UNIVERSITY OF ALBERTA

INTRODUCTION TO APPLIED STATISTICS II

STAT 252

LECTURE NOTES

Dr. Greg M. Wagner

SECTION ONE: RESEARCH DESIGN & REVIEW OF 1ST LEVEL STATS

1.1 Objectives of the course

- Learning statistical methods and their application in research (practical approach)
- Review the concepts of summarizing and describing data with graphs and numbers, as well as basic probability theory, which forms the basis for statistical inference
- Performing statistical inference
  - Provides meaningful interpretation of data
  - Enables us to draw sound scientific research conclusions, based on data collected during a study
  - Also tells us the probability that we are wrong when drawing each conclusion, that is, measures the chance of error
- Learning to do statistical analysis with a computer

1.2 Best Approach to Learning Statistics

- STATS is a systematic approach to presenting research findings and drawing conclusions
- Therefore, BE systematic and organized in solving problems
- Understand the PROCESSES involved in statistical analysis
- See the logic (no need to memorize)
- Learn by doing – Practice makes perfect; nowhere is that truer than in statistics
- Participate – ask questions, answer questions, give ideas
- Active learning
- Revise throughout the course; do not wait until the last minute
- You cannot "cram" for a statistics exam, because then you will not understand the processes
- There is no need to be afraid of statistics—it is enjoyable
- Fear is a block to learning—consciously put it out of your mind

- **Attendance in ALL lectures IS ESSENTIAL** in order to perform well in this course. If a student's attendance is poor, even just passing may be unlikely. Also, during lectures you **must be fully attentive, interactive and write detailed notes.**

- **My lecture notes** posted on blackboard contain basic theory, definitions, graphs and formulas that are tedious to write in class.
  - It is important that you print these and bring them to class so that you can annotate them while I am explaining them.
  - My notes also contain blank spaces (demarcated with green and red arrows), especially later in the course, in which you will be required to write additional notes by hand in class.
  - The latter will mainly include illustrations and step-by-step examples, which we will work through together as a way of promoting active learning.

1.3 What is Statistics?

Statistics = the science of collecting, classifying, analyzing, describing and presenting data as well as drawing scientific conclusions about the phenomena being studied.

Statistics is the science of **learning from data**

Statistics is a **way of reasoning** in order to help us understand the world around us (both society and nature)

Statistics involves 3 main aspects:

1. **Research Design** = planning and designing appropriate ways of collecting data for the investigation of a particular scientific problem

2. **Descriptive Statistics** = description, summarization and presentation of data using both numerical and graphical methods (sometimes called exploratory data analysis)

3. **Inferential Statistics** = drawing scientific conclusions and making predictions about a population (as well as measuring the reliability of those conclusions), based on data obtained from a sample from that population. It involves:
    - **Hypothesis tests**
    - **Confidence intervals**
    - Making an **estimate** about a population, based on a sample

- In general, a researcher applies descriptive statistics to his/her data first and then applies inferential statistics to the same data

Statistic(s) = a calculated or estimated statistical quantity, such as mean, t statistic, correlation coefficient (r), F-statistic, etc.

Purpose of Statistics is to have an **objective, unbiased** approach to learning from data:
- To see the bigger picture (so you can see the forest instead of the trees)
- To compare treatments or groups in order to see which one is better, bigger, more effective, etc.
- To look for causation (cause-and-effect relationships) or association between variables

Application of Statistics

- In all branches of natural and social sciences, **particularly where variation occurs**, eg.:

Biological sciences	Sociology
Agricultural sciences	Economics
Medical sciences	Commerce
Earth sciences	Education
Engineering	Psychology

- Less application in exact sciences such as some branches of physics, where there is no variation in phenomena because they follow precise laws of physics.
    - E.g., if you throw a ball of a certain weight, in a certain direction with a certain velocity, you know exactly what its path will be and where it will fall, with no variation

- **Statistical analysis** forms the basis for scientific papers, government publications, education surveys, etc.

- Therefore, **Statistics** is extremely important and indispensable.

- For you as undergraduates, this course will:
    - Empower you to properly conduct undergraduate research projects in several of your other courses
    - Prepare you for later employment in research jobs
    - Help you to understand reports or any literature that presents research results, even if you yourself are not required to conduct research
    - Thus, enhance your performance in many types of jobs
    - Provide you with the fundamentals you require to do a Master's degree or Ph.D.
    - Facilitate planning, making the best decisions and taking the most appropriate actions, even in our everyday lives

1.4 Populations and Samples

Population (Target population)
= the entire collection of all individuals or items under consideration in a statistical study
Population size (N) = total number of individuals or items in the population under study.
Census = collecting data about the entire population

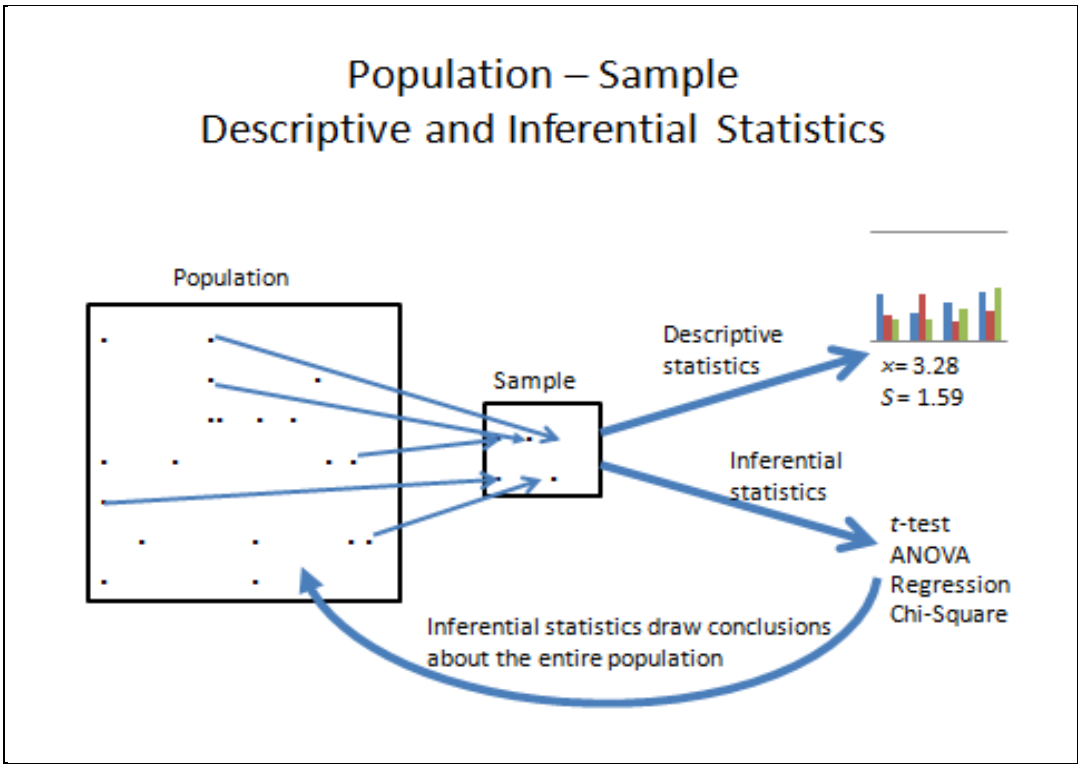
- Often too expensive or even impossible to undertake

Sample = a subset of the population from which the information is obtained

- A sample is a relatively small number of observations from the population being investigated
- Usually it is impossible or too expensive to measure a variable for an entire population (census), so a sample of individuals is measured.
- **Use of the word "sample":**
  - In some types of research, the term "sample" is used to mean one observation
  - In statistics, the term "sample" means a collection of measurements from a population

Sample size (n) = number of observations or measurements in a single sample

Inferential statistics - uses information from a sample to make decisions, conclusions and predictions about the entire population



Parameter versus Statistic

Parameter = a descriptive measure of a population (symbolized by Greek letters), e.g., population mean μ, population standard deviation σ and slope of the population regression line β_1

Statistic = a descriptive measure of a sample, used to estimate a parameter
e.g., sample mean $\overline{y}$, sample standard deviation S and slope of the sample regression line b_1

1.5 The Components of Research Design

- Research design is advance planning.
- Improves the reliability of the results.
- Appropriate research design ensures getting the most reliable and conclusive results.
- Forward planning helps to avoid mistakes or oversights in the research.

Components of Research Design are list in this textbox and explained in detail below

Components of a research design that must be planned in advance, described in a research proposal, and included in the methods section of the final report:

- Study units (within the context of the target population and study area/sites)
- Variables
- Description of the type of research design and sampling strategy
- Spatial Aspects of Design
- Temporal Aspect of Design
- Techniques and Methods of Data Collection
- Forward Planning for Data Analysis
- Repeatability

- These are sometimes referred to as the "W's" of research: "Who, What, Where, When, Why and How", though this classification is often too simplistic.
- The "Why" is the importance, significance and relevance of the study.

1.5.1 Study Units (Within the Context of the Target Population and Study Area/Sites)

Study Units

- Study units are the individuals or subjects (people, animals, objects, or things) about which information is required or on which measurements are recorded.
- These are sometimes referred to as the units of analysis or cases.
- In an experimental study, these can be referred to as experimental units and in an observation study, they may be referred to as units of observation.
- In agricultural research, these are the pre-determined plots where different treatments are applied.
- Study units are sometimes considered as the "Who" component of research, though this term only applies to social sciences.

1.5.2 Variables

Variable

- A variable is a characteristic that varies from one study unit (individual, subject, person, or thing) to another.
- There is natural variation in everything, so when you measure several objects or specimens, you will get different values, thus the term "variable".
- The characteristic being measured on the study units depends upon the particular research.
- Sometimes regarded as the "What" component of research design (what is measured or recorded).
- If possible, include a plan of the scales and units to be used for recording variables.
- In social sciences, these may be opinions, behavior, attitudes, perception, etc.
- In physics – weight, force, energy, light.
- In biology – growth rate, chlorophyll content, height, density, color.

Distribution of a variable = all the values that a variable takes on.
Data = the values of a variable, i.e., actual measurements/observations recorded of a variable under study.
Datum = individual piece of data (an observation) or a single measurement.

1.5.2.1 Types of Variables: Data Scales and Recording Levels

- The type of variables recorded affects the statistical tests to be applied in the analysis.
- The terms below also apply to types of data.

Categorical Variables (Also called **Qualitative Variables)** Recorded on a **Nominal Scale**

- A categorical variable is a nonnumerically valued variable and does not follow an ordered scale.
- Values of the variable are classified by some quality or attribute, i.e., the values are put into categories.
- A categorical variable cannot be measured, but rather, the frequencies of individuals in the categories are counted.
- Examples include: color, gender (male or female); marital status; like or dislike a certain hobby or activity; yes, no or indifferent to something; types of animals; types of items to purchase; languages.
- Data are recorded on a **nominal scale** (nominal refers to "names").
- Nominal scales may sometimes be assigned numbers for recording purposes, but it is still categorical (not quantitative), for example, 1 = single, 2 = common-law, 3 = married, 4 = separated, 5 = widowed, 6 = divorced.

Ordinal Scale Variables/Data

- Data or observations which can be put in order from lowest to highest, but which do not have a constant interval between successive units, i.e., the data can be ranked.
- Relative magnitudes are known, so many types of statistical analysis can be applied.
- For example, when assessing the quality of a product, an ordinal scale of 1 – 5 can be used, where 1 = very poor, 2 = poor, 3 = moderate, 4 = good, 5 =very good.

Quantitative Variables

- A quantitative variable is a numerically valued variable.
- Constant interval size between successive units.
- **Discrete versus continuous quantitative variables**
  - **Discrete or discontinuous quantitative variable** – a quantitative variable whose possible values only take on specific values, usually whole numbers.
    - a countable variable.
    - e.g. number of people, animals, stars must be whole numbers.
  - **Continuous quantitative variable** = a quantitative variable that may have an infinite number of values between any observed range.
    - a measureable variable.
    - e.g., the weight of a person may be 71 kg or 72 kg or an infinite number of values in between, such as 71.42 kg or 71.42893 kg, depending upon the accuracy of the scale or balance.
    - time, distance and height (regardless of units) are always continuous variables.
- **Ratio scale versus interval scale for quantitative variables**
  - **Ratio scale** has a true zero point, which makes it possible to establish a ratio, e.g., the Kelvin scale is a ratio scale for measuring temperature: absolute 0°K is -273°C, so comparing 313°K (40°C) to 293°K (20°C), the ratio is 313/293 = 1.068. This is a real ratio.
  - **Interval scale** has no true zero point, making it impossible to establish a ratio, e.g. Temperature on the Celsius scale, 0 has no real meaning; it is arbitrary, so 40°C is not twice as hot as 20°C. Another example: 0 on a scale for IQ doesn't mean 0 intelligence

Indicator Variables (Dummy Variables)

- Categorical variables that are coded in order to obtain quantitative variables that can be analyzed using hypotheses tests like ANOVA and regression.
- Two categories are coded as 0 and 1 (e.g., 0 = male, 1 = female).
- Three categories are coded as combinations of 0 and 1.
  (E.g. $z_1 = z_2 = 0$ for downtown; $z_1 = 1$, $z_2 = 0$ for inner suburbs; $z_1 = 0$, $z_2 = 1$ for outer suburbs).

1.5.2.2 Types of Variables: Their Roles in Research

Explanatory and Response Variables

- **Explanatory or Predictor variables (sometimes referred to as independent variables)** = variables of interest that are hypothesized to explain or affect other variables in the study, but which are not likely to be affected by those other variables.
- **Response variable (sometimes referred to as a dependent variable)** = variable that is hypothesized to be affected by the explanatory or independent variables.
- E.g., age and height – height does not affect age, but age affects height.
- Explanatory and response variables must be defined in the statement of the research problem and are the basis for formulating the research objectives and hypotheses.
- Generally, the application of explanatory variables must either precede or occur during the same time period as the expected reaction of the response variable.

Extraneous variables

- Explanatory variables that are NOT of interest or are NOT related to the purpose of the study, though they could be of interest in a different study.
- These may potentially affect the response variable, interfering with the study and leading to "experimental error".
- These are sometimes variables that are not measured or cannot be measured (confounding, hidden or lurking variables).

Factors

- When explanatory variables are applied as treatments in an experiment or considered as levels in an observational study, they are usually referred to as factors.
- The researcher tries to determine the effects that the different levels of the factor have on the responses of the study units.

1.5.3 Spatial Aspects of Design

- The "Where" component of research design
- Involves the way the observations or replicates are arranged in space (distance, area, or volume)
- Linked to the study unit – where are they sampled and measured

1.5.4 Temporal Aspects of Design

- The "When" component of research design **–** the way observations or replicates are arranged in time
- Time period (year, month, time of day) and frequency of observations
- Start, end, frequency of recording the variables

1.5.5 Techniques and Methods of Data Collection

- The "How" component of research design
- Specific methods and techniques used to take measurements of the variables or to record data
- The specific techniques to be applied will differ from one field of natural or social science to the other.

1.6 Types of Research Designs and Sampling Strategies

This involves describing or specifying the following:

- What type of sampling will be done, e.g., simple random sampling, systematic random sampling, stratified random sampling, etc.
- Integrates the "where" and "when" components of research design.
- Whether it is an observational or experimental study

1.6.1 Sampling Strategies and Randomness

- Study units or individuals are randomly sampling from the target population so that they represent the population.
- **Random sampling** = the selection of individuals or units from a population without bias, such that:
  1. All individuals have an equal chance of selection (or, each possible sample of a given size is equally likely to be the one obtained).
  2. The selection of individuals is independent, i.e., the selection of one does not affect the selection of others.
- **Random sampling** ensures that the sample is as **representative** as possible of the entire population.
- All statistical tests assume that samples are obtained randomly from a population.
- Sampling strategy and research design specify the way observations are recorded in space and time.
- Sampling must eliminate bias as much as possible, because bias over-emphasizes or under-emphasizes some characteristics of the population.
- Randomness is applied differently in observational and experimental studies, as explained below.

- Random sampling involves the use of **probability sampling** and can be done by:
  - Tossing a coin,
  - Drawing numbers from a box,
  - Using a random number table,
  - Simulation with a computer program (random number generator), or
  - Using some procedure such as throwing a quadrat without looking.

- When possible, the researcher should have a list of the population and this list should be numbered (a numbered list of the population is called a **sampling frame**).

Computer Programs

- Do not generate truly random numbers since, if they start at the same place, they will give the same numbers; thus, simulations are not always independent and not completely random.
- Nevertheless, it is random enough for most purposes.
- Very commonly used.

Sampling With and Without Replacement

- **Sampling with replacement** = an individual has a chance of being selected more than once.

- **Sampling without replacement** = an individual can only be selected once.
  - Strictly speaking, this violates the requirement for independent selection of individuals (condition #2 above).
  - Makes little difference, however, when sample size is small relative to populations size.
  - This type is actually very commonly applied.
  - For example, it is common in social surveys; if you randomly select individuals to interview, it would be awkward (and boring) if the same individual was selected twice.
  - Also, used in biology if recording data requires destroying the test organism.

Types of random sampling

- **Simple random sampling (SRS)**
  - Every individual is selected completely randomly and independently.
  - Every group or area of the population has an equal chance of selection.
  - All the statistical tests dealt with in this course require simple random sampling.

- **Systematic random sampling**
  - The first sample is selected randomly, then all other samples are selected sequentially, e.g., every 30 seconds of swimming over a coral reef, every 10 m, every 5 min, every 5th person, etc.
  - E.g. sampling plots in a forest.
  - Good system unless there is a rhythmic cycle in the data.

- **Stratified random sampling**
  - The population is divided into strata, based on a pilot study or some prior information.
  - Items within each subpopulation are considered relatively homogeneous.
  - **Proportional allocation** = sampling intensity in each stratum is proportional to the estimated density of the items in the stratum or size of the stratum.
  - Within each stratum, do simple random sampling (SRS).
  - This gives the most accurate results if there are definite strata in the study area.

- **Multistage random sampling**
  - Example of sampling leaves on trees of a certain species:
    - Randomly sample trees, then
    - Randomly sample the branches on the selected trees, then
    - Randomly sample some of the leaves from the selected branches and take them for analysis.

- **Cluster Random Sampling**
  - For example, if a company with a large number of apartment blocks (e.g. 100) want to get the opinions of their tenants about some proposed changes.
    - Randomly select a few apartment blocks and then interview all the tenants in the selected blocks. Each selected block would then be considered as a cluster.

Sample size (n) = number of observations or measurements in a single sample

- **Plan for adequate sample** size because increasing sample size:
  - Reduces the standard error of the estimate,
  - Increases accuracy and precision, and
  - Increases the power of a hypothesis test.

- **Adequate sample size depends upon:**
  - The characteristic you are measuring,
  - How frequently it occurs in the population.
  - Degree of variability of the material or objects being studied
    - large sample size is required in some studies in biology and earth sciences where variation in the natural environment is often very large.
    - in some branches of chemistry and physics, variability of the material is very small, so smaller sample size is adequate.
  - Magnitude of the difference you expect to find between groups.
  - Precision of the techniques used.

Sample size versus sample fraction

- Sample fraction = n/N = the proportion of the population that is included in the sample.
- Sample size is the absolute number of observations (n).

- Sample size is more important than sample fraction because sample size determines the power of the hypothesis test and the type of hypothesis test to be used.

Sampling variability = the differences or variation in the characteristics of interest from one sample to the other from the same population.
- No matter how well sampling is done, there will always be differences from one sample to the other
- Increasing sample size decreases sampling variability.

Problems or Bias due to Poor Sampling Procedures

Convenience sampling = selecting individuals for recording data simply because they are easily accessible or are convenient to observe or question. This leads to bias because it does not provide a random sample
- E.g. interviewing people in a shopping mall about products sold there. They may be there just because they favor those products sold there.
- E.g., observing the behavior of wildlife within a convenient distance from a highway (their behavior might be very different from animals that live far away in the wilderness).

Voluntary response bias = asking for volunteers to participate in a social survey (people who favor something are more likely to volunteer).
- Very common in telephone surveys.
- Always leads to serious bias since it does not provide a random sample from the target population.

Response bias = questions in a social survey that appear to suggest or prompt a particular response favored by the researcher.
- May also result from a poorly worded question.

Nonresponse bias = occurs when a large fraction of those sampled fail to respond to some or any of the questions
- Sometimes a result of the questionnaire not being adequate.

Incomplete sampling frame = some individuals or groups who actually belong to a certain population are not included in the sampling frame.
- May be homeless people, long-term travelers or people with only cell phones (in a telephone survey).

Undercoverage = some portion of the population not being included or given smaller representation
- May be a result of an incomplete sampling frame.
- May also be a result of conducting SRS instead of stratified random sampling.

1.6.2 Overall Type of Design: Observational Research versus Experimental Research
- Unless it is obvious from the research problem, it should be mentioned in the introduction and/or methods sections whether it is an observational or experimental study.

Observational studies or surveys
- Observational studies may be applied in almost all fields, but when applied in order to get opinions from people it is often called a **sample survey** or **social survey.**
- Tries to estimate population parameters.
- Researcher collects data about a particular phenomenon <u>as it occurs in nature or in society</u>.
- Randomness
  - <u>Random sampling (or random selection)</u> of study units (individuals) from the target population.
- Variables of interest are measured or recorded for the study units.
- <u>No imposing</u> of treatments on the subjects or individuals.
- <u>No manipulation or control</u> of any variables or conditions.
- Extraneous variables cannot be controlled.

- **Population inferences** – can be made if there is random selection from the target population.
- **Causal interferences** – can NOT be made**, that is,** causation or cause-and-effect relationships among variables can NOT be established because there are so many unmeasured factors (or **extraneous variables**) that may affect the variable being measured.
- May suggest possible cause-and-effect relationships or correlations among variables that can later be tested by setting up an experiment.
- Two types of Observational Studies:
  - **Prospective =** subjects identified beforehand and data are recorded as the study proceeds.
  - **Retrospective =** subjects identified and data recorded after events have already occurred.
    These are sometimes difficult to implement unless there are accurate historical records.

Experimental Studies

- Researcher "sets up" an experiment (This is an active process).
- **Randomness** = Randomization (an active process).
  - First, study units (experimental units) are randomly selected from the target population.
  - Secondly, the experimental units are randomly assigned to treatment and control groups.
- **Manipulation of predictor or explanatory variables (factors)** – the researcher changes them deliberately.
  - **Treatment groups –** exposed to new conditions, that is, one or more levels of the predictor variable or factor being manipulated; the treatments are imposed upon these groups
  - **Control groups –** exposed to the usual level of the manipulated variable or not exposed to it at all
  - E.g., In an experiment where one is testing the effects of fertilizers on plant growth
    - Control group is subjected to the usual conditions of soil, water, light, etc., but no fertilizer.
    - Experimental groups are subjected to the same conditions + different types of fertilizers or different amounts of the same fertilizer.
  - In the case of human subjects, the control group receives a **placebo** so that they don't know whether they are receiving a treatment or not (e.g. in medical research it is an inert substance).
  - In some studies, it is difficult to have a control, but it is best to have a control when possible.
- **Extraneous variables (or lurking variables)** are controlled or made constant for all treatment and control groups.
  - E.g., if you are testing the effect of fertilizer on plant growth, then light, soil type, water, etc. are extraneous variables and must be kept constant for all treatments.
- **Response variable** is measured or recorded for all experimental units in the treatment and control groups to see if these variables are affected by the predictor variables.
- Can establish causation (cause-and-effect relationships) among variables.
- **Population inferences** – can be made if there is random selection from the target population.
- **Causal interferences** – can be made if there is random assignment to treatment and control groups.
- **Both population and causal inferences** – can be made if there is random selection and random assignment.
- Although experimental studies lead to more definite conclusions than observational studies, it may be impossible or unethical to set up an experiment for some types of studies.

Replication

- Replication of each treatment and control is important.
- Good experimental results should be repeatable and replicable. One way to replicate the results is to have several samples or replicates in the same experiment.
- Replication is required to:
  - Check or confirm the results,
  - Apply statistical analysis – analysis is based on replicates,
  - Estimate the precision (e.g., calculate standard deviation) or state the probability that the conclusion is correct, and
  - Increase the power of the test.
- No. of replicates = no of samples or sample size (n).

Blinding:

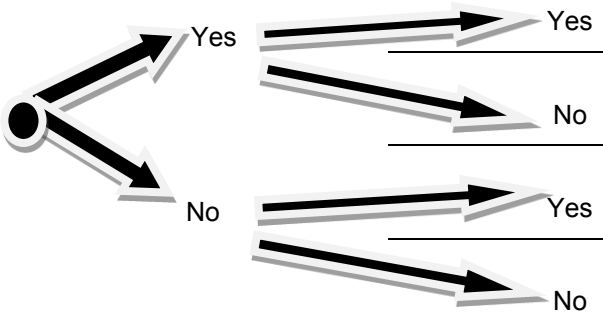
- Those who could affect the results (the subjects, treatment administrators or technicians).
- Those who evaluate the results (judges, treating physicians, researchers).
- **Single-blind experiment** – all individuals of one or the other of the above groups are blind.
- **Double blind experiment** – all individuals of both of the above groups are blind.

Placebo effect

- Psychological effect of receiving a placebo, which may result in a subject responding to a treatment when, in fact, they only received an inert placebo.

Example: Comparison of observational and experimental studies on the possible effect of Vitamin E in preventing/controlling heart disease

Dichotomous Tree on Types of Inferences Possible for Observation and Experiment



Random selection from the target population	Random assignment to treatments	Inferences Possible
Yes	Yes	Both population inferences and causal inferences can be made
	No	Population inferences can be made, but causal inferences cannot
No	Yes	Causal inferences can be made, but population inferences cannot
	No	Neither type of inference can be made

A Proper Observational Study

- Involves random selection (sampling) from the target population

A Proper Experiment

- Involves random selection from the target population, followed by random assignment to the treatments.

1.6.3 Specific Types of Research Designs With Respect to the Variables involved and the Types of Hypotheses Tested

- Depends upon whether the overall purpose of the study is
  - to investigate the relationship between two or more variables, or
  - to determine differences between populations or groups.
- Paired sample design or independent sample design should be specified.
- **The following specific types of research design will be discussed** in detail as we proceed with this course:
  - **Completely randomized single-factor design (independent sample design)**
  - **Paired design**
  - **Randomized block design**
  - **Completely randomized two-factor design**
  - **Simple linear regression**
  - **Multiple linear regression**

1.7 Descriptive Statistics: Categorical Data

- Descriptive Statistics, both graphical or numerical methods, summarize the data and present them in a way that can be understood at a glance
    - Gives you the overall, "bigger picture"
- The first step in drawing graphs is to first group the data into frequency distribution tables
- The type of data being presented will determine the types of graphs that can be used

1.7.1 Grouping Qualitative Data

Frequency (fi) = the counts or number of observations that fall into a given class/category of a variable or that have a given value of the variable

Frequency distribution = a listing/presentation of all classes/categories or values of a variable, together with the number of observations (frequency) for each class or value

- May be presented in a table or a graph

Relative frequency = the ratio of the frequency of a class/category (or certain value) to the total number of observations = $\frac{fi}{\sum fi}$

Relative percent frequency = $\frac{fi}{\sum fi}$ x 100

Table: Frequency distribution table, including relative frequency and relative percent frequency, showing the favorite sports people in a certain area of Edmonton liked to watch on TV in 2000.

Sport	Frequency (fi)	Relative frequency	Relative percent frequency (%)
Hockey	142	$\frac{142}{320} = 0.444$	44.4
Baseball	55	0.172	17.2
Football	78	0.244	24.4
Soccer	45	0.141	14.1
Total	**320**	**1.001**	**100.1**

1.7.2 Pie Charts

- Pie charts require the calculation of degrees of the circle that represent each category, though computer programs calculate the degrees automatically

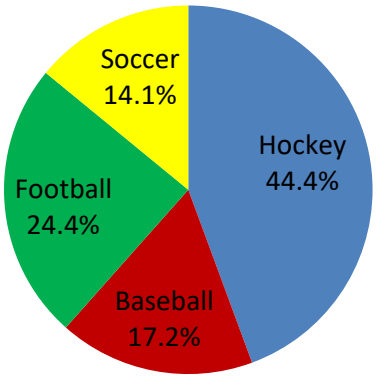


Fig: Favorite sport to watch on TV amongst people in one area of Edmonton in 2000.

1.7.3 Bar Graphs and Contingency Tables

Simple bar graph

- Shows the frequencies of categories for one variable
- Leave spaces (gaps) between the bars
- Can show the same information as a pie chart

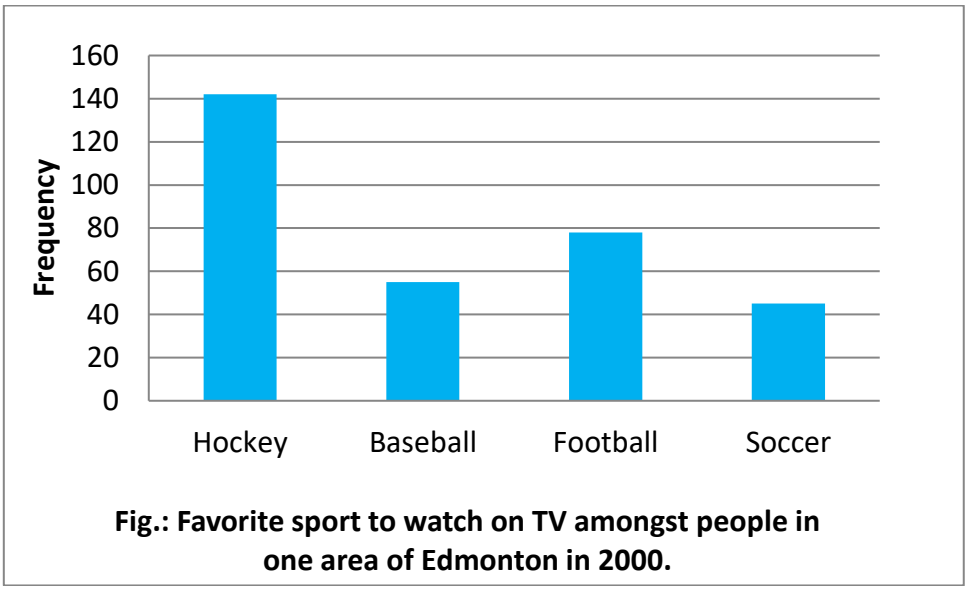


Fig.: Favorite sport to watch on TV amongst people in one area of Edmonton in 2000.

Area Principle

- Area under the graph must equal the value (frequency, percentage) being presented

Contingency Tables

- Tables that give frequencies for **two variables** at the same time (called **bivariate data) – both are qualitative variables**
- Sometimes called **two-way tables** or **cross-tabulation table**
- Show how the number of observations of one variable is "contingent" on the other variable
- Consists of: rows, columns and cells

Table: Frequency distribution table showing which sports people in a certain area of Edmonton liked to watch on TV the most in 1990, 2000 and 2010.

	Frequency (f)			
Sport	**1990**	**2000**	**2010**	**Total**
Hockey	169	142	114	**425**
Baseball	53	55	54	**162**
Football	72	78	90	**240**
Soccer	26	45	62	**133**
Total	**320**	**320**	**320**	**960**

Multiple Bar Graph (= Side-by-Side Bar Graph)

- Shows the frequencies of categories for two variables at the same time
- Thus, can show more information than a pie chart

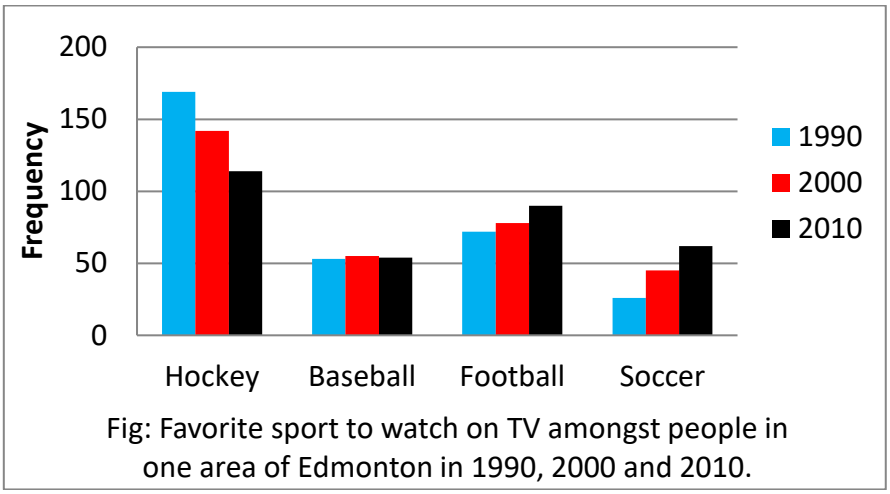


Fig: Favorite sport to watch on TV amongst people in one area of Edmonton in 1990, 2000 and 2010.

- A multiple bar graph can show more information then a pie char.
- These graphs indicate possible changes over time, but inferential statistics are required to show whether the changes over time are statistically significant

Segmented Bar Graph (= Component Bar Graph)

- Similar to multiple bar graph, but the segments are piled up on top of each other
- Displays the conditional distribution of a categorical variable within each category of another variable
- Frequencies may be converted to percentages of the whole total for a given category, so the segments add up to 100%

Segmented, Stacked or Component Bar Graph

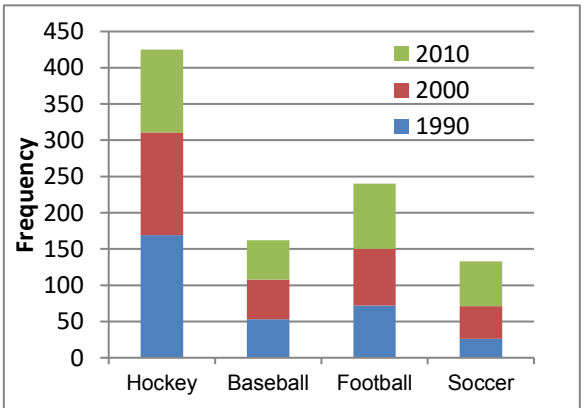


Fig.: Favorite sport to watch on TV amongst people in one area of Edmonton in 1990, 2000 and 2010.

