order to increase our understanding of the phenomenon about which we are concerned or interested".**

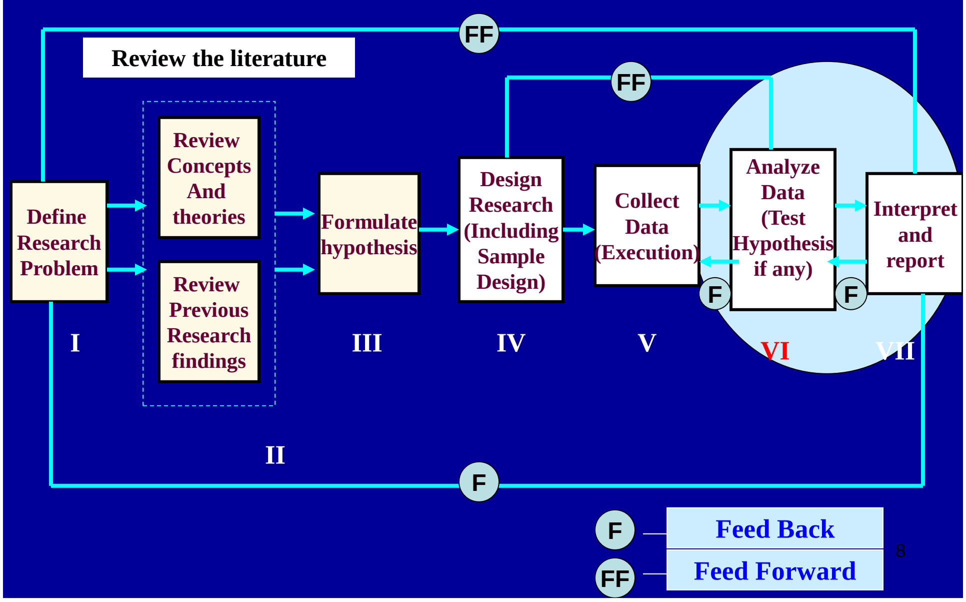
Motivation in conducting Research

The possible motivation for doing research may be either one or more of the following:

- Desire to get a **research degree** along with its consequential benefits;
- Desire to face **the challenge** in solving the unsolved problems, i.e., **concern over practical problems** initiates research;
- Desire to get **intellectual joy** of doing some creative work;
- Desire to be of **service to society** (Bio fuel & bio degradable plastics, control of mosquitos)

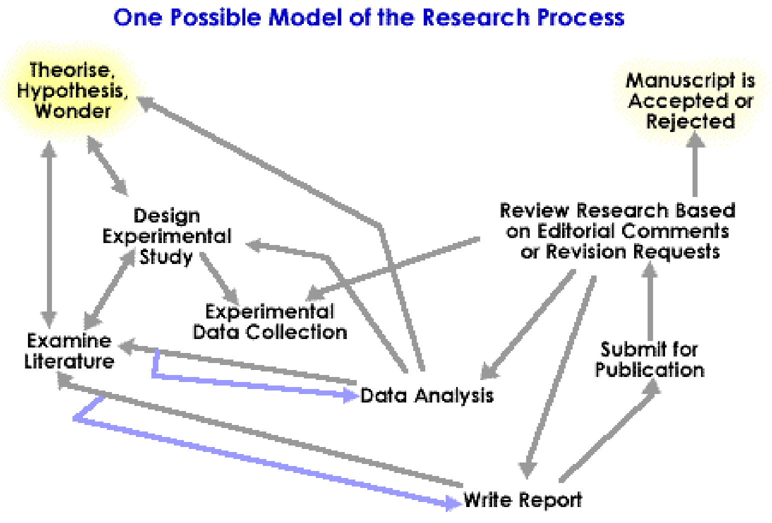
7

RESEARCH PROCESS



8

RESEARCH PROCESS



9

Why do we need to have statistical knowledge in research?

- To understand the statistical portions of most scientific articles.
- To plan and design your own research projects and experiment
- To know which types of statistical tests to apply to a given data set and your research…..

10

What is Statistics?

Statistics is a field of study concerned with collection, organization, summarization and analysis of data.

- **Statistics has three primary components:**
  - **How best can we collect data?**
  - **How should it be analyzed?**
  - **And what can we infer from the analysis?**

Biostatistics

- **Statistical methods used to analyze data in various field of study, including human biology, life sciences, medicine, public health, and business.**

- Statistics applied to Life sciences – biological sciences - **biostatistics** or **biometry.**

Important Terms Used in Statistics

POPULATION *VERSUS* SAMPLE

Population:

- The complete collection of all elements (scores, people, measurements, and so on) to be studied.
  - Examples: Animals; a fish species; human beings; SP citizens; who are high school students; Aimst students, males and females etc.,
  - **Example of Study:**
    - Rate of obesity and smoking habits in male and females in Malaysia;
    - Birth weights of new born babies in Malaysia

Sample:

- A portion of a population selected for further analysis/ study.
- Example: Birth weights of 100 babies born in a certain hospital

13

Important Terms Used in Statistics

❖Census

- the collection of data from every member of the population.

❖Sample

- a sub-collection of elements drawn from a population.
  - Example: Birth weights of 100 babies born in a certain hospital

14

Important Terms Used in Statistics

- **Parameter**: Numerical characteristic of the whole population.
- **Statistic**: Numerical characteristic of a sample.
- **Variable:** **A *variable* is a measurable property or attribute associated with each subject of a population or sample**
  - **Examples: blood group; age; body weight and height of patient**
- **Data:** Observations (such as measurements, genders, survey responses) that have been collected.

15

Types of variables

Categorical variables: record the data into several categories

Examples:

- Blood type: A, B, AB, O
- Sex of a fish: Male, female
- Race: Malay, Chinese, Indian and others

Ordinal variables: Some categorical variables can be arranged in a rank

Examples:

- Body pain: mild, moderate and severe
- Quality of Beef meat: tough, slightly tough, tender
- Likert scale agreement in a questionnaire: Strongly agree, agree, neutral, disagree, strongly disagree

16

Quantitative & Qualitative variables

Quantitative variables:

- Measurements made on quantitative variables convey information regarding amount.
  - Example: heights of adult males; body weights; age of patients; hemoglobin levels, etc..

Qualitative (categorical) variables:

- Some characteristics are not capable of being measured in the sense that height, weight, and age are measured.
- These characteristics are categorized only
  - a person is designated as belonging to an racial group,

    Race: **Malay, Chinese, Indian, etc.**

    Sex: **Male or female**

    Smoking : **Yes or no**

17

Independent Variables VS Dependent Variables

Independent variables:

- Experimental.
- Manipulated.
- Controlled.

Dependent Variables:

- Effects.
- Measured variables.
- Outcome variable
- Dependent variables are functions of independent variables.

Example : Effect of temperature on hatching rate of fish eggs

- Temperature – **Independent variable**
- Hatching rate of fish eggs- **Dependent variable**
- **Blood cholesterol and glucose levels** in **different age groups**.
- Effect of **salinity** on **hatching rate of fish eggs**.

18

Extraneous Variables

- Independent variables that are irrelevant to the focus of the study.
- It may affect the dependent variable and affect interpretation of results.
- Examples:

  time of day

  sex of investigator

Measuring glucose level in blood

19

Extraneous Variables

Example of study:

The relationship between background music and task performance among employees at a packing facility in a factory.

Independent variable:
Background music (a **nominal variable** because employees are either provided **with** or **without** background music)

Dependent variable:
Task performance (a **continuous variable**, measured in terms of the number of tasks employees perform correctly per hour)

Extraneous Independent variables:

- Type of background music (e.g., chart music, dance/electronic music, easy listening, classical music, etc.)
- Loudness of background music (e.g., low, medium, high volumes, etc.)
- Time of day when the background music was played (e.g., morning, afternoon, night, etc.)

20

Confounding variable

- A confounding variable is a variable, **other than the independent variable that you're interested in, that may affect the dependent variable**.
- This can lead to erroneous conclusions about the relationship between the independent and dependent variables.

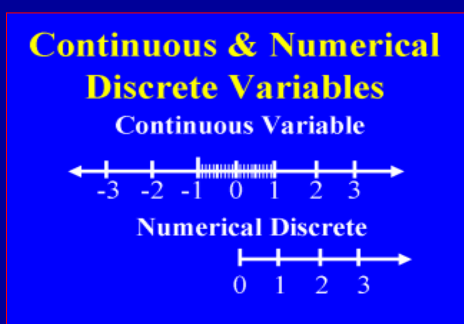
Controlling confounding variables

- Designing an experiment to eliminate differences due to confounding variables is critically important.
- Provide homogenous conditions and environment to control the confounding variables effect of dependent variable.

21

Types of Quantitative Data

Quantitative data can further be distinguished between discrete and continuous types.



22

Discrete & Continuous

Discrete - count (How many)

❖ Data result when the number of possible values is either a finite number or a 'countable' number of possible values. Values are invariably whole numbers

0, 1, 2, 3, . . .

Example:

- The number of eggs that hens lay.
- Number of children in a family..
- Number of bacterial colonies in a petri dish.
- Number of phone calls received per day

23

Discrete & Continuous

Continuous - **measure or (how much)**

(numerical) data result from infinitely many possible values (fractional values) that correspond to some continuous scale that covers a range of values without gaps, interruptions.

Example:

- **The amount of milk that a cow produces; e.g. 12.343115 Liters per day.**
- **Body weight in kg 75.589**
- **Blood Cholesterol, Blood glucose**

24

Scales of measurements

- Another way to classify the variables is to assign number to the objects or events according to a set of rules.