Percent Segmented (Component) Bar Graph

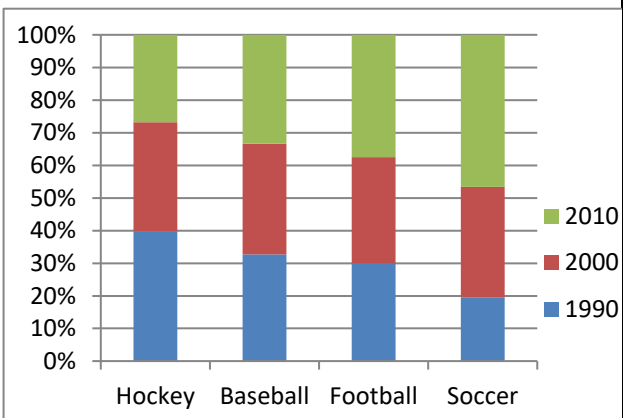


Fig.: Favorite sport to watch on TV amongst people in one area of Edmonton in 1990, 2000 and 2010.

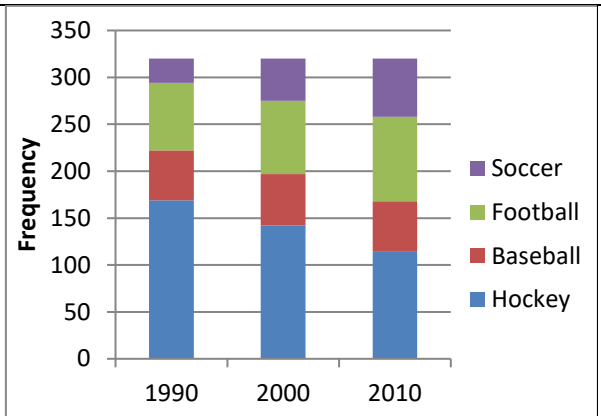


Fig.: Favorite sport to watch on TV amongst people in one area of Edmonton in 1990, 2000 and 2010.

Stacked (component) bar graph

(Row/column switched)

1.8 Descriptive Statistics: Quantitative Data

- **Describing** the distribution of a quantitative variable involves **3 aspects:**
  1. **Shape**
  2. **Center** – the middle of the distribution
  3. **Spread** – variation or dispersion of the distribution
- These 3 aspects can be determined approximately by looking at **graphs** and more exactly by **numerical calculations**, e.g., mean, median, standard deviation, variance

1.8.1 Histograms

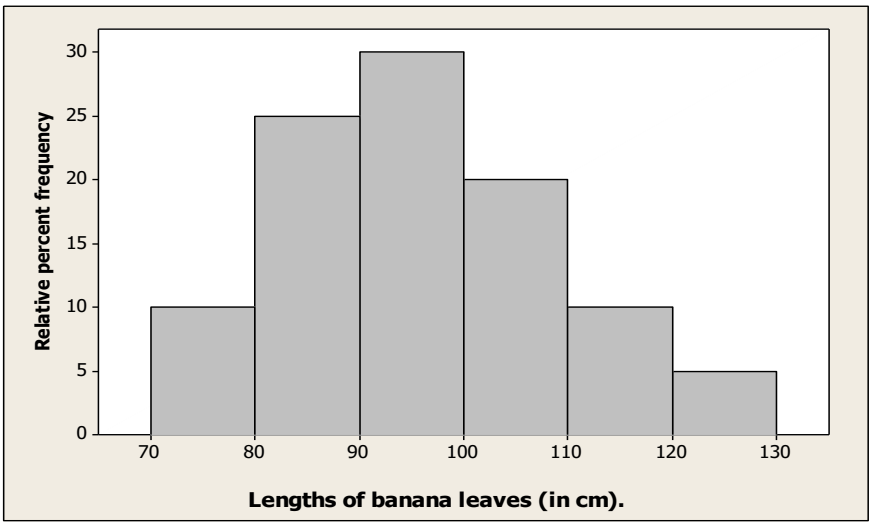
- Like a bar graph, but **no space** between bars
- Y-axis can show frequency, relative frequency or relative percent frequency

Example for Graphing Quantitative Data
Raw data for lengths of 20 banana leaves (in cm).

107	104	118	74	95	123	71	88	96	98
113	98	83	87	91	102	85	108	97	82

Table: Frequency distribution table for the lengths of banana leaves, including relative frequencies and midpoints.

Length of leaf (cm)	Number of leaves (frequency)	Relative frequency	Midpoint
70 – 79	2	2/20 = 0.10	(70 + 80) / 2 = 75
80 – 89	5	0.25	85
90 – 99	6	0.30	95
100 – 109	4	0.20	105
110 – 119	2	0.10	115
120 – 129	1	0.05	125
Total	**20**	**1.00**	



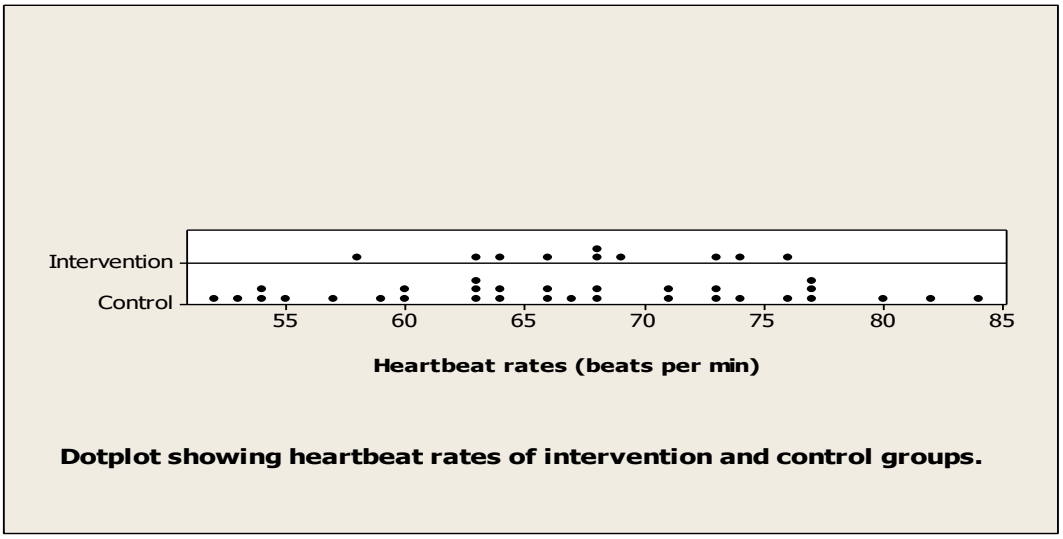
1.8.2 Dotplots

- Useful for showing the **relative positions** of the data, which are not shown in a histogram
- Can be used to compare two or more populations or groups

Example: Bus drivers are often under a lot of stress due to icy and snowy streets, bad drivers on the road and angry passengers. This can affect their blood pressure and heartbeat rates. A study was conducted whereby drivers were randomly allocated to two groups: one group was subjected to the usual conditions (control) and another group was placed into a program where improved conditions were provided (intervention group). The heartbeat rates of the two groups are shown below.

Table 2.9: Heartbeat rates of bus drivers in intervention and control groups.

Intervention		Control						
68	66	74	52	67	63	77	57	80
74	58	77	53	76	54	73	54	
69	63	60	77	63	60	68	64	
68	73	66	71	66	55	71	84	
64	76	63	73	59	68	64	82	



Note: the dotplot above can show more detailed information than a histogram with respect to the relative position of numbers within groups, e.g., you can see that in the control group, between 75 and 80, there is 76, 77, 77, 77, which are closer to 75 than 80.

Compare the two distributions shown in the dotplot: Shape, Centre and Spread

- Shape: nearly the same
- Center: nearly the same (They both have a mean of about 68 beats per min)
- Spread: Bus Drivers Exposed to Stress (Control Group)
    - Shows a very wide spread of heartbeats.
    - This is likely due to the fact that stress affects people in different ways. Stress causes some people to become depressed, those heartbeat rates may go down. Stress causes other people to become hyper, thus increasing heartbeat rates
- Spread: Intervention Group
    - Shows a narrower range, with most of them having heartbeat rates closer to the average of 68, which is normal and healthy.

1.8.3 Stem-and-Leaf Diagrams (or Stemplots)

- The first one or two digits of the observations are considered as the stems, while the last digit is considered as the leaf
    - Thus, 137, 135 and 130 all have the same stem (13), but different leaves: 7, 5 and 0.

Example on lengths of banana leaves (same example as above)
Table: Raw data for lengths of 20 banana leaves (in cm) (data repeated).

107	104	118	74	95	123	71	88	96	98
113	98	83	87	91	102	85	108	97	82

One Line Per Stem Diagram (stems not split)

(Stem)	(Leaves)
7	1 4
8	2 3 5 7 8
9	1 5 6 7 8 8
10	2 4 7 8
11	3 8
12	3

Two Lines Per Stem Diagram (also called a **Split Stem Diagram)**

- For **Line One** of the stem: put leaves with digits 0 – 4
- For **Line Two** of the stem: put leaves with digits 5 – 9

7	1 4
7	
8	2 3
8	5 7 8
9	1
9	5 6 7 8 8
10	2 4
10	7 8
11	3
11	8
12	3
12	

Which gives a better presentation of the distribution for this data, one line per stem or two lines per stem?

- For some distributions the one line per stem is better; for others the two lines per stem is better

Note: if you turn a stemplot around 90º, it looks like a histogram, but provides more detail about the distribution of the values within groups.

Other variations in stemplots

- For some data sets, it is better to have even more than 2 lines per stem, e.g., 5 lines per stem. In that case:
  - Line 1 shows leaves 0-1
  - Line 2 shows leaves 2-3
  - Line 3 shows leaves 4-5
  - Line 4 shows leaves 6-7
  - Line 5 shows leaves 8-9
- Sometimes, you may truncate (drop) the last digit and use the second last digit as the leaf.

Back-to-back Stem-and-Leaf Diagram

- Used to compare two populations or groups
- Construct one common stem for the two groups and put a vertical line on each side
- Then, put the eaves for one group on the left side (starting from the stem and going in ascending order to the left) and
- Put the leaves for the second group on the rights side (starting from the stem and going in ascending order to the right)

Example of Back-to-back Stem-and-Leaf Diagram

The number of patents a university receives is an indication of their research level. The table below shows the number of patents received by 15 randomly selected universities in each of two countries. Construct back-to-back stemplots illustrating the data, using one line per stem.

Country A						Country B				
13	24	15	56	58		17	22	46	33	28
4	43	46	37	54	48	37	42	30	47	35
49	40	36	38			27	19	30	34	36

Stemplot comparing the number of patents in universities in Country A and Country B

Country A		Country B
4	0	
53	1	79
4	2	278
876	3	0034567
98630	4	267
864	5	

Number of patents received by universities in two countries.

1.8.4 Other Types of Graphs for Quantitative Data

- **Boxplots –** Dealt with at the end of this section because a numerical summary is required
- **Normal Probability plots:** used for assessing normality (dealt with later)
- **Scatter diagrams (xy graphs)** – Dealt with under regression and correlation

1.8.5 Shape of a Distribution

Population distribution = the distribution of population data

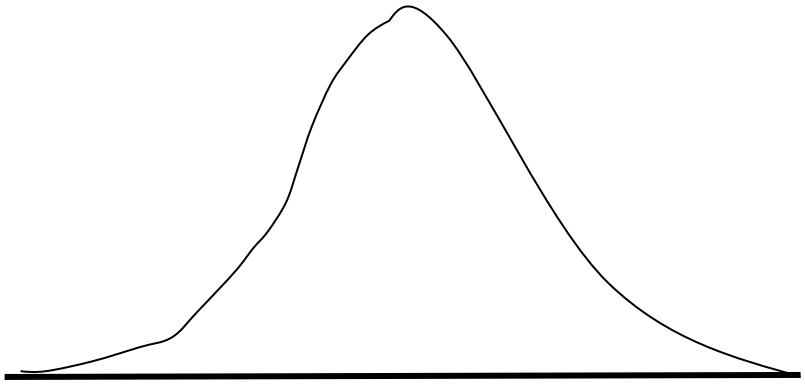
Sample distribution = the distribution of sample data

- If you take several samples from the same population, every sample will have a slightly **different shape or distribution**
- The **larger the sample size**, the better will be its approximation to the population distribution

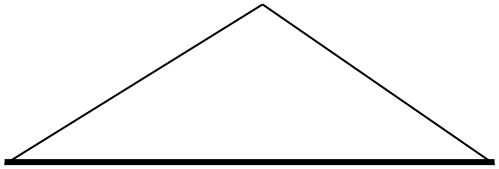
Symmetrical versus Skewed

Symmetrical = distribution that can be divided into two parts such that one is a mirror image of the other

Types of symmetrical distributions



Bell-shaped



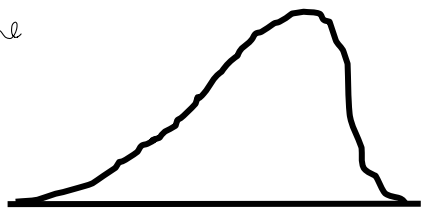
Triangular

Uniform

Skewed = distribution that has one tail of the distribution longer than the other (therefore not symmetrical)

- May be either:
  - **Left skewed (negatively skewed)** = left tail is longer than right tail
  - **Right skewed (positively skewed)** = right tail is longer than left tail
- Use of the terms negatively or positively skewed is actually better than using left or right, because boxplots are usually drawn vertically

Skewness follow the tail.



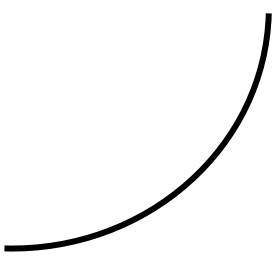
Left skewed

Right skewed

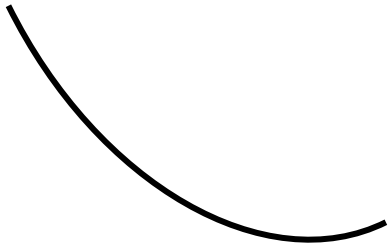
J-shaped = special type of negatively skewed distribution that has no right tail

- In ecology, shows population growth in an unlimited environment

Reverse J-Shaped = special type of positively skewed distribution that has no left tail



J-Shaped

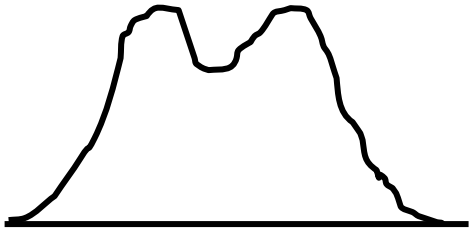


Reverse J-Shaped

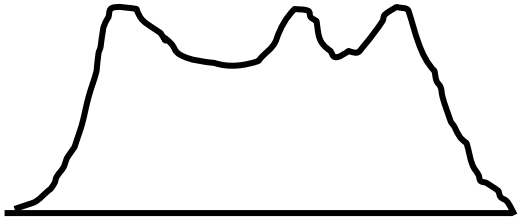
Modality

- Refers to the number of peaks in the distribution

Unimodal = one peak (or one mode) - **All** of the above distributions are unimodal
Bimodal = two major peaks
Multimodal = three or more peaks



Bimodal



Multimodal

1.8.6 Measures of Central Tendency (= Measures of Center)

Mean

Population Mean and Sample Mean

Population mean (μ) = summation of all items in the population / population size

$$\mu = \frac{y_1 + y_2 + \ldots + y_N}{N} = \frac{\sum y_i}{N}$$

Where ***μ*** is the Greek letter "mew", Subscript "i" indicates the i^{th} observation, ***N*** = population size

Sample mean ($\bar{y}$) = summation of all observations in a sample / sample size

$$\bar{y} = \frac{y_1 + y_2 + \ldots + y_n}{n} = \frac{\sum y_i}{n}$$

Where $\bar{y}$ is read "***y*** bar", ***n*** = sample size

- The sample mean is considered as the best estimate of the population mean
- There can be only one population mean, but every sample from that population will have a different sample mean, though they may be close to each other
- Increasing sample size will increase the closeness of the sample means to each other and two the population mean

Median = the middle observation in a distribution

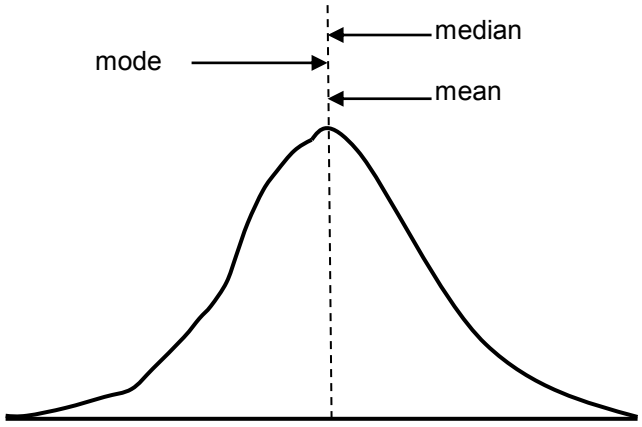
Median class = the class in a frequency distribution in which the median is found

Mode = one or more points in a frequency distribution that have the greatest frequency

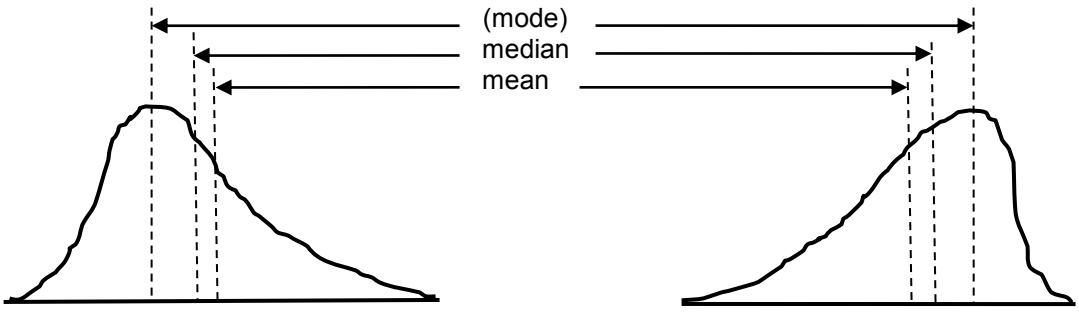
- As already discussed, distributions may be unimodal, bimodal or multimodal

Comparison of Mean, Median and Mode

- Mean is the center of gravity of the distribution (histogram)
- Median divides the area under the curve into two equal halves
- **Median** is a **resistant measure** because it is more robust to extreme values or skewness than the mean and therefore is a better measure of centre for a very skewed distribution
- **Mean** is **not a resistant measure** of centre, because it is seriously influenced by skewness (pulled in the direction of a few extreme observations)
- **Mode** is the only measure of center that can also be used for qualitative data
- For skewed distributions, the best measure of center is the median
- For symmetric distributions, the best measure of center is the mean



Symmetric distribution



Right-skewed

Left-skewed

1.8.7 Measures of Variation (= spread)

Range = Max – Min **=** difference between the highest and lowest observations in a data set

- A biased measure of variation (because outliers can give a much wider range than is the real spread of the main data set)

Sample Variance and Sample Standard Deviation

- Sample standard deviation is the best estimate of population standard deviation

Sample Variance and Sample Standard Deviation

Sample variance (s^2) = sum of squared deviations from the mean / Sample size – 1

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

Sample standard deviation (s) = positive square root of the sample variance

$$s = \sqrt{\text{sample variance}} = \sqrt{s^2} = \sqrt{\frac{\sum(y - \bar{y})^2}{n-1}}$$

Example: Calculate the range, standard deviation and variance of the lengths (cm) of a sample of 5 gastropods of a certain species.

1.4 0.8 0.9 1.3 0.7

Range = 1.4 – 0.7 = 0.7

Sample mean = $\bar{y} = \frac{\sum y_i}{n} = \frac{5.1}{5} = 1.02$

	Lengths (cm) y_i	**Deviation from mean** $(y_i - \bar{y})$	**Squared deviation** $(y_i - \bar{y})^2$
	1.4	1.4 – 1.02 = 0.38	$0.38^2 = 0.1444$
	0.8	-0.22	0.0484
	0.9	-0.12	0.0144
	1.3	0.28	0.0784
	0.7	-0.32	0.1024
Totals	**5.1**	**0**	**0.3880**

$$s = \sqrt{\frac{\sum(y_i - \bar{y})^2}{n-1}} = \sqrt{\frac{0.3880}{5-1}} = 0.31145 = 0.31$$

Sample variance = $s^2 = (0.31145)^2 = 0.10$ or 0.097

Rounding Rules

1. Do not perform any rounding until all calculations are complete; otherwise substantial rounding errors can occur. @ least 3 decimal precision or the precision on the A-table
2. When giving the final answer, keep at least one more decimal place than is given in the raw data.

Degrees of Freedom (df)

- df = the number of independent observations
- For standard deviation, df = n – 1 because the sample mean is used as an estimate of the population mean in calculating it (using defining formula). You first calculate the mean and include that as part of the formula. When the sample mean is used and you know (n – 1) observations, the n^{th} observation is fixed and is therefore not independent
- For standard deviation, df is the denominator of the formula

Population standard deviation and Population variance

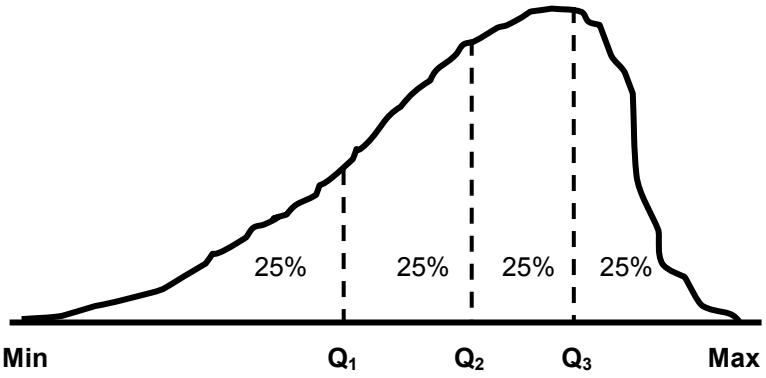
Population standard deviation (σ) (pronounced sigma)

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$

Population variance = (population standard deviation)^2 = σ^2

1.8.8 The Five-Number Summary and Boxplots

Percentiles – divide a data set into 100 equal parts
Deciles – divide a data set into 10 equal parts
Quartiles – divide a data set or distribution into 4 equal parts



First quartile (Q_1) = the median of that part of the data set that lies below the median of the entire data set (Note: that means if there is an odd number of observations, the median is NOT included when determining Q_1)

Second quartile (Q_2) = the median of the entire data set

Third quartile (Q_3) = the median of that part of the data set that lies above the median of the entire data set (Note: that means if there is an odd number of observations, the median is NOT included when determining Q_3)

Interquartile range (IQR) = the difference between the first and third quartiles
$= Q_3 - Q_1$

Note: Quartiles are **more resistant** to extreme observations and skewness than standard deviation

For skewed distributions, quartiles are the best measure of spread
For symmetric distributions, standard deviation is the best measure of spread

Five-Number Summary

Five-Number Summary = Min, Q_1, Q_2, Q_3, Max

The 1.5 x IQR Rule for Calculating Lower and Upper Limits

Lower limit = $Q_1 - 1.5 \times IQR$
Upper limit = $Q_3 + 1.5 \times IQR$

- Used to determine outliers and adjacent values

Outliers = observations that lie outside the overall pattern of the data
- May be due to
  - recording error
  - may belong to a different population
  - may just be unusually extreme observations
- try to determine its cause
- observations that lie outside the lower and upper limits are **potential outliers**

Adjacent values = the most extreme observations that still lie within the lower and upper limits
- There is always a lower adjacent value and an upper adjacent value
- If a data set has no potential outliers,
  - the adjacent values = **Minimum** and **Maximum**

Boxplots
(Also called a **box-and-whisker diagrams)**
- ends of the box are **Q1 and Q3**
- **Median** is indicated by a line across the box
- **Whiskers** are Lines extending from the box to the **maximum** and **minimum** observations (**or** to the **adjacent values**, if there are potential outliers)
- **Potential outliers** are usually marked as **asterisks**
- Sometimes boxplots are drawn vertically

Example of a Five-Number Summary When *n* is Odd
Determine the five-number summary of the following (n is odd):

3 6 7 10 12 13 18

Five-number summary = 3, 6, 10, 13, 18

Example of determining Five-Number Summary, Potential Outliers and constructing Boxplot
The table below shows the cost per night (in US dollars) for a room in a random sample of beach resorts around the island of Phuket in Thailand.

109	126	147	177	224	105	119	141	169	209	349	113	135	159	191	259

(a) Determine the five-number summary of the cost per night for rooms in beach resorts in Phuket.

The observations re-arranged in order: (n = 16)
105 109 113 119 126 135 141 147 159 169 177 191 209 224 259 349

Median = (147 + 159)/2 = 153 US dollars
Q_1 = (119 + 126)/2 = 122.5 US dollars
Q_3 = (191 + 209)/2 = 200 US dollars

Five-number summary: Min, Q_1, Median, Q_3, Max
= 105, 122.5, 153, 200, 349 (in US dollars)

(b) Calculate the lower and upper limits in order to determine the adjacent values and find any potential outliers (if they occur).

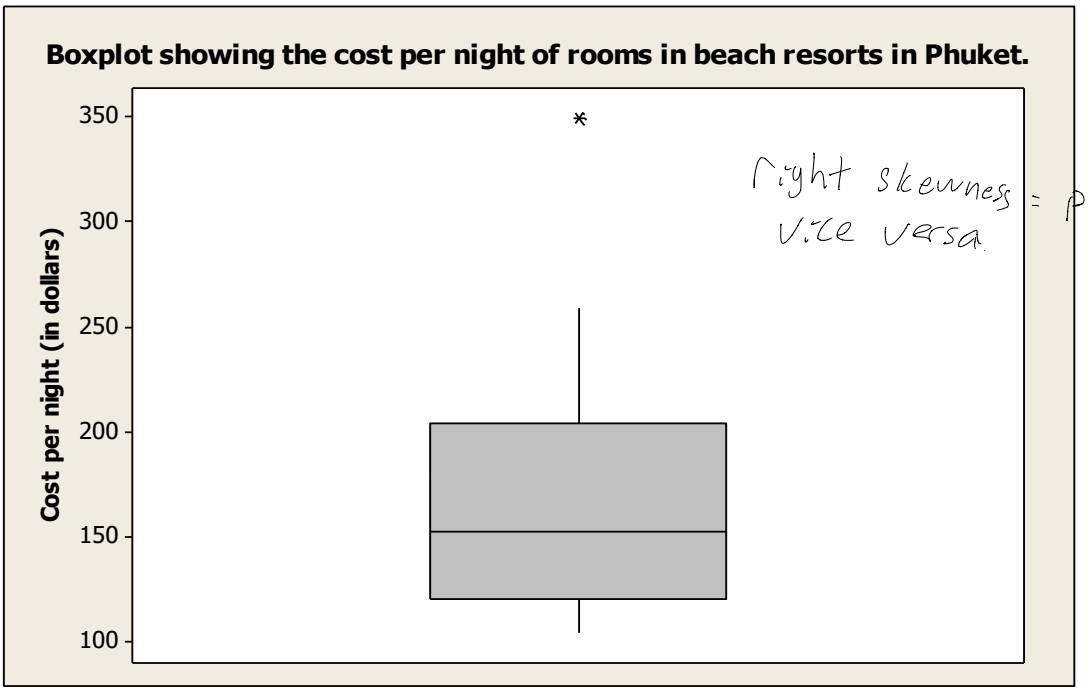
Lower limit = Q1 – 1.5 x IQR = 122.5 – 1.5(200 – 122.5) = 122.5 – 116.25 = 6.25 US dollars
Upper limit = Q3 + 1.5 x IQR = 200 + 1.5(77.5) = 200 + 116.25 = 316.25 US dollars

Since 349 is higher than the upper limit of 316.25, this observation is a potential outlier. The maximum is 349. There are no observations less than the lower limit of 6.25; therefore there are no potential outliers on the lower end of the distribution.

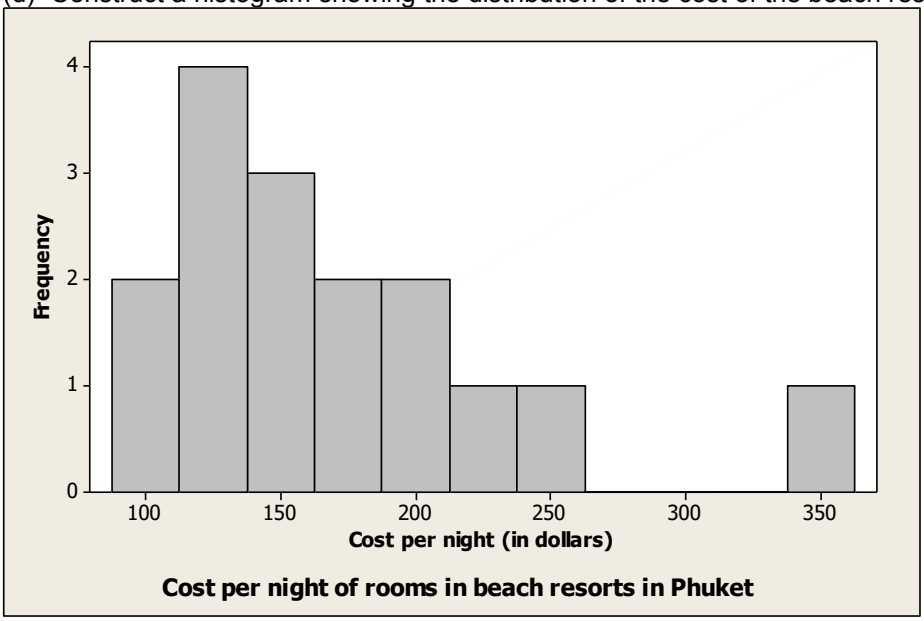
Adjacent values = 105 (on the lower end of the distribution) (also is the min)
= 259 (on the upper end of the distribution)

(c) Construct a boxplot showing the distribution of the cost of the beach resort rooms.

Explanation: To construct the boxplot, use Q1, Q2 and Q3 to make the box. The lower whisker should extend down to the lower adjacent value (105), which is also the minimum, and the upper whisker should extend up to the upper adjacent value (259), which is not the maximum since there is a potential outlier. The potential outlier (349) should be marked with an asterisk.



(d) Construct a histogram showing the distribution of the cost of the beach resort rooms.



Histograms: Show Outliers as a bar that is separated by a gap from all of the other bars.

(e) Describe the shape of this distribution.

Summary of Characteristics of Graphs and Numerical Summaries Used for Quantitative data

Comparing Groups with Graphs

- **Dotplots** – can give good visual comparison of two or many groups
- **Boxplots** – can give good visual comparison of two or many groups, using side-by-side boxplots
- **Stemplots** – can give good visual comparison of two groups back-to-back, but not more than two
- **Histograms** – possible to compare several groups, but visually they are not so easy to compare because they must be done separately

Other Advantages/Disadvantages of Types of Graphs

- **Histograms**
  - Can summarize large amounts of data
  - Cannot show detail, that is, each number or observation
- **Boxplots**
  - Can summarize large amounts of data
  - Cannot show detail, that is, each number or observation
- **Stemplots**
  - Cannot summarize large amounts of data
  - Can show detail, that is, show every number or observation
- **Dotplots**
  - Cannot summarize large amounts of data
  - Can show detail, that is, show every number or observation

Determining Shape of a Distribution

- There are two ways to determine shape:

1. Compare mean and median
   - If mean > median $\Rightarrow$ right skewed
   - If mean < median $\Rightarrow$ left skewed
   - If mean = median $\Rightarrow$ possibly (but not definitely) symmetric

2. Compare quartiles
   - If $Q_3 - Q_2 > Q_2 - Q_1 \Rightarrow$ right skewed
   - If $Q_3 - Q_2 < Q_2 - Q_1 \Rightarrow$ left skewed
   - If $Q_3 - Q_2 = Q_2 - Q_1 \Rightarrow$ possibly (but not definitely) symmetric

Choice of Measures of Center and Spread

- For symmetric distributions, the best measures of center and spread are mean and standard deviation, respectively
- For skewed distributions, the best measures or center and spread are median and IQR, respectively

1.9 The Normal Distribution

Density Curve = a model for a frequency distribution whereby the areas (or density) under the curve represents relative frequencies as well as probabilities

Area under curve = Relative frequency = Probability = Percentage of observations

Continuous probability model

- Form a **smooth curve**
- Used for continuous quantitative variables
- By contrast, a discrete quantitative variable (covered later) is presented in graphs with "steps"

The Normal Model

- The normal distribution is the most important distribution in statistics
- Many variables in both social and natural sciences are normally distributed
- The **normal distribution** is a specific type of **continuous density curve**.

Normally Distributed Variable = a variable that follows a normal, bell-shaped distribution and forms a normal curve

Approximately normally distributed population = population that approximately follows a normal curve

- Most populations are approximately normal, rather than completely normal

Characteristics of the normal curve (normal distribution):

- Bell shaped (a special type of symmetrical shape)
- Centered at the mean (μ)
- Is completely defined by its **mean** and **standard deviation**, which are called the **parameters** of the normal curve
- **Notation:** N(μ, σ) defines a given normal distribution
- The total area under the normal curve = 1
- The measures of center (mean, median and mode) all coincide.
- The normal curve extends indefinitely in both directions, approaching, but never touching, the horizontal axis.
- The normal curve follows the empirical rule

The Empirical Rule for the Normal Model (Also known as the 68.26 – 95.44 – 99.74 Rule)

Any normally distributed variable is distributed according to these properties:

- 68.26% of all observations lie within **one** standard deviation on either side of the mean.
- 95.44% of all observations lie within **two** standard deviations on either side of the mean
- 99.74% of all observations lie within **three** standard deviations on either side of the mean

Note: This rule, in fact, gives any normal curve its characteristic **bell-shape**.

Application of the z-score formula to a normal distribution

Standard normal distribution Or standard normal curve: N(μ, σ)

The standardized version of a normally distributed variable y is given by:

$$z = \frac{y - \mu}{\sigma}$$ [Note: the formula is the same as the z-score formula above]

A standardized normal variable always has:

- **Shape:** bell-shaped
- **Center:** Mean = 0
- **Spread:** Standard deviation = 1

Solving Problems Involving Normally Distributed Variables

Applying the Empirical Rule to a Normally Distributed Variable
Example: Suppose the weights of adults of a certain breed of chickens are normally distributed, with a mean of 1.36 kg and a standard deviation of 0.17 kg. [N(1.36, 0.17)] Calculate the weights of the chickens that are plus and minus 1, 2 and 3 standard deviations away from the mean and give the percentages of the population of chickens that have weights between these weights.

One standard deviation:
One standard deviation to the left = μ - σ = 1.36 – 0.17 = 1.19 kg
One standard deviation to the right = μ + σ = 1.36 + 0.17 = 1.53 kg.
So, 68.26% of the chickens have weights between 1.19 kg and 1.53 kg.

Two standard deviations:
Two standard deviation to the left = 1.36 – (2)(0.17) = 1.02 kg
Two standard deviation to the right = 1.36 + (2)(0.17) = 1.70 kg
So, 95.44% of the chickens have weights between 1.02 kg and 1.70 kg.

Three standard deviations:
Three standard deviation to the left = 1.36 – (3)(0.17) = 0.85 kg
Three standard deviation to the right = 1.36 + (3)(0.17) = 1.87 kg
So, 99.74% of the chickens have weights between 0.85 kg and 1.87 kg.

Determining Percentages or Probabilities for a Normally-Distributed Variable using the z-Score Formula

Return to the Previous Example: Suppose the weights of a certain breed of chickens are normally distributed, with a mean of 1.36 kg and a standard deviation of 0.17 kg.

(a) Find the percentage of chickens with weights between 1.0 kg and 1.5 kg.

Given y $\Rightarrow\Rightarrow$ **Find** z $\Rightarrow\Rightarrow$ **Find Area (%)**

For ***y*** = 1.0 kg $\Rightarrow\Rightarrow$ $z = \dfrac{y-\mu}{\sigma}$

$z = \dfrac{1.0-1.36}{0.17} = -2.1176 \approx -2.12$ $\Rightarrow\Rightarrow$ Area to the left is 0.0170

For ***y*** = 1.5 kg $\Rightarrow\Rightarrow$ $z = \dfrac{1.5-1.36}{0.17} = 0.8235 \approx 0.82$ $\Rightarrow\Rightarrow$ Area to the left is 0.7939

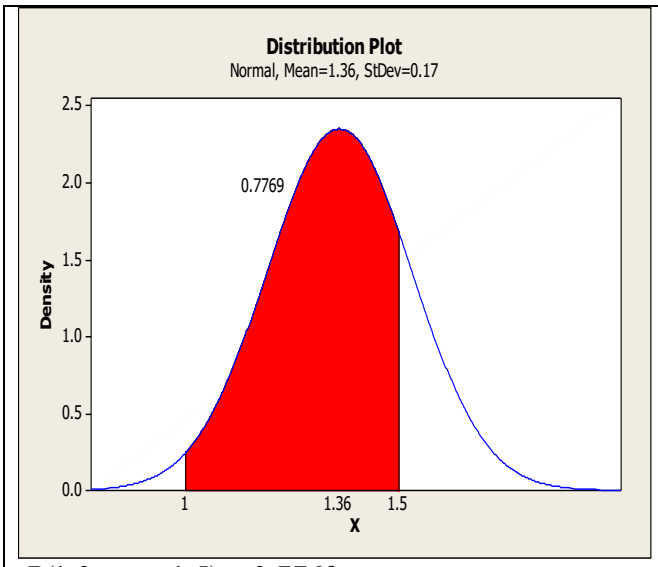
Area between is 0.7769

Interpretation: The percentage of chickens with weights between 1.0 kg and 1.5 kg is 77.69%.

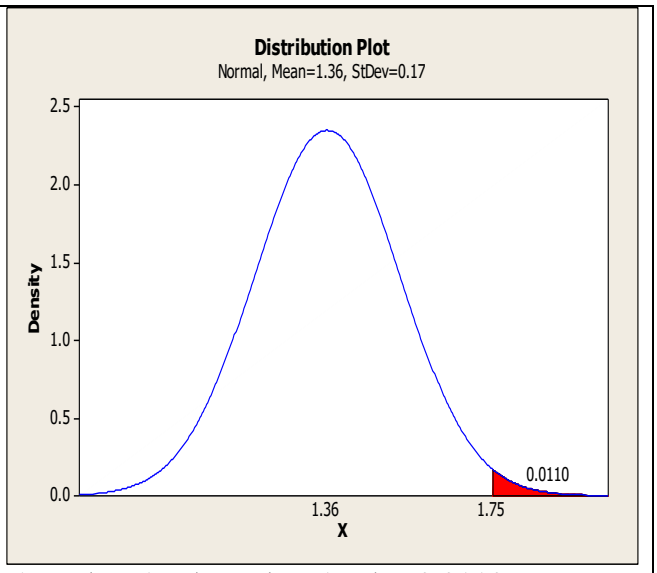
(b) Find the percentage of chickens with weights greater than 1.75 kg.

$$z = \frac{y-\mu}{\sigma} = \frac{1.75-1.36}{0.17} = 2.29$$

For z = 2.29, area to the left = 0.9890
So, area to the right = 1 – 0.9890 = 0.0110
Interpretation: The percentage of chickens with weights greater than 1.75 kg is 0.0110 or 1.1%.



$P(1.0 < y < 1.5) = 0.7769$



Area $(y > 1.75) = P(y > 1.75) = 0.0110$

Graph for part (a) **Graph for part (b)**

Finding the Observations for a Specified Percentage (Percentiles) for a Normally Distributed Variable

- This is the reverse of the previous example

Step 1: Use the standard normal table <u>in the reverse way</u> to find the z-value for a given area under the curve (or specified percentage)

Step 2: Substitute the obtained z-value into the z-score formula and **solve for y**.

Solving for *y* in the *z*-score formula:

If we solve the formula $z = \dfrac{y - \mu}{\sigma}$ for y, we get:

$$y = \mu + (z)(\sigma)$$

Return to the Example on Weights of Chickens: (mean = 1.36 kg, standard deviation = 0.17 kg)
Find the 90th percentile (P_{90}) for these weights of chickens.
[In other words, find the weight (***y***-value) that is higher than the weights of 90% of all chickens of this variety.] Note: The ***z***-score for P_{90} is the one having an area of 0.90 to its left under the standard normal curve.

Find y	$\Leftarrow\Leftarrow$	**Find** z	$\Leftarrow\Leftarrow$	**Given Area (%)**
$y = \mu + (z)(\sigma)$	$\Leftarrow\Leftarrow$	$z = 1.28$	$\Leftarrow\Leftarrow$	0.90 (≈ 0.8997)

$y = 1.36 + (1.28)(0.17) = 1.58$ kg

Interpretation: The 90th percentile for these weights of chickens is 1.58 kg.
Note: This also means that the <u>top 10% heaviest chickens</u> are those with weights greater than 1.58 kg.

1.10 The Sampling Distribution of the Sample Mean

1.10.1 Sampling Error and Sampling Distributions

Population distribution = the distribution of all values of a variable in a population
Sampling distribution of the sample mean = distribution of the values of the variable $\bar{y}$, for a variable ***y*** and a given sample size ***n.*** In statistics, this term is equivalent to the terms:

- Distribution of the variable $\bar{y}$, and
- Distribution of all possible sample means of a given sample size

Sampling Error = the error resulting from using a sample to estimate a population characteristic, e.g. mean, standard deviation

Let's demonstrate this statement with an example:
Example: In a certain hospital, 5 baby girls were born on a particular day and their birth weights are shown in the table below. The mean birth weight (***μ***) of this small population was 3.06 kg.

Table: Birth weights of a small population of 5 baby girls born on the same day.

Baby	Ann	Bev	Carol	Deb	Eva
Weight (kg)	2.8	3.3	3.1	2.5	3.6

Table: All possible samples (10) and sample means for samples of size 2.

Sample	Weights (kg)	Sample mean ($\bar{y}$)
AB	2.8, 3.3	3.05
AC	2.8, 3.1	2.95
AD	2.8, 2.5	2.65
AE	2.8, 3.6	3.20
BC	3.3, 3.1	3.20
BD	3.3, 2.5	2.90
BE	3.3, 3.6	3.45
CD	3.1, 2.5	2.80
CE	3.1, 3.6	3.35
DE	2.5, 3.6	3.05

Table: All possible samples (10) and sample means for samples of size 3.

Sample	Weights (kg)	Sample mean ($\bar{y}$)
ABC	2.8, 3.3, 3.1	3.07
ABD	2.8, 3.3, 2.5	2.87
ABE	2.8, 3.3, 3.6	3.23
ACD	2.8, 3.1, 2.5	2.80
ACE	2.8, 3.1, 3.6	3.17
ADE	2.8, 2.5, 3.6	2.97
BCD	3.3, 3.1, 2.5	2.97
BCE	3.3, 3.1, 3.6	3.33
BDE	3.3, 2.5, 3.6	3.13
CDE	3.1, 2.5, 3.6	3.07

Check to see if we got the correct number of samples (combinations) for all possible samples of size 3 (n = 3) from a population of size 5 (N = 5):

$$ {}_N\mathrm{C}_n = {}_m C_r = \frac{m!}{r!(m-r)!} == \frac{5!}{3!(5-3)!} = \frac{5\times 4\times \cancel{3!}}{\cancel{3!}\times 2\times 1} = 10 $$

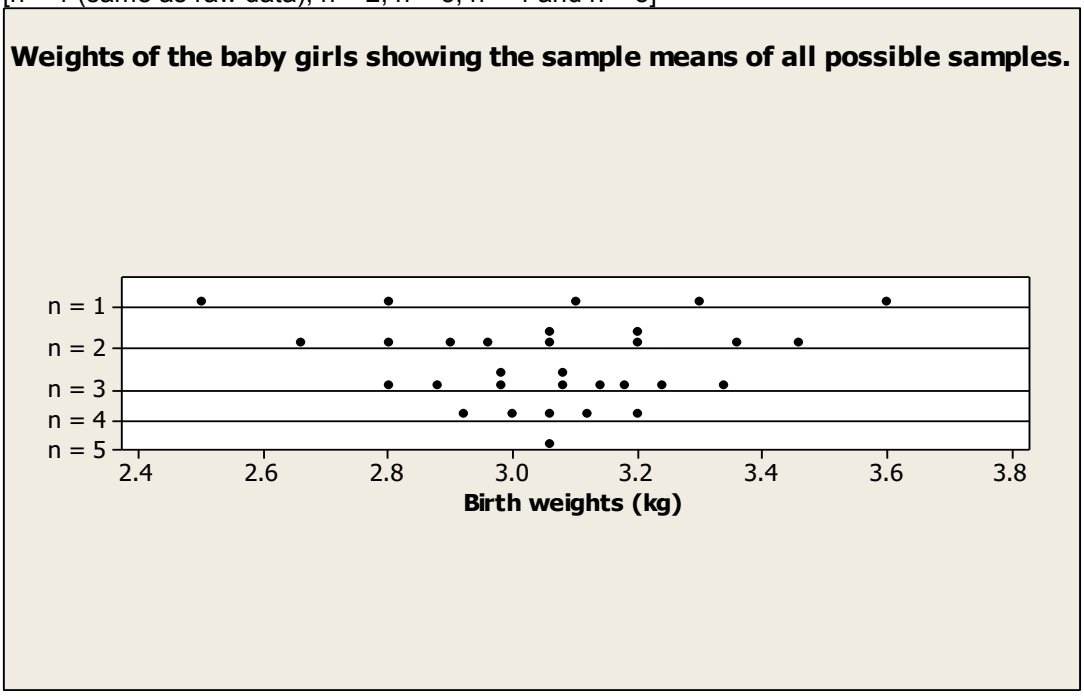
Table: All possible samples (5) and sample means for samples of size 4.

Sample	Weights (kg)	Sample mean ($\bar{y}$)
ABCD	2.8, 3.3, 3.1, 2.5	2.925
ABCE	2.8, 3.3, 3.1, 3.6	3.200
ABDE	2.8, 3.3, 2.5, 3.6	3.050
ACDE	2.8, 3.1, 2.5, 3.6	3.000
BCDE	3.3, 3.1, 2.5, 3.6	3.125

Table: All possible samples (1) and sample means for samples of size 5.

Sample	Weights (kg)	Sample mean ($\bar{y}$)
ABCDE	2.8, 3.3, 3.1, 2.5, 3.6	3.06

Dotplot Showing the Sample Means from the Tales Above
[n = 1 (same as raw data), n = 2, n = 3, n = 4 and n = 5]



What do these dotplots show about the sampling distributions of the sample means?

1. **Shape –** all distributions are normally distributed (though this is difficult to see with so few data points)
2. **Center –** the same for all sample sizes
3. **Spread –** decreases as sample size increases

1.10.2 The Mean and Standard Deviation of the Sample Mean

Mean of the Sample Mean

Mean of the Sample Mean

For samples of size ***n***, the mean of the sample means (i.e., the mean of the variable $\bar{y}$) equals the mean of the population or variable under study, i.e.:

$$\mu_{\bar{y}} = \mu$$

- This means that the mean of all possible sample means equals the population mean
- This holds true for any sample size ***n***
- When we refer to the mean of the sample mean, we change the symbol from $\bar{y}$ to $\mu_{\bar{y}}$ because it actually becomes a variable of its own

- **<u>Note the difference between the following:</u>**
  - **When taking any <u>one</u> sample that is small in size, the sample mean will vary from the population mean (called sampling error), but this error will decrease with larger sample size;**
  - **However, when you take the mean of <u>all possible sample</u> <u>means (even if they are small samples)</u>, it will equal the population mean, regardless of sample size**

Returning to the Example: (Population of the weights of the 5 newborn baby girls)
Find the mean of the sample means of all possible outcomes of sample size ***n*** and compare each to the population mean.

First, find the population mean:

Population mean (***μ***) = $\dfrac{\sum y_i}{N} = \dfrac{2.8+3.3+3.1+2.5+3.6}{5}$ = 3.06 kg

Then, calculate the mean of the sample means for all possible samples of size ***n*** using the formula:

$$\text{Mean of the sample mean} = \frac{\text{Summation of all possible sample means}}{\text{Number of possible samples}}$$

Mean of the sample mean for n = 1 (mean of the variable $\bar{y}$ for n = 1):

$\mu_{\bar{y}} = \dfrac{2.8+3.3+3.1+2.5+3.6}{5}$ = 3.06 kg [Same as the calculation of the population mean]

Mean of the sample mean for n = 2 (mean of the variable $\bar{y}$ for n = 2):

$\mu_{\bar{y}} = \dfrac{3.05+2.95+2.65+3.20+3.20+2.90+3.45+2.80+3.35+3.05}{10}$ = 3.06 kg

Mean of the sample mean for n = 3 (mean of the variable $\bar{y}$ for n = 3):

$\mu_{\bar{y}} = \dfrac{3.07+2.87+3.23+2.80+3.17+2.97+2.97+3.33+3.13+3.07}{10} = 3.06$ kg

Mean of the sample mean for n = 4 (mean of the variable $\bar{y}$ for n = 4):

$$\mu_{\bar{y}} = \frac{2.925 + 3.200 + 3.050 + 3.000 + 3.125}{5} = 3.06 \text{ kg}$$

Mean of the sample mean for n = 5 (mean of the variable $\bar{y}$ for n = 5):

- There is only one possible sample of size 5, so:

$$\mu_{\bar{y}} = \frac{3.06}{1} = 3.06 \text{ kg}$$

Conclusion: Regardless of sample size, the mean of the sample means
= mean of the variable $\bar{y}$
= the mean of the population

Standard Deviation of the Sample Mean

Formula for Calculation of the Standard Deviation of the Sample Mean

For samples of size ***n***, the standard deviation of the sample mean (or the standard deviation of the variable $\bar{y}$) equals the standard deviation of the variable under study divided by the square root of the sample size, i.e.:

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$$

This is also referred to as the **standard error (SE) of the sample mean** because it determines the amount of sampling error to be expected when a population mean is estimated by a sample mean.

Note: This formula applies to:

- sampling **with replacement from a finite population**
- Or, sampling **from an infinite population (or very large population)** with or without replacement

- When sample size is small relative to population size (n ≤ 0.05N), which is the usual case in most practical applications, there is little difference between sampling with replacement and without replacement; therefore, this formula is usually applied (though sometimes, the equality may be somewhat approximate)

As sample size *n* gets larger and larger,

- The standard deviation of the sample means (or the standard deviation of $\bar{y}$) gets smaller and smaller until, when ***n*** = ***N***, the standard deviation of the sample means = 0.

Example: Find the standard deviation of the sample means of the weights of the newborn baby girls for all possible outcomes of sample size ***n*** and compare each to the population standard deviation. (Recall: μ = 3.06 kg)

Using the defining formula for population standard deviation:

$$\sigma = \sqrt{\frac{\sum (y_i - \mu)^2}{N}}$$

Considering the entire population (5 babies), population standard deviation is:

$$\sigma = \sqrt{\frac{(2.8-3.06)^2 + (3.3-3.06)^2 + (3.1-3.06)^2 + (2.5-3.06)^2 + (3.6-3.06)^2}{5}} = 0.383kg$$

Standard deviation of the sample means for n = 2:

$$\sigma_{\bar{y}} = \sqrt{\frac{\begin{array}{l}(3.05-3.06)^2 + (2.95-3.06)^2 + (2.65-3.06)^2 + (3.20-3.06)^2 + (3.20-3.06)^2 \\ +(2.90-3.06)^2 + (3.45-3.06)^2 + (2.80-3.06)^2 + (3.35-3.06)^2 + (3.05-3.06)^2\end{array}}{10}} = 0.234kg$$

Standard deviation of the sample means for n = 3:

$\sigma_{\bar{y}}$ = 0.155 kg

Standard deviation of the sample means for n = 4:

$\sigma_{\bar{y}}$ = 0.096 kg

Standard deviation of the sample means for n = 5:

$\sigma_{\bar{y}}$ = 0 kg (There is only one sample mean ($\bar{y} = 3.06$) so deviation equals 0.)

Conclusion: As sample size gets larger, the standard deviation of the sample means gets smaller until, when n = N, the standard deviation of the sample means = 0.

Summary of Results

Sample size (n)	Standard deviation of $\bar{y}$ (based on actual calculations above, **without replacement**)	Using the formula: $\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$ (with replacement)	Use formula for sampling **without replacement** from **finite populations** $\sigma_{\bar{x}} = \sqrt{\frac{N-n}{N-1}} \times \frac{\sigma}{\sqrt{n}}$
1	0.383	0.383	0.383
2	0.234	0.271	0.234
3	0.155	0.221	0.155
4	0.096	0.192	0.096
5	0.000	0.171	0.000

- The discrepancy here is because this example is based on sampling **without replacement** from a **finite (very small) population**
- Usually population size is much larger than that, so the discrepancy would be negligible

1.10.3 The Sampling Distribution of the Sample Mean

Describing the sampling distribution of the sample mean (or the mean of all possible sample means) involves the following 3 aspects:

1. The shape of the distribution
2. The mean (center)
3. The standard deviation (spread)

Sampling Distribution for a Normally Distributed Variable

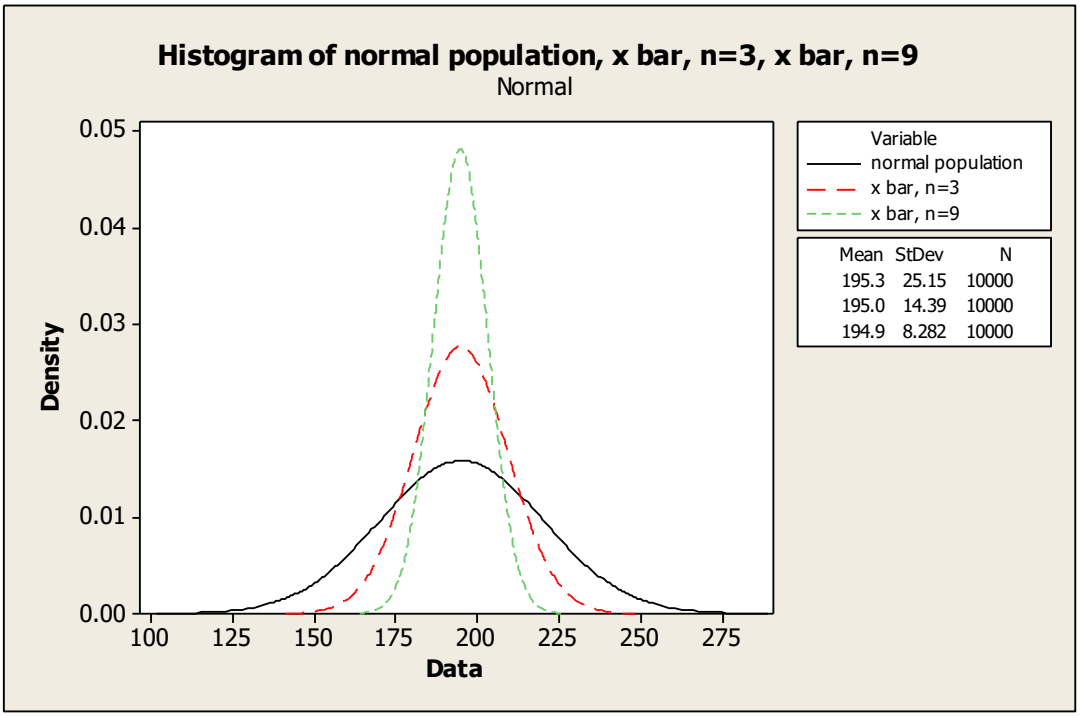
Sampling Distribution of the Sample Mean for a Normally Distributed Variable

If a variable ***x*** of a population is normally distributed with mean ***μ*** and standard deviation σ , then, for samples of size ***n*** (even if ***n*** is small):

1. **Shape:** The sampling distribution of all possible sample means (known as variable $\bar{x}$) is also normally distributed
2. **Center:** The mean of the sampling distribution is: $\mu_{\bar{x}} = \mu$
3. **Spread:** The standard deviation of the sampling distribution is: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

Example: There are millions of wildebeests in Serengeti Game Park in Tanzania. The body weights of adult wildebeests are normally distributed, with a mean of 195 kg and standard deviation of 25 kg. (Range = about 120 – 270 kg.)

Graph showing Sampling Distributions of Samples of Different Sizes
Three curves in the graph: one for the population data, one for the sampling distribution of the sample mean when n = 3 and one for n = 9. (The sampling distribution of the sample mean = the means of all possible sample means.)



[This graph was done in Minitab by generating 10,000 rows of normally distributed data (mean = 195, sigma = 25) in 9 columns. One column of data was used to draw the curve for the normal population, the means of 3 columns were calculated and used to make the graph for n = 3 and the means of all 9 columns were used to make the graph for n = 9.]

Note:

- The mean is almost exactly the same for all of them
- The standard deviation of the sampling distributions = $\frac{\sigma}{\sqrt{n}}$
- The larger the sample size, the smaller will be the sampling error of the sampling distribution (as indicated by the decreasing standard deviation)
- Sampling distributions of all possible sample means from a normally distributed population are also normally distributed

The Standardized Version of the variable $\bar{y}$ (the sample mean)

Standardized version of the variable $\bar{y}$ (the sample mean):

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

Example of a Normally Distributed Population: Calculations based on the entire population and also based on a sampling distribution

Coelacanths are lobed-fin fish that were thought be have become extinct at least 65 million years ago, but that were re-discovered in 1938 in the deep sea off the coast of South Africa, though they are very rare. The body weights of this population are normally distributed with a mean of 80 kg and a standard deviation of 9 kg.

(a) Determine the percentage of coelacanths that have body weights between 75 and 90 kg.

For *x* = 75 kg: $z = \frac{y - \mu}{\sigma} = \frac{75 - 80}{9} = -0.5556 \approx -0.56$

For *x* = 90 kg: $z = \frac{90 - 80}{9} = 1.1111 \approx 1.11$

Using the Table for the Standard Normal Curve, find the following areas:

Area to the left of z = – 0.56 is 0.2877
Area to the left of z = 1.11 is 0.8665

So, $P(75 < Y < 90) = 0.8665 - 0.2877 = 0.5788$

Interpretation: The percentage of coelacanths having body weights between 75 and 90 kg is 0.5788 or 57.88%.

(b) Determine the 90th percentile of the body weights of coelacanths.

Area of 0.90 (approximately 0.8997) under the standard normal curve corresponds to a *z*-score of 1.28.

$y = \mu + (z)(\sigma)$ = 80 + (1.28)(9) = 91.52 kg

Interpretation: The 90th percentile of the body weights of coelacanths is 91.52 kg or 90% of the body weights of coelacanths are less than 91.52 kg.

(c) Describe the sampling distribution of the sample mean for random samples of 10 coelacanths and explain the logic of your answer.

1. Shape: Since the population of coelacanths have normally distributed body weights, even if a small sample size is taken where n = 10, the sampling distribution is approximately normally distributed.
2. Center: The mean of the sampling distribution is:
$$\mu_{\bar{y}} = \mu = 80 \text{ kg}$$
3. Spread: The standard deviation of the sampling distribution is:
$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}} = \frac{9}{\sqrt{10}} = 2.8460 \approx 2.8 \text{ kg}$$

(d) Suppose that you randomly sample 10 coelacanths, determine the percentage of all samples of 10 coelacanths that have mean body weights between 75 kg and 90 kg.

Although the sample size small (n = 10), since the population is normally distributed, the sampling distribution is also normal and therefore the standardized version of variable ($\bar{y}$) can be applied.

For $\bar{y}$ = 75 kg: $$z = \frac{\bar{y} - \mu}{\sigma / \sqrt{n}} = \frac{75 - 80}{9 / \sqrt{10}} = \frac{-5}{2.8460} = -1.7568 \approx -1.76$$

For $\bar{y}$ = 90 kg: $$z = \frac{90 - 80}{9 / \sqrt{10}} = \frac{10}{2.8460} = 3.5137 \approx 3.51$$

Using the Table for the Standard Normal Curve, find the following areas:
Area to the left of z = – 1.76 is 0.0.0392
Area to the left of z = 3.51 is 0.9998

So, $P(75 < \bar{y} < 90) = 0.9998 - 0.0392 = 0.9606$

Interpretation: The percentage of all samples of 10 coelacanths that have mean body weights between 75 kg and 90 kg is 0.9606 or 96.06%.

(e) Note the difference between the answers in:
Part (a) – Entire population [$P(75 < Y < 90) = 0.8665 - 0.2877 = 0.5788$]
and Part (d) – Sampling distribution[$P(75 < \bar{y} < 90) = 0.9998 - 0.0392 = 0.9606$]

The Sampling Distribution for <u>ANY</u> Type of Distribution

The Central Limit Theorem (CLT)

- One of the most important theorems in Statistics

> **The Central Limit Theorem (CLT)**
>
> Regardless of the distribution of the variable under study, for a relatively <u>large sample size</u>, the variable $\bar{y}$ is approximately normally distributed. The approximation becomes better with increasing sample size.

How Large is Relatively Large???

- The farther the variable under study is from being normally distributed, the larger the sample size must be in order for variable $\bar{y}$ to be approximately normally distributed

> **Simple Rule for Relatively Large Sample Size**
>
> - **Usually, a sample size of 30 or more ($n \geq 30$) is large enough**

>>>>>>>>>>

Sketch graphs of normal, reverse J-shaped and uniform variables for entire population, n = 2, n = 10 and n = 30 (after Weiss, p. 313)

>>>>>>>>>>

Sampling Distribution for ANY Variable that is NOT Normally Distributed

Sampling Distribution of the Sample Mean for a Variable that is NOT Normally Distributed

1. **Shape:** Regardless of the distribution of ***x***, if the sample size is large (n ≥ 30), the sampling distribution of all possible sample means (i.e., the distribution of the variable $\overline{x}$) is approximately normally distributed.sampling distribution of all possible sample means (known as variable $\overline{x}$) is also normally distributed.

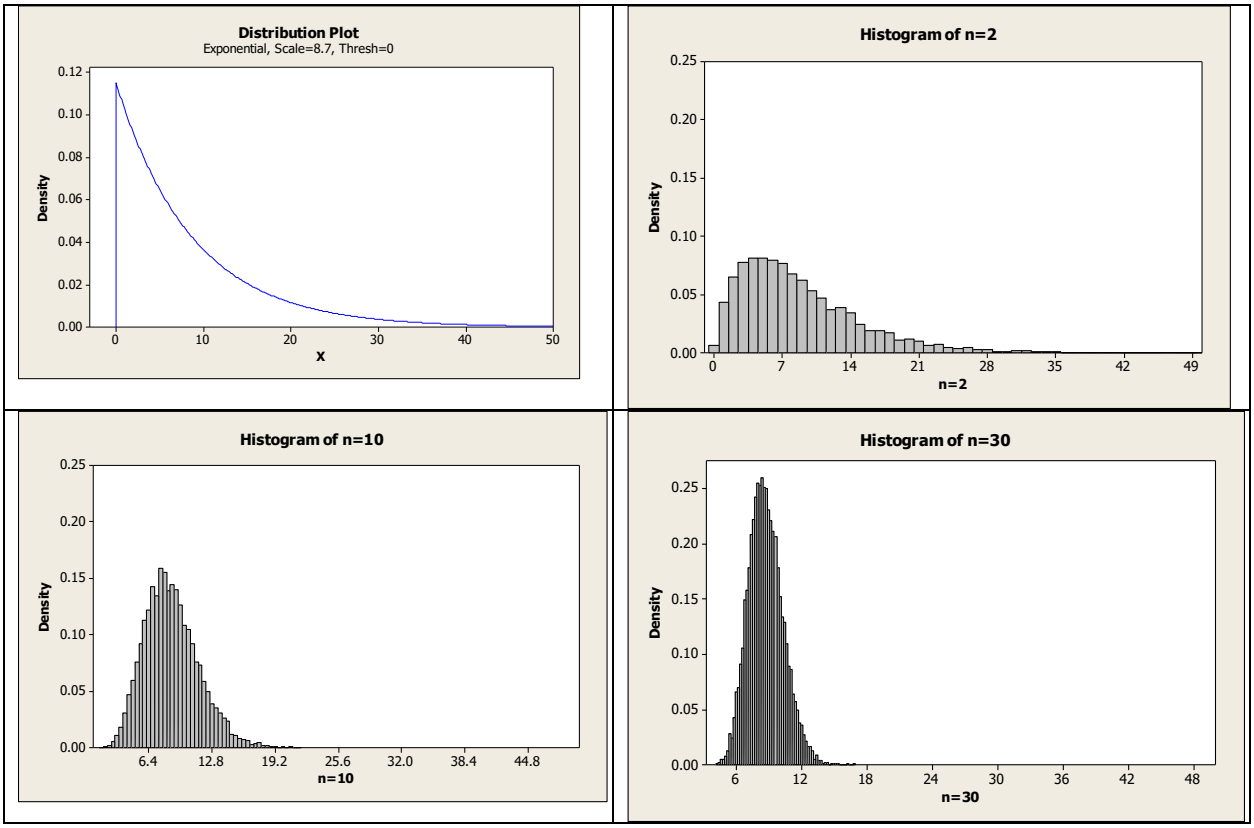
2. **Center:** The mean of the sampling distribution is: $\mu_{\bar{x}} = \mu$

3. **Spread:** The standard deviation of the sampling distribution is: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

Example of an Exponential Distribution

At a certain emergency hospital, the time from the arrival of one patient to the next (known as interarrival time) has an exponential distribution with a mean of 8.7 minutes and a standard deviation of 8.7 minutes.

(a) Graph the exponential distribution of the population and the sampling distributions for n = 2, n = 10 and n = 30.



[Compare these graphs based on the exponential distribution with the Reverse-J shape on the previous page]

(b) Determine the sampling distribution of the sample mean for samples of size 30 and explain the logic of your answer. (Note this means taking all possible sample or at least a large number of samples, e.g., 10000 samples, each with sample size 30.)

1. Shape: According to the CLT, since sample size is large (n = 30), the sampling distribution of the sample mean is approximately normally distributed (even though the population follows an exponential distribution)
2. Center: The mean of the sample means is:

$$\mu_{\bar{y}} = \mu = 8.7 \text{ minutes}$$

3. Spread: The standard deviation of the sample means is:

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}} = \frac{8.7}{\sqrt{30}} = 1.588 \text{ minutes}$$

(c) If we randomly select 36 interarrival times, find the probability that the average interarrival time is more than 10 minutes.