- They are commonly broken down into four types:
  - **Nominal**
  - **Ordinal**
  - **Interval (numerical)**
  - **Ratio (numerical)**

Nominal

- Observations are grouped by name into categories e.g. sex – male/female; Ethnicity: Political Party Affiliation; Colors of marbles
- The data we collect often has to be converted to numbers for statistical or tabulation purposes.
- So when we have nominal categorical data we often arbitrarily assign a numerical value for tabulation purposes such as
  - Male =1 and Female = 2
  - Malay = 1, Chinese = 2 and Indian = 3
  - Green = 1, Blue = 2, Red = 3
- No ordering,
  - it makes no sense to state that M > F

Ordinal

Similar to nominal except that the categories can be put in a certain order
e.g. pain - mild, moderate, severe

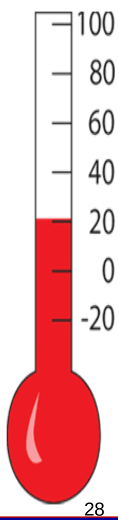
- A scale that expresses data as rankings, rather than scores:
- Examples:
  - Course grades A, B, C, D, F
  - Socio Economic Status
    - high, medium, low
    - first, second, third

- The distance between the categories is not equal; the difference between grades F and D is probably not the same as the difference between A and B.

27

Interval

- Measurement scales expressed in equal number units, but not having a true zero point
  - IQ score (there is no such thing as zero intelligence)
  - Temperature (there is a zero but it has meaning, **it does not represent "nothingness")**
- When we record this data, we use the actual numbers



28

Interval

- **Interval scale** is next higher order among scales and contains all the characteristics of nominal and ordinal scale but with an added characteristic of equal distance or interval between observations
- If three items a, b, and c have the numerical value of 1, 2, and 3, we can say that the interval or distance between a and b is 1, and between b and c is also 1 and between a and c is 2

29

Ratio

- Measurement scales expressed in equal number units, but having a true zero point
  - Test scores
  - Salary
  - Weight & Height
  - Distance
- Highest order measurement scale and contains the characteristics of all other scales
- Numbers on the scale indicate the actual amount of the property is measured
- Based on true zero; if a measurement is zero in a ratio scale **then the object has none of the property** being measured
- In a ratio scale, a score of 8 has twice as value as that of a score of 4
  - Example: age, weight, income, rainfall, price, etc.

30

Summary - Levels of Measurements

- Nominal - categories only
- Ordinal - categories with some order
- Interval - differences but no natural starting point
- Ratio - differences and a natural starting point

31

Sampling

- Is it possible to work out what 50 million people think by asking only 1000?
  - **YES**

  The small group chosen for study is called sample.

  If you were to study everybody then it is called a census.

32

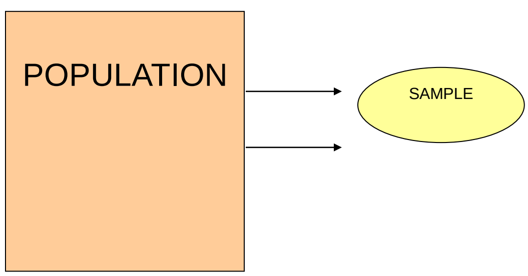
Methods of Sampling

- **Sampling** is the fundamental method of inferring information about an entire population without going to the trouble or expense of measuring every member of the population.
- Developing the proper sampling technique will **greatly affect the accuracy** of your results.

33

How samples are taken? Random.

The **sampling technique** is said to be random if each member of the population has the same chance of being chosen.



34

Reasons for using samples

There are many good reasons for studying a sample instead of an entire population:

- Samples can be **studied more quickly than** populations.
- **Speed** can be important if a physician needs to determine something quickly, such as a vaccine or treatment for a new disease.

- A study of a **sample is less expensive** than a study of an entire population because a smaller number of items or subjects are examined.

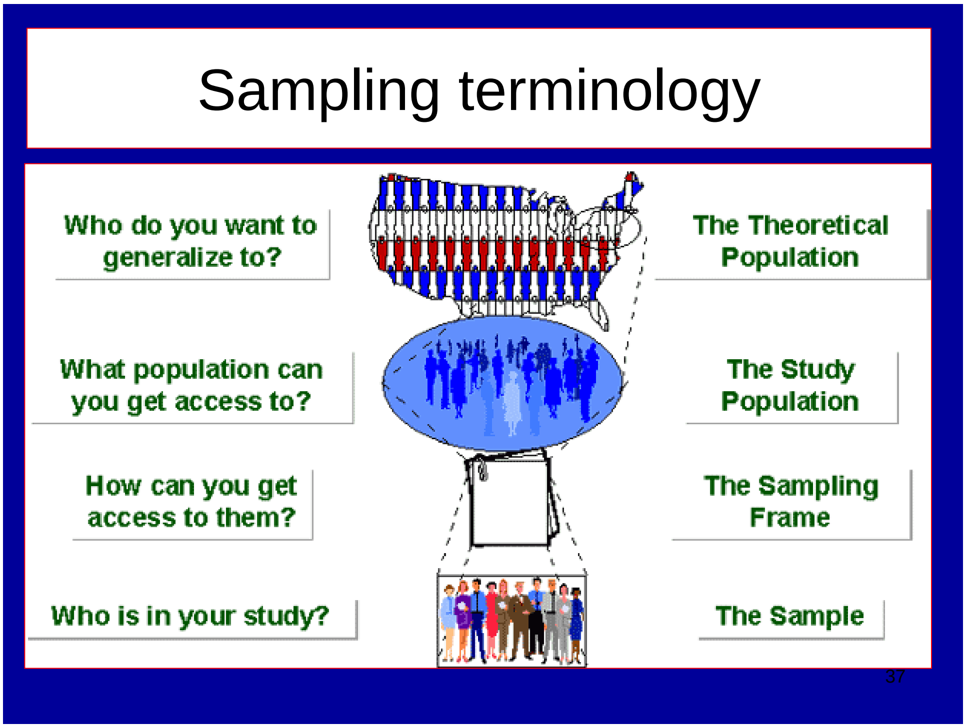
- A study of the entire **populations is impossible** in most situations.
  Example (Reproductive biology of shark fish; number of cancer patients in Malaysia)

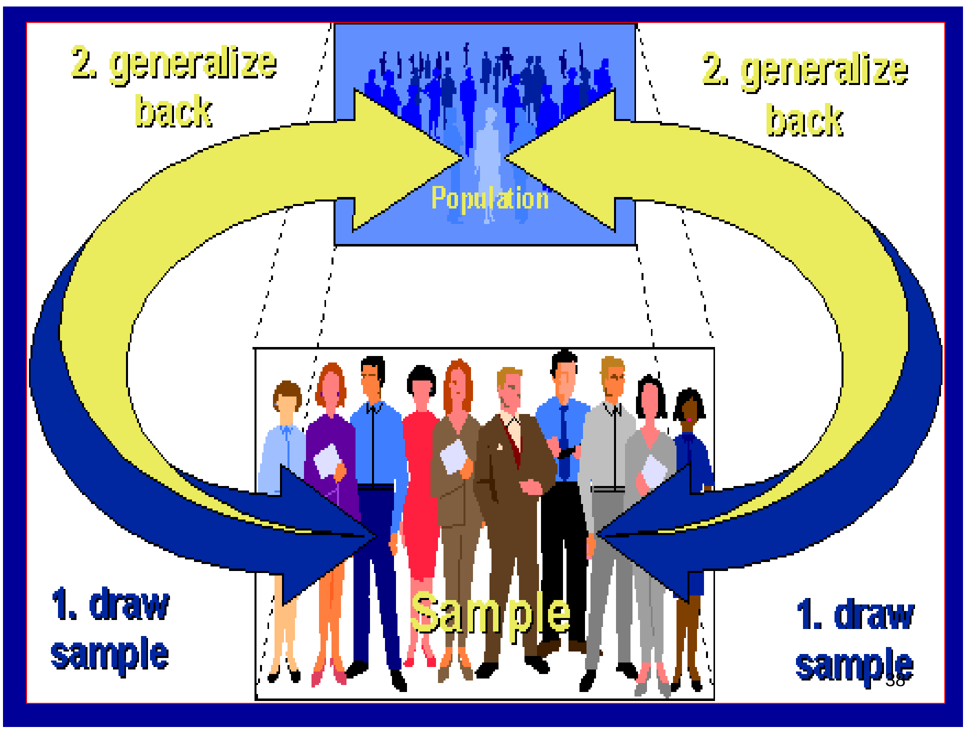
35

Sampling Terminology

- Population- all possible cases or elements that meet criteria of study (content/characteristic, unit of study, extent/geographic region, time)
- Sampling frame- list of all elements
- Sampling error- estimate of how the values of the sample differ from those of the population

36


Sampling terminology
Who do you want to generalize to?
The Theoretical Population
What population can you get access to?
The Study Population
How can you get access to them?
The Sampling Frame
Who is in your study?
The Sample
37


2. generalize back
2. generalize back
Population
Sample
1. draw sample
1. draw sample
38

Types of Sampling Methods

- Sampling
  - Non-Probability Samples
    - Consecutive
    - Convenience
    - Judgemental
  - Probability Samples
    - Simple Random
    - Systematic
    - Stratified
    - Cluster

39

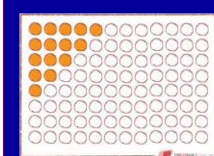
Sampling Methods
Non-probability samples

1. Convenience sampling:

- It is the process of taking those members of the accessible population who are easily available.
- It is widely used in clinical research because of its obvious advantages in **cost and logistics**.

2. Consecutive sampling:

- It involves taking every patient who meets the selection criteria over a specified time interval or number of patients. "first-come, first-chosen" basis..



3. Judgemental sampling:

- It involves hand-picking from the accessible population those individuals judged most appropriate for the study.

40

Simple Random Sampling

- Every **individual or item** from the target population has an *equal chance* of being selected.
- One may use table of random numbers or computers programs for obtaining samples.

41

Systematic Sampling

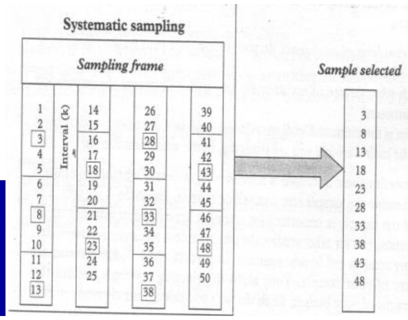
Select some starting point and then select every K^{th} element in the population

42

Systematic Sampling Procedure

Estimate HIV prevalence in children born during a specified period at a hospital

1. Impossible to construct sampling frame in advance
2. Select a random number between some pre-specified bounds
3. Beginning with the random number chosen, take every 5th birth and measure for HIV infection.

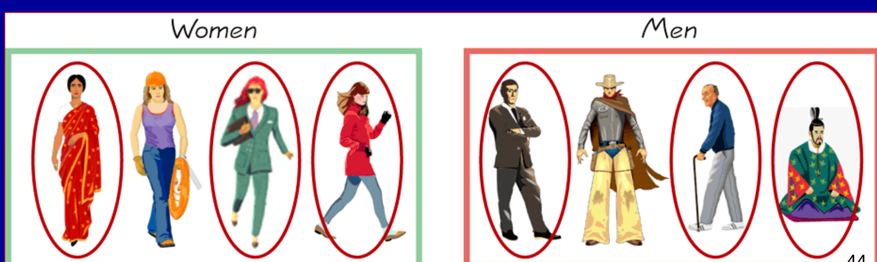


43

Stratified Sampling

subdivide the population into at least two different subgroups that share the same characteristics, then draw a sample from each subgroup (or stratum).

Examples of variables like sex, race, education, income, social class, religion, rural/urban residence.



44

Stratified Random Sampling

Assess dietary intake in adolescents

1. Define three age groups: 11-13, 14-16, 17-19
2. Stratify age groups by sex
3. Obtain list of children in this age range from schools
4. Randomly select children from each of the strata until sample size is obtained
5. Measure dietary intake

Advantage:

- Allows investigator to estimate parameters in different strata.
- If strata are homogeneous, this method is as “precise” as simple random sampling but with a smaller total sample size.

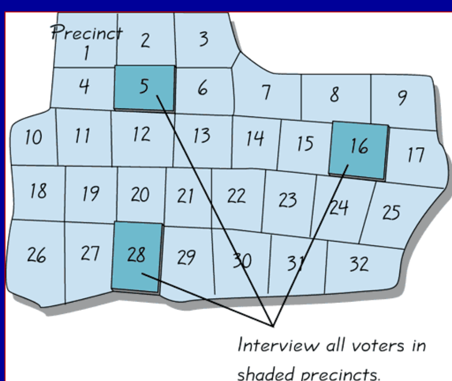
Disadvantages

- Loss of precision if small number of units is sampled from strata………..

45

Cluster Sampling

divide the population into sections
(or clusters); randomly select some of those clusters;
choose all members from selected clusters



46

Observational & Experimental Studies

- In an *observational study*, measurements of variables of interest are observed and recorded, without controlling any factor that might influence their values.

Examples: Field study, animal and plant biodiversity, disease prevalence, number of diabetics patients in SP, a survey of smoking or drinking habits among students

- An *experiment*, on the other hand, deliberately imposes some treatment on individuals in order to observe their responses. The researcher intervenes to change something (e.g., gives some patients a drug) and then observes what happens. **In an observational study there is *no* intervention.**

In principle, only experiments can give good evidence for causation.

47

Types of observational studies

- **Case series/case reports**
- **Cross - sectional study**
- **Retrospective (or case control) study**
- **Prospective (or longitudinal or cohort) study**

48

Types of Observational Study

1. Case series: A simple descriptive account of interesting characteristics observed in a group of subjects. Example: Study a group of patients with certain illness.

2. Cross Sectional Study : Data are observed, measured, and collected at one point in time. An observational study that examines a characteristic in a set of subjects at one point in time; a "snapshot" of a characteristic or condition of interest also called survey.

Example: what is the prevalence of diabetes in this community?

3. Retrospective (or Case Control) Study: An observational study that begins with patients cases who have the outcome or disease being investigated and control subjects who do not have the out come or disease. It then looks backward to identify the possible precursors or risk factors. Data are collected from the past by going back in time.

Example, study a group of patients with brain cancer and do not have brain cancer.

4. Prospective (or Longitudinal or Cohort) Study: An observational study that begins with a set of subjects who have a risk factors (or have been exposed to an agent) and as second set of subjects who do not have the risk factors or exposure. Both sets are followed prospectively through time to learn how many in each set develop the outcome or consequences of interest. Data are collected in the future from groups (called cohorts) sharing common factors.

Example: Whether using a cell phone leads to brain cancer

49

Experimental Study

- **Does the use of stents reduce the risk of stroke?**
- The researchers who asked this question collected data on 451 at risk patients. Each volunteer patient was randomly assigned to one of two groups:
  - **Treatment group:** Patients in the treatment group received a stent and medical management. The medical management included medications, management of risk factors, and help in lifestyle medication.
  - **Control group:** Patients in the control group received the same medical management as the treatment group, but they did not receive stents.

50

Experimental study

- Researchers studied the effect of stents at two time points: 30 days after enrollment and 365 days after enrollment.

	0-30 days		0-365 days	
	stroke	no event	stroke	no event
treatment	33	191	45	179
control	13	214	28	199
Total	46	405	73	378

Table 1.2: Descriptive statistics for the stent study.

51

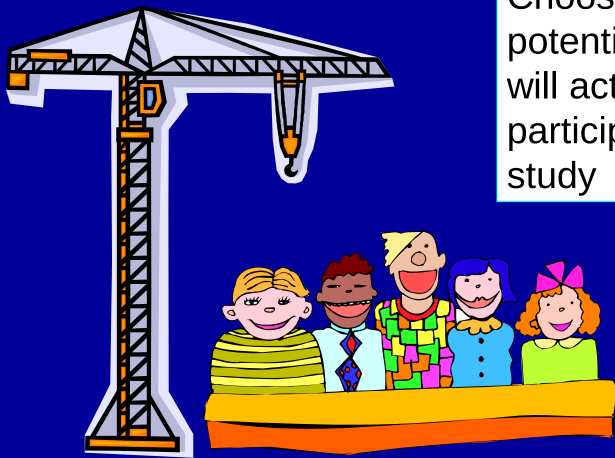
Random Selection vs. Random Assignment

– Random Selection = every member of the population has an equal chance of being selected for the sample.

– Random Assignment = every member of the sample (however chosen) has an equal chance of being placed in the experimental group or the control group.

52

Subject Selection (Random Selection)



Choosing which potential subjects will actually participate in the study

53

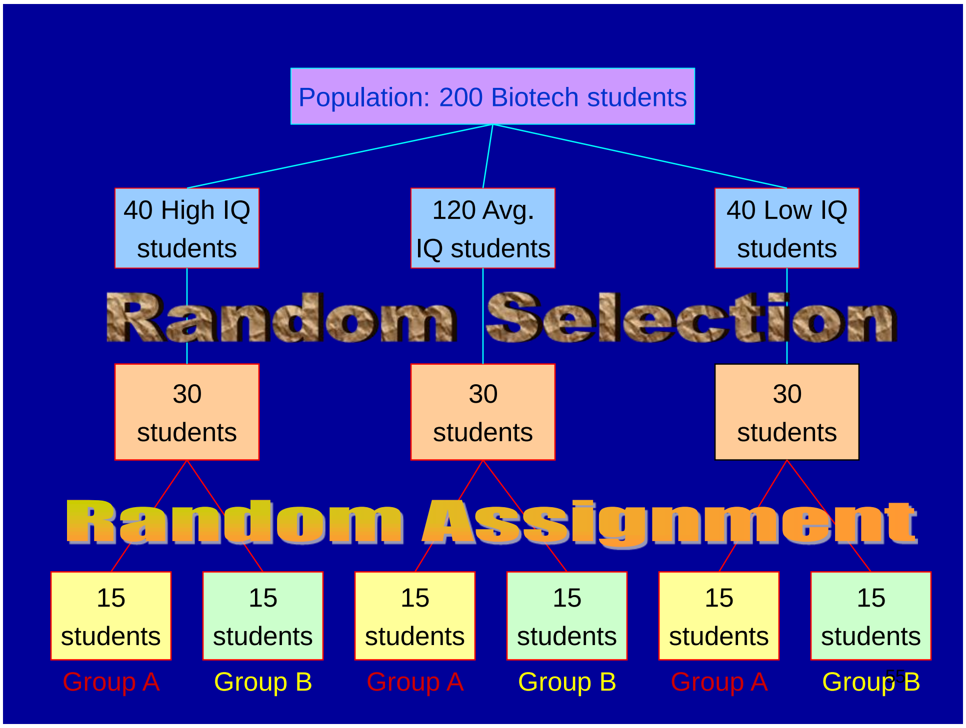
Subject Assignment (Random Assignment)

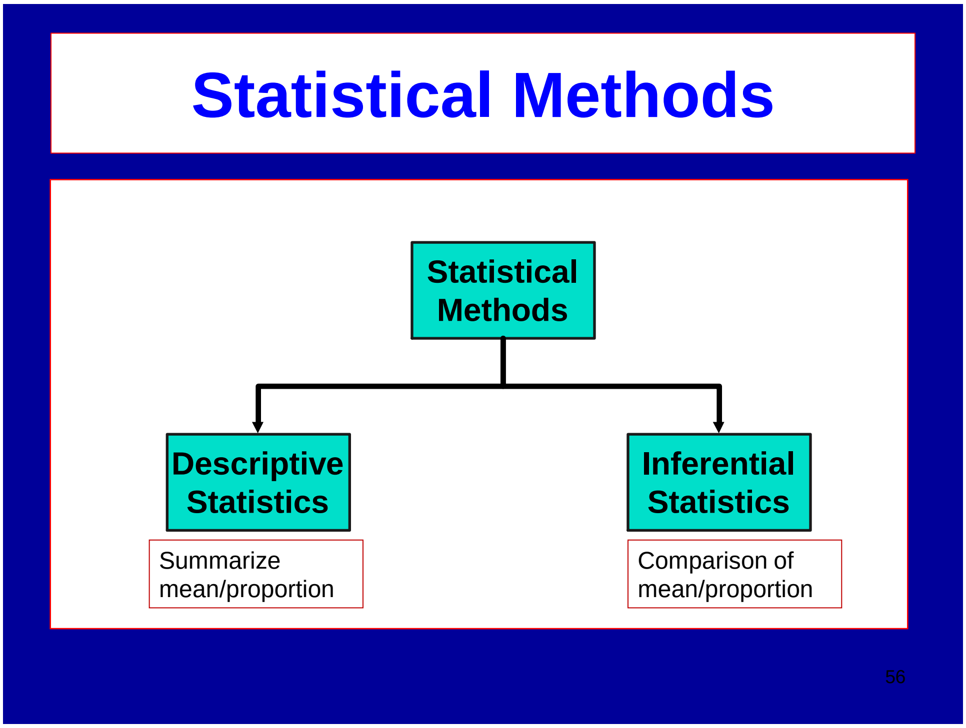
Deciding which group or condition each subject will be part of