According to the CLT, when sample size is large as it is in this case (n = 36, which is > 30), any distribution approaches the normal distribution, so we use the normal probability distribution in these calculations.

$$z = \frac{\bar{y} - \mu}{\sigma / \sqrt{n}} = \frac{10 - 8.7}{8.7 / \sqrt{36}} = \frac{1.3}{1.45} = 0.8966 \approx 0.90$$

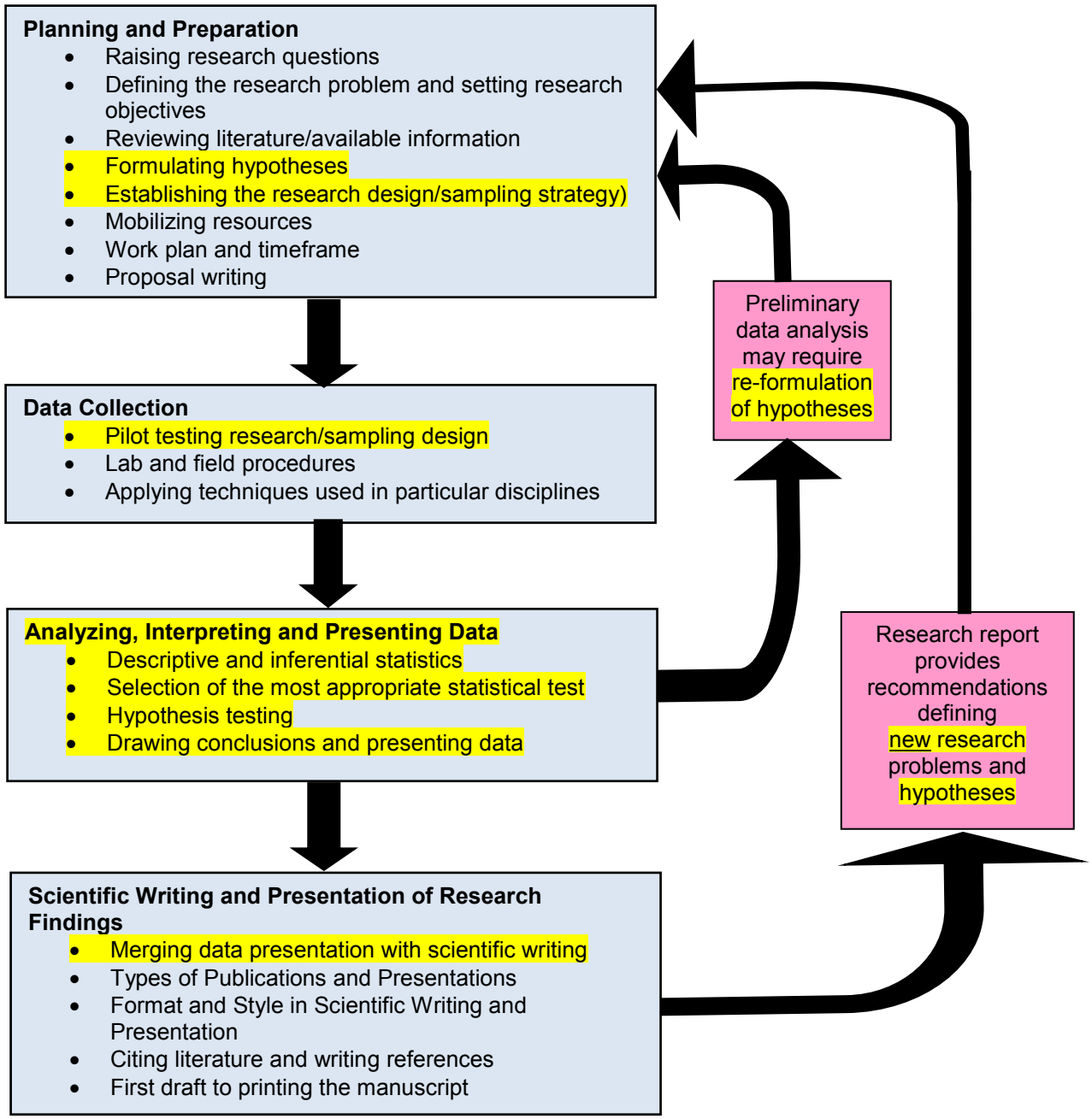
For z = 0.90, area to the left = 0.8159
So, area to the right = 1 – 0.8159 = 0.1841

Interpretation: The probability that the average interarrival time is more than 10 minutes is 0.1841.

1.11 Inferential Statistics: The Concept and Processes of Hypothesis testing and Determining Confidence Intervals

- This is an introduction to Inferential Statistics
- Statistical analysis is an essential and integral part of research and the scientific process

Flow Chart of the Scientific Process



- Statistics is required every step of the way (all components highlighted in yellow)

- Inferential Statistics involves drawing conclusions about an entire population based on analysis of data obtained from a sample. It includes two main aspects:
    - Hypothesis Testing
    - Confidence Intervals
- For any given research problem, there is an appropriate hypothesis test and a corresponding confidence interval that can be applied to solve the problem
- The appropriate hypothesis test and the corresponding confidence interval give the same conclusion

1.11.1 Research Objectives and Formulation of Hypotheses

- Discussion of this topic will help to connect the study of statistics with its important application in research.

Research objectives

- Scientists raise **research questions** which lead to objectives
- Research objectives are the **focal point** of any research project
- Generally, each objective has **two possible outcomes**
    - These form **two possible hypotheses:**
        - The **null hypothesis** and
        - The **alternative hypothesis**
- **Example**:
    - **Research Question:** Which of two varieties of rice give better yield?
    - **Research objective:** To determine whether there is a difference in the yield of two varieties of rice
    - Each objective has two possible outcomes or hypotheses
        - **Null hypothesis:** There is no significant difference in yield between the two varieties of rice
        - **Alternative hypothesis:** There is a significant difference in yield between the two varieties of rice

Null and Alternative Hypotheses

Null hypothesis (symbolized as $\mathbf{H_0}$): A statement that says that there is no difference (among groups, etc.) or no relationship (among variables)

- Null refers to "nothing"
- This is actually the hypothesis that is being tested in an inferential statistical test

Alternative hypothesis (symbolized as $\mathbf{H_a}$): A statement which gives the alternative to the null hypothesis, i.e., it states that there is a difference or there is a relationship.

Research Hypothesis versus Statistical Hypotheses

Research hypothesis

- The researcher makes a prediction about the one outcome that he/she expects will be verified by the study, based on knowledge of the field of study and review of the relevant literature on the topic
- This **predicted outcome** is the **research hypothesis**
- **Usually it is the alternative hypothesis**, but occasionally it may be the null hypothesis
- After completing the study, the researcher applies inferential statistical analysis to data from a **sample** in order to test his/her hypothesis to determine whether it is true
    - Then the researcher draws a conclusion or makes inferences about the entire **population** under study

Statistical hypotheses

- These are the null and alternative hypotheses
- When performing statistical analysis on the data, it is actually the null hypothesis that is tested

Two types of objectives/hypotheses encountered in research

1. **Differences** between two or more groups/treatments of **one variable**
    - Analyzed with such inferential statistical tests such as the two-sample ***t*** test and ANOVA

    Example:
    Null hypothesis: There is no significant difference in the effectiveness of four types of drugs in the treatment of malaria.

    Alternative hypothesis: There is a significant difference in the effectiveness of four types of drugs in the treatment of malaria.

2. **Relationships** between **two or more variables**, which could be positive or negative (inverse) relationships
    - Analyzed with inferential statistical procedures such as correlation and regression analysis

    Example:
    Null hypothesis: There is no significant relationship between ongoing mental activity and Alzheimer's disease.

    Alternative hypothesis: There is a significant relationship between ongoing mental activity and the occurrence of Alzheimer's disease

    (There is evidence that there is an inverse relationship between these two variables, i.e., active use of the mental faculties appears to decrease the chances of getting Alzheimer's disease)

Two-tailed and One-tailed hypotheses

- A hypothesis test may have two-tailed hypotheses or one-tailed hypotheses
    - Then the test may be referred to as a two-tailed test or a one-tailed test
- A one-tailed test, may be a left-tailed test or a right-tailed test
- Thus, for any given research objective, there are 3 possible types of tests or hypotheses, each with a null hypothesis and an alternative hypothesis

Example:
Research Objective: To determine which variety of rice gives the highest yield, a new variety (Variety N) or the commonly used variety in a certain area (Variety C).

- **Two-tailed hypotheses (for a two-tailed test)**
  Null hypothesis: $H_0 : \mu_N = \mu_C$
  There is no difference in mean yield between the new rice variety and the common variety.

  Alternative hypothesis: $H_a : \mu_N \neq \mu_C$
  There is a difference in mean yield between the new rice variety and the common variety.

- **One-tailed hypotheses (left-tailed)**
  Null hypothesis: $H_0 : \mu_N = \mu_C$
  There is no difference in mean yield between the two varieties.
  OR: The yield of the new rice variety is not less than the yield of the common rice variety.

  Alternative hypothesis: $H_a : \mu_N < \mu_C$
  The yield of the new rice variety is less than the yield of the common rice variety.

- **One-tailed hypotheses (right-tailed)**
  Null hypothesis: $H_0 : \mu_N = \mu_C$
  There is no difference in mean yield between the two varieties.
  OR: The yield of the new rice variety is not greater than the yield of the common rice variety.

  Alternative hypothesis: $H_a : \mu_N > \mu_C$
  The yield of the new rice variety is greater than the yield of the common rice variety.

Which of the above 6 hypotheses is an agricultural researcher likely to choose as his/her research hypothesis??? Why???

1.11.2 Test Statistic, Critical Values, Rejection Region and Nonrejection Region

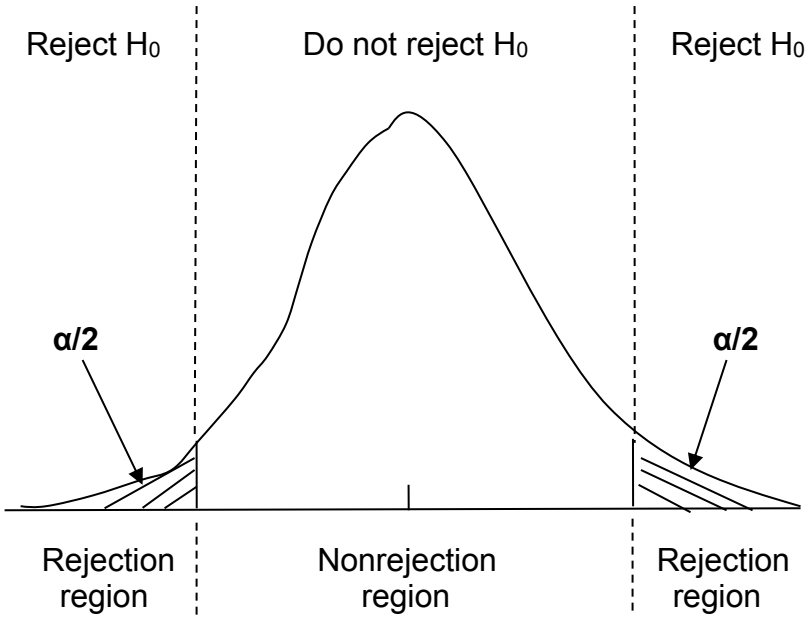
Test statistic = the statistic used as a basis for deciding whether the null hypothesis should be rejected, e.g., t statistic, z statistic or correlation coefficient, r.

- **Calculated value (observed value) of the test statistic**
  - calculated from the data collected
- **Critical value of the test statistic**
  - obtained **from a table** showing its theoretical distribution, which is compared with the calculated value in order to make a decision about the hypotheses
  - Forms the border separating the rejection and nonrejection regions (the critical value itself is considered to be part of the rejection region)

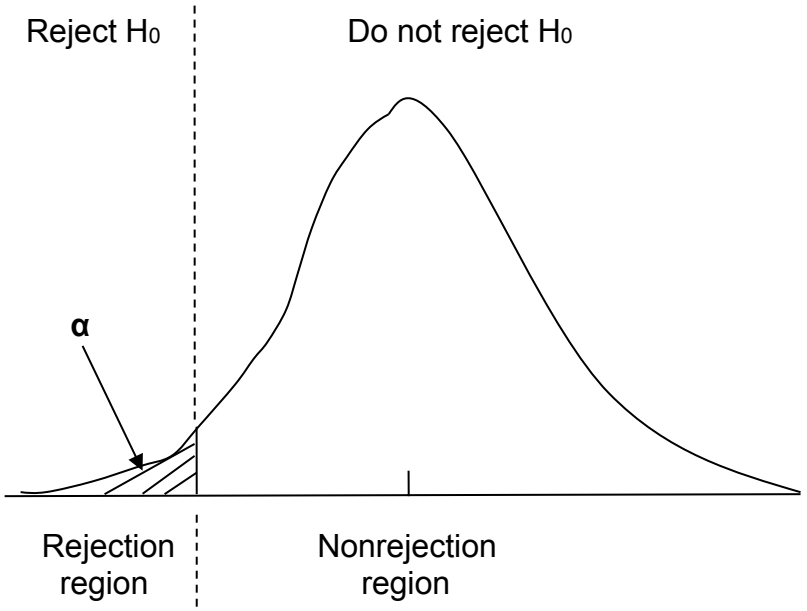
Rejection region = the set of values for the test statistic that leads to rejection of the null hypothesis

Nonrejection region = the set of values of the test statistic that leads to nonrejection of the null hypothesis

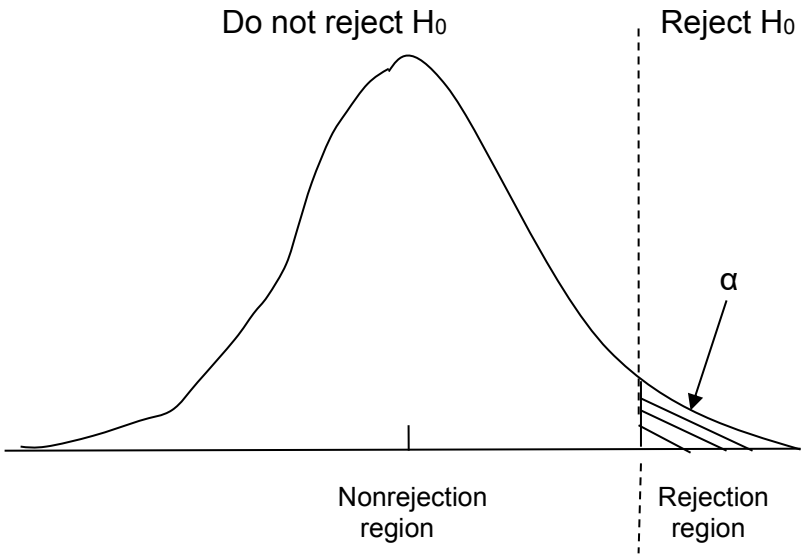
Alpha (α) = Type I error (defined below)



Rejection/Nonrejection regions for two-tailed tests



Rejection/Nonrejection region for left-tailed tests



Rejection/Nonrejection regions for right-tailed tests

1.11.3 Type I and Type II Errors and the Power of the Test

		Null Hypothesis (H_0) (in reality) is:	
		True	**False**
Decision	**Do not reject H_0**	Correct decision (no error)	Type II error
	Reject H_0	Type I error	Correct decision (no error)

[Give 2 examples of M vs. F: intelligence; physical strength]

Type I error = rejecting H_0 when it should not be rejected because it is in fact true. [You concluded there is a difference when actually there isn't.]

- P (Type I error) = α (significance level),
  i.e., the probability of committing a Type I error is α (alpha).
- The Type I error can be determined by examining the theoretical distribution of the test statistic.
- The risk factor

Type II error = not rejecting H_0 when it should be rejected because it is in fact false. [You concluded there is no difference when actually there is.]

- p (Type II error) = β (Beta)
- The Type II error is generally not determined in a hypothesis test

Power of a statistical test = 1-β = a correct decision
= probability of rejecting a H_0 when it should be rejected because it is in fact false

- The power of a test can be increased [or the p (Type II error) can be decreased] by:
  - Increasing sample size (n)
  - Selecting the most powerful statistical test for the situation

Relationship between Type I and Type II Error Probabilities

- For a fixed sample size, the smaller we specify the significance level, α, the larger will be the probability, β, of not rejecting a false hypothesis.
- Balancing Type I and Type II Error probabilities is important
  - Assess the risks involved in each
  - The only way of decreasing both types of error simultaneously is increasing sample size

1.11.4 Steps in Testing Hypotheses Statistically

Steps in Hypothesis Testing

Step 1: Choose appropriate statistical test based on purpose and assumptions
[e.g., t test, analysis of variance, correlation, etc.]

- Consider: Type of hypothesis you are testing (difference between groups of one variable or relationship between variables)
- Consider: types of variables, number of populations, etc.
- Consider: what is given and what is asked for (purpose)
  [See "Diagram for Selection of Hypothesis Tests"]

Step 2: State Hypotheses
[Null hypothesis (H_0) and Alternative hypothesis (H_a)]
Also, identify the significance level (α).

- Generally should be set by the researcher in advance
- In this course, it will usually be specified in the question
- If a question does not specify alpha (α), a common alpha to assume is $\alpha = 0.05$ or 5%, (which means that there should be less than a 5% chance of making a mistake)

Step 3: Calculate the test statistic.

- Gives the calculated value (or observed value) of the test statistic.

Step 4: Decide to reject H_0 or not reject H_0 and state the strength of the evidence against H_0

- Find the P-value (probability of a Type I error) by examining the table of the theoretical distribution showing the critical values of the test statistic (e.g., t-table, F-table, etc.) at the appropriate ***n*** or ***df***
- **Apply rules for rejecting H_0 or not rejecting H_0 (P-value approach)**

  1. If the P-value ≤ α, we reject H_0.
     [and conclude that the alternative hypothesis is true]

  2. If the P-value > α, we do not reject H_0.
     [We conclude that the data do not provide sufficient evidence to reject the null hypothesis (or support the alternative hypothesis)]

Step 5: Interpretation (conclusion) in words in terms of the research problem being investigated.

Critical-Value Approach

- Can be used in Step 4 of Hypothesis Testing as a way of deciding to to reject H_0 or not reject H_0 (in place of the P-value approach)
- Gives same conclusion as the P-value approach
- **Disadvantage:** Does not give the strength of the evidence against H_0.
- It is not necessary to use both approaches, but you should understand both. (Preferably use the P-value approach)
- **Rules for rejecting H_0 or not rejecting H_0 (Critcal-value approach):**

  1. If the **|calculated test statistic| ≥ |critical value|** of the test statistic in the table (at the stated α or α/2 and the appropriate n or df), **we reject H_0.**
  2. If the **|calculated test statistic| < |critical value|, we do not reject H_0**.

Guidelines for Using P-values as Criteria for Rejection of H_0 and Statistical Significance

P-value (Risk factor)	Evidence for Rejection of H_0
P > 0.10	Weak
0.05 < P ≤ 0.10	Moderate
0.01 < P ≤ 0.05	Strong
0.001 < P ≤ 0.01	Very strong
P ≤ 0.001	Extremely strong

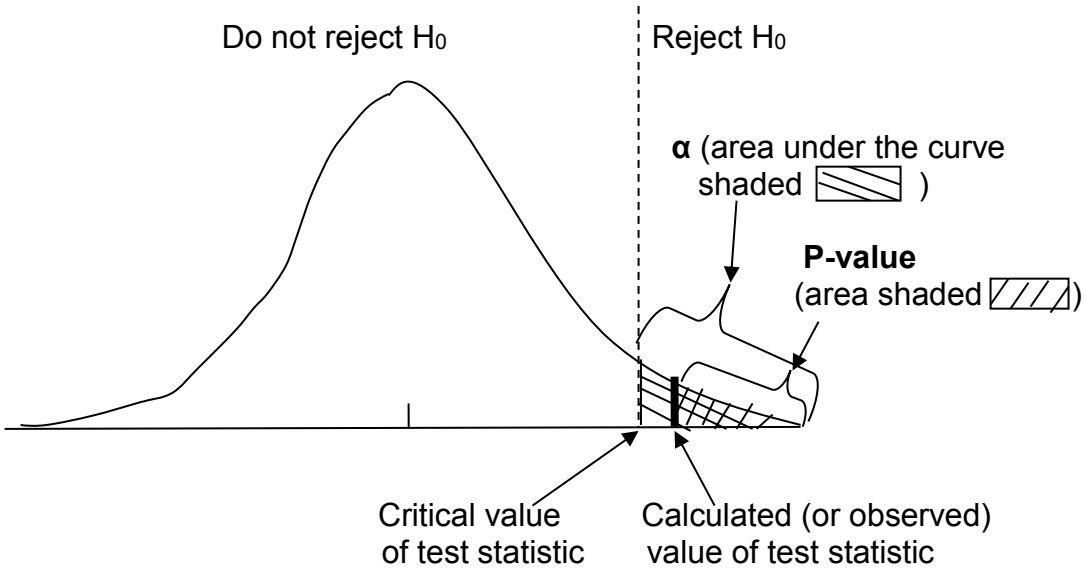
Difference between alpha (α) and P-value

- **Alpha (α)** (significance level) = the maximum probability of the Type I error that you will allow when rejecting Ho (used as a cutoff or criterion for making the decision)
- **P-value** = the observed probability of the Type I error that you find based on the data obtained, calculation of the observed test statistic and examination of the appropriate statistical table

Hypothesis Testing is Conservative (Scientists Don't Jump to Conclusions)

- You might wonder why we don't reject H_0 if P < 0.50 because then we have more than a 50% chance that we will be correct in rejecting H_0
- But in science we never want to jump to conclusions, so we want to be at least 90% certain (P ≤ 0.10) or 95% certain (P ≤ 0.05) before we draw our conclusion

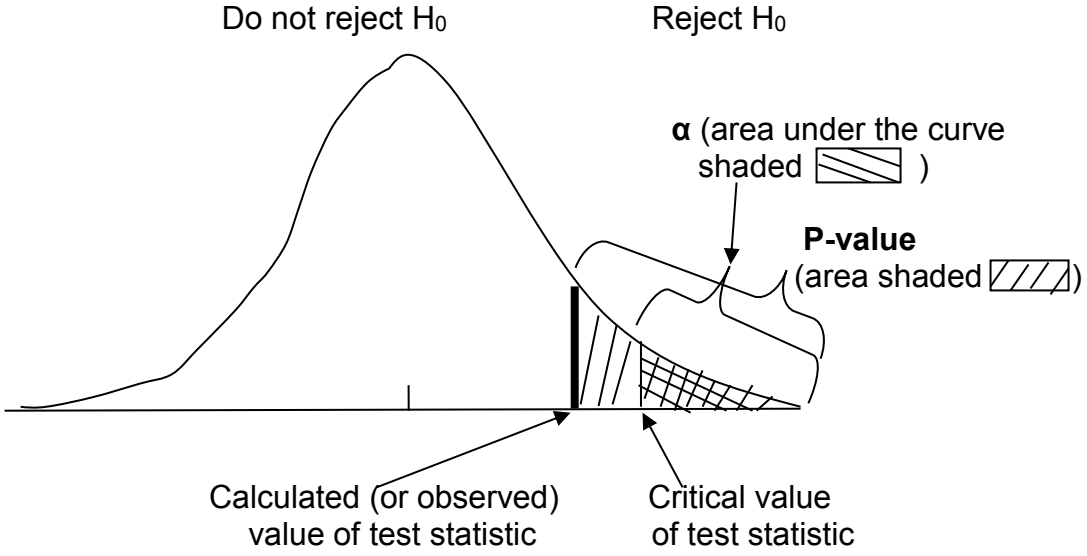
Case of Rejecting H_0



Note:

- Calculated value of the test statistic ≥ Critical value
  AND
- P-value ≤ α

Case of Not Rejecting H_0



Note:

- Calculated value of the test statistic < Critical value
  AND
- P-value > α

1.11.5 Confidence Intervals

- **Point Estimate** of a parameter = the value of the corresponding sample statistic used to estimate the parameter
- **Point Estimate of a population mean, *μ*** (which is a parameter) = the value of the sample mean $\overline{x}$ (which is a statistic) used to estimate the parameter

Confidence-Interval Estimate

Confidence-Interval Estimate

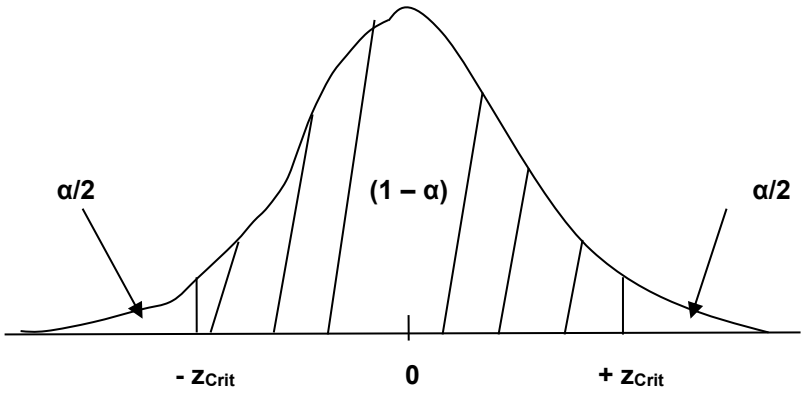
Confidence interval (CI) = an interval or range of numbers derived from a point estimate of a parameter

Confidence level = the confidence (usually expressed in percentage) we have that the parameter lies within the confidence interval (i.e., that the confidence interval contains the parameter).

Confidence-interval estimate = the confidence level and confidence interval.

- These terms can apply to any parameter of a population, but quite commonly they are applied to a population mean, which we discuss here

- **The meaning of a confidence interval:**
  - For a certain percentage (the confidence level) of all samples of size ***n***, the population mean ***μ*** lies within the confidence interval of the sample mean $\overline{x}$



Example (One-mean confidence interval): A certain company produces thousands of size C batteries. The lifetimes of these batteries form a normally distributed population, having a mean of 22 hours and a standard deviation of 4 hours. We randomly sample 30 batteries at a time, taking a total of 20 samples.

Calculate the sample mean and 95.44% confidence interval (CI) for each sample (i.e., the confidence level is set at 95.44%). [Recall from the Empirical Rule regarding ± 2 standard deviations from the mean]

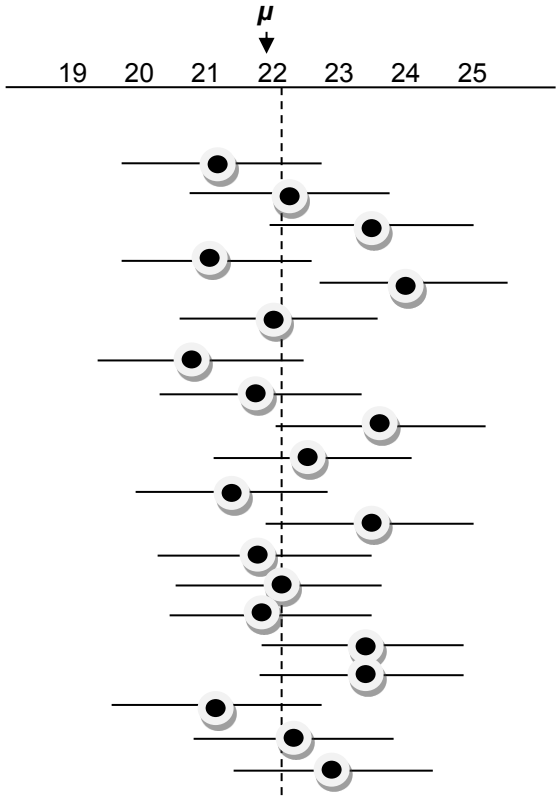
So for each sample, we calculate the confidence interval (CI) as follows:

$\bar{x} \pm 2\sigma_{\bar{x}}$ where $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{4}{\sqrt{30}} = 0.73$

Then we add and subtract (2)(0.73) = 1.46 to each $\bar{x}$ to get its CI.

Table: Twenty samples of lifetimes of batteries: their means, CI, and whether the ***μ*** is included in their CI.

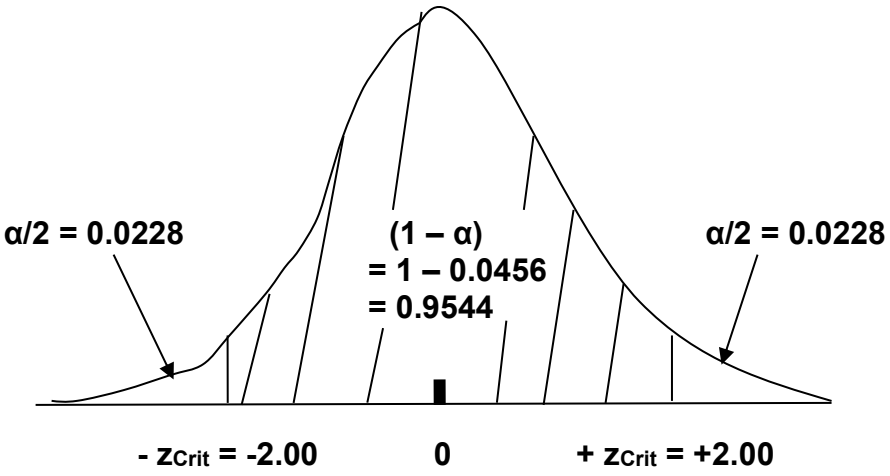
Sample	$\bar{x}$	**95.44% CI**	***μ*** in CI
1	21.04	19.58-22.50	Yes
2	22.15	20.69-23.61	Yes
3	23.34	21.88-24.80	Yes
4	20.92	19.46-22.38	Yes
5	23.86	22.40-25.32	No
6	21.86	20.40-23.32	Yes
7	20.75	19.29-22.21	Yes
8	21.67	20.21-23.13	Yes
9	23.40	21.94-24.86	Yes
10	22.38	20.92-23.84	Yes
11	21.31	19.85-22.77	Yes
12	23.27	21.81-24.73	Yes
13	21.74	20.28-23.20	Yes
14	22.02	20.56-23.48	Yes
15	21.80	20.32-23.24	Yes
16	23.25	21.79-24.71	Yes
17	23.08	21.62-24.54	Yes
18	21.12	19.66-22.58	Yes
19	22.24	20.78-23.70	Yes
20	22.87	21.41-24.33	Yes



- **Note:**
  - Every sample has a different mean
  - μ is not within the confidence interval for all samples
  - 19/20 = 95% of the samples have μ falling within their confidence interval
  - If we took many samples (e.g., 1000 or 10,000), we would find that μ falls within their confidence intervals of 95.44% of the samples.

The above distribution can be illustrated as follows:

For a 95.44% confidence level, α = 1 – 0.9544 = 0.0456



(z_{Crit} can be found in the table for the standard normal curve)

1.11.6 Parametric versus nonparametric methods in Inferential Statistics

Parametric Statistical methods:

- Involve the estimation of population parameters, e.g., mean, variance, etc.
- Have certain underlying assumptions about the populations being tested, such as
  - Random sampling
  - Populations normally distributed (if sample size is small)
  - Homogeneity of variances: when comparing 2 or more samples, the variances of all samples must be approximately equal
- More powerful than nonparametric tests: i.e., greater chance of rejecting Ho when in fact it is false (that means is less chance of making a Type II error)
- Examples: t tests, analysis of variance and regression

Nonparametric statistical methods:

- Do not use estimates of population parameters in their calculations
- Have fewer assumptions about the nature of the distribution of the populations being studied
- Only assumption or requirement is that the samples must be selected randomly
- Therefore, they can be applied in many cases when the parametric methods are not valid
- Can be applied to categorical data, whereas parametric tests cannot.
- Slightly less powerful
- Examples of nonparametric tests: Chi-square test, Mann-Whitney U test, Kruskal-Wallis test, Spearman rank correlation

1.11.7 General Approach to Hypothesis Tests and Confidence Intervals

Parameter = characteristic of the population being investigated

Estimate = sample statistic used to estimate the parameter of the population being investigated

Standard Error of the Estimate = standard deviation of the sampling distribution, taking into consideration the sample size (calculations of this will vary, depending upon the hypothesis test being performed)

H_0 value = the value that the parameter being investigated would have, assuming that the null hypothesis is true

The general formula for a test statistic is:

Test statistic: $$\frac{Estimate - H_0 value}{SE(Estimate)}$$

The general formula for a confidence interval is:

Confidence Interval: $$Estimate \pm Critical\ Value \times SE(Estimate)$$

Or: $$Estimate \pm Margin\ of\ Error$$

Critical Value = value obtained from a table showing the theoretical distribution of the test statistic, at a given level of confidence

SECTION 2: ONE OR TWO POPULATION MEANS

2.1 Inferences for One Mean (The One-Sample Case)

- We can make inferences about one population, based on one sample
- One-sample inferences may be of two types:
  - Inferences about one mean
  - Inferences about one proportion
- In this course we only do inferences for means (not proportions)
- Recall the z-score formula for a population and for a sampling distribution

$$z = \frac{y - \mu}{\sigma} \quad \text{and} \quad z = \frac{\bar{y} - \mu}{\sigma / \sqrt{n}}$$

- Occasionally, sigma (**σ**) is known and, thus, the standardized version of the sample mean ($\bar{y}$) can be converted to the formula for the one-sample z-test as follows:

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

- However, usually sigma (**σ**) is not known, so we estimate the population standard deviation using the sample standard deviation (***s***) and obtain the Student ***t*** (studentized) version of the sample mean ($\bar{y}$)
- The Student ***t*** version of the sample mean can be converted to the one-sample t-test, which will be discussed in detail below.

Student *t* (studentized) version of the Sample Mean, $\bar{y}$

Suppose that a variable ***y*** of a population is normally distributed with mean $\bar{y}$, then for samples of size ***n***, the variable

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

has the ***t***-distribution with ***n*** – 1 degrees of freedom (i.e., ***df = n – 1***)

Introduction to the t-distribution

- The ***t***-distribution is very important in statistics and is applied, for example, in:
  - One-sample ***t***-test, two-sample ***t***-test, paired-sample ***t***-test
  - ***t***-test for the significance of the slope of a regression line
- The ***t***-distribution was developed by William Gosset in 1908. He published it under the name of 'student' so it became known as the student ***t*** distribution
- The ***t***-curve is more spread out than the normal distribution, particularly at the base, and especially for small sample sizes.

Properties of the *t*-curve
(Properties 1 – 3 are in common with the standard normal distribution)

Property 1: The total area under a ***t***-curve = 1.
Property 2: A ***t***-curve is symmetrical about 0.
Property 3: A ***t***-curve extends indefinitely in both directions, approaching, but never touching, the horizontal axis.
Property 4: There is a different ***t***-curve for each sample size, identified by its number of degrees of freedom (df), whereby ***df = n – 1***.
Property 5: As ***df*** increases, the **t**-curve approaches the normal curve until, at ***df*** = ∞, the ***t***-curve coincides with the normal curve.