Group A

Group B

54


Population: 200 Biotech students
40 High IQ students
120 Avg. IQ students
40 Low IQ students
Random Selection
30 students
30 students
30 students
Random Assignment
15 students
15 students
15 students
15 students
15 students
15 students
Group A
Group B
Group A
Group B
Group A
Group B
55


Statistical Methods
Statistical Methods
Descriptive Statistics
Summarize mean/proportion
Inferential Statistics
Comparison of mean/proportion
56

Statistical Methods

- **Descriptive statistics** generally characterizes or describes a set of data elements by **graphically displaying the information** or describing its **central tendencies** and how it is distributed.
- It describes patterns and general trends in a data set.
  - Typically the data are **reduced** down to one or two descriptive summaries like the **mean** and **standard deviation** or **correlation**, or by **visualization** of the data through various graphical procedures like histograms, frequency distributions, and scatterplots

- **Inferential statistics** tries to infer information about a population by using information gathered by sampling. Use sample data to study associate, or to compare differences or predictions about a larger set of data.

57

Tables & Graphs

- **Tables & graphs used to summarize data to communicate information.**

- **Tables -** good for showing exact values, small amounts of data and/or multiple localized comparisons.

- **Graphs –** used to show and present **qualitative trends and large amounts of data**.
  - It is a mathematical picture.
  - Graphic representation of data proves quite an effective and economical device for the presentation, understanding and interpretation of statistical data

58

Principles of Tabulation

- Every table should have a clear, concise and adequate title.
- Every table should be given a distinct number to facilitate easy reference.
- The column heading and row heading of the table should be clear and brief.
- Abbreviation should be avoided to the extent possible. If you give need to explain below the table as foot note.
- Units of measurements should be given, (µg/ml, % or g, mg)
- Source from which the data in the table have been obtained must be indicated just below the table.

59

Sample of Table

Table 1
Comparison of enzyme activities (mean±S.E.M.) of fish species Arctic char, brook trout, koi carp, striped bass, haddock, cod and hagfish (n=15 except koi carp where n=6)

Fish species	Lysozyme activity (U/mg protein)	Alkaline phosphatase activity (U/mg protein)	Cathepsin B activity (U/mg protein)	Protease activity (U/mg protein)
Freshwater species				
Arctic char	24.1±1.1	0.17±0.03	3.04±0.18	65.5±13.5
Brook trout	16.8±0.5	0.5±0.1	4.9±0.6	216.0±61.0
Koi carp	42.0±1.1	2.9±0.4	1176.0±60.0	99.4±14.7
Striped bass	47.7±2.6	0.6±0.1	2.4±0.2	40.5±4.3
Seawater species				
Haddock	88.1±2.6	0.9±0.1	0.61±0.02	10.8±0.6
Cod	63.1±2.4	0.32±0.05	8.4±0.6	15.1±0.7
Hagfish	124.7±5.3	1.32±0.36	65.0±3.2	818.0±105.0

60

Table 4
Screening of active acidic mucus extracts of brook trout, haddock and hagfish against human and fish pathogens

Microbial strains	MBC (µg protein/mL)		
	Brook Trout	Haddock	Hagfish
Human pathogens			
Escherichia coli D31	19.0	14.0	6.1
Salmonella enterica Serovar Typhimurium C610	10.0	14.0	8.3
Staphylococcus epidermis C621	136.5	192.5	82.5
Pseudomonas aeruginosa Z61	34.1	96.2	21.0
P. aeruginosa K799	273.0	192.5	41.2
Candida albicans C627	136.5	192.5	41.2
Fish pathogens			
Aeromonas salmonicida sub sp. salmonicida A449	19.0	27.0	8.3
Listonella anguillarum 02-11	39.0	27.0	16.0
Yersinia ruckeri 96-4	19.0	14.0	6.1

Minimal bactericidal concentration (MBC) is the minimum concentration of mucus protein (µg) that exhibited no observable bacterial growth.

61

Gel image

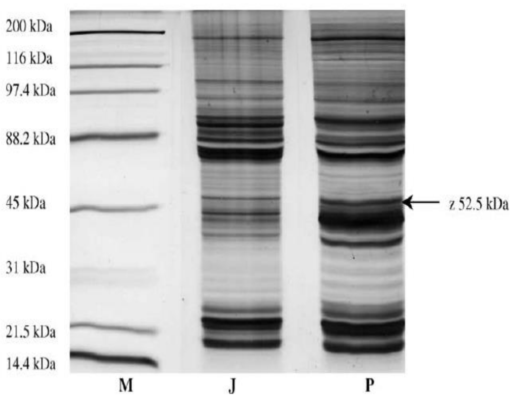


Fig. 2. SDS-PAGE gel showing protein profile of discus mucus from juvenile (J) and female parent during day 10–15 of free-swimming larvae (P). Molecular weight marker (M) used were Broad Range (BIORAD®).

62

Graph

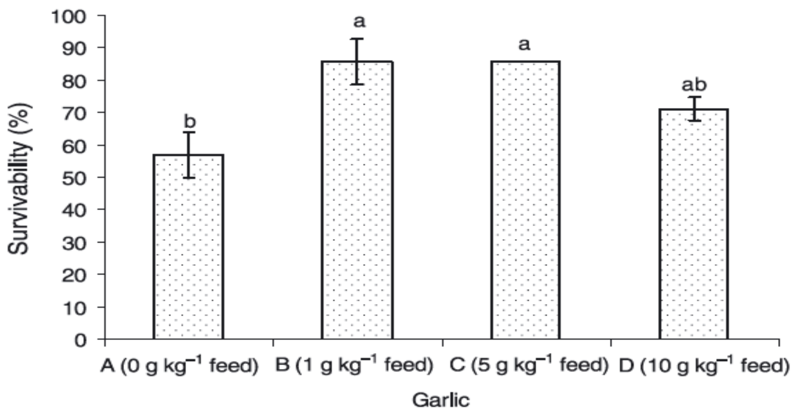


Fig. 7. Effect of garlic on survivability of *L. rohita* after bacteria challenge (values are mean ± SE). Mean values bearing same superscript are not statistically significant, $P > 0.05$ ($n = 28$)

63

Important Characteristics of Data

1. **Center:** A representative or average value that indicates where the middle of the data set is located
2. **Variation:** A measure of the amount that the values vary among themselves
3. **Distribution:** The nature or shape of the distribution of data (such as bell-shaped, uniform, or skewed)
4. **Outliers:** Sample values that lie very far away from the vast majority of other sample values

64

Frequency Distributions

❖ Frequency Distribution

- ❖ Lists data values (either individually or by groups of intervals), along with their corresponding frequencies or counts..

Example: Survey of blood group; body weight; height etc..

Table 1. Distribution of Blood Group of 62 Students of AIMST

Blood Group	Frequency	Relative Frequency (%)
A	8	13.0
B	24	38.7
AB	3	4.8
O	27	43.5
Total	**62**	**100.0**

65

Frequency Distributions

Table 2-1 Measured Cotinine Levels in Three Groups

Smoker: The subjects report tobacco use.

ETS: (Environmental Tobacco Smoke) Subjects are nonsmokers who are exposed to environmental tobacco smoke ("secondhand smoke") at home or work.

NOETS: (No Environmental Tobacco Smoke) Subjects are nonsmokers who are not exposed to environmental tobacco smoke at home or work. That is, the subjects do not smoke and are not exposed to secondhand smoke.

Smoker:	1	0	131	173	265	210	44	277	32	3
	35	112	477	289	227	103	222	149	313	491
	130	234	164	198	17	253	87	121	266	290
	123	167	250	245	48	86	284	1	208	173
ETS:	384	0	69	19	1	0	178	2	13	1
	4	0	543	17	1	0	51	0	197	3
	0	3	1	45	13	3	1	1	1	0
	0	551	2	1	1	1	0	74	1	241
NOETS:	0	0	0	0	0	0	0	0	0	0
	0	9	0	0	0	0	0	0	244	0
	1	0	0	0	90	1	0	309	0	0
	0	0	0	0	0	0	0	0	0	0

Table 2-2 Frequency Distribution of Cotinine Levels of Smokers

Cotinine	Frequency
0–99	11
100–199	12
200–299	14
300–399	1
400–499	2
Total	40

66

Relative Frequency Distribution

$$\text{Relative frequency} = \frac{\text{class frequency}}{\text{sum of all frequencies}}$$

Cotinine	Frequency
0–99	11
100–199	12
200–299	14
300–399	1
400–499	2

Total Frequency = 40

Table 2-3
Relative Frequency Distribution of Cotinine Levels in Smokers

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

11/40 = 28%

12/40 = 40%

etc.