Illustration of t-curve for df = 1, df = 6 and df = 1000 (Using MINITAB)

Using the *t*-Table

- A ***t***-table can be either one-tailed or two-tailed or both.
- The table used in this course (NOT from the textbook) is one-tailed (right-tailed)
- For cases where a required ***df*** is not shown, use the next lower ***df*** (to be conservative)

Examples:

1. For a one-tailed test, when α = 0.0005 and df = 12, $t_{0.0005} = 4.318$
2. For a one-tailed test, when α = 0.025 and df = 75, $t_{0.025} \approx 2.000$

2.1.1 Hypothesis Test for One Population

The One-Mean *t*-Test (also called the one-sample *t*-test)

Step 1: Check the purpose and assumptions (to see if this is the appropriate test for the research problem)

Purpose of the test: To test for the difference between a population mean (by taking a sample and calculating a sample mean) and some hypothesized (theoretical) mean or value.

Assumptions of the test:

1. Simple random sample
2. The population under study is normally distributed or sample size is large
3. σ is unknown

Step 2: State the null and alternative hypotheses

The null hypothesis is H_0: $\mu = \mu_0$ and the alternative hypothesis may be one of the following:

H_a: $\mu \neq \mu_0$	or	H_a: $\mu < \mu_0$	or	H_a: $\mu > \mu_0$
(two-tailed)		(left-tailed)		(right-tailed)

Step 3: Obtain the Calculated Value (or Observed Value) of the test statistic as follows:

$$t = \frac{\bar{y} - \mu_0}{s\ /\sqrt{n}} = \frac{\bar{y} - \mu_0}{SE(\bar{y})} \qquad \text{df} = n - 1$$

where μ_0 = some hypothesized (theoretical) mean or value

Or, the formula can be broken down into the following components:

Parameter	Estimate	SE(Estimate) of the mean	H_0 value	Reference Distribution
μ	$\bar{y}$	$\frac{s}{\sqrt{n}}$	μ_0	t_{n-1}

Step 4: Decide to reject H_0 or not reject H_0 and state the strength of the evidence against H_0
Examine the t-table at df = n – 1
If the P-value ≤ α, we reject H_0 (otherwise do not reject H_0)

Step 5: Interpretation (Conclusion in Words)

The general formula for a test statistic is:

Test statistic: $\dfrac{Estimate - H_0 value}{SE(Estimate)}$

Degrees of Freedom (df) = number of independent observations

- For standard deviation, ***df = n – 1*** (found in the denominator of the formula) because the sample mean is used as an estimate of the population mean in calculating it (using defining formula).
- When the sample mean is known, only **n – 1** observations are independent. One observation is not independent; it is fixed by knowing the other observations
- For the one-sample t-test, also ***df = n – 1*** because standard deviation is in the denominator of the formula

Example of a One-Sample t-test: Comparing Two-tailed and Left-tailed Hypothesis Tests

Acropora formosa (Staghorn Coral) forms large colonies. On mature reefs in Tanzania the average height of these colonies is about 75 cm, but they may reach 150 cm. (DSM, pp. 53, 55)

Research problem: A researcher measures the height of randomly selected *Acropora formosa* colonies along the reef crest of Mbudya Island (Dar es Salaam), obtaining measurements (in cm) as shown in the table below. Summary statistics describing the data are also shown below. The population of heights of these colonies is known to be normally distributed, so even a small sample size is adequate. At the 5% significance level, test whether the mean height of colonies of *Acropora formosa* colonies found on this reef crest of Mbudya Island is different from the mean height of 75 cm for such colonies throughout the country.

Heights (in cm) of random samples colonies of *A. formosa* at Mbudya																		
Mbudya	81	48	74	69	79	56	59	64	51	61	72	74	84	81	67	57	69	68

Descriptive Statistics: *A.formosa*-reef crest

Variable	N	Mean	SE Mean	StDev	Min	Q1	Median	Q3	Max
Height (cm)	18	67.4444	2.4921	10.5731	48.00	58.50	68.50	75.25	84.00

Step 1: The one-sample t-test is selected

Purpose of the research problem: To test for the difference between one population mean (based on the sample mean) and the hypothesized mean of 75 cm.

Assumptions:

1. Random sample
2. Population normally distributed
3. σ is not known

>>>>>>>>>>

H_0: 75cm = μ

H_A: 75cm ≠ μ

0.05 p-value

N = 18, df = 17

$\bar{Y}$ = 67.444

SE(7) = 2.4921

$t^* = -3.0318 = \frac{67.444 - 75.00}{2.4921}$

$(0.005 < p < 0.01) \times 2$

p-value two-tailed

df(17) @ 0.025 = -2.110 : 2.110

$p < \alpha$ of 0.05

$t^* \leq df$

therefore we reject H_0

>>>>>>>>>>

An Experienced Coral Reef Researcher's Approach

This researcher knows that corals in shallow water, such as the reef crest, are generally shorter in height than average due to exposure to stronger wave action. Therefore, he re-formulates the **research question** as follows: At the 5% significance level, test whether the mean height of colonies of *Acropora formosa* colonies found on this reef crest of Mbudya Island is shorter than the mean height of 75 cm for such colonies throughout the country.

>>>>>>>>>>

$H_0: \mu = 75cm$

$H_A: \mu < 75cm$

$N = 18$

$df = 17$

P-value $(0.005 > P > 0.0025)$

$P < \alpha$ of 0.05

Still reject H_0

>>>>>>>>>>

Note: In the above example, the evidence against H_0 is stronger than when performing the two-tailed hypothesis, but it is still within the same range of $0.001 < P \leq 0.01$ and considered as very strong evidence.

Exact probabilities for a Type I Error

- When we use statistical tables to find the Type I error, we can only say the p is greater than or less than or between certain values
  - E.g., $P < 0.001$ or $P > 0.10$ or $0.05 < P < 0.01$, etc.
- However, when we use a computer program it will tell us the exact P
  - E.g. $P = 0.1332$ or $P = 0.00167$, etc.

Significance Level (α) versus P-Value

Distinction between Significance Level (α) and P-Value

- When planning a hypothesis test, the significance level (α) = Probability of a Type I error (the maximum that you can accept for a given question)
- P-Value read from a statistical table based on the calculated value of the test statistic = the observed Probability of the Type I error
- If you reject H_0, the P-value = the probability of making a mistake by committing a Type I error
- If you do not reject H_0, the P-value = the probability of the mistake you would have made if you had rejected H_0. That is why you did not reject H_0, because the chance of error would have been too large, i.e., greater than alpha (α).

2.1.2 Confidence Intervals for One Population Mean

- The general formula for a confidence interval is:

Confidence Interval: $Estimate \pm Critical\ Value \times SE(Estimate)$

One-Mean *t*-Interval Procedure (OR one-sample *t*-interval procedure)

Step 1: Find the Critical value: For a given confidence level (1 – α), use the t-table showing the critical values of the t-distribution to find $t_{\alpha/2}$(=t_{Crit}) (=t*) in the row for the appropriate df, where ***df = n – 1*** and ***n*** is the sample size.

Step 2: Two-sided confidence interval for ***μ*** is given by the endpoints:

$$\bar{y} - t_{\alpha/2} \times \frac{s}{\sqrt{n}} \quad \text{to} \quad \bar{y} + t_{\alpha/2} \times \frac{s}{\sqrt{n}}$$

this means a two-tail test

OR $\bar{y} \pm t_{\alpha/2} \times \frac{s}{\sqrt{n}} \Rightarrow \bar{y} \pm t_{\alpha/2,n-1} \times SE(\bar{y})$ ← df

where **n** is the sample size; $\bar{y}$ and ***s*** are computed from the sample data.

Or, the formula can be broken down as follows:

Parameter	Estimate	Critical value	SE(Estimate) of the mean
μ	$\bar{y}$	$t_{\alpha/2,n-1}$ (or t*)	$\frac{s}{\sqrt{n}}$

Step 3: Interpret the confidence interval in terms of the research problem being investigated.

Note: **Margin of Error (E) =** $t_{\alpha/2} \times \frac{s}{\sqrt{n}}$

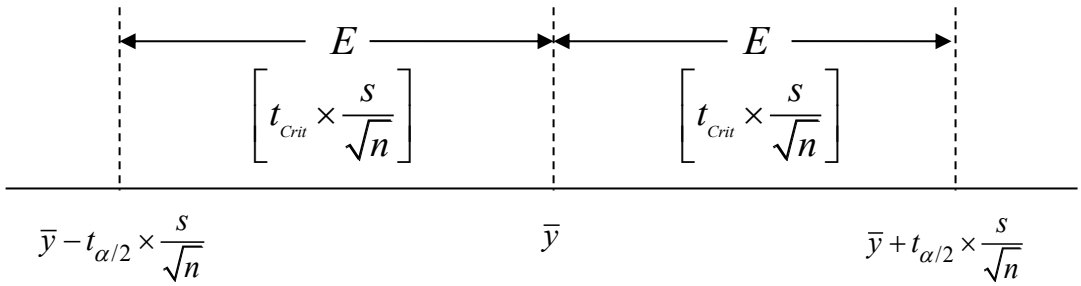
One-Sided t-intervals for One-Sample

Lower bound t-interval: $\bar{y} - t_{\alpha,n-1} \times \frac{s}{\sqrt{n}}$ $\Rightarrow$ Consistent with a right-tailed test

Upper bound t-interval: $\bar{y} + t_{\alpha,n-1} \times \frac{s}{\sqrt{n}}$ $\Rightarrow$ Consistent with a left-tailed test

Note: $t_{Crit} = t_{\alpha,n-1}$

Two-sided Confidence Interval and Margin of error can be **illustrated** as follows:



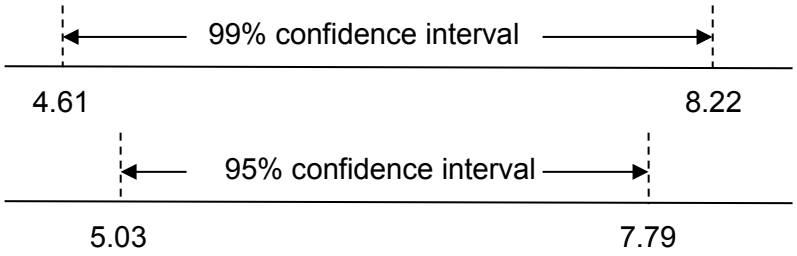
Margin of error = half the length of the confidence interval)

- The margin of error determines the precision with which $\bar{x}$ estimates ***μ***
- Increasing sample size:
  - Decreases the margin of error, and
  - Increases precision

Confidence and Precision

- The length of the confidence interval is inversely proportional to precision
  - For a given confidence level, a wide confidence interval indicates poor precision of the data
- Thus, for a fixed sample size, decreasing the confidence level decreases the confidence interval and improves the precision, and vice versa

Compare the two confidence intervals and determine which has the greatest precision.



Conclusion about comparative precision: For the same sample size, since the 95% confidence interval is shorter, it has greater precision than the 99% confidence interval.

Example: Using the data and information given for the hypothesis test above, calculate a 95% confidence interval for the mean height of colonies of *Acropora formosa* found on the reef crest of Mbudya Island (Dar es Salaam).

>>>>>>>>>>

95% CI, $\alpha = 0.05$ @ $df = n-1$, $18-1 = 17$

$t_{\alpha/2, n-1} = t_{0.05/2, 18-1} = t_{0.025, 17} = 2.110$

Estimate $= \bar{y} = 67.444$

SE of estimate $= SE(\bar{y}) = 2.4921$

$\bar{y} \pm t^* \times SE(\bar{y})$

$67.444 \pm 2.110 \times 2.4921$

67.444 ± 5.2583

$(62.19, 72.70)$ cm

>>>>>>>>>>

SPSS is designed to only do two-tailed test and C.I.

Relation between Hypothesis Tests and Confidence Intervals

Relation between Hypothesis Tests and Confidence Intervals: Inferences for One Population Mean

For a two-tailed hypothesis test for comparing one population mean (μ) with some theoretical mean or value (μ_0) at the significance level α:

- **The case of rejecting H_0:** The null hypothesis will be rejected if and only if the (1-α) confidence interval for μ does not contain the theoretical mean or value (μ_0).
- **The case of not rejecting H_0:** The null hypothesis will not be rejected if the (1-α) confidence interval for μ does contain the theoretical mean or value (μ_0).

  [If μ_0 is within the interval, there is no significant difference between μ and μ_0]

For a right-tailed hypothesis, the null hypothesis will be rejected in favour of H_a: $\mu > \mu_0$ if and only if the (1 – α)-level lower confidence bound for the population mean μ is greater than μ_0

For a left-tailed hypothesis, the null hypothesis will be rejected in favour of H_a: $\mu < \mu_0$ if and only if the (1 – α)-level upper confidence bound for the population mean μ is less than μ_0

Note: Confidence level = 1 – α = 1 – significance level

Two conditions that must be met to ensure that the conclusions will be the same for a hypothesis test and a confidence interval performed on the same data:

1. The confidence level must be a compliment of the significance level (α) applied in the hypothesis test.
2. They must be the same "sided", that is, both two-sided or both one-sided.

   - If these two conditions are not met, the hypothesis test and confidence level may still give the same conclusion, but there is no guarantee

Example: Compare the conclusions of the hypothesis tests and confidence interval for the research problem on *Acropora formosa* on the reef crest at Mbudya Island.

Two-tailed hypothesis test for $\mu \neq 75$ cm (at $\alpha = 0.05$):

H_0 was rejected at $0.01 > P > 0.005$ (exact P-value = 0.008). So, we concluded that there was a difference.

Confidence Interval (at 95% level):

The interval (62.19, 72.70) cm does not contain 75 cm and thus confirms that the mean height of the population of colonies of Acropora formosa on the reef crest is different from the hypothesized value or mean height of this species countrywide.

One-tailed hypothesis test for $\mu < 75$ cm (at $\alpha = 0.05$):

H_0 was rejected at $0.005 > P > 0.0025$ (exact P-value = 0.004). So, again we concluded that there was a difference. Therefore, this is even greater confirmation that the two-sided confidence interval would not contain the hypothesized value of 75 cm.

SPSS Output for Hypothesis Test

One-Sample Test

	Test Value = 75					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Mbudya	-3.032	17	.008	-7.55556	-12.8134	-2.2977

Interpretation:
Note: "Sig." is actually the P-value (not significance level). The significance level is not input into SPSS. The P-value given is always the 2-tailed value and cannot be changed.

Two-tailed hypothesis test for the difference between the mean height of *Acropora formosa* on the reef crest at Mbudya Island and the hypothesized mean height of 75 cm for such colonies throughout the country

Result: t = -3.032, df = 17, P = 0.008
Note: The confidence interval shown when doing a hypothesis test is meaningless.

One-tailed (left-tailed) hypothesis test to determine whether the mean height of *Acropora formosa* on the reef crest at Mbudya Island is less than the hypothesized mean height of 75 cm for such colonies throughout the country

Result: t = -3.032, df = 17, P = (0.008)/2 = 0.004

→ you have to divide by 2

SPSS Output for Confidence Interval: For two-sided 95% confidence interval

One-Sample Test

	Test Value = 0					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Mbudya	27.063	17	.000	67.44444	62.1866	72.7023

- **Note:** The test value must be changed from 75 to 0

Set the test value to 0 to get $\bar{y}$ and the CI.

Two-sided confidence interval

- The two-sided 95% confidence level is (62.19, 72.70) cm
- Though we don't input the test value of 75, since it is outside this interval, the mean height of Acropora at Mbudya is different from the hypothesized 75 cm
- **Note:** the t-statistic (27.063) and P-value (0.000) for the hypothesis test are now meaningless.

SPSS Output for Confidence Interval: For one-sided (Upper bound 95% confidence interval)

One-Sample Test

	Test Value = 0					
	t	df	Sig. (2-tailed)	Mean Difference	90% Confidence Interval of the Difference	
					Lower	Upper
Mbudya	27.063	17	.000	67.44444	63.1092	71.7797

One-sided confidence interval (upper-bound)

- SPSS does not do one-sided confidence intervals or hypothesis tests, so to obtain this 95% upper bound, we had to "trick" SPS by determining alpha as 0.05 x 2 = 0.10. Then, we input 90% (100% – 10%)
- So, the 95% upper bound is 71.78 cm, which can also be written as $(-\infty, 71.78)$ cm
- Since 75 is higher than the upper bound, that means that the mean height of Acropora on the reef crest at Mbudya is less than the hypothesized value of 75.
- **Note:** the lower bound shown (63.11) is meaningless

>>>>>>>>>>

Illustration:

for 95% C.I. $\alpha = 1 - 0.95 = 0.05$ @ $df = n - 1 = 17$ $t_\alpha = t_{0.05} = 1.740$

Hand calculations for a 95% upper bound confidence interval

UB $\bar{Y} + t_{0.05,17} \times SE(\bar{Y})$

$67.444 + 1.740 \times 2.4921$

$64.444 + 4.336$

$(-\infty, 71.78)$ cm

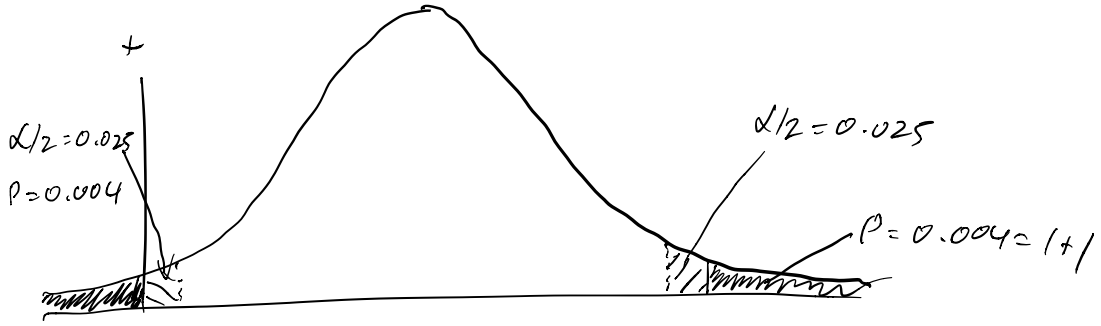
we can be 95% confident that ... mean is shorter than 71.98 cm

<<<<<<<<<<

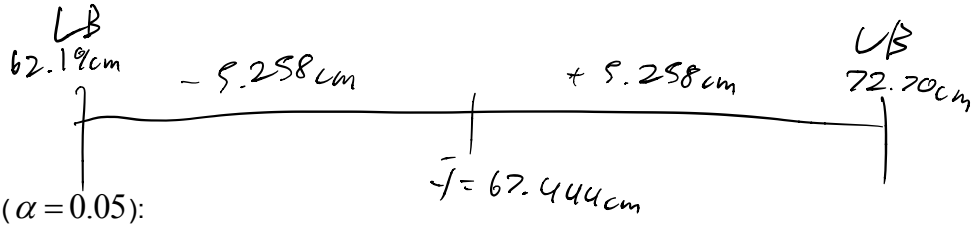
Illustrations

>>>>>>>>>>

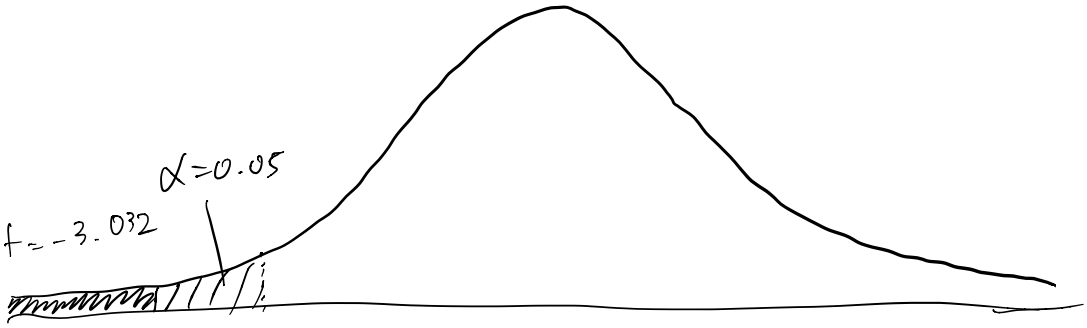
Two-tailed test ($\alpha = 0.05$)



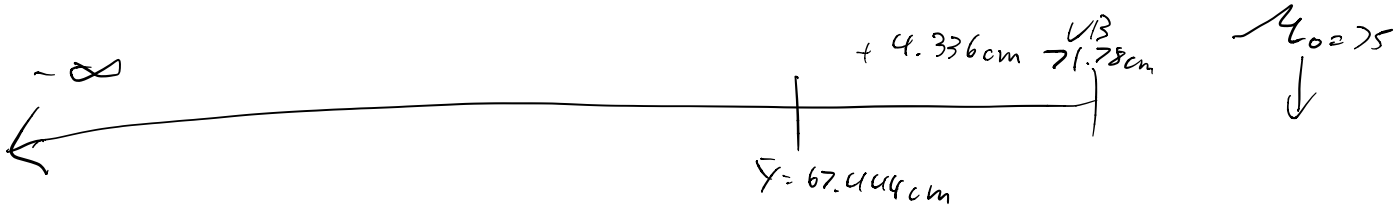
Two-sided confidence interval (95%):



Left-tailed test ($\alpha = 0.05$):



Upper bound confidence interval (95%):



A one-sided confidence interval will be shorter on one side than the two-sided confidence interval and it will go to infinity on the other side.

>>>>>>>>>>

2.2 Inferences for Two Means (The Two-Sample Case)

In this Section we are talking about the case where:

- We have **one variable**
- We are comparing **two populations** or groups by taking **two samples**
- Two sample inference:
  - Comparing **two population means: Independent samples**
  - Comparing **two population means: Paired samples**
  - Comparing **two proportions** (not covered in this course)

2.2.1 Inferences for Two Population Means: Based on Two Independent Samples

2.2.1.1 The Sampling Distribution of the Difference between Two Sample Means for Independent Samples

Table: Notation for parameters and statistics when comparing two populations.

Parameter/Statistic	Population 1	Population 2
Population mean	μ_1	μ_2
Population standard deviation	σ_1	σ_2
Sample mean	$\bar{y}_1$	$\bar{y}_2$
Sample standard deviation	s_1	s_2
Sample size	n_1	n_2

The Sampling Distribution of the Difference Between Two Sample Means, $\bar{y}_1 - \bar{y}_2$, for Independent Samples

Suppose that x is a normally distributed variable on each of two populations; then, for independent samples of size n_1 and n_2 from the two populations,

- The mean of all possible differences between the two sample means equals the difference between the two population means:

$$\mu_{\bar{y}_1 - \bar{y}_2} = \mu_1 - \mu_2$$

- The standard deviation of all possible differences between the two sample means equals the square root of the sum of the population variances, each divided by the corresponding sample size:

$$\sigma_{\bar{y}_1 - \bar{y}_2} = \sqrt{(\sigma_1^2 / n_1) + (\sigma_2^2 / n_2)}$$

- $\bar{y}_1 - \bar{y}_2$ is assumed to be normally distributed.

2.2.1.2 Inferences for Two Population Means Using Independent Samples, Standard Deviations Not Assumed Equal

- This is the general case of comparing two populations (independent samples) and can be applied to all situations, whether standard deviations are equal or not equal, and regardless of sample sizes

Nonpooled Two-Mean t-test (or Nonpooled Two-Sample t-test)

Purpose: To test for the difference between two population means (μ_1 and μ_2) based on two sample means ($\bar{y}_1$ and $\bar{y}_2$).

Assumptions:

1. Simple random samples (also implies also independent sampling within samples)
2. Both populations are normally distributed or both samples are large
3. Samples are independent

Test statistic:

$$t = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{(s_1^2 / n_1) + (s_2^2 / n_2)}}$$

Where $df = \dfrac{[(s_1^2 / n_1) + (s_2^2 / n_2)]^2}{\dfrac{(s_1^2 / n_1)^2}{n_1 - 1} + \dfrac{(s_2^2 / n_2)^2}{n_2 - 1}}$ (rounded down to nearest integer)

- You will not need to perform this hypothesis test in this course, but you need to understand the difference between this test and the pooled t-test described below and to understand the assumptions of each.

Example of the Two-Mean t-test (Nonpooled t-test):
The weights of random samples of chickens (in kg) of two different breeds (A and B) are shown below. Both populations are normally distributed, so even small samples are adequate. At the 5% significance level, determine if there is a difference in the weights of Breed A and Breed B.

	Sample 1 (Breed A)	Sample 2 (Breed B)
	1.5	1.4
	1.0	1.1
	1.3	1.2
	1.9	1.3
	1.7	1.2
	1.5	1.1
		1.2
		1.4
		1.3
		1.2
		1.3
Mean	$\bar{x}_1 = 1.483333$	$\bar{x}_2 = 1.245455$
Standard deviation	$s_1 = 0.312517$	$s_2 = 0.103573$

Note: Standard deviations are very different and sample sizes are different, but the nonpooled t-test can still be applied.

2.2.1.3 Inferences for Two Population Means Using Independent Samples, Standard Deviations Assumed Equal

- This is the special case of comparing two populations (independent samples), which can be applied when the standard deviations of the two populations are similar and samples sizes are nearly equal
- Where the assumptions of the pooled t-test are met, it is slightly more powerful than nonpooled t-test

Pooled t-test (Two-sample t test assuming equal variances; also called two-mean t-test assuming equal variances)

Purpose: To test for the difference between two population means (μ_1 and μ_2) based on two sample means ($\bar{y}_1$ and $\bar{y}_2$).

Assumptions:

1. Simple random samples
2. Both populations are normally distributed or both samples are large
3. Samples are independent
4. Equal population standard deviations (A rule of thumb: if the ratio of the larger to the smaller sample standard deviation < 2, we can say the assumption has been met) [OR apply Levene's Test for Equality of Variances]
5. Sample sizes should be roughly equal

Null and Alternative Hypotheses:
The null hypothesis is H_0: $\mu_1 = \mu_2$ and the alternative hypothesis may be one of the following:

H_a: $\mu_1 \neq \mu_2$ (two-tailed) or H_a: $\mu_1 < \mu_2$ (left-tailed) or H_a: $\mu_1 > \mu_2$ (right-tailed)

Test statistic: $$t = \frac{\bar{y}_1 - \bar{y}_2 - \Delta_0}{s_p\sqrt{(1/n_1)+(1/n_2)}} = \frac{\bar{y}_1 - \bar{y}_2 - \Delta_0}{SE(\bar{y}_1 - \bar{y}_2)}$$

Where s_p (pooled standard deviation) is given by:

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

$$df = n_1 + n_2 - 2$$

Or, the formula can be broken down as follows:

Parameter	Estimate	SE(Estimate) of the difference between the means	H_0 value	Reference Distribution
$\mu_1 - \mu_2$ (Assume $\sigma_1 = \sigma_2$)	$\bar{y}_1 - \bar{y}_2$	$s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$ Where $s_p = \sqrt{\frac{(n_1-1)s_1^{\,2} + (n_2-1)s_2^{\,2}}{n_1+n_2-2}}$	$0 (or \Delta_0)$	$t_{n_1+n_2-2}$

Note: Δ_0 = some hypothesized difference, which is almost always 0

Note: If you obtain a *df* that is not shown in the *t*-table, always go to the *df* below that.

If the P-value ≤ α, we reject H_0 (otherwise do not reject H_0)

Confidence Intervals for the Difference Between the Means of Two Populations, Using Independent Samples, Standard deviations Assumed Equal

Two-Mean t-Interval Procedure (=Pooled t-Interval)

Purpose: To find a confidence interval for the difference between two population means, μ_1 and μ_2 based on two sample means ($\bar{y}_1$ and $\bar{y}_2$).

Assumptions: Same as for the Pooled t-test

Step 1: For a given confidence level (1 – α) at **df** = $n_1 + n_2 - 2$

Step 2: The endpoints of the confidence interval of $\mu_1 - \mu_2$:

Two-sided interval:

$$(\bar{y}_1 - \bar{y}_2) \pm t_{\alpha/2} \times s_p \sqrt{(1/n_1) + (1/n_2)} \quad \textbf{or} \quad (\bar{y}_1 - \bar{y}_2) \pm t_{\alpha/2, n_1+n_2-2} \times SE(\bar{y}_1 - \bar{y}_2)$$

Lower bound t-interval: $(\bar{y}_1 - \bar{y}_2) - t_{\alpha} \times SE(\bar{y}_1 - \bar{y}_2)$ $\Rightarrow$ Consistent with a right-tailed test

Upper bound t-interval: $(\bar{y}_1 - \bar{y}_2) + t_{\alpha} \times SE(\bar{y}_1 - \bar{y}_2)$ $\Rightarrow$ Consistent with a left-tailed test

Where s_p = pooled standard deviation

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

Or, the formula can be broken down as follows:

Parameter	Estimate	Critical value	SE(Estimate) of the difference between the means
$\mu_1 - \mu_2$ (Assume $\sigma_1 = \sigma_2$)	$\bar{y}_1 - \bar{y}_2$	Two-sided $t_{\alpha/2, n_1+n_2-2}$ One-sided t_{α, n_1+n_2-2}	$s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ Where $s_p = \sqrt{\frac{(n_1 - 1)s_1^{\,2} + (n_2 - 1)s_2^{\,2}}{n_1 + n_2 - 2}}$

Step 3: Interpret the confidence interval in terms of the research problem.

Margin of Error (E) (two-sided) = $t_{\alpha/2} \times s_p \sqrt{(1/n_1) + (1/n_2)}$

Note: Confidence level = 1 – α = 1 – significance level

Example of the Pooled t-test (also a Two-Mean t-test)
Research problem: The heights of randomly selected *Acropora formosa* colonies along the reef crests of Mbudya Island and Bongoyo (Dar es Salaam) are shown below, along with summary statistics. Both populations have colony heights that are normally distributed, so even small sample sizes are adequate. At the 5% significance level, test whether there is a difference in the mean heights of colonies of *Acropora formosa* colonies found on the reef crests of these two islands.

Heights (in cm) of random samples colonies of A. formosa at Mbudya and Bongoyo																		
Mbudya	81	48	74	69	79	56	59	64	51	61	72	74	84	81	67	57	69	68
Bongoyo	86	87	70	62	73	71	85	57	82	74	81	60	63	59	63	78		

Descriptive Statistics

	N	Minimum	Maximum	Mean		Std. Deviation
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic
Mbudya	18	48.00	84.00	67.4444	2.49211	10.57312
Bongoyo	16	57.00	87.00	71.9375	2.59562	10.38248
Valid N (listwise)	16					

Step 1: Pooled-sample t-test is selected because
Purpose of the study: To test for a difference between two population means
Assumptions:
1. Two independent random samples
2. Both populations are normally distributed
3. The two standard deviations are nearly equal (10.57/10.38 < 2) and sample sizes are nearly equal.

Step 2:
H_0: $\mu_1 = \mu_2$ (There is no difference in the mean heights of colonies of *Acropora formosa* colonies found on the reef crests of these two islands.)
Ha: $\mu_1 \neq \mu_2$ (There is a difference in the mean heights of colonies of *Acropora formosa* colonies found on the reef crests of these two islands.)
Parameter: $\mu_1 - \mu_2 = \mu_{Mbudya} - \mu_{Bongoyo}$
(mean height of *A. formosa* at Mbudya – mean height at Bongoyo)

>>>>>>>>>>

Estimate of the diff between means $= \bar{Y}_1 - \bar{Y}_2 = 67.444 - 71.9375 = -4.4931\text{cm}$

Estimate of pooled std dev

$$S_p = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}$$

$$= \sqrt{\frac{(18-1)(10.57312)^2 + (16-1)(10.38248)^2}{16+18-2}} = 10.48418$$

SE of the estimate of the diff between means

$$SE = 10.484189\sqrt{\frac{1}{18}+\frac{1}{16}}$$

$$SE_{\bar{Y}_1-\bar{Y}_2} = 3.60228$$

$$t = \frac{-4.49}{3.60228} = -1.247 = t$$

$$df = 18+16-2 = 32$$

these are 1-tailed test double the p-value for 2-tailed test

p-value $(0.15 > p > 0.10) \times 2$

$0.30 > p > 0.20$

there is weak evidence against H_0 $p > \alpha$ of 0.05

>>>>>>>>>>

Example of Calculating a Pooled t-interval

Using the data and information given for the pooled t-test above, calculate a 95% confidence interval for the difference between mean heights of colonies of *Acropora formosa* found on the reef crests at Mbudya Island and Bongoyo Island (Dar es Salaam).

>>>>>>>>>>

95% C.I $\alpha = 0.05$

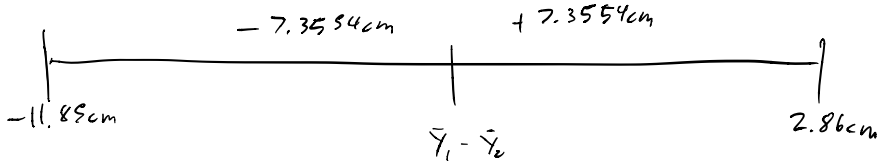
@ $df = n_1+n_2-2 = 32$ $t_{\alpha/2} = t_{0.05/2} = t_{0.025} = 2.042$

$$(\bar{Y}_1-\bar{Y}_2) \pm t_{0.025} \times SE(\bar{Y}_1-\bar{Y}_2)$$

$$-4.4931 \pm 2.042 \times 3.60228$$

$$-4.4931 \pm 7.35585576 \rightarrow ME$$

$(-11.8489, 2.8628)$ cm



>>>>>>>>>>

Relation between Hypothesis Tests and Confidence Intervals

Relation between Hypothesis Tests and Confidence Intervals: Inferences for Two Population Means

For a two-tailed hypothesis test for comparing two population means at the significance level α:

- The case of rejecting H_0**:** The null hypothesis will be rejected if and only if the (1-α) confidence interval for the difference between the population means ($\mu_1 - \mu_2$) does not contain 0, i.e. either both endpoints are negative OR both endpoints are positive.
- The case of not rejecting H_0**:** The null hypothesis will not be rejected if the (1-α) confidence interval for the difference between the population means ($\mu_1 - \mu_2$) does contain 0, i.e. one endpoint is negative and the other is positive.

  [If 0 is within the interval, there is 0 difference OR no significant difference.]

For a right-tailed hypothesis, the null hypothesis will be rejected in favour of Ha: $\mu_1 > \mu_2$ if and only if the (1 – α)-level lower confidence bound for the difference between $\mu_1 - \mu_2$ is positive (i.e., > 0)

For a left-tailed hypothesis, the null hypothesis will be rejected in favour of Ha: $\mu_1 < \mu_2$ if and only if the (1 – α)-level upper confidence bound for the difference between $\mu_1 - \mu_2$ is negative (i.e., < 0)

Two conditions that must be met to ensure that the conclusions will be the same for a hypothesis test and a confidence interval performed on the same data:

1. The confidence level must be a compliment of the significance level (α) applied in the hypothesis test.
2. They must be the same "sided", that is, both two-sided or both one-sided.

- If these two conditions are not met, the hypothesis test and confidence level may still give the same conclusion, but there is no guarantee

Example of Relating Hypothesis Test and Confidence Interval for the Difference Between Two Means

Return to the example of comparing the heights of Acropora formosa colonies at Mbudya and Bongoyo Islands.

Results from the hypothesis test showed that, at a significance level 5% (α = 0.05), there was no difference in the mean heights of *Acropora formosa* colonies at Mbudya and Bongoyo Islands.

Results from calculating the confidence interval showed that the (1 – α)% or 95% confidence interval for the difference between the two means (─ 11.85, 2.86 cm) includes 0. Therefore, we can be 95% confident that the difference between the two means is 0 (not significantly different from 0).

Therefore, the two types of inferential statistics give the same conclusion.

SPSS Output: Hypothesis test and Confidence Interval for the Difference in Mean Height of *Acropora formosa* on the Reef Crests at Mbudya and Bongoyo Islands

is diff from the t-test

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
A.formosa	Equal variances assumed	.026	.873	-1.247	32	.221	-4.49306	3.60228	-11.83067	2.84456
	Equal variances not assumed			-1.249	31.662	.221	-4.49306	3.59831	-11.82565	2.83954

>>>>>>>>>>

Levene's test

$H_0: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$

H_a: variance are diff for at least 2 pop.

using the default α of 0.05

P of sig. is > than 0.05, there is no sig diff.

>>>>>>>>>>

2.2.2 Inferences for Two Population Means: Using Two Paired Samples

- Applies when two populations or measurements are **paired** in space or in time or by some relationship or paired on the same subject (study unit)
- Examples of pairing:
  - **Ecological monitoring** (coral reefs, mangroves, etc.) – same plots observed over time.
  - Taking measurements at the same time in different sites (pairing in time)
  - Taking measurements on the same patient, such as blood pressure, before and after treatment
  - Measuring something like educational level of husbands and their wives
  - Heights of fathers and their oldest sons

Paired-Sample t-test (or Paired t-test) [Sometimes called **Matched Pairs t-test**]

Step 1: Check the purpose and assumptions (to see if this is the appropriate test)

Purpose: To test for the difference between two populations means, μ_1 and μ_2 (based on the differences between two paired samples).

Assumptions:

1. Simple random sample
2. Samples are paired (random paired sample) (not independent)
3. Differences between paired observations are normally distributed or sample size is large

Step 2: State the null and alternative hypotheses

The null hypothesis is H_0: $\mu_1 = \mu_2$ and the alternative hypothesis may be one of the following:

H_a: $\mu_1 \neq \mu_2$	or	H_a: $\mu_1 < \mu_2$	or	H_a: $\mu_1 > \mu_2$
(two-tailed)		(left-tailed)		(right-tailed)

Step 3: Obtain the Calculated Value (or Observed Value) of the test statistic as follows:

Mean difference = $\bar{d} = \dfrac{\sum d_i}{n}$ [Note: $d = y_1 - y_2$ for each pair]

Standard deviation of the mean difference = $s_d = \sqrt{\dfrac{\sum d_i^2 - (\sum d_i)^2 / n}{n-1}}$

$$t = \frac{\bar{d} - \Delta_0}{s_d / \sqrt{n}} = \frac{\bar{d} - \Delta_0}{SE(\bar{d})}$$

n = number of paired observations and $df = n - 1$

The formula can be broken down as follows:

Parameter	Estimate	SE(Estimate) of the mean difference	H_0 value	Reference Distribution
μ_d $(or \mu_1 - \mu_2)$	$\bar{d}$	$SE(\bar{d}) = \dfrac{s_d}{\sqrt{n}}$	$0 (or \Delta_0)$	t_{n-1}

Note: Δ_0 = some hypothesized difference, which is almost always 0

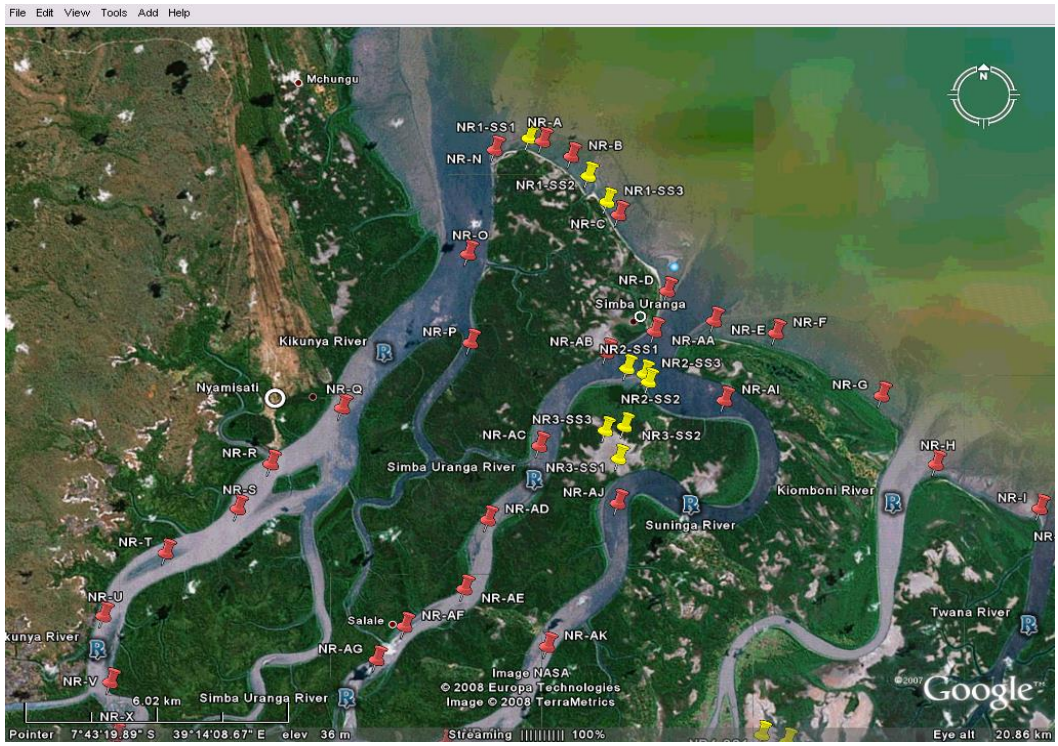
Step 4: Decide to reject H_0 or not reject H_0 and state the strength of the evidence against H_0
If you obtain a *df* that is not shown in the *t*-table, go to the next lower *df*.
If the P-value ≤ α, we reject H_0 (otherwise do not reject H_0)

Step 5: Interpretation (conclusion) in words in terms of the research problem being investigated.

Note on Power of the Paired-sample t-test:

- Where there is strong pairing (significant correlation) between paired observations, this paired-sample t-test is more powerful than the pooled or nonpooled t-test for independent samples
- Where correlation between pairs is not significant, the t-tests for independent samples are more powerful

Example of Paired Design:
Research on Impacts of Climate Change on Mangroves in Tanzania



Research design:

- Permanent plots were established
- Girth at breast height of mangroves was measured to get mangrove basal area in permanent plots in 2007 and 2009.
- Thus, **the measurements are paired in space (same plots) for two time periods.**

Two Trends:

1. **On seaward edges** of mangroves, there is drastic erosion in many areas due to sea level rise combined with increase in storms and wave activity
2. **On landward edges** there is landward migration of mangroves in some areas due to sea level rise into the open area (sea level is rising approximately 4 mm per year vertically, which can mean as much as 1 m per year influx of water horizontally in a very flat area)

Long stretch of coastline at Subsite NR-SS3 being eroded away by increased wave action and sea level rise due to climate change.

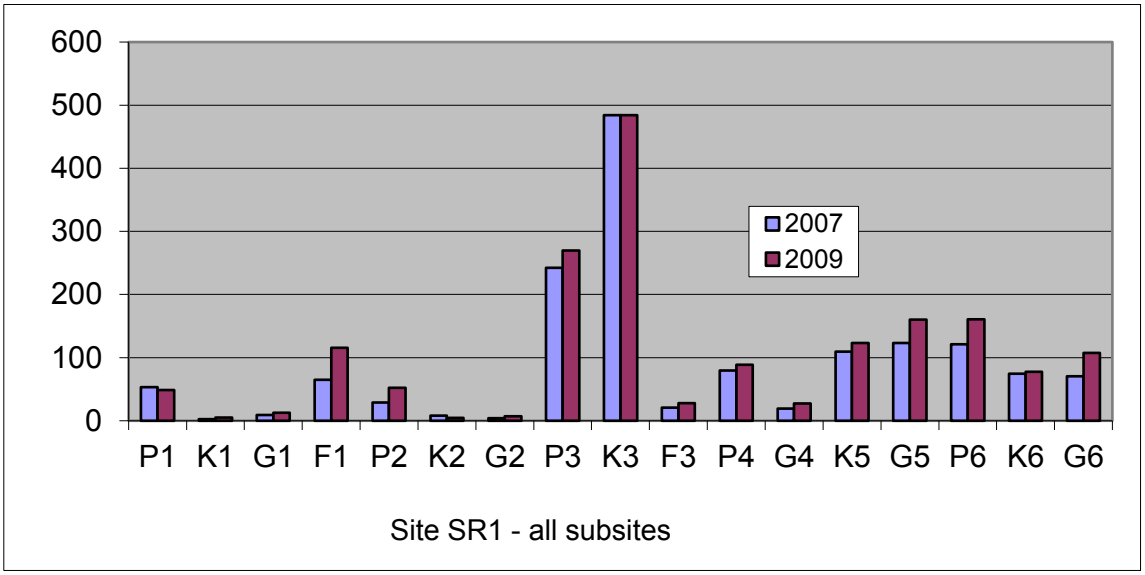
Avicennia mangroves invading into a saline flat area due to sea level rise, but some are also under stress due to decreased rainfall and humidity.

Research Hypothesis for landward edges: There was a difference/change in mangrove basal area (a measure of mangrove abundance) in Site SR1 (around the periphery of a saline flat area) between 2007 and 2009. It may have:

- **Increased** due to sea level rise into the open area, or
- **Decreased** due to decreased rainfall and humidity and increased temperatures, which cause desiccation and stress on the mangroves

Therefore, since the direction of change was not predicted,

- this is a **<u>two-tailed test</u>**



Research Problem: At the 5% significance level, test whether mangrove basal area in Site SR1 <u>changed</u> between 2007 and 2009. (Reduced to 10 plots) Assume that the required assumptions are met.

	Basal area (cm^2/25-m^2 plot)	
Plot	**2007**	**2009**
1	3	5
2	9	13
3	29	52
4	8	5
5	4	7
6	242	270
7	484	484
8	21	28
9	80	89
10	19	28

Step 1: Paired-sample t-test is selected because:
Purpose of the study: To test for a difference between two population means, based on two samples that were paired in space (in the same permanent plots) in 2007 and 2009.
Assumptions:
1. Random paired samples
2. The differences between the paired observations are approximately normally distributed

Step 2: H_0: $\mu_1 = \mu_2$ (Mean mangrove basal area did not change between 2007 and 2009)

H_a: $\mu_1 \neq \mu_2$ (Mean mangrove basal area changed between 2007 and 2009)

Parameter: $\mu_d = \mu_{2007} - \mu_{2009}$

>>>>>>>>>>

Step 3:

	Basal area (cm^2/25-m^2 plot)			
Plot	**2007**	**2009**	**d**	**d^2**
1	3	5	−2	4
2	9	13	−4	16
3	29	52	−23	529
4	8	5	3	9
5	4	7	−3	9
6	242	270	−28	784
7	484	484	0	0
8	21	28	−7	49
9	80	89	−9	81
10	19	28	−9	81
Sums			−82	1562

$\bar{d} = \frac{-82}{10} = -8.2$

SD of the diff

$S_d = \sqrt{\frac{\sum d^2 - \frac{(\sum d)^2}{n}}{n-1}} = \sqrt{\frac{1562 - \frac{(82)^2}{10}}{9}}$

$= 9.94205$

$t = \frac{\bar{d}}{S_d/\sqrt{n}} = \frac{-8.2}{9.94205/\sqrt{10}}$

$= 3.143$

$df = 9$

P-value $(0.01 < P < 0.025) \times 2 = -2.608$

$0.02 < P < 0.04$

<<<<<<<<<<

Comparison of the Power of the Paired-sample t-test and Pooled t-test in this case where there is very strong pairing (correlation)

[Based on all 17 plots in this site]

- Correlation coefficient: r = 0.9898 (this is very strong correlation)

	Paired-sample t-test	**Pooled t-test**
Calculated t	t = - 3.637	t = - 0.369
Exact P-value	P = 0.002271	P = 0.7147
Decision	Reject H_0	Do not Reject H_0
Evidence to reject H_0	Very strong	Very weak
Power	Very powerful	Not powerful

Therefore, where pairing is very strong, such as in this example, the paired-sample t-test is much more powerful than a t-test that is applied for independent samples.

Note:

1. Paired design made the study very sensitive to small changes over time.
2. If you had not paired these in fixed plots, the variation among plots within the same time period would have masked (been much greater than) the difference over time (2007 to 2009)

Paired t-Interval Procedure

Paired t-Interval Procedure

Purpose: To find a confidence interval for the difference between two population means, μ_1 and μ_2 based on paired observations.

Assumptions: Same as for the Paired t-test

Step 1: For a given confidence level (1 – α), use the t-table to find $t_{\alpha/2}$ in the row for the appropriate ***df***, where **df** = n – 1

Step 2: The endpoints of the confidence interval of $\mu_1 - \mu_2$ are defined by:

$$\bar{d} \pm t_{\alpha/2} \times \frac{s_d}{\sqrt{n}} \qquad \textbf{or} \qquad \bar{d} \pm t_{\alpha/2,n-1} \times SE(\bar{d})$$

Or, the formula can be broken down as follows:

Estimate	**Critical value**	**SE(Estimate) of the mean difference**
$\bar{d}$	$t_{\alpha/2,n-1}$ (or t*)	$\frac{s_d}{\sqrt{n}}$

Step 3: Interpret the confidence interval in terms of the research problem being investigated.

Margin of Error (E) = $t_{\alpha/2,n-1} \times \frac{s_d}{\sqrt{n}}$

Example
Based on the mangrove data shown above, calculate a 95% confidence interval for the difference in mangrove basal area in Site SR1 between 2007 and 2009.

>>>>>>>>>>

For 95% C.I, and an α of 0.05 @ $df = n-1 = 10-1 = 9$ $t_{0.025} = 2.262$

Parameters: $\mu_d = \mu_{2007} - \mu_{2009}$

$SE(\bar{d}) = 3.14395$

$\bar{d} \pm t_{\alpha/2} \times SE(\bar{d})$

$-8.2 \pm 2.262 \times 3.14395$

$(-15.31, -1.09)$

It is estimated with a 95% C.I that the diff in ... between 2007 and 2009 was somewhere between -15.31 and -1.09. Note: ME = 7.11

Because 0 is not in the C.I. we reject H_0.

>>>>>>>>>>

SPSS Output: Paired t-test and Confidence Interval for the Difference in Mangrove Basal Area in Site SR1 between 2007 and 2009

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	2007	89.9000	10	156.42282	49.46524
	2009	98.1000	10	157.54396	49.81978

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	2007 & 2009	10	.998	.000

Paired Samples Test

	Paired Differences					t	df	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
2007 - 2009	-8.20000	9.94205	3.14395	-15.31212	-1.08788	-2.608	9	.028

Research Problem on Exercise Program to Reduce Weight
It is claimed that a certain exercise program will reduce body weight by 20 kg or more within 6 months in seriously overweight people. The table below shows the body weights of a random sample of 15 people before and after undertaking this program. At the 1% significance level, test whether this claim is true.

Person	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Before	88	94	121	160	138	115	105	112	109	99	123	135	150	155	142
After	65	68	94	136	118	95	88	92	89	73	100	114	134	133	120
Diff.	23	26	27	24	20	20	17	20	20	26	23	21	16	22	22

Step 1: Paired-sample t-test is selected because the purpose of the study is to test for a difference between two population means based on the difference in body weights that are paired on the same person (before and after).

>>>>>>>>>>

$H_0: \mu_{before} - \mu_{after} = 20\ kg$

$H_A: \mu_{before} - \mu_{after} > 20\ kg$

Parameter: $\mu_d = \mu_{before} - \mu_{after}$

hypothesised diff: $\Delta_0 = 20\ kg$

>>>>>>>>>>

Step 3:
Minitab Output
Paired T-Test and CI: Body wt-before, Body wt-after

```
                 N    Mean  StDev  SE Mean
Body wt-before  15  123.07  22.68     5.86
Body wt-after   15  101.27  23.61     6.10
Difference      15  21.800  3.167    0.818


99% lower bound for mean difference: 19.654
T-Test of mean difference = 20 (vs > 20): T-Value = 2.20  P-Value = 0.022
```

By Hand Calculations

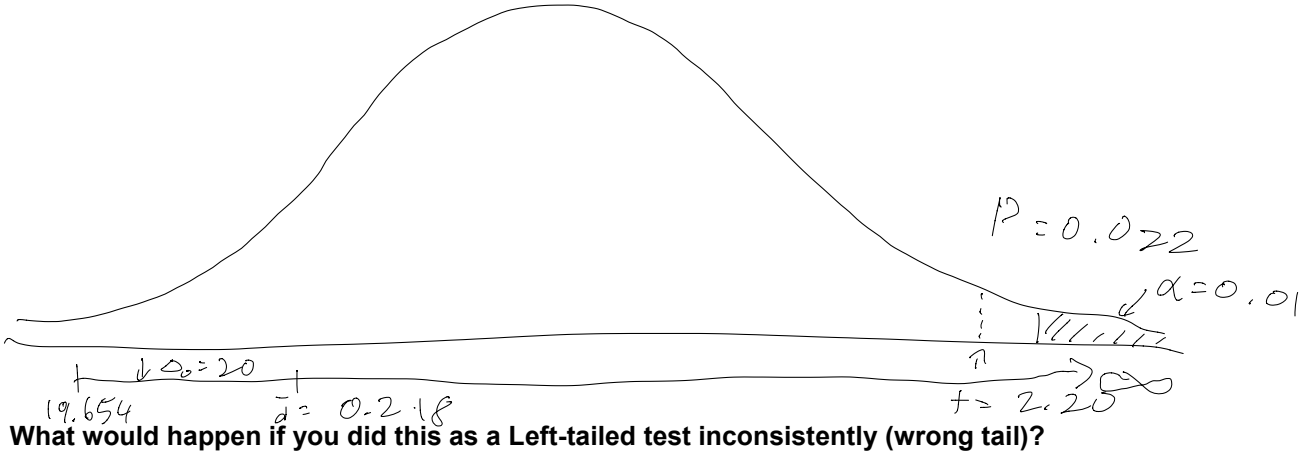
$$t = \frac{\bar{d} - 20}{s_d / \sqrt{n}} = \frac{21.8 - 20}{3.167 / \sqrt{15}} = \frac{1.8}{0.818} = 2.200$$

Step 4: $df = n - 1 = 15 - 1 = 14$

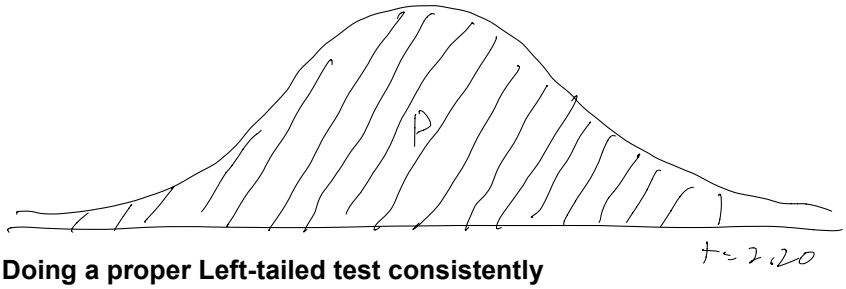
At df = 14, P-value = 0.02 < P < 0.025 (Minitab gives exact P-value = 0.022)
Since P-value > α (0.01), do not reject H_0. There is strong evidence against H_0 but not very strong evidence as required by this research problem.

Step 5:
At the 1% significance level, the data do not provide sufficient evidence to prove the claim that the exercise program reduces body weight by 20 kg or more in seriously overweight people.

Sketch graph of Right-tailed test and Lower-bound *t*-interval



What would happen if you did this as a Left-tailed test inconsistently (wrong tail)?



Doing a proper Left-tailed test consistently

for a left tail test you need a negative diff

What would happen if you did this test without setting a hypothesized difference (in other words the Ho value = 0)?

>>>>>>>>>>

IMPORTANT NOTE: P-value of One-tailed and Two-tailed tests

- A two-tailed can sometimes have a P-value > 0.50
- However, for a one-tailed test, if you get a P-value > 0.5, you have tested the wrong tail

Research problem on corals: The heights of randomly selected *Acropora formosa* colonies at Mbudya Island and Bongoyo are shown below, along with summary statistics. However, this time the research design is planned such that it measures colonies at different depths (below mean sea level) along the slope of the reef. At the 5% significance level, test whether there is a difference in the mean heights of colonies of *Acropora formosa* colonies found on these two reefs.

Depth (m)	1	1	2	2	3	3	4	5	6	7	8	9	10
Mbudya (cm)	68	67	70	73	88	83	82	74	83	77	84	79	90
Bongoyo (cm)	70	71	69	71	79	74	77	73	69	72	72	70	82
Diff. In height	-2	-4	1	2	9	9	5	1	14	5	12	9	8

Step 1: Paired-sample t-test is selected because the purpose of the study is to test for a difference between two population means, based on samples that are paired in space (paired at the same vertical depths on the reef).

Step 2: H_0: $\mu_{Mbudya} - \mu_{Bongoyo} = 0$

Ha: $\mu_{Mbudya} - \mu_{Bongoyo} \neq 0$

Parameter: $\mu_1 - \mu_2 = \mu_{Mbudya} - \mu_{Bongoyo}$ (mean heights of *A. formosa* colonies)

Step 3: (Minitab Output)

Paired T-Test and CI: Mbudya, Bongoyo

	N	Mean	StDev	SE Mean
Mbudya	13	78.31	7.49	2.08
Bongoyo	13	73.00	4.02	1.12
Difference	13	5.31	5.45	1.51

95% CI for mean difference: (2.01, 8.60)
T-Test of mean difference = 0 (vs not = 0): T-Value = 3.51 P-Value = 0.004

Step 4:
$df = n - 1 = 13 - 1 = 12$
At df = 12, P-value = (0.001 < P < 0.0025) x 2 = (0.002 < P < 0.005)
(Minitab gives exact P-value = 0.004)
Since P-value < α (0.05), reject H_0. There is very strong evidence against H_0 (< 0.01)

Note: the 95% confidence interval for the mean <u>difference</u> does <u>not</u> contain 0, which agrees with the conclusion of rejecting Ho.

Step 5: Interpretation (Conclusion in Words)
At the 5% significance level, there is very strong evidence that there is a difference in the mean heights of *Acropora formosa* colonies along the reef slopes at Mbudya and Bongoyo Islands.

Cora reef researcher's interpretation: The greater height of *Acropora formosa* colonies along the reef slope at Mbudya compared to Bongoyo is likely explained by (or related to) two environmental factors:

1. Faster ocean currents at Mbudya due to its being more open to the ocean. Thus, coral polyps have access to a greater abundance of prey items and the symbiotic zooxanthellae (microscopic plants embedded in the coral tissue) can absorb nutrients more quickly.
2. Greater pollution of the seawater at Bongoyo due to a coastal sewage pipe. This pollution increases levels of phytoplankton, thus blocking light required by the zooxanthellae.

2.3 More on Assumptions of Statistical Inference (Including ANOVA)

2.3.1 Summary of the Assumptions for Various Hypothesis Tests/Confidence Intervals

- Generally, the assumptions are the same for any given hypothesis test and its corresponding confidence interval (e.g. Pooled t-test and Pooled t-confidence interval)

Simple Random Sampling

- Required by all statistical inference

Independence of Sampling

- Required only for the **two-sample t-test for independent samples (pooled and nonpooled)**
- And **One-way ANOVA**

Normally Distributed Data

- **One-sample t-test:** The one sample must come from a population that is normally distributed
- **Two-sample t-test for independent samples (Pooled or Nonpooled):** Both samples must come from populations that are normally distributed
- **Paired-sample t-test:** Differences between paired observations must be normally distributed (the two separate populations may not necessarily be normally distributed)
- **One-way ANOVA:** All samples being compared must come from normally distributed populations

Equal Standard Deviations (or Variances)

- **Two-sample t-test for Independent Samples (Pooled t-test only):** The standard deviations of the two samples must be equal (or approximately so)
  - Along with this the two sample sizes must be nearly equal
- **ANOVA:** All samples being compared must have equal variances (or approximately so)

2.3.2 Assessing/Testing for Violation of the Assumptions

Planning for and Assessing Simple Random Sampling

- Must be planned as a basic part of the research design
- No transformation can correct the sampling design once the research has been conducted

Planning for and Assessing Independence of Sampling

- Required only for the **two-sample t-test for independent samples (pooled and nonpooled)** and **One-way ANOVA**
- Cannot be corrected by any transformation, but if it is realized later that there is some kind of pairing of the observations, you can switch to doing a paired test

Assessing/Testing for Normality

- **Histograms, Stem-and-leaf diagrams and Dotplots**
  - Perform a visual assessment by comparing the distribution of the population with a bell-shaped curve
  - Very subjective
- **Normal probability plot (also called the Q-Q Plot)**
  - Used to to assess each data set separately
  - Perform a visual assessment by comparing the distribution of the population with a straight line
  - Normality assumption is not violated if the all data points fall approximately in a straight line
  - Normality assumption is violated if there are **serious departures** from a straight-line pat
  - Easier to evaluate than histograms, etc.
  - Still somewhat subjective, but much easier to determine a straight line than a bell-shaped curve

Assessing Normality Using a Normal Probability Plot (Similar to Q-Q Plots in interpretation)

- Normal probability plots can also be used to assess outliers

Guidelines for Assessing Normality Using a Normal Probability Plot

- If the plot is roughly linear, you can assume that the variable is approximately normally distributed.
- If the plot is not roughly linear, you can assume that the variable is not approximately normally distributed.

These guidelines should be applied:

- loosely for small samples, but
- strictly for large samples

- **Hypothesis Tests**
  - Can make an objective decision
  - **Chi-square Goodness of Fit Test:** Expected frequencies are calculated using the proportions of the normal distribution and these are compared with the observed frequencies
  - **Anderson-Darling Test (AD Test)**
    - Very powerful for testing normality

>>>>>>>>>> Hypotheses of AD test:

H_0: The distribution is not diff from the normal distribution.

H_1: The distribution is diff from the normal distribution.

Default is $\alpha = 0.05$

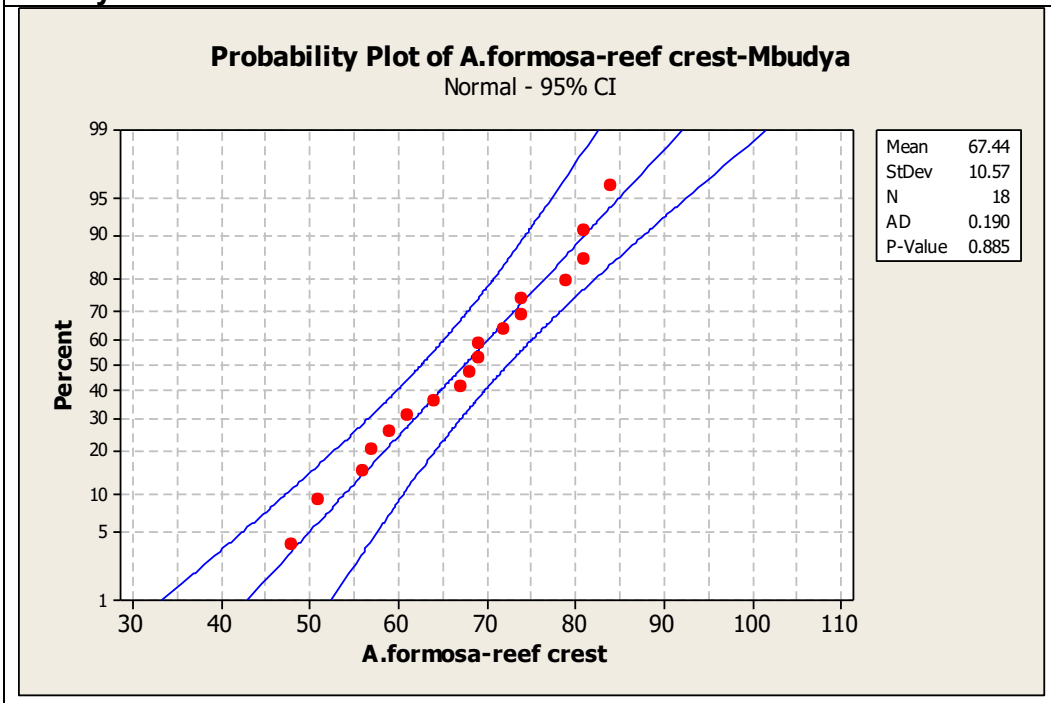
>>>>>>>>>>

Assessing/Testing for Equal Standard Deviations

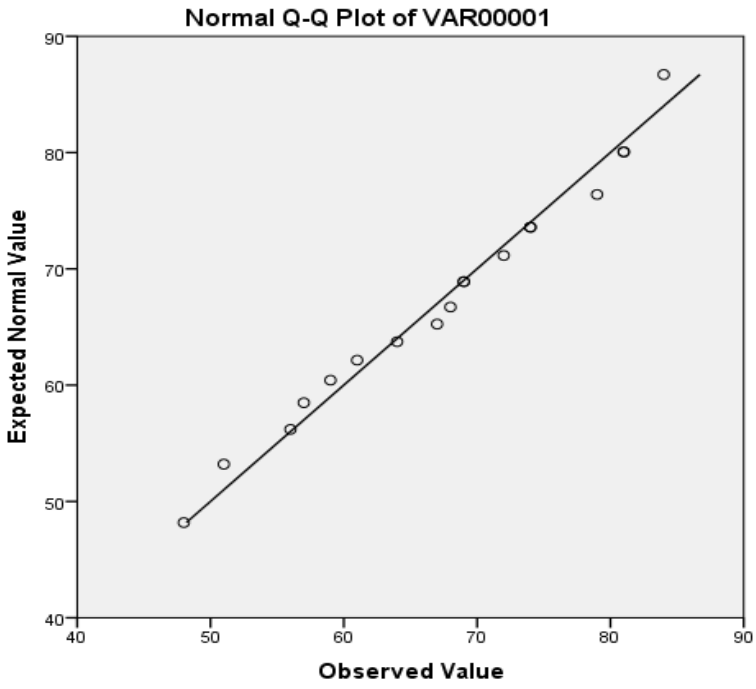
- **Boxplots**
  - Construct side-by-side boxplots on the same scale and compare their spread
  - Subjective method
- **Ratio of the standard deviations**
  - Calculate the ratio of the largest standard deviation divided by the smallest standard deviation. If this ratio is ≤ 2, we may consider the standard deviations to be equal enough to perform the pooled t-test
  - Cannot be applied for ANOVA
- **Levene's Test for the Equality of Variances**
  - Perform in SPSS
  - Very accurate way of assessing equality
  - Can be applied for ANOVA

Testing for Normality in the Previous Examples

One-sample of data on the heights of Acropora formosa colonies on the reef crest at Mbudya Island