67

Cumulative Frequency Distribution

Cotinine	Frequency
0–99	11
100–199	12
200–299	14
300–399	1
400–499	2

Table 2-4
Cumulative Frequency Distribution of Cotinine Levels in Smokers

Cotinine	Cumulative Frequency
Less than 100	11
Less than 200	23
Less than 300	37
Less than 400	38
Less than 500	40

Cumulative Frequencies

68

Frequency Tables

Table 2-2
Frequency Distribution of Cotinine Levels of Smokers

Cotinine	Frequency
0–99	11
100–199	12
200–299	14
300–399	1
400–499	2

Table 2-3
Relative Frequency Distribution of Cotinine Levels in Smokers

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

Table 2-4
Cumulative Frequency Distribution of Cotinine Levels in Smokers

Cotinine	Cumulative Frequency
Less than 100	11
Less than 200	23
Less than 300	37
Less than 400	38
Less than 500	40

Descriptive Statistics

Measures of *central tendency &* Measures of *dispersion*

- **Measures of central tendency or centre are summary statistics that summarize the average value of a set of measurements.**
- **Measures of dispersion or spread are summary statistics that indicate the spread of the data**
- **Average body weight of biotech students**

Mean ± SD

65 ± 10 kg

Measures of Central Tendency

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:

Mean, Median, and Mode

Mean = $\frac{\text{sum of all values}}{\text{total number of values}}$

Median = *middle value (when the data are arranged in order)*

Mode = *most common value*

Measures of Dispersion

- **Range:** is the difference between the largest and smallest observations in the sample
- **Variance: s^2,** is the arithmetic mean of the squared deviations from the sample mean

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

- **Standard deviation: How far each and every observation deviates from the mean and is** the square-root of the variance

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

$$S = \sqrt{\frac{\Sigma (X - \bar{X})^2}{N}}$$

where S = the standard deviation of a sample,
Σ means "sum of,"
X = each value in the data set,
$\bar{X}$ = mean of all values in the data set,
N = number of values in the data set.

- **Coefficient of variation: is** the sample standard deviation expressed as a percentage of the mean. CV is a useful way of comparing the dispersion of variables measured on different scales

$$\text{CoV} = \frac{s}{\bar{x}} \times 100$$

$$Coefficient\ of\ Variation = \frac{Standard\ Deviation}{Mean} \times 100$$

$$Coefficient\ of\ Variation = \frac{\sigma}{\mu} \times 100$$

73

Measures of Dispersion

A measure of dispersion conveys information regarding the amount of variability present in a set of data.

Note:

1. If all the values are the same
   → There is no dispersion .
2. If all the values are different
   → There is a dispersion:
3. If the values close to each other
   →The amount of Dispersion small.
4) If the values are widely scattered
   → The Dispersion is greater.

74

Difference between standard error and standard deviation

Standard deviation (SD): This describes the spread of values in the sample. The sample standard deviation, *s*, is a random quantity -- it varies from sample to sample.

$$S = \sqrt{\frac{\Sigma(X - \overline{X})^2}{N}}$$

where S = the standard deviation of a sample,
Σ means "sum of,"
X = each value in the data set,
$\overline{X}$ = mean of all values in the data set,
N = number of values in the data set.

Standard error of the mean (SE): This is the **standard deviation of the sample mean** and describes its accuracy as an estimate of the population mean. The SEM quantifies how precisely you know the true mean of the population.

$$SE_{\overline{x}} = \frac{s_x}{\sqrt{n}}$$

75

Coefficient of Variation

Mean height and body weight of 50 final year medical students.

	Height	Weight
Mean	**176.57**	**72.63**
SD	**10.91**	**11.94**
C.V.	**6.18%**	**16.44%**

$$CoV = \frac{s}{\overline{x}} \times 100$$

Weight has variation than height

76

Locating Extreme Outliers: Z-Score

- To compute the Z**-score** of a data value, subtract the mean and divide by the standard deviation.
- The Z-score is the number of standard deviations a data value is from the mean.
- **A data value is considered an extreme outlier if its Z-score is less than -3.0 or greater than +3.0.**
- The larger the absolute value of the Z-score, the farther the data value is from the mean.

77

Locating Extreme Outliers: Z-Score

$$Z = \frac{X - \overline{X}}{S}$$

where X represents the data value

$\overline{X}$ is the sample mean

S is the sample standard deviation

78

Locating Extreme Outliers: Z-Score

- Suppose the mean math SAT score is 490, with a standard deviation of 100.
- Compute the Z-score for a test score of 620.

$$Z = \frac{X - \bar{X}}{S} = \frac{620 - 490}{100} = \frac{130}{100} = 1.3$$

A score of 620 is 1.3 standard deviations above the mean and would not be considered an outlier.

79

Shape of a Distribution

- Describes how data are distributed
- Two useful shape related statistics are:
  - **Skewness**
    - Measures the **amount of asymmetry** in a distribution
  - **Kurtosis**
    - Measures the **relative concentration of values** in the center of a distribution as compared with the tails

80

Shape of a Distribution

Skewness: Indicator used in distribution analysis as a sign of asymmetry and deviation from a normal distribution.

Interpretation:

- Skewness > 0 - Right skewed distribution - most values are concentrated on left of the mean, with extreme values to the right.
- Skewness < 0 - Left skewed distribution - most values are concentrated on the right of the mean, with extreme values to the left.
- Skewness = 0 - mean = median, the distribution is symmetrical around the mean.

Kurtosis: Indicator used in distribution analysis as a sign of flattening or "peakedness" of a distribution.

Interpretation:

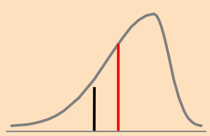
- Kurtosis > 3 - Leptokurtic distribution, sharper than a normal distribution, with values concentrated around the mean and thicker tails. **This means high probability for extreme values.**
- Kurtosis < 3 - Platykurtic distribution, flatter than a normal distribution with a wider peak. The probability for extreme values is less than for a normal distribution, and the values are wider spread around the mean.
- Kurtosis = 3 - Mesokurtic distribution - normal distribution for example.

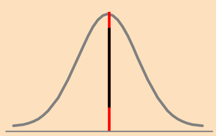
81

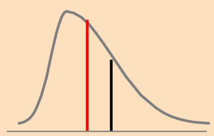
Shape of a Distribution (Skewness)

- Describes the amount of asymmetry in distribution
  - Symmetric or skewed

Left-Skewed	Symmetric	Right-Skewed
Mean < Median	**Mean = Median**	**Mean > Median**





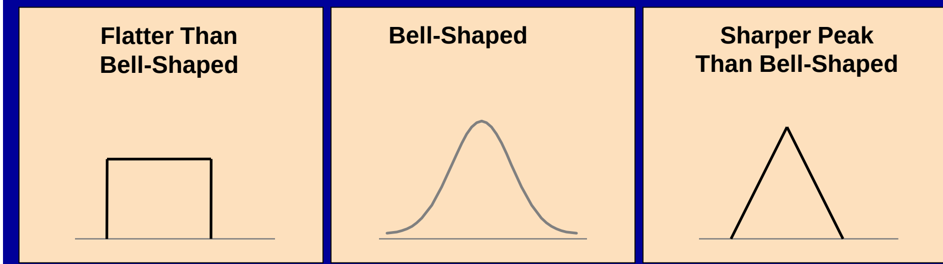


Skewness Statistic	< 0	0	>0

82

Shape of a Distribution (Kurtosis)

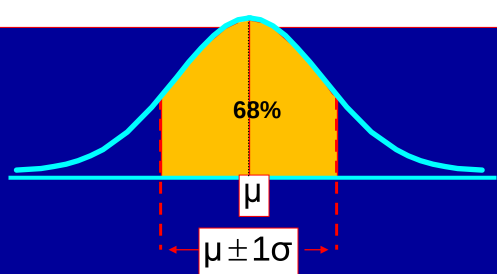
- Describes relative concentration of values in the center as compared to the tails



Kurtosis Statistic	< 0	0	>0

The Empirical Rule

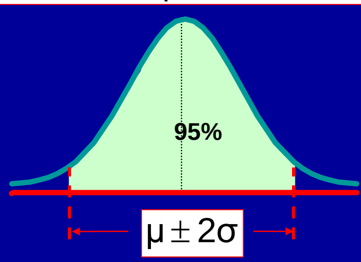
- The empirical rule approximates the variation of data in a bell-shaped distribution
- Approximately 68% of the data in a bell shaped distribution is within ± one standard deviation of the mean or $\mu \pm 1\sigma$

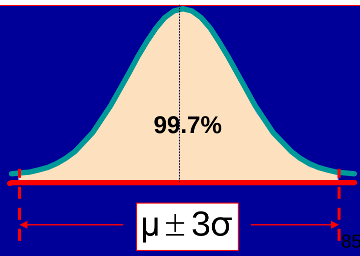


The Empirical Rule

- Approximately 95% of the data in a bell-shaped distribution lies within ± two standard deviations of the mean, or $\mu \pm 2\sigma$

- Approximately 99.7% of the data in a bell-shaped distribution lies within ± three standard deviations of the mean, or $\mu \pm 3\sigma$





85

Using the Empirical Rule

- Suppose that the variable exam scores is bell-shaped with **a mean of 500** and a standard deviation of 90. Then,
  - **68%** of all test takers scored between 410 and 590 **(500 ± 90).**
  - **95%** of all test takers scored between 320 and 680 **(500 ± 180).**
  - **99.7%** of all test takers scored between 230 and 770 **(500 ± 270).**

86

Hypothesis testing

What is hypothesis testing

- In statistics, a hypothesis is a claim or statement about a property of a population.
- A hypothesis test (or test of significance) is a standard procedure for testing a claim about a property of a population.

87

Hypothesis Testing

For example:

- Students who receive counseling will show a greater increase in creativity than students not receiving counseling"

Or

- "the automobile *A* is performing as well as automobile *B*."
- Treatment group perform better than control group..

88

Hypothesis Testing

– population mean

Example: The mean monthly cell phone bill of AIMST student is μ = RM 50

– population proportion

Example: The proportion of male population in AIMST university is = 0.68

89

BASIC CONCEPTS CONCERNING TESTING OF HYPOTHESIS

- **Null hypothesis and alternative hypothesis**
- **The level of significance**
- **Type I and Type II errors**
- **Two-tailed and One-tailed tests**

90

Null hypothesis and alternative hypothesis

The process of choosing between the null and alternative hypotheses is called **hypothesis testing.**

$$H_0 : \mu \leq 115$$

Null Hypothesis

$$H_a : \mu > 115$$

Alternative Hypothesis

91

Null hypothesis

- The null hypothesis is generally symbolized as $\boldsymbol{H_0}$ and the alternative hypothesis as $\boldsymbol{H_a}$.
- Suppose we want to test the hypothesis that the population mean (μ) is equal to the hypothesized mean **(μ_{H0}) = 100.**

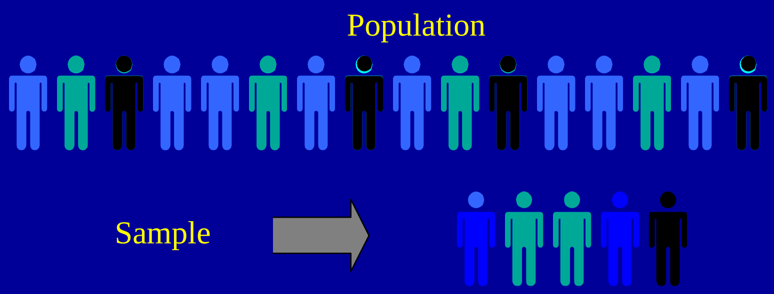
$H_0 : \mu = \mu_{H0} = 100$

If our sample results do not support this null hypothesis we should conclude rejecting the null hypothesis and accepting alternative hypothesis.

92

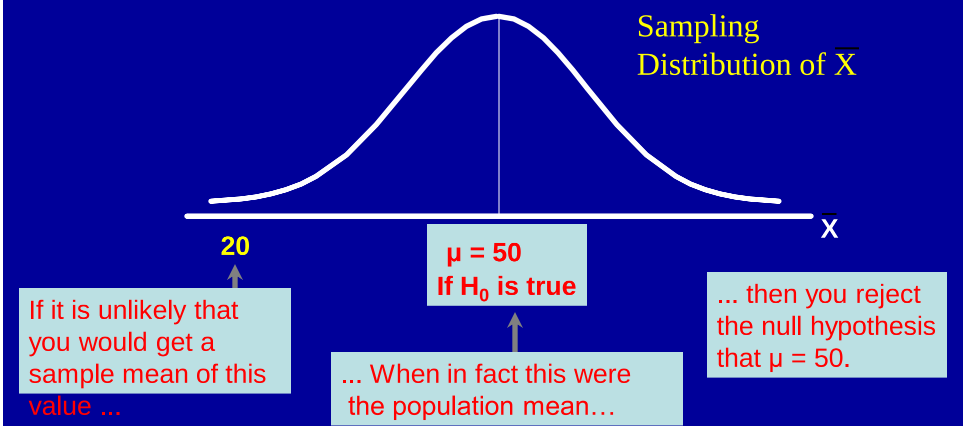
The Hypothesis Testing Process

- Claim: The population mean age is 50.
  H_0: μ = 50, H_1: μ ≠ 50
- Sample the population and find sample mean.



93

The Hypothesis Testing Process



94

Null Hypotheses and Alternate Hypotheses

- State Null Hypotheses and Alternate Hypotheses
- Positive relationship use **>** sign (one tail) **Right tailed test**

$H_0: \mu = \mu_{H_0}$ and $H_a: \mu > \mu_{H_0}$

- Negative relationship use **<** sign (one tail) **left tailed**

$H_0: \mu = \mu_{H_0}$ and $H_a: \mu < \mu_{H_0}$

- No clear relationship use **≠** sign (two tail) **Two sided**

$H_0: \mu = \mu_{H_0}$ and $H_a: \mu \neq \mu_{H_0}$

- **Null:** The mean birth weight of 100 CMV infected babies is equal to **3060.75g**
- **Alternate:** The mean birth weight of 53 CMV infected babies is not equal to 3060.75g

95

Type I and Type II errors:

- There are basically two types of errors we can make.
- **Type I error** means rejection of true null hypothesis which should have been accepted and **Type II error** means accepting the hypothesis which should have been rejected.
- Type I error is denoted by alpha error, and Type II error is denoted by (beta) known as b error.

96

Possible Errors in Hypothesis Test Decision Making

- **Type I Error**
  - Reject a true null hypothesis
  - Considered a serious type of error (punishing a innocent)
  - The probability of a Type I Error is α
    - Called level of significance of the test
    - Set by researcher in advance (0.05; 0.01, 0.001)
- **Type II Error**
  - Failure to reject a false null hypothesis
  - The probability of a Type II Error is β

97

Type I and Type II errors:

Table 7-1 Type I and Type II Errors

		True State of Nature	
		The null hypothesis is true.	The null hypothesis is false.
Decision	We decide to reject the null hypothesis.	**Type I error** (rejecting a true null hypothesis) α	Correct decision
	We fail to reject the null hypothesis.	Correct decision	**Type II error** (failing to reject a false null hypothesis) β

- If type I error is fixed at 5 per cent, it means that there are about 5 chances in 100 that we will reject *H*0 (null hypothesis) when *H*0 is true.
- For instance, if we fix it at 1 per cent, we will say that the maximum probability of committing Type I error would only be 0.01.

98

The level of significance – P value

- **The probability of the null hypothesis is TRUE**
- We take the significance level at 5 per cent, then this implies that H_0 will be rejected when the sampling result (i.e., observed evidence) has a less than 0.05 probability of occurring if H_0 is true.

99

Confidence interval and P value

- **P value is the probability of your null hypothesis is TRUE.**
- **Confidence interval** (CI) is a type of **interval** estimate of a population parameter.
  - **How confident you are about your null hypothesis is true. Whether 95% or 99%**

100

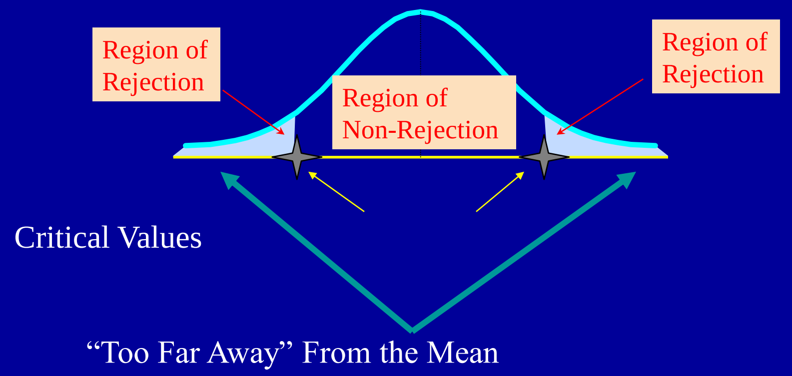
The Test Statistic and Critical Values

- If the sample mean is close to the stated population mean, the null hypothesis is not rejected.
- If the sample mean is far from the stated population mean, the null hypothesis is rejected.

101

The Test Statistic and Critical Values

Sampling Distribution of the test statistic



102

Confidence interval and P value

Example

- The mean birth weight of 53 CMV (**Cytomegalovirus)** infected babies was 3060.75g (standard deviation = 601.03g, standard error = 82.57g).
- A 95% confidence interval for the population mean birth weight of CMV infected babies is therefore will be
- (3060.75 ± **1.96**(82.57)g) = (2898.91, 3222.59g)
- Similarly, the 99% confidence interval for the mean is:
- (3060.75 ± **2.58**(82.57)g) = (2847.72,3273.78g)
- We are 95% confident that the true mean is somewhere **between 2898.91 and 3222.59g.**
- We are 99% confident that the true mean is between **2847.72 and 3273.78g**

Confidence level	Z value
90%	1.65
95%	1.96
99%	2.58
99,9%	3.291

103

Reporting Significance

Report p values as being less than .05, .01, or .001.

If a result is not significant, report p as being greater than .05 ($p > .05$)

Here are some examples…

if $p = .017$	report $p < .05$	We conclude that group means are significantly different
if $p = .005$	report $p < .01$	We conclude that group means are significantly different
If $p = .24$	report $p > .05$	We conclude that group means are NOT significantly different

104

Hypotheses & Significance

- If *p* value is <u>significant</u> ($p < .05$)
  - Reject the Null hypothesis
- If *p* value is <u>not significant</u> ($p > .05$)
  - Failure to reject the Null hypothesis

105

SAMPLE SIZE

- The sample size of a statistical sample is the number of observations that constitute it.
- Determining the sample size to be selected is an important step in any research study.
  - Example: If you want to determine prevalence of eye problems in school children and wants to conduct a survey.
  - **"How many participants should be chosen for a survey"?**

106

What Should Be the Sample Size?

- The choosing of sample size depends on non-statistical and statistical considerations.
- The non-statistical considerations may include
  - availability of resources, manpower, budget, ethics and sampling frame.
- The statistical considerations will include the desired **precision of the estimate of prevalence and the expected prevalence** of eye problems in school children.

107

What Should Be the Sample Size?

1. **The Level of Precision**
   - **The range in which the true value of the population is estimated to be. The % of prevalence of eye problem in population. Whether high or low?**

2. The Confidence Level

- **If a confidence interval is 95%, it means 95 out of 100 samples will have the true population value within range of precision.**

3. Degree of Variability

- **The more heterogeneous a population is, the larger the sample size is required to get an optimum level of precision.**

108

Determining Sample Size

Formula:

Means $n = (ZS/E)^2$

Proportions $n = Z^2 pq / E^2$

Z at 95% confidence = 1.96

Z at 99% confidence = 2.58

Let n = sample size

S = standard deviation

Z = confidence level (ex. 95% confidence = 1.96),

E = range of possible random error (how much error you are willing to accept)

p = estimated proportion of successes

q = 1 – p, or estimated proportion of failures

pc = percentage

109

Standard deviation VS Standard error of the mean

- The term **"standard deviation"** refers to the variability in individual observations in a single sample (s) or population (σ)
- The **standard error of the mean** is also a measure of standard deviation, but not of individual values, rather variation in multiple sample means computed on multiple random samples of the same size, taken from the same population

110

Parametric tests VS Nonparametric tests

- Parametric tests **have requirements about the nature or shape of the populations involved.**
- **Nonparametric tests do not require that samples come from populations with normal distributions or have any other particular distributions. Hence, nonparametric tests are called distribution-free tests.**

111

Parametric Test Procedures

1. Involve Population Parameters (Mean)
2. Have Stringent Assumptions (Normality)
3. Examples: t Test, ANOVA

112

Nonparametric Test Procedures

1. Do Not Involve Population Parameters
   Example: Probability Distributions,
2. Data Measured on Any Scale (Ratio or Interval, Ordinal or Nominal)
3. When the Outcome is a Rank
4. When there are definite outliers

Example: Wilcoxon Rank Sum Test

113

Some Commonly Used Statistical Tests

Normal distribution based test	Corresponding nonparametric test	Purpose of test
t test for independent samples	Mann-Whitney U test; Wilcoxon rank-sum test	**Compares two independent samples**
Paired *t* test	Wilcoxon matched pairs signed-rank test	**Examines a set of differences**
Pearson correlation coefficient	Spearman rank correlation coefficient	**Assesses the linear association between two variables.**
One way analysis of variance (*F* test)	Kruskal-Wallis analysis of variance by ranks	**Compares three or more groups**
Two way analysis of variance	Friedman Two way analysis of variance	**Compares groups classified by two different factors**

114

Comparison of means
t- tests

- **One-sample t-test:**
  - Used to compare **one sample mean** to a **population mean** or some other known value.