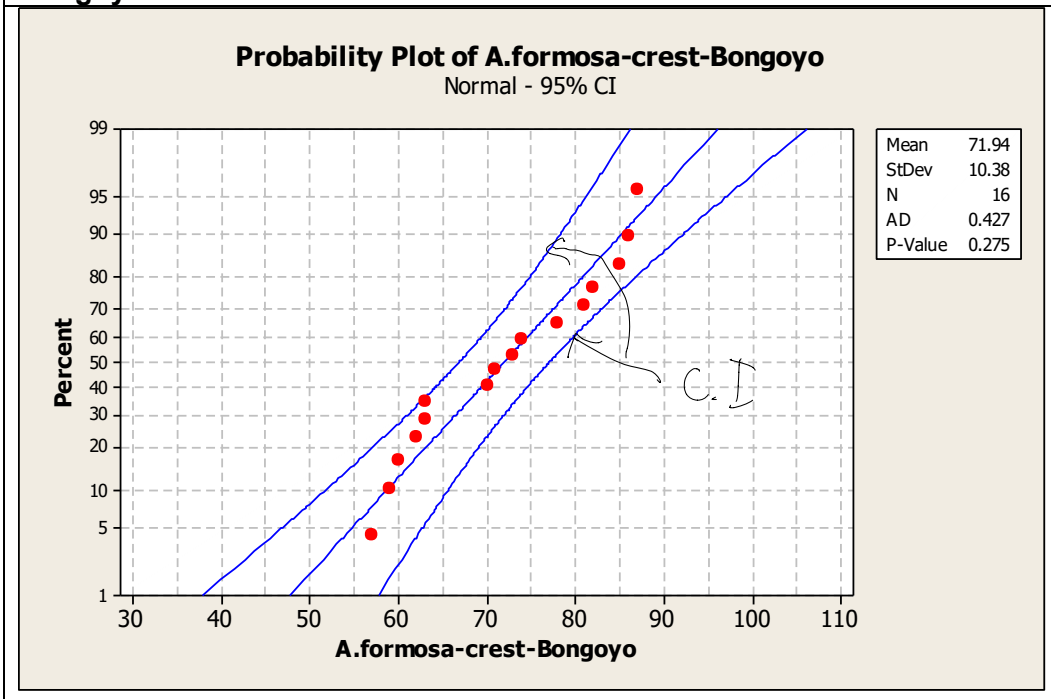
Minitab output



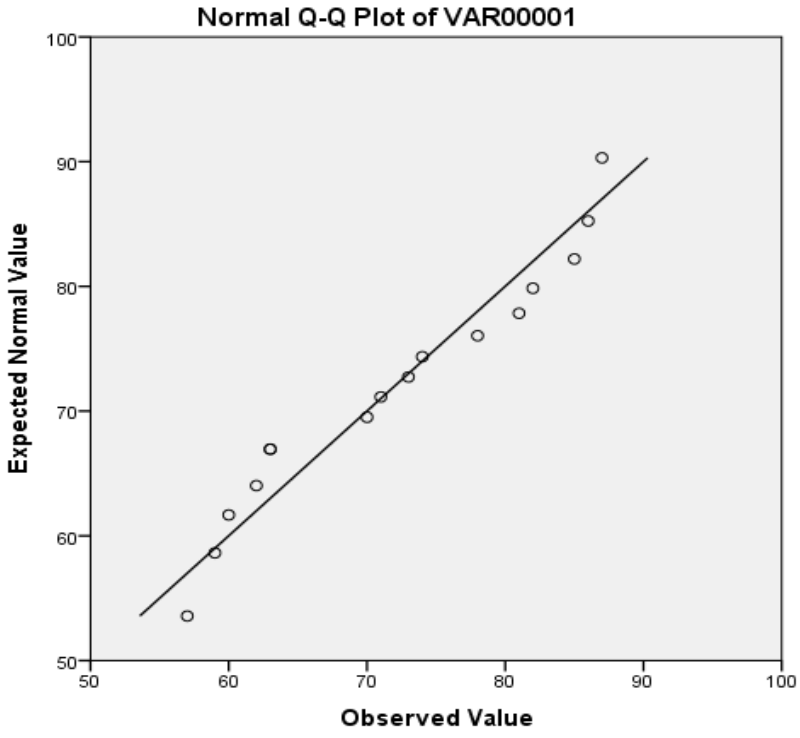
SPSS output

Testing for Normality in the Previous Examples

One-sample of data on the heights of Acropora formosa colonies on the reef crest at Bongoyo Island



Minitab output

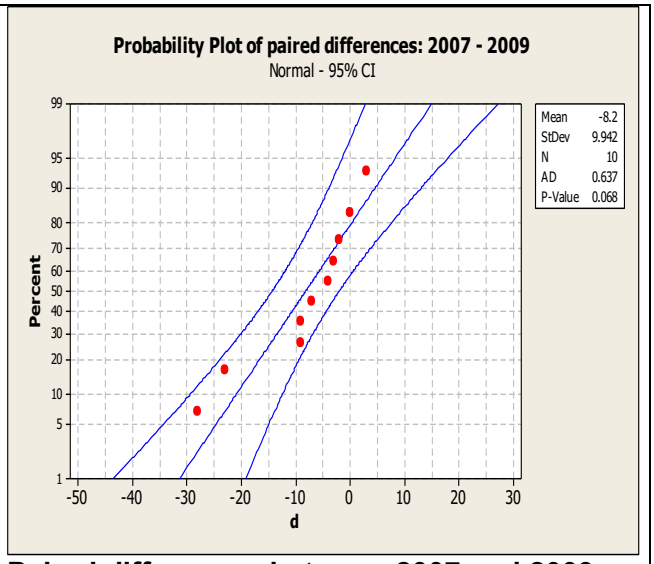


SPSS output

	Basal area (cm^2/25-m^2 plot)		
Plot	**2007**	**2009**	**d**
1	3	5	-2
2	9	13	-4
3	29	52	-23
4	8	5	3
5	4	7	-3
6	242	270	-28
7	484	484	0
8	21	28	-7
9	80	89	-9
10	19	28	-9
Sum			**-82**

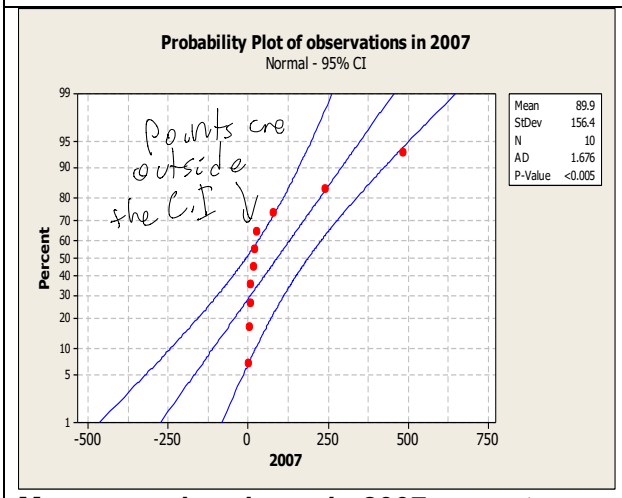
Paired design applied:

1. Much more powerful due to pairing in space
2. Separate samples for 2007 and 2009 seriously violate the assumption of normality



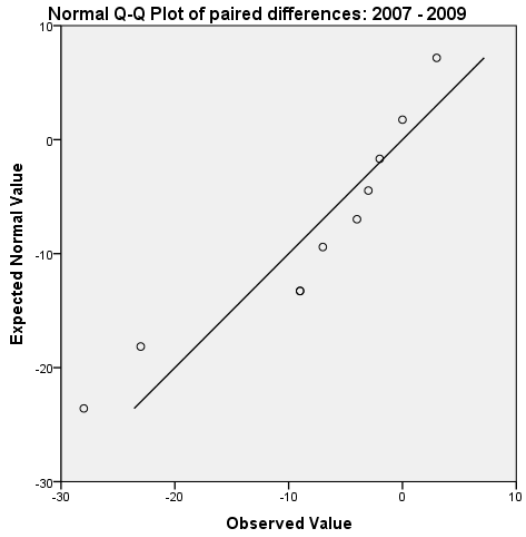
Paired differences between 2007 and 2009 are normally distributed because:

1. All data points fall within 95% CI lines
2. Anderson-Darling (AD) test statistic = 0.637; P-value = 0.068. Thus H_0 not rejected; no difference between this data set and a normal distribution



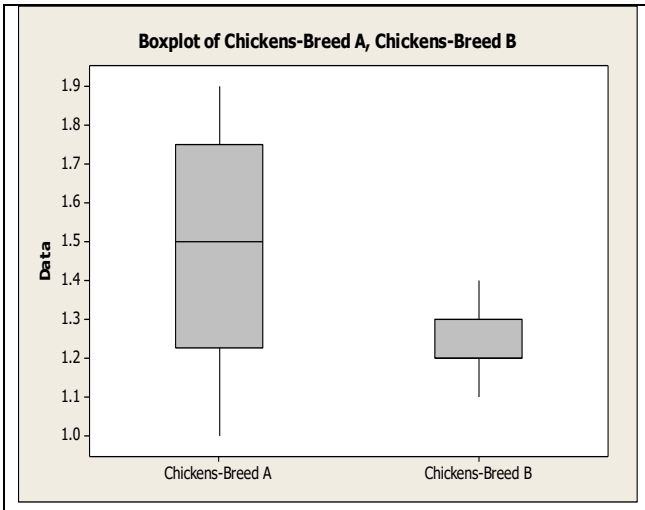
Mangrove abundance in 2007 are not normally distributed because:

1. The plot is curved, and some data points fall outside the 95% CI lines
2. Anderson-Darling (AD) test statistic = 1.676; P-value < 0.005. Thus, H_0 is rejected; that is, there is a significant difference between this data set and a normal distribution



SPSS output:
Paired differences between 2007 and 2009

Assessing/Testing for Equal Standard Deviations



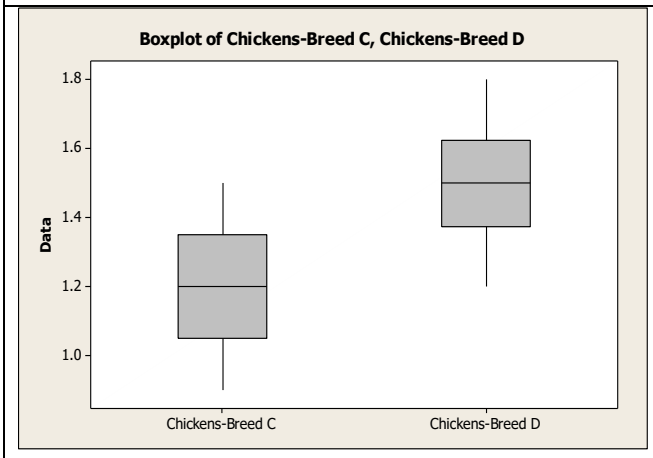
Side-by-side boxplots:
Very different spread

Ratio of standard deviations:

$$\frac{s(larger)}{s(smaller)} = \frac{0.3125}{0.1036} = 3.016 > 2$$

[Also sample sizes very different (6 and 11)]

Levene's test for equality of variances:
F = 4.960, P = 0.042



Side-by-side boxplots:

Ratio of standard deviations:

$$\frac{s(larger)}{s(smaller)} = \frac{0.1900}{0.1826} = 1.041 < 2$$

[Sample sizes nearly the same (9 and 10)]

Levene's test for equality of variances:
F = 0.013, P = 0.911

2.3.4 Robustness and Resistance of the t-Tools

- Virtually no data set will fit all the assumptions perfectly
- How much can a data set depart from the assumptions without the statistical procedures being significantly affected

Robustness = the ability of a statistical procedure to withstand departures or violations of the assumptions without being seriously or significantly affected

Departures from normality

- The t-tools are relatively robust to departures from normality
- Only if normality is seriously violated, will the t-test give unreliable results
- If normality is seriously violated, data transformation should be performed, or one can apply nonparametric tests

Departures from equal standard deviations

- This is the most crucial assumption
- If standard deviations are significantly unequal, then the pooled estimate of the population standard deviation does not result in an accurate SE(Estimate) and the results of the pooled t-test are then invalid

Departures from independence

- The SE(Estimate) in a t-test is calculated based on the independence assumption, so if the samples were not obtained independently, the t-test analysis may give very misleading results
- The experiment may have to be repeated or, if significant pairing of the data occurs, the paired-sample t-test may be performed in place of the t-test for independent samples

Resistance of t-tools to Outliers

- Drastic outliers can often affect the t-tools more seriously than violations of some of the other assumptions
- Consider the effects of outliers on the sample mean:
  Data set I: 10, 20, 30, 50, 70 Sample mean = 36
  Data set II: 10, 20, 30, 50, 700 Sample mean = 162

- Sample mean is not resistant to outliers
- Standard deviation also not resistant to outliers
- However, the median is resistant to outliers: For both samples, median = 30
- Since t-tools are based on calculation of the sample mean, these tools are not resistant to serious outliers
- Thus, one or two drastic outliers can greatly affect the confidence interval as well as the t-statistic, changing the P-value and sometimes completely changing the conclusion
- Transformations or nonparametric methods can be used when there are serious outliers

2.4 Transformations

- Nonrandomness and Nonindependence - Cannot be corrected by transformations
- Nonnormality, outliers and unequal standard deviations can often be corrected by transformations
- Often one transformation can correct several violations of the assumptions at the same time

Types of Transformations

Logarithmic transformation

- The Log transformation is the most common type of transformation and will be emphasized in this course
- Both natural logarithms and log base 10 (common) can be used, but in this course we will primarily use natural logarithms to the base ***e*** (where ***e*** = 2.7183)
- Data that becomes normally distributed after this transformation is referred to as lognormal

Square Root Transformation

- Take the square root of all observations
- Useful for data recorded in counts

Reciprocal transformation

- Useful for waiting times

Arcsine Square Root Transformation

- Useful for proportions

Analysis of Data After Transformations Have Been Applied

- After transformation, the t-tools are applied to the transformed data
- The analysis must be back-transformed in order to interpret the research problem and draw a conclusion
- Back-transformation is often the most difficult part
- However, with log transformed data, back transformation is relatively easy; therefore we will mainly be applying log transformations

Demonstration on Effects of Natural Log (base *e*) Transformations on Various Scenarios:

1. **Data sets (A and B) that fit all assumptions**
   - The demonstration shows that, the two data sets were normally distributed, with no outliers, and had nearly equal standard deviations before transformation
   - The natural log transformation has little effect on the extent of normality and the equality of standard deviations
2. **Data sets (C and D) with unequal standard deviations, but both were normal**
   - Before transformation, the ratio of larger SD to smaller SD = 29.31/1.912 = 15.33 (much greater than 2, therefore unsuitable of the pooled t-test
   - After transformation, the ratio is 0.4001/0.3626 = 1.10 (suitable for the pooled t-test)
3. **Data set that was very nonnormal (curved, right skewed) and had a very large SD**
   - Before transformation, the Anderson-Darling test statistic is 1.068 with P-value = 0.006, meaning that the distribution is very significantly different from normality
   - After transformation, AD is 0.202 with a P-value of 0.848, making it very normal (not significantly different from normality)
   - At the same time, the transformation changes SD from being huge (79.64) to being moderate 1.557, making it likely to be similar to other data sets with which you might want to compare it
   - Note: one natural log transformation corrected the data for 2 assumptions at the same time
4. **Data set with two extreme outliers**
   - Before transformation, the AD test gave P-value < 0.005 and several data points are outside the 95% CI lines
   - After transformation, the AD test gave P-value = 0.028, which is just under 5%, making the transformed data not completely normal. However, the data points are almost within the CI lines and, as such the transformed data does not vert seriously violate normality. Apart from the two extreme outliers, the rest of the data are close to being normal.
   - This points to the fact that outliers are sometimes more difficult to correct than data that are nonnormal as indicated by a curved probability plot

2.5 Inference after a (Natural) Log Transformation

Steps in Transforming Data and Making Inferences

1. Transform the data
2. Check whether the transformed data fit the assumptions of the required test
3. Perform the hypothesis test/ the confidence interval calculations on the transformed data
4. Back-transform the estimate and the confidence interval
5. State the conclusions on the original scale

If the transformation is successful, the log-transformed data will be approximately symmetric such that:

Mean [ln(Y)] = Median [ln(Y)]

And since the log preserves ordering,

Median [ln(Y)] = ln[Median(Y)]

$\overline{LnY}_1$ and $\overline{LnY}_2$ represent the averages of the logged values of Sample 1 and Sample 2

Thus, $\overline{LnY}_1 - \overline{LnY}_2$ estimates $\ln[Median(Y_1)] - \ln[Median(Y_2)]$

And $\overline{LnY}_1 - \overline{LnY}_2$ estimates $= \ln\left[\frac{Median(Y_1)}{Median(Y_2)}\right]$

And $e^{\overline{LnY}_1 - \overline{LnY}_2}$ estimates $\left[\frac{Median(Y_1)}{Median(Y_2)}\right]$

Example: Students were randomly allocated to two groups, one group to a new program and the other group to a standard program. At the end of the experiment, their test scores were recorded (scale = 0 – 700) as shown below. Since the test scores are not normally distributed and the standard deviations are very different, the log (natural) transformed data are also shown as well as summary statistics. At the 10% significance level, test whether there is a difference in the test scores of students undertaking the two programs and calculate 90% confidence limits.

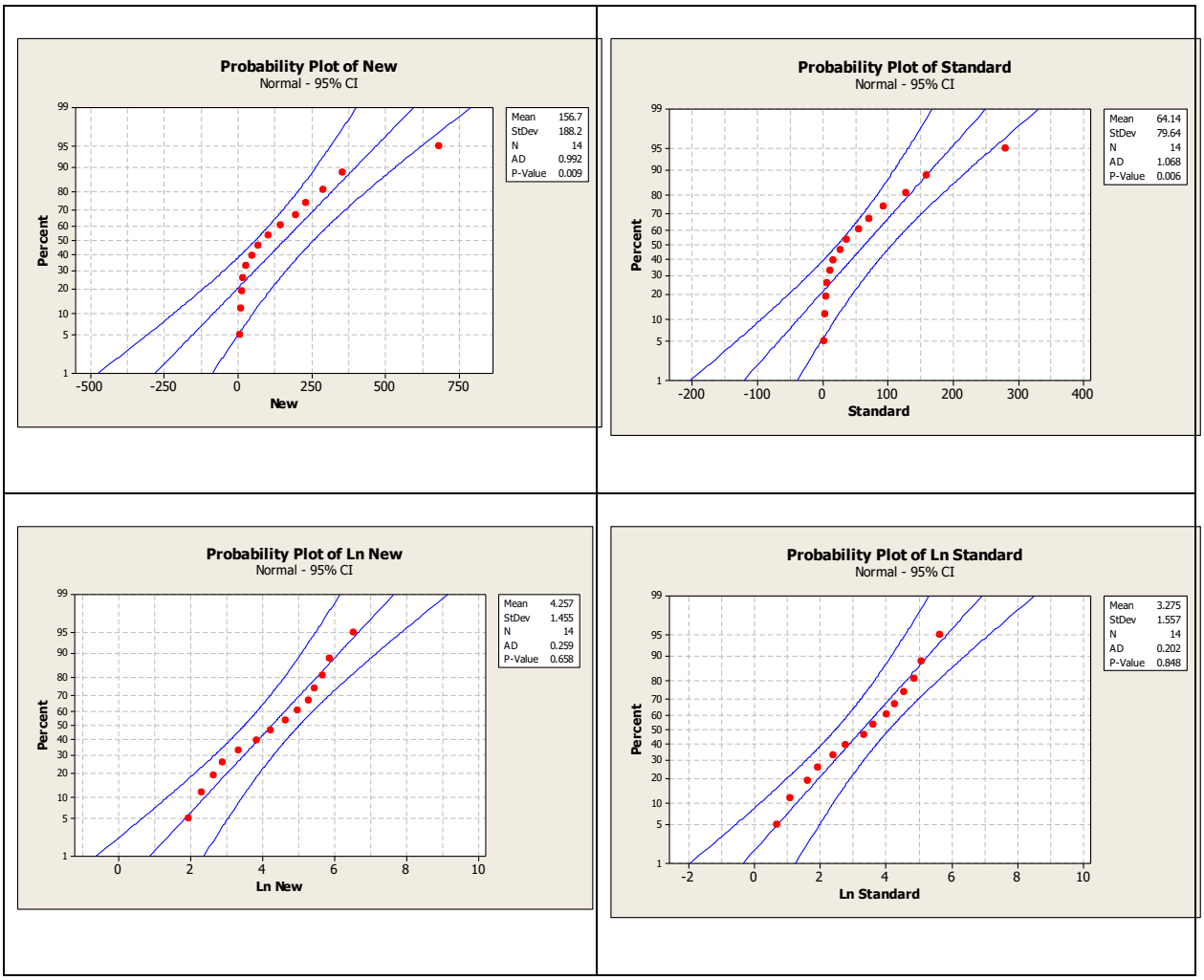
New	Standard	Ln(New)	Ln(Standard)
7	2	1.94591	0.693147
10	3	2.302585	1.098612
14	5	2.639057	1.609438
18	7	2.890372	1.94591
28	11	3.332205	2.397895
47	16	3.850148	2.772589
69	28	4.234107	3.332205
104	37	4.644391	3.610918
145	55	4.976734	4.007333
198	72	5.288267	4.276666
230	94	5.438079	4.543295
288	128	5.66296	4.85203
356	160	5.874931	5.075174
680	280	6.522093	5.63479

Descriptive Statistics: New, Standard, Ln New, Ln Standard

Variable	Total Count	Mean	SE Mean	StDev	CoefVar	Minimum	Q1	Median
New	14	156.7	50.3	188.2	120.11	7.0	17.0	86.5
Standard	14	64.1	21.3	79.6	124.15	2.0	6.5	32.5
Ln New	14	4.257	0.389	1.455	34.18	1.946	2.828	4.439
Ln Standard	14	3.275	0.416	1.557	47.55	0.693	1.862	3.472

Variable	Q3	Maximum	Skewness
New	244.5	680.0	1.85
Standard	102.5	280.0	1.78
Ln New	5.494	6.522	-0.17
Ln Standard	4.620	5.635	-0.21

Checking normality and equal standard deviations before and after transformations



Two-Sample T-Test and CI: Ln New, Ln Standard (Performed on log transformed data)

```
              N  Mean  StDev  SE Mean
Ln New       14  4.26   1.45     0.39
Ln Standard  14  3.28   1.56     0.42


Difference = mu (Ln New) - mu (Ln Standard)
Estimate for difference:  0.982
90% CI for difference:  (0.011, 1.954)
T-Test of difference = 0 (vs not =): T-Value = 1.72  P-Value = 0.096  DF = 26
Both use Pooled StDev = 1.5070
```

Conclusions based on the log transformed data:

Hypothesis test:
At the 10% significance level, there is moderate evidence that there is a difference in the means of the logged test scores between the new program and the standard program (Two-sample pooled t-test: t = 1.72, df = 26, P = 0.096).

Confidence interval:
The estimate of the difference between the means of the logged test scores of the new program and the standard program is 0.982 and the 90% confidence interval for the additive effect of the new program on the test scores is between 0.011 and 1.954. [Also, we can be 90% confident that there is a difference between the means of new and standard programs because 0 is not inside this confidence interval.]

Back Transformation of the Estimate and Confidence Int. and Interpretation on the original scale (Done by taking antilogs)

undo everything >>>>>>>>>>

Estimate of diff $= e^{0.982} = 2.6698$

LB of C.I. $= e^{0.011} = 1.0111$

UB of C.I. $= e^{1.954} = 7.0569$

Conclusions on the original scale (indicating the multiplicative effect of the treatment):

Conclusion of the hypothesis test:
@ the 10% sig level: the median test score of those who took the new prog. is estimated to be 2.6698 times the median score of those in the test prog.

Conclusion of the confidence interval:

At 90 % C.I for the ratio of the median in the original scale is

$\left|\frac{Med(new)}{Med(old)}\right|$ is $C(e^{0.011}, e^{1.954}) = C(1.0111, 7.0569)$

Or it is estimated with 90% confidence that the median of the test program score of the new program is in between ✓ times the median test score for the std prog.

>>>>>>>>>>

OR, we can say that the median test score for the new program is between 1.1% [(1.011 – 1) x 100)] and 605.7% [(7.057 – 1) x 100)] higher than the median test score for the standard program.

[**NOTE:** Also, this means that we can be 90% confident that there is a difference between the medians of new and standard programs because 1 (NOT 0) is not inside this confidence interval (1.011, 7.057).
This is because ln 1 = 0 and the antilog of 0 = 1(that is, $e^0 = 1$)

Note that:
The log-transformed data for both distributions are approximately symmetric
Mean [ln(Y)] = Median [ln(Y)]
For new program: 4.257 (mean) ≈4.439 (median)
For standard program: 3.275 (mean) ≈3.472 (median)
[Please **note** that, for symmetric distributions, the mean will always be approximately equal to the median; however, if a distribution has mean = median, this does not guarantee that it is symmetric.]
And since the log transformation preserves ordering,
Median [ln(Y)] = ln[Median(Y)]
For new program: The median of the logged values (4.439)
≈ln of the median of the original data (ln 86.5 = 4.460)
For standard program: The median of the logged values (3.472)
≈ln of the median of the original data (ln 32.5 = 3.481)

Furthermore:

$$e^{(\overline{LnY}_1 - \overline{LnY}_2)} = e^{(4.257-3.275)} = e^{0.982} = 2.670$$

estimates $\left[\dfrac{Median(Y_1)}{Median(Y_2)}\right]$ (population parameters)

estimated by $\dfrac{Median(sample1)}{Median(sample2)} = \dfrac{86.5}{32.5} = 2.662$

This points to the multiplicative interpretation of the ratio of the population medians. Recall that the median is a better measure of the center of a skewed distribution than the mean.

Back Transformation in Reverse

[Reverse means: Subtracting Standard minus New, instead of New minus Standard (as above)]

Back Transformation of the estimate and confidence interval to the original data:

>>>>>>>>>>

Estimate of the diff $= e^{-0.482} = 0.3746$

UB of the C.I. $= e^{-0.011} = 0.9891$

LB of the C.I $= e^{-1.954} = 0.1417$

Conclusions on the original scale (indicating the multiplicative effect of the treatment)::

Conclusion based on the hypothesis test:

@ the 10% sig level the median test score of those who took the std program is estimated to be 0.3746 times the median test score of those in the new program .

Conclusion based on the confidence interval:

A ... ratio of the median is the orginal scale $\left|\dfrac{med(std)}{med(new)}\right|$

is $(e^{-1.954}, e^{-0.011}) = (0.1417, 0.9891)$

>>>>>>>>>>

Comparing the two results

Estimate of the difference for New vs. Standard = 2.670

Estimate of the difference for Standard vs. New = 0.3746

2.670 is the inverse of 0.3746

Ratio of the endpoints of the confidence interval for New vs. Standard $= \dfrac{1.011}{7.057} = 0.1433$

Ratio of the endpoints of the confidence interval for Standard vs. New $= \dfrac{0.1417}{0.9891} = 0.1433$

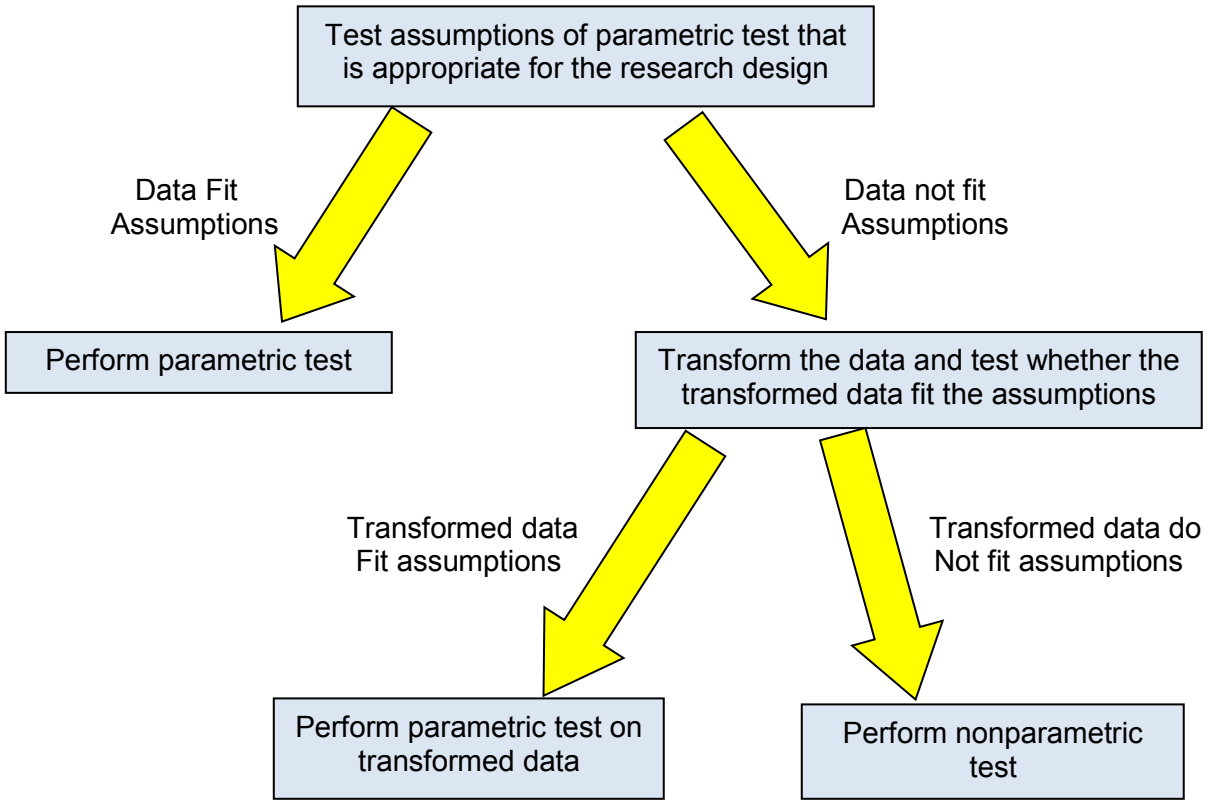
2.6 Nonparametric Methods

- If the data do not fit the assumptions of parametric tests, even after applying a transformation, then nonparametric methods should be performed
- Nonparametric methods do not use estimates of population parameters in their calculations
- Do not make any assumptions about the nature of the distribution of the populations being investigated
- Only assumption or requirement is that the samples must be selected randomly
- Therefore, they can be applied in many cases when the parametric methods are not valid
- Can be applied to categorical data, whereas parametric tests cannot.
- Slightly less powerful (approximately 95% as powerful)
- Most nonparametric tests convert the data to ranks and then calculations are performed on the ranks

Table: Parametric tests and their nonparametric equivalents

Parametric test	**Nonparametric test**
One-sample t-test	Wilcoxon signed-rank test
Two sample t-test (independent samples)	Mann-Whitney U test
Paired-sample t-test	Wilcoxon paired-sample test
One-factor ANOVA	Kruskal-Wallis test

2.6 Best Approach to Selecting Statistical Methods



Null Distribution

- If asked, "What is the distribution of the test-statistic under the null hypothesis?", you are just required to state the hypothesis test that you selected, the test statistic and degrees of freedom.
- For example: One-way ANOVA, "F(3, 28)" OR a two-sample t-test, "t(16)"

Supplementary Example

A fuel manufacturer wanted to test the effectiveness of a new gasoline additive. A random sample of 6 cars were driven one week without the additive and one week with the additive, obtaining summary statistics as shown in the table below (in miles per gallon). Note: You might not need all of the statistics shown.

Summary statistics	Without additive	With additive	Difference
Average	23.40	25.12	-1.72
Standard Deviation	5.42	5.87	1.43

Suppose that the numbers highlighted in yellow are not given.

(a) The confidence interval for the difference in mileage without and with the additive is $(-3.22094, -0.21906)$. Determine the confidence level at which this interval was calculated.

>>>>>>>>>>

$$ME = m = \frac{[UB - LB]}{2}$$

$$\text{paired t-interval} = \frac{[-0.21906 - (-3.22094)]}{2} = 1.50094$$

$$\bar{d} \pm t_{\alpha/2} \times \frac{s_d}{\sqrt{n}} \Rightarrow m = ME$$

$$t_{\alpha/2} = m \times \frac{\sqrt{n}}{s_d} = 1.50094 \times \frac{\sqrt{6}}{1.43}$$

$$t_{\alpha/2} = 2.571 = t_{0.025}$$

$$\Rightarrow \alpha = (0.025) \times 2 = 0.05$$

>>>>>>>>>>

(b) Calculate a 99% confidence interval for the difference in mileage without and with the additive.

>>>>>>>>>>

you get the idea....

>>>>>>>>>>

(c) Compare the confidence interval given in part (a) and the confidence interval you calculated in part (b). Based on each of these confidence intervals, is there a difference in mileage without and with the additive. Explain why you either got the same conclusion or different conclusions from the two intervals.

The confidence interval given in part (a), $(-3.22094, -0.21906)$ is shorter and more precise than the confidence interval calculated in part (b), $(-4.074, 0.634)$

Based on the 95% confidence interval given in part (a), we would conclude that there is a difference in mileage without and with the additive, it does not contain 0.
However, based on the 99% confidence interval calculated in part (b), we would conclude that there is no difference, because it does contain 0.

SECTION 3: SEVERAL POPULATION MEANS

- The purpose of Analysis of Variance (ANOVA) is to compare several (more than two) groups or means – that is, to find the **difference** among **more than two groups**
- Measurements of **only one variable** are recorded, but they come from different populations, treatments or groups
- The one variable being measured can be considered as the **response variable**

3.1 F-Distribution

- All types of ANOVA utilize the F-distribution
- The **F-statistic** is arrived at by calculating two types of variations and dividing one by the other
- The F-statistic has two numbers for degrees of freedom:
  - The **numerator degrees of freedom**, which corresponds to the type of variation placed in the numerator when calculating the test statistic, and
  - The **denominator degrees of freedom**, which corresponds to the variation placed in the denominator of the F-statistic
- These degrees of freedom are denoted, for example

df = (12, 35)

Numerator df denominator df

- There are an infinite number of F-distributions, each identified by the two degrees of freedom.

Basic Properties of F-Curves

Property 1: The total area under an F-curve equals 1.

Property 2: An F-curve starts at 0 on the horizontal axis and extend indefinitely to the right, approaching, but never touching, the horizontal axis.

Property 3: An F-curve is right skewed.

Property 4: At df = (∞, ∞), F = 1.000 at all significance levels.