  Examples:
  Average birth weight of new born baby in Malaysia
  Average daily energy intake over 10 days of healthy women.

- **Independent sample t test:** **used to** test the **means of two normally distributed** populations are equal or not.

  - Example: Hemoglobin levels in male and female is same or not?
  - Body fat content in pig fed with two different diets
  - Birth weight of children born to 15 non smoking mother with heavy smoking mothers

- **Paired sample t test for Repeated measures:** **Same individuals are studied more than once in different circumstances**
  - Blood glucose levels before and after fasting
  - Weight loss for dieting

115

Analysis of variance (ANOVA)

- Analysis of variance **(abbreviated as ANOVA)** is an extremely useful technique concerning researches in the fields of **economics, biology, education, psychology, sociology, business/industry**.
- Analysis of variance compares three or more populations/treatment data.
- Specifically, we are interested in determining whether differences exist between the population/treatment means.
- The procedure works by analyzing the sample variance.

116

Examples - Research Problem

- Comparing the yield of crop from several varieties of seeds,
- Total phenolic contents in five different plants
- Effect of different temperature on hatchability of eggs
- **The gasoline mileage of four automobiles.**
- **Effect of different temperature on hatchability of eggs**
- **Effect of different cooking methods on proximate and mineral composition in snakehead fish**

117

One way ANOVA

- In a one-way anova, there is one measurement variable (dependant) and one nominal variable (independent).
- Multiple observations of the measurement variable are made for each value of the nominal variable.
  - For example, you could measure the amount of protein for multiple samples taken from arm muscle, heart muscle, brain, liver, and lung.
  - The amount of protein would be the **measurement variable,** and the tissue type (arm muscle, brain, etc.) would be the **nominal variable.**
  - **You can test which tissue type has more amount of protein.**

118

Hypotheses of One-Way ANOVA

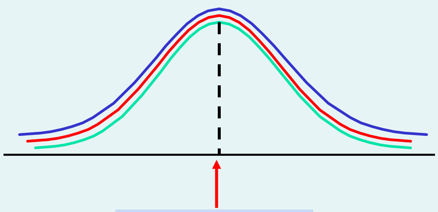
- $H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_c$
  - All population means are equal
  - i.e., no factor effect (no variation in means among groups)
- $H_1 :$ Not all of the population means are the same
  - At least one population mean is different
  - i.e., there is a factor effect
  - Does not mean that all population means are different (some pairs may be the same)

One-Way ANOVA

$H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_c$

$H_1 :$ Not all μ_j are the same

The Null Hypothesis is True
All Means are the same:
(No Factor Effect)



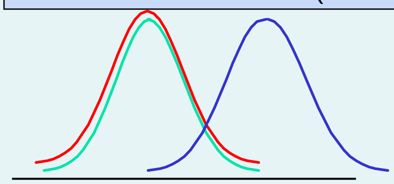
$\mu_1 = \mu_2 = \mu_3$

One-Way ANOVA

$$H_0: \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_c$$

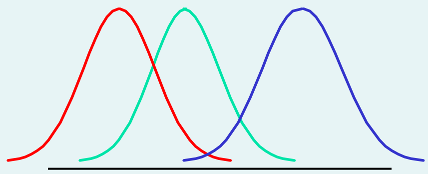
$$H_1: \text{Not all } \mu_j \text{ are the same}$$

The Null Hypothesis is NOT true
At least one of the means is different
(Factor Effect is present)



$\mu_1 = \mu_2 \neq \mu_3$

or



$\mu_1 \neq \mu_2 \neq \mu_3$

Chap 11-121

ANOVA versus *t* tests

- *t* test used for comparing means for 2 groups

- ANOVA used for comparing means for more than 2 groups

122

EXAMPLE - One way ANOVA

- **Effect of dietary protein level on the reproductive performance of female swordtails *Xiphophorus helleri* (Poeciliidae).**
- Five isocaloric semi-purified diets containing 20%, 30%, 40%, 50% and 60% dietary protein were used.
- **Data and statistical analysis:**
  - Comparison of various growth and reproductive parameters from different dietary treatments was carried out using analysis of variance (ANOVA) with **Tukey's test** was used to test the effect on the treatment.

Table 2
Mean values (± S.E.) of various growth parameters of female swordtail fed different levels of dietary protein

	Diet				
	20P	30P	40P	50P	60P
Initial weight (g)	1.17 ± 0.04	1.13 ± 0.07	1.15 ± 0.08	1.20 ± 0.09	1.19 ± 0.08
Final weight (g)	2.95 ± 0.05a	3.52 ± 0.04ab	3.93 ± 0.19b	4.14 ± 0.10bc	4.35 ± 0.24b
Weight gain (g)	1.79 ± 0.04a	2.39 ± 0.05b	2.78 ± 0.15b	2.94 ± 0.09c	3.16 ± 0.17c
SGR (%)	0.94 ± 0.01a	1.16 ± 0.02b	1.25 ± 0.11bc	1.26 ± 0.09bc	1.32 ± 0.27c
FCR	2.45 ± 0.23a	2.28 ± 0.28a	2.07 ± 0.09b	2.02 ± 0.05b	2.22 ± 0.17ab

Mean values in similar row with different letters are significantly different (Tukey's HSD, $P < 0.05$).

123

How to interpret the ANOVA results output in graph?

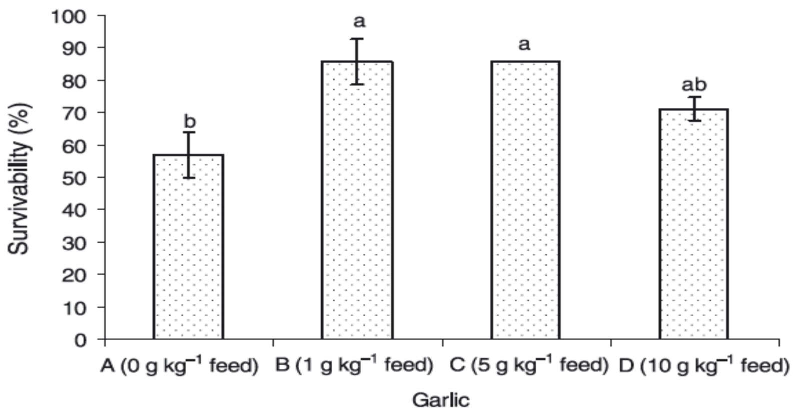


Fig. 7. Effect of garlic on survivability of *L. rohita* after bacteria challenge (values are mean ± SE). Mean values bearing same superscript are not statistically significant, $P > 0.05$ ($n = 28$)

124

Two-way ANOVA

- Two-way ANOVA technique is used when the data are classified on the basis of two factors.
  - For example, the agricultural output may be classified on the basis of different **varieties of seeds** and also on the basis of **different varieties of fertilizers** used.

Per Acre Production Data of Wheat

(*in metric tonnes*)

Varieties of seeds	*A*	*B*	*C*
Varieties of fertilizers			
W	6	5	5
X	7	5	4
Y	3	3	3
Z	8	7	4

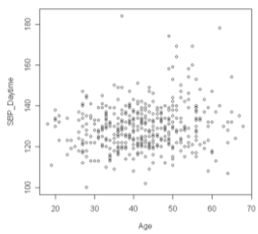
Also state whether variety differences are significant at 5% level.

125

When to use Correlation

- Correlation is used when you have two measurement variables, such as
  - **Food intake and weight,**
  - **Drug dosage and blood pressure,**
  - **Age and blood pressure**
  - **Body length and body weight**
  - **Body length and fecundity**
  - **Body weight and GSI**

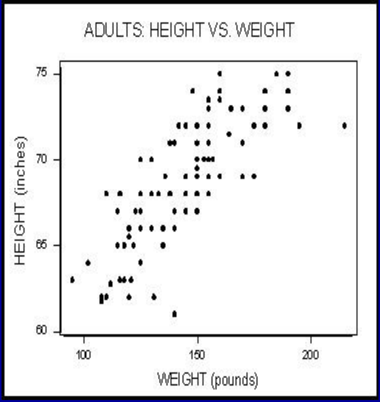
Age versus Systolic Blood Pressure in a Clinical Trial



126

Correlation

- A correlation can indicate:
  - Whether **is there any relationship** between the two variables.
  - The direction of the relationship, **i.e. whether it is positive or negative**.
  - The **strength, or magnitude of the relationship.**



127

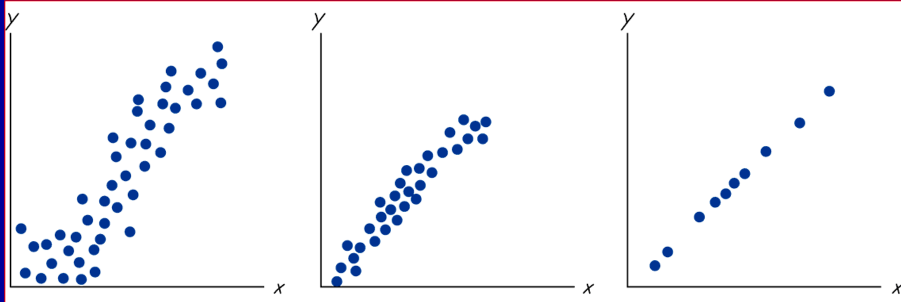
Correlation Coefficients

Definition: A correlation coefficient is a statistic that indicates the **strength & direction** of the relationship b/w 2 variables.

- Correlation coefficients provide a single numerical value to represent the relationship b/w the 2 variables
- Correlation coefficients **ranges -1 to +1**
  - -1.00 (negative one) a perfect, inverse relationship
  - +1.00 (positive one) a perfect, direct relationship
  - 0.00 indicates no relationship

128

Positive Linear Correlation



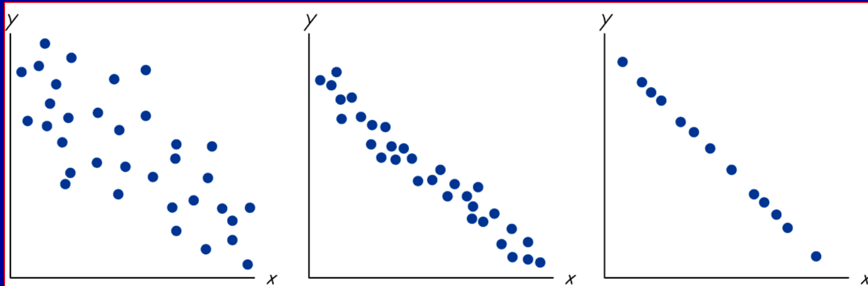
(a) Positive correlation between x and y

(b) Strong positive correlation between x and y

(c) Perfect positive correlation between x and y

129

Negative Linear Correlation



(d) Negative correlation between x and y

(e) Strong negative correlation between x and y

(f) Perfect negative correlation between x and y

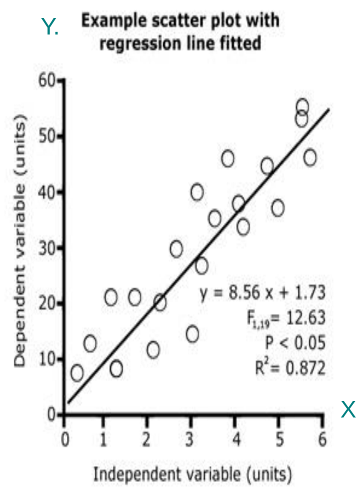
130

Regression analysis

Regressions look for functional relationships between two continuous variables.

A regression assumes that a change in X causes a change in Y.

E.g. Does an increase in light intensity cause an increase in plant growth?



131

Linear Regression

- The Linear Regression model postulates that two random variables X and Y are related by a straight line as follows:

$Y = a + bX$

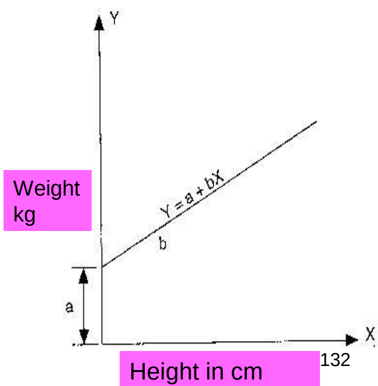
Where

Y is the *dependent variable (weight)*

X is the *independent variable (height)*

a is the Y *intercept*

b is the *slope*

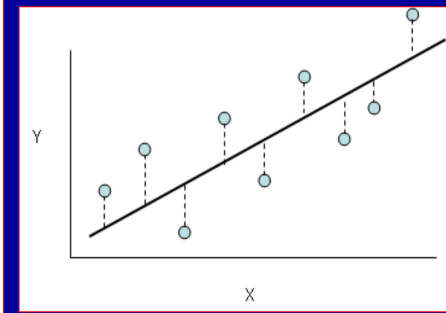


132

Linear Regression

Scatter plots

- In order to perform regression analysis visually, need to do scatter plot for the 2 variables
- A visual relationship can often be observed when looking at these plots.
- Need to draw the ***line of best fit.***
- ***Best fit means that the sum of the squares of the vertical distances from each point to the line is at minimum.***
- ***You can predict 1 cm increase in height and corresponding weight increase***



133

Chi-square statistic

- The Student's t-test and Analysis of Variance are used to analyze **measurement** data (quantitative data), in theory, are continuous variable.
  - **Between a measurement of, say, 1 *mm* and 2 *mm* there is a continuous range from 1.0001 to 1.9999 mm.**
- But in some types of experiment we wish to record how many individuals fall into a particular category, such as blue eyes or brown eyes, motile or non-motile cells, etc.
  - These counts, or **enumeration data**, are discontinuous data or discrete data

134

When Chi-square test used

- Used to test categorical data
- Nominal variables
  - Examples: gender, blood group
- Ordinal Variables
  - Birth order
  - Severity of diseases (absent, mild moderate, severe)

135

Chi-square statistic

- The chi-square test is used to determine whether there is a significant difference between the expected frequencies and the observed frequencies in one or more categories.
- Do the number of individuals or objects that fall in each category differ significantly from the number you would expect?

136

Chi-square statistic

- Two **non-parametric hypothesis tests** using the chi-square statistic:
  - **the chi-square test for goodness of fit and**
    - Goodness of fit refers to how close the observed data are to those predicted from a hypothesis
  - **the chi-square test for independence.**

137

Distribution of Blood Group of Students

Blood Group	Frequency	Relative Frequency (%)
A	8	13.0
B	24	38.7
AB	3	4.8
O	27	43.5
Total	**62**	**100.0**

Example:
Suppose we wish to test the null hypothesis that Dr. Hisham gives equal numbers of A's, B's, C's, D's, and F's as final grades in his Enviro biotechnology classes with 100 students.
The observed frequencies are: **A: 6, B: 24, C: 50, D: 10, F: 10.**

138

Sign test

The sign test is a nonparametric (distribution free) test that uses plus and minus signs to test different claims, including:

- **Claims involving matched pairs of sample data;**
- **Claims involving nominal data;**
- **Claims about the median of a single population.**

The "paired-samples sign test", is used to determine whether there is a median difference between paired or matched observations.

The test is considered as an alternative to the **dependent t-test** (also called the paired-samples t-test) or **Wilcoxon signed-rank test**

139

SIGN TEST

Presented weights of students measured in two times.

Weight (kg) Measurements of Students in Their Freshman Year

September weight	67	53	64	74	67	70	55	74	62	57
April weight	66	52	68	77	67	71	60	82	65	58
Sign of difference	+	+	–	–	0	–	–	–	–	–

140

Wilcoxon signed-ranks test

- The Wilcoxon signed-rank test is the nonparametric test equivalent to the dependent t-test (paired sample t test)
- Wilcoxon signed-rank test does not assume normality in the data, it can be used when this assumption has been violated and the use of the dependent t-test is inappropriate.
- It is used to compare two sets of scores that come from the same participants
- **To test difference between paired data**

141

Example - Wilcoxon signed-rank test

Laureysens et al. (2004) measured metal content in the different types of wood growing in a polluted area, once in August and another one in November. Concentrations of aluminum (in mg of aluminum per gram of wood).

Types of wood	August	November
Balsam Spire	8.1	11.2
Beaupre	10	16.3
Hazendans	16.5	15.3
Hoogvorst	13.6	15.6
Raspalje	9.5	10.5
Unal	8.3	15.5
Columbia River	18.3	12.7
Fritzi Pauley	13.3	11.1
Trichobel	7.9	19.9
Gaver	8.1	20.4
Gibecq	8.9	14.2
Primo	12.6	12.7
Wolterson	13.4	36.8
	12.8	12.8

142

Mann-Whitney U test

- This test is used to compare differences between two independent groups when the dependent variable is either ordinal or continuous, but not normally **distributed**
- The most commonly used alternative to the **independent-samples *t* test.**
- **Example:**
  - To understand whether salaries differed based on educational level (i.e., dependent variable is "salary" and independent variable is "educational level", which has two groups: "high school" and "university").

143

Example 1: **Mann-Whitney U test-**

Table shows the clinical attachment level of two groups of patients (smokers and non-smokers) at the end of a period of periodontological treatment.

CAL = **The amount of space between attached periodontal tissues and a fixed point, usually the cementoenamel junction. A measurement used to assess the stability of attachment as part of a periodontal maintenance program.**

We want to know if there is a difference between the groups.

Non-smoker	CAL (mm)	Smoker	CAL (mm)
1	1.0	14	2.8
2	0.6	15	0.0
3	1.1	16	4.2
4	1.2	17	1.3
5	0.7	18	3.6
6	1.3	19	1.6
7	0.9	20	0.9
8	0.4	21	1.3
9	0.9	22	1.0
10	0.2	23	1.5
11	1.4	24	2.8
12	0.9	25	2.8
13	0·8	26	2·0

144

Example 2 : Mann-Whitney test

- Bicep skinfold thickness has been measured in patients with two different types of intestinal disease.

- **Research question:**
- **Is there a difference in the median skinfold thickness between the two groups of patients?**

- H0= skinfold thickness between the two groups of patients is same
- Ha= skinfold thickness between the two groups of patients is not same

145

When to use Kruskal–Wallis test

- The Kruskal–Wallis test is most commonly used when there is one nominal variable and **one measurement variable (dependent variable)**, and the measurement variable does not meet the normality assumption of an ANOVA.

- It is the non-parametric analogue of a one-way ANOVA.

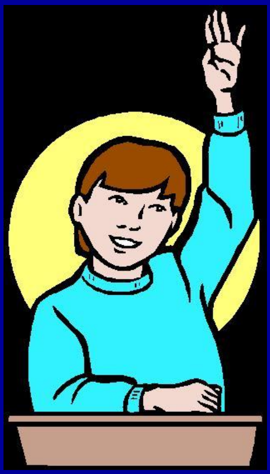
146

Kruskal–Wallis test

A study was conducted to examine the clinical efficacy of a new antidepressant. Depressed patients were randomly assigned to one of three groups: **a placebo group, a group that received a low dose of the drug, and a group that received a moderate dose of the drug**. After four weeks of treatment, the patients completed the **Beck Depression Inventory**. The higher the score, the more depressed the patient.

Placebo	Low Dose	Moderate Dose
38	22	14
47	19	26
39	8	11
25	23	18
42	31	5

147



THANK YOU

148