Sketch of Two Different F-curves, df = (3, 40) and df = (15, 100)

The F-Table

- The F-table gives the areas (or probabilities) under the curve to the right of given values of F
- **Numerator degrees of freedom** (indicated as ***dfn***) are shown along the top of each page
- **Denominator degrees of freedom** (indicated as ***dfd***) are shown along the sides of each page.
- For any given combination of dfn and dfd, the F-values are given in a cluster and their significance levels are indicated along the sides.
- The critical values of F are always ≥ 1, though the calculated (observed) values may be < 1

Examples

- Find $F_{0.05}$ at df = (5,23) = 2.64
- Find $F_{0.025}$ at df = (11,180) ≈ (10, 100) = 2.18

Guidelines for Using P-values as Criteria for Rejection of H_0 and Statistical Significance

P-value	Strength of Evidence Against H_0
P > 0.10	Weak
0.05 < P ≤ 0.10	Moderate
0.01 < P ≤ 0.05	Strong
0.001 < P ≤ 0.01	Very strong
P ≤ 0.001	Extremely strong

3.2 ANOVA: Assumptions and Logic

- While the pooled t-test is used to compare one variable measured in two populations, ANOVA is used to compare one variable measured in more than two populations
- **One-Way ANOVA** (also called **Single-Factor ANOVA**)
  - Used to compare the values of one variable between (among) several groups or populations that are affected by **one factor**
  - This one factor may also be considered as one explanatory variable
  - The different values of the factor are called levels of that factor or treatment
- **Two-Way ANOVA** (also called **Two-Factor ANOVA**)
  - Used to compare the values of one variable among populations that are classified or grouped according to **two factors**
  - So, we can consider these as two explanatory variables
- Factors may be categorical variables or quantitative variables
- **Multiway Factorial ANOVA** deals with comparisons where more than two factors affect the populations
- **Randomized-block ANOVA** is an extension of the Paired-sample t-test, where you have more than two samples "blocked" in time or space or by some relationship.
- The Meaning of Analysis of Variance is that **we analyze and compare variances among populations with variance within the populations**
- The following terms are used synonymously:
  Groups = Treatments = Samples (taken from Populations)

- The **F-statistic** is:

  F = Between Groups (Samples) Variability
  Within Groups (Samples) Variability

  OR

  F = Treatment Mean Square (variation between samples)
  Error Mean Square (variation within samples)

- **"Error" = "Residual" = "Within Groups Variability"**

- Although the purpose of ANOVA is to compare several population means and the sample means are calculated during the analysis; in the end, the F-statistic only makes a comparison of variability (among and within), thus the term "Analysis of Variance"

One-Way ANOVA: Three Sources of Variation

Three Sources of Variation and Sums of Squares in One-Way ANOVA

For one-way ANOVA of ***k*** population means,

Total Sum of Squares (SS_{Total} **)** = total variation between and within samples or groups

Treatment Sum of Squares ($SS_{Treatment}$ **)** = variation between treatments or groups

Error (or Residual) Sum of Squares (SS_{Error} **)** = variation within treatments or groups

One-Way ANOVA Identity:

$$SS_{Total} = SS_{Treatment} + SS_{Error}$$

Mean Squares and F-Statistic in One-Way ANOVA

Treatment mean square ($MS_{Treatment}$)

= treatment sum of squares divided by treatment degrees of freedom

$$MS_{Treatment} = SS_{Treatment} / (k-1)$$

Where *k* = number of populations being compared

Error mean square (MS_{Error})

= error sum of squares divided by error degrees of freedom

$$MS_{Error} = SS_{Error} / (n-k)$$

Where *n* = total number of observations

F-Statistic (F)

= the ratio of the variation between groups to the variation within groups

F = Between Groups Variability / Within Groups Variability

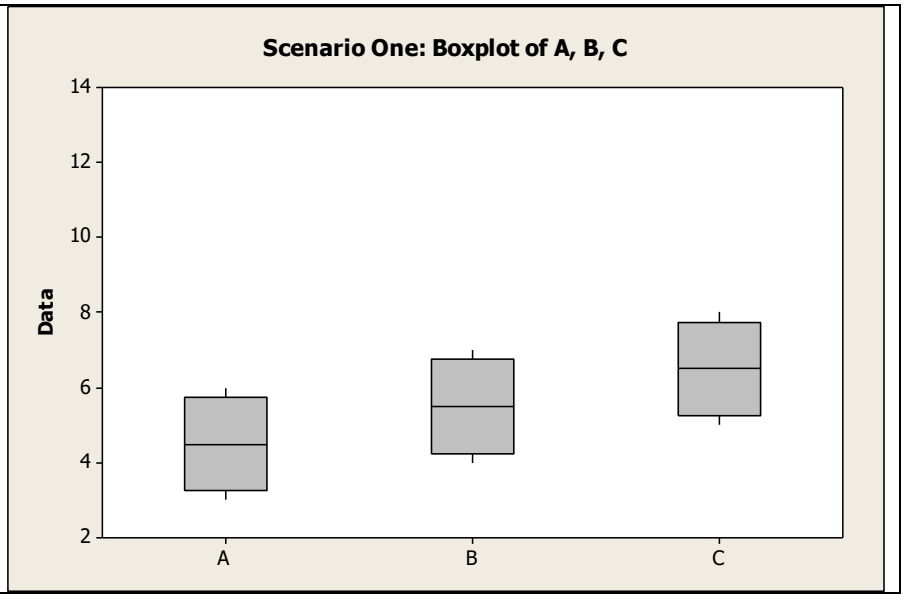
$$F = \frac{MS_{Treatment}}{MS_{Error}} = \frac{SS_{Treatment} / (k-1)}{SS_{Error} / (n-k)}$$

Three Scenarios to Explain the Logic of One-Way ANOVA

Scenario One

- Small variation among groups and small variation within groups
- No significant difference at α = 0.05

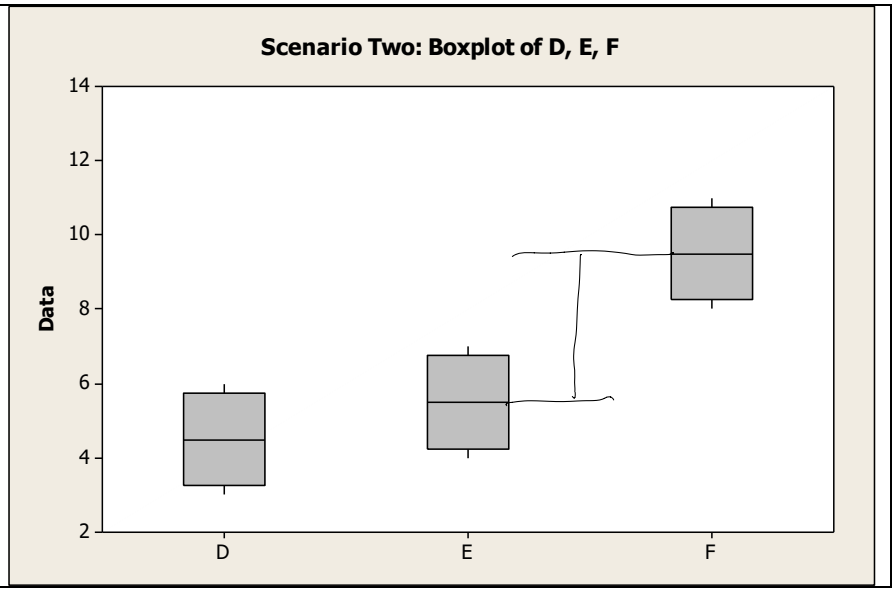
A	B	C
4	6	8
3	5	7
5	7	5
6	4	6



Scenario Two

- Larger variation among groups than within
  [This was done by adding 3 to each observation in Treatment C above]
- There is an extremely significant difference at α = 0.05

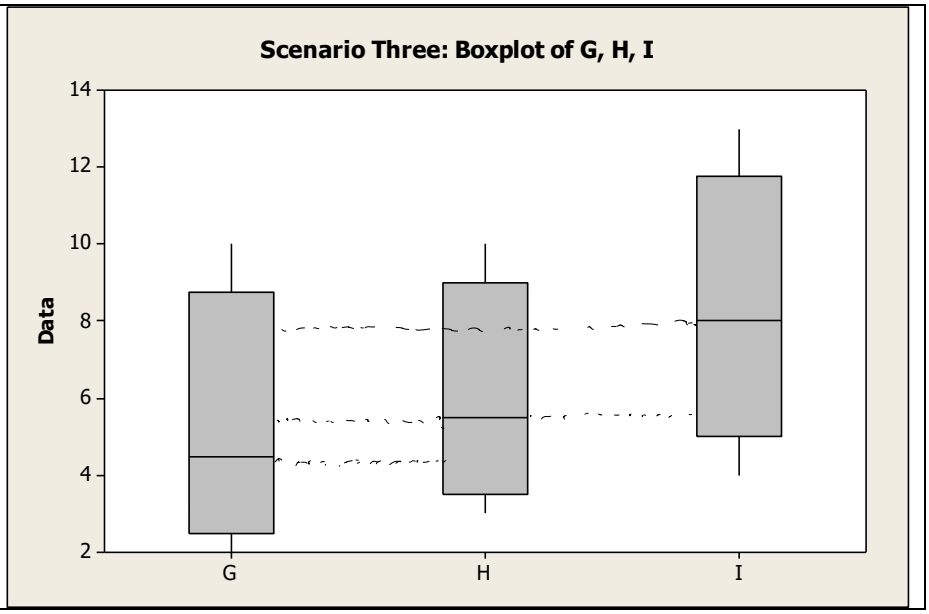
D	E	F
4	6	11
3	5	10
5	7	8
6	4	9



Scenario Three

- Large variation among groups (means), but even larger variation within groups
- No significant difference at α = 0.05

G	H	I
4	5	13
5	10	8
2	3	4
10	6	8



Computer Output: Scenario One

Summary Statistics

Groups	*Count*	*Sum*	*Average*	*Variance*
A	4	18	4.5	1.666667
B	4	22	5.5	1.666667
C	4	26	6.5	1.666667

ANOVA Table

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	8	2	4	2.4	0.146095	4.256495
Within Groups	15	9	1.666667			
Total	23	11				

Computer Output: Scenario Two

Summary Statistics

Groups	*Count*	*Sum*	*Average*	*Variance*
D	4	18	4.5	1.666667
E	4	22	5.5	1.666667
F	4	38	9.5	1.666667

ANOVA Table

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	56	2	28	16.8	0.000916	4.256495
Within Groups	15	9	1.666667			
Total	71	11				

Computer Output: Scenario Three

Summary Statistics

Groups	*Count*	*Sum*	*Average*	*Variance*
G	4	21	5.25	11.58333
H	4	24	6	8.666667
I	4	33	8.25	13.58333

ANOVA Table

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	19.5	2	9.75	0.864532	0.453485	4.256495
Within Groups	101.5	9	11.27778			
Total	121	11				

3.3 One-Way ANOVA Hypothesis Test

One-Way ANOVA Hypothesis Test

Purpose: To test for the difference between several (***k***) population means.

Assumptions:

1. Simple random samples from each population (implies independent sampling within populations)
2. Independent samples (All ***k*** samples are sampled independently of each other)
3. All populations being compared are normally distributed
4. Equal population standard deviations

Step 1: Check the purpose and assumptions

Step 2: State the null and alternative hypotheses:
H_0: $\mu_1 = \mu_2 = \ldots \mu_k$ (One-mean model)
H_a: Not all the means are equal. (***k***-mean model)

$\exists \mu_1, \mu_2 \in M$ s.t. $\mu_1 \neq \mu_2$

Step 3: Obtain the three sums of squares (SS_{Total}, SS_T and SS_E) and construct a **One-way ANOVA table** to obtain the calculated value of the F-statistic

$$SS_{Treatment} = \sum\sum(\bar{y}_j - \bar{\bar{y}})^2 = \sum n_j(\bar{y}_j - \bar{\bar{y}})^2$$

$$SS_{Error} = \sum\sum(y_{ij} - \bar{y}_j)^2$$

$$SS_{Total} = \sum\sum(y_{ij} - \bar{\bar{y}})^2$$

Nemerator

One-Way ANOVA Table

Source of variation	***SS***	***df***	***MS = SS/df***	***F*-statistic**
Treatment (Between groups)	$SS_{Treatment}$	$k - 1$	$MS_{Treatment} = \frac{SS_{Treatment}}{k-1}$	$F = \frac{MS_{Treatment}}{MS_{Error}}$
Error (Within groups)	SS_{Error}	$n - k$	$MS_{Error} = \frac{SS_{Error}}{n-k}$	
Total	SS_{Total}	$n - 1$		

$$F = \frac{SS_{Treatment}/(k-1)}{SS_{Error}/(n-k)} = \frac{MS_{Treatment}}{MS_{Error}}$$

denominator

Step 4: Decide to reject or not reject Ho
***df* =** (numerator degrees of freedom, denominator degrees of freedom)
$df = (k-1, n-k)$ or F_{n-k}^{k-1}
If the P-value ≤ α, we reject H_0 (otherwise do not reject H_0)

Never double the p-value

Step 5: Conclusion in terms of the research problem

Note: ANOVA is a **one-tailed test** and the ANOVA **table is one-tailed**.

Example of One-Way ANOVA: Experiment on Yield of Different Varieties of Sorghum
An experiment was conducted to compare the yield of three varieties of sorghum by planting them in plots in a completely randomized design in a uniform field, obtaining data as shown below. The data are normally distributed and the three samples have equal variances. At the 5% significance level, test whether there is a difference in the mean yield of the three varieties.

Variety A	Variety B	Variety C
5	6	10
8	5	8
7	7	11
6	8	10
	9	8

Step 1: Check purpose and assumptions

- Purpose: To compare ***k*** population means
- The three populations are normally distributed, with equal variance
- The three samples are random and independent

Step 2: H_0: $\mu_1 = \mu_2 = \mu_3$ (There is no difference in mean yield among the three varieties)
(One-mean model)
H_a: Not all the means are the same for the yield of the three varieties. (Three-mean model)

Step 3: Obtain the three sums of squares and construct a one-way ANOVA table to obtain the F-statistic

Quantity	Variety A	Variety B	Variety C	**Grand**
Total	26	35	47	108
Sample size	4	5	5	14
Mean	6.5	7	9.4	**7.7143**

mean of means

Calculate Treatment Sum of Squares (Measures variation between groups):

Quantity	Variety A	Variety B	Variety C	Totals
$(\bar{y}_j - \bar{\bar{y}})$	6.5 – 7.7143 = -1.2143	7 – 7.7143 = -0.7143	9.4 – 7.71428 = 1.6857	
$n_j(\bar{y}_j - \bar{\bar{y}})^2$	4 x (-1.214)^2 = 5.898	5 x (-0.7143)^2 = 2.551	5 x (1.6857)^2 = 14.208	$\sum n_j(\bar{y}_j - \bar{\bar{y}})^2 = 22.657$

$$SS_{Treatment} = \sum n_j(\bar{y}_j - \bar{\bar{y}})^2 = 22.657$$

Variation from the grand mean.

Calculate Error (or Residual) Sum of Squares (Measures variation within groups):

	Variety A	**Variety B**	**Variety C**	
	(5-6.5)^2 =2.25	(6-7)^2 =1	(10-9.4)^2 =0.36	
	(8-6.5)^2 =2.25	(5-7)^2 =4	(8-9.4)^2 =1.96	
	(7-6.5)^2 =0.25	(7-7)^2 =0	(11-9.4)^2 =2.56	
	(6-6.5)^2 =0.25	(8-7)^2 =1	(10-9.4)^2 =0.36	
		(9-7)^2 = 4	(8-9.4)^2 =1.96	
$\sum(y_{ij} - \bar{y}_j)^2$	5	10	7.2	$\sum\sum(y_{ij} - \bar{y}_j)^2 = 22.2$

$$SS_{Error} = \sum\sum(y_{ij} - \bar{y}_j)^2 = 22.2$$

Sum of the variations

$$SS_{Total} = \sum\sum(y_{ij} - \bar{\bar{y}})^2 = SS_{Treatment} + SS_{Error} = 22.657 + 22.2 = 44.857$$

[The Total Sum of Squares is not actually required in order to calculate the F-statistic.]

The value of the Total Sum of Squares can be verified as follows:

	Variety A	**Variety B**	**Variety C**	
	$(5-7.7143)^2$ =7.367	$(6-7.7143)^2$ =2.939	$(10-7.7143)^2$ = 5.224	
	$(8-7.7143)^2$ =0.082	$(5-7.7143)^2$ =7.367	$(8-7.7143)^2$ = 0.082	
	$(7-7.7143)^2$ =0.510	$(7-7.7143)^2$ =0.510	$(11-7.7143)^2$ =10.796	
	$(6-7.7143)^2$ =2.939	$(8-7.7143)^2$ =0.082	$(10-7.7143)^2$ = 5.224	
		$(9-7.7143)^2$ = 1.653	$(8-7.7143)^2$ = 0.082	
$\sum(y_{ij} - \bar{\bar{y}}_j)^2$	10.898	12.551	21.408	$\sum\sum(y_{ij} - \bar{\bar{y}}_j)^2$ =44.857

Excel Output

Summary Statistics

Groups	*Count*	*Sum*	*Average*	*Variance*
Var. A	4	26	6.5	1.666667
Var. B	5	35	7	2.5
Var. C	5	47	9.4	1.8

One-Way ANOVA Table

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	22.65714	2	11.32857	5.613256	0.020887	3.982298
Within Groups	22.2	11	2.018182			
Total	44.85714	13				

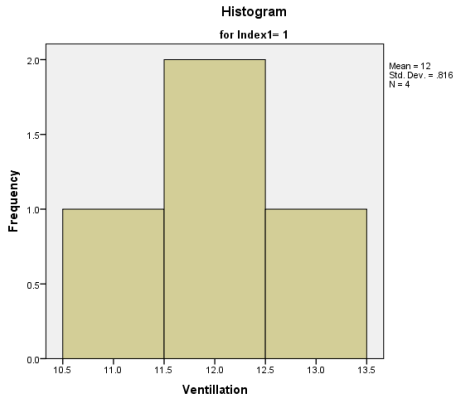
Step 4:
df = (k – 1, n - k) = (2, 11) From the F-table: 0.025 > P > 0.01. The exact P-value = 0.020887.
So, there is strong evidence against H_0.
Since P-value < α (0.05), reject H_0.

Step 5:
At the 5% significance level, the data provide sufficient evidence to conclude that there is a difference in the mean yield of the three varieties (that is, at least two means are different).

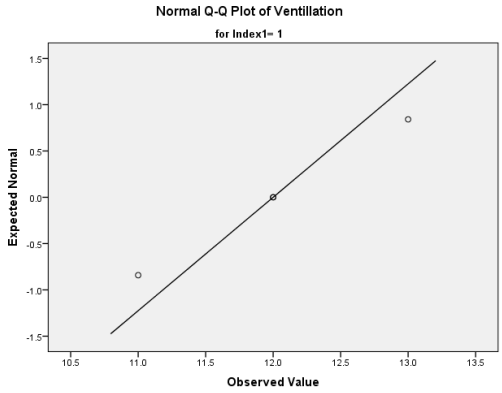
Example: Effect of Certain Diseases on Human Ventilation Rates
The normal resting ventilation rate is about 6 liters per minute (L/min) in healthy people, but is higher in people with a disease. The table below shows the ventilation rates of random samples of patients suffering from three different diseases. At the 1% significance level, determine whether there is a difference in the mean ventilation rates of people suffering from these three diseases.

Ventilation rate (L/min)		
Cancer	**Heart disease**	**Diabetes**
12	14	12
11	13	15
13	16	13
12	16	13
	13	14
	15	14
	15	

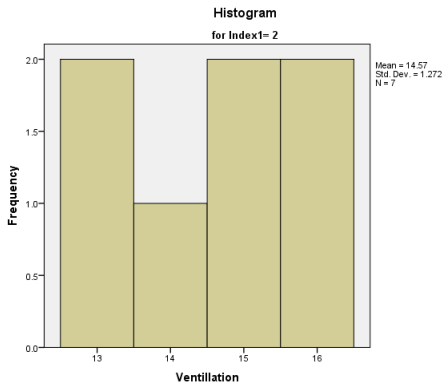
Checking Assumptions (SPSS Output)



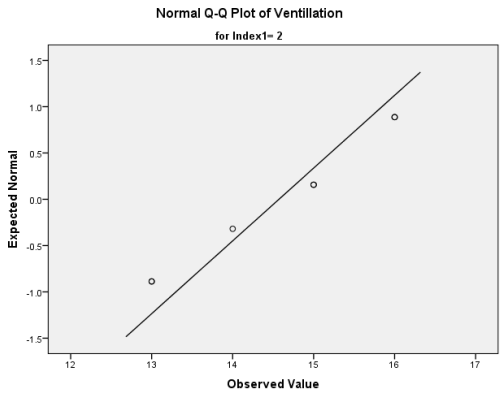
Histogram for Cancer



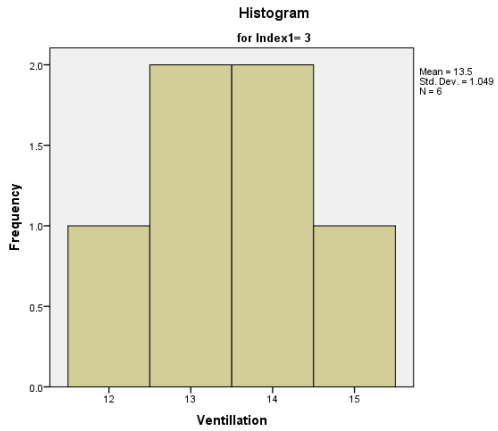
Normal Probability Plot for Cancer



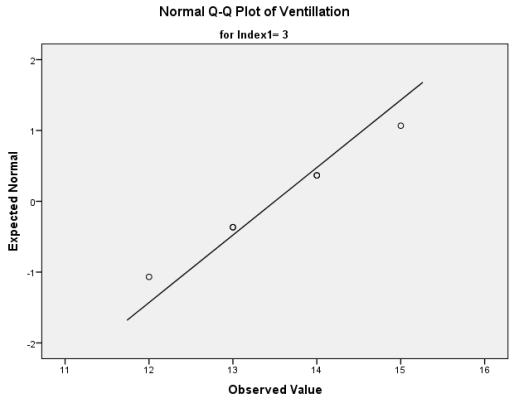
Histogram for Heart Disease



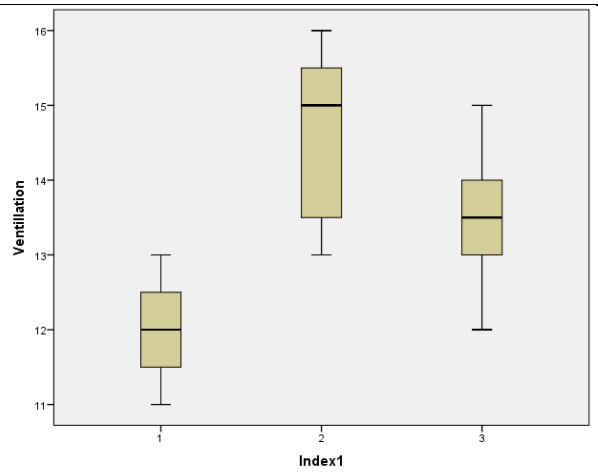
Normal Probability Plot for Heart Disease



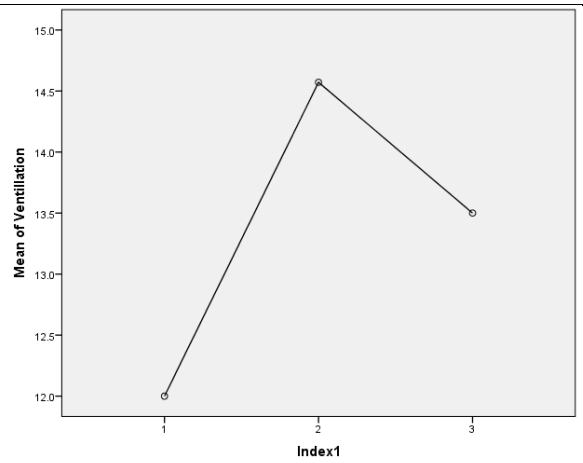
Histogram for Diabetes



Normal Probability Plot for Diabetes



Side-by-side Boxplots for Cancer, Heart Disease and Diabetes



Means Plot comparing mean ventilation rates for patients with cancer, heart disease and diabetes

Test of Homogeneity of Variances

Ventilation

Levene Statistic	df1	df2	Sig.
1.351	2	14	.291

SPSS Output

Descriptives

Ventilation

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
1	4	12.00	.816	.408	10.70	13.30	11	13
2	7	14.57	1.272	.481	13.39	15.75	13	16
3	6	13.50	1.049	.428	12.40	14.60	12	15
Total	17	13.59	1.460	.354	12.84	14.34	11	16

$SS_{Treatment}$

ANOVA

Ventilation

Source of variation	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	16.903	2	8.452	6.874	.008
Within Groups	17.214	14	1.230		
Total	34.118	16			

SS_{Error}

Suppose that only partial ANOVA output is given, so the numbers highlighted in yellow are not given.

>>>>>>>>>>

$H_0: \mu_1 = \mu_2 = \mu_3$ (no diff between means)

$H_A: \mu_1, \mu_2, \mu_3$ (at least 1 diff between means)

k = # of pop being compared

n = total num of observations $\sum_{i=1}^{k} n_i$

$$F = \frac{SS_T/(k-1)}{SS_E/(n-k)} = \frac{MS_T}{MS_E} = \frac{16.903/(3-1)}{17.214/(17-3)} = \frac{8.452}{1.230} = 6.874$$

$df = (2, 14) \Rightarrow 0.005 < p\text{-value} < 0.01$

$p\text{-value} < \alpha$ of 0.05

Very strong evidence against H_0

$P < \alpha = 0.01$ $\therefore$ we reject H_0

for medical experiment

>>>>>>>>>>

Experiment to test the ultimate strength of stainless steel, steel alloy and titanium alloy
An experiment was conducted to test the ultimate strength (in MPa's) of random samples of stainless steel, steel alloy and titanium alloy. Below is incomplete output of one-way ANOVA obtained from SPSS.

Summary Statistics				
Groups	*Count*	*Sum*	*Average*	*Variance*
Stainless Steel	5	4320	864	1930
Steel alloy	7	5740	820	1633.333
Titanium alloy	7	6240	891.4286	1347.619

Calculate the F-statistic by filling in missing values.
>>>>>>>>>>

ANOVA

Ultimate strength

Source of variation	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	18110.08	2	9055.0400	5.66	0.014
Within Groups	25605.17	16	1600.3569		
Total	43715.79				

dfC(2,16) = P < 0.001

>>>>>>>>>>
At the 5% significance level, what conclusion can you draw regarding the ultimate strength of the three materials?

(a) There is no significant difference in ultimate strength of the three materials.
(b) Ultimate strength of titanium alloy is greater than that of steel alloy, but is not greater than that of stainless steel.
(c) Ultimate strength of Titanium alloy is greater than that of both steel alloy and stainless steel.
(d) All the means for ultimate strength of the three materials are different.
(e) At least two of the means for ultimate strength of the three materials are different.

Answer: e

Filling in Missing Values in an ANOVA Table
>>>>>>>>>>

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	12409.96		111.4	5.546	0.001 < P < 0.005
Within Groups	461.9	23	20.08		
Total	12871.86	28			

>>>>>>>>>>

3.4 Multiple Comparisons (= Unplanned comparisons)

- If, and only if, one-way ANOVA results in rejecting the null hypothesis, then it is often desirable to do multiple comparisons in order to determine which means are different the which other means
- Known as pairwise comparisons
- The number of pairwise comparisons that are possible for a given question is given by:
  k(k-1)/2, where k = number of means (groups) being compared

- There are several types of multiple comparisons, including:
  1. Tukey multiple-comparisons (also called HSD = honest significant difference)
     - Requires that the sample sizes for all groups be the same (or very similar)
     - When sample sizes are equal, the CIs are shorter than other methods and therefore more likely to show differences
  2. Bonferroni method
     - Can be used for a general case where sample sizes are different
     - Can control the overall error rate
  3. Fisher method
  4. Scheffe method - results in wider CIs than Tukey's test and therefore more conservative
  5. Least significant difference (LSD) - not suitable if the number of groups being compared is large
  6. Student-Newman-Keuls (SNK) test

3.4.1 Tukey Multiple Comparisons

Tukey Multiple Comparisons

Purpose: To determine pairwise differences between ***k*** population means when the null hypothesis has been rejected in one-way ANOVA.

Assumptions (same as for One-way ANOVA):

Step 1: At the given confidence level, 1 – α, find the critical value q_α at
df = (k, n – k) in the appropriate statistical table.

Step 2: Obtain the endpoints of the confidence interval for the difference, $\mu_i - \mu_j$

$$(\bar{y}_i - \bar{y}_j) \pm \frac{q_\alpha}{\sqrt{2}} \times \sqrt{MSE}\sqrt{(1/n_i) + (1/n_j)}$$

Where, ***MSE*** = Error mean square from one-way ANOVA table

Do so for all possible pairs of means with ***i < j*** and summarize the confidence intervals in a table.

[**Note:** There will be k(k – 1)/2 pairwise differences.]

Step 3: Compile the results in a matrix and declare two population means different if the confidence interval for the difference does not contain 0; otherwise, do not declare the two population means different.

Step 4: Conclusion
Summarize the results in a **means comparisons diagram** by ranking the sample means from smallest to largest and by connecting with lines those whose population means were not declared different.

And: Interpret the results of the multiple comparisons **in words**

Example for Tukey Multiple comparisons: Experiment on Yield of Different Varieties of Sorghum
An experiment was conducted to compare the yield of three varieties of sorghum by planting them in plots in a randomized design in a uniform field. One-way ANOVA resulted in rejecting the null hypothesis and thus drawing the conclusion that not all means are equal (or at least two means are different). Perform Tukey multiple comparisons to determine which pairs of means are different at the 95% confidence level.

Information already known based on ANOVA is as follows:

Groups	Mean	Sample size
Variety A	6.5	4
Variety B	7	5
Variety C	9.4	5

Error mean square (MSE) = 2.0182

At ***df = (k, n – k)*** = (3, 11) and **α** = 0.05, the critical value q_{α} = 3.82
>>>>>>>>>>

$q_{\alpha} = 3.82$

$m = \binom{3}{2} = 3 = \frac{3(3-1)}{2}$

$(\bar{Y}_1 - \bar{Y}_2) \pm \frac{q_{\alpha}}{\sqrt{2}} \times \sqrt{MS_E} \sqrt{\left(\frac{1}{n_i}\right) + \left(\frac{1}{n_j}\right)}$

$(-0.5) \pm \frac{3.82}{\sqrt{2}} \times \sqrt{2.0182} \sqrt{\frac{1}{4} + \frac{1}{5}}$

Calculate this first

$(-3.07, 2.07)$

mat	Var A	Var B	Var C
Var A	—	—	—
Var B		—	—
Var C			—

Var A 6.5 — Var B 7 — Var C 9.4

>>>>>>>>>>

3.4.2 Bonferroni's Method of Multiple Comparisons

Bonferroni's Method of Multiple Comparisons

Purpose: To determine pairwise differences between ***k*** population means when the null hypothesis has been rejected in one-way ANOVA.

Step 1: Find the number of multiple comparisons (m) that are possible:

$$m = \frac{k(k-1)}{2}$$, where k = number of means (groups) being compared

Step 2: Calculate the individual comparison-wise error rate (α_I) based on the family-wise (experiment-wise) error rate (α_F) or confidence level (1 – α_F) given:

$$\alpha_I = \frac{\alpha_F}{m}$$

Step 3: Find the Critical value of ***t*** at ***df = n – k*** for α_I/2: $t_{df,\alpha_I/2}$

Step 4: Calculate the margin of error (ME) for each comparison (group ***i*** vs. Group ***j***):

$$ME_{ij} = Crit.value \times S.E.(\bar{y}_i - \bar{y}_j)$$

$$ME_{ij} = t_{n-k,\alpha_I/2} \times \sqrt{MSE}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

Step 5: Declare two population means different if the absolute value of the difference between their sample means is greater than or equal to the corresponding margin of error

$$\mu_i - \mu_j \neq 0$$ **, if** $$\left|\bar{y}_i - \bar{y}_j\right| \geq ME_{ij}$$

Present the results in a matrix

Step 6: Summarize the results in a **means comparisons diagram** by ranking the sample means from smallest to largest and by connecting with lines those whose population means were not declared different and **state the conclusion in words**.

ANOVA: Effect of Certain Diseases on Human Ventilation Rates

It was concluded with very strong evidence that there is a difference in the mean ventilation rates of people suffering from the three diseases tested (cancer, heart disease and diabetes) [One-way ANOVA: F = 6.87, df = (2, 14), P-value = 0.008324]. Perform Bonferroni's method of multiple comparisons to determine which pairs of means are different at the 95% confidence level.

Disease	Cancer	Heart disease	Diabetes
Sample mean	12	14.57	13.5
Sample size	4	7	6

ANOVA Table						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	16.90336	2	8.451681	6.873566	0.008325	6.514884
Within Groups	17.21429	14	1.229592			
Total	34.11765	16				

>>>>>>>>>>

$m = 3$

ind. error rate $= \alpha_I = \frac{\alpha_F}{m} = \frac{0.05}{3} = 0.0167$

Crit. Val $= t_{\alpha/2,14} = t_{0.0083,14} = t_{0.005,14} = 2.977$

↑ always go for the largest crit. val.

$ME_{ij} = t_{\alpha_I} \times \sqrt{MS_E} \times \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$

$= 2.977 \sqrt{\frac{17.21429}{14}} \times \sqrt{\frac{1}{4} + \frac{1}{7}}$

$M_{12} = 2.129$

$M_{23} = 1.836$

	Cancer	HD	Diabetes
Can	—	—	—
HD	$7.5 > 2.064$	—	—
Diabetes	$1.5 < 2.131$	$1.07 < 1.86$	—

Cancer 12 — Diabetes 13.5 — HD

Conclusion is basically the same.

>>>>>>>>>>

3.5 Linear Combinations (Contrasts) (=Planned Comparisons)

- The multiple comparisons discussed above are sometimes called "unplanned comparisons"
- Linear combinations, on the other hand, are planned comparisons
- Ideally, the means to be compared should be planned before collecting the data

Linear Combinations (Contrasts)

Step 1: Develop the linear combination by deciding which means or groups of means you want to compare.

$$\gamma_{D-E} = \frac{(\mu_{1,1} + \mu_{1,2} + ... + + \mu_{1,d})}{d} - \frac{(\mu_{2,1} + \mu_{2,2} + ... + + \mu_{2,e})}{e}$$

Where ***D*** and ***E*** are combinations of means to be compared and ***d*** and ***e*** are the number of means within those combinations, respectively

Then, define the parameter of the contrast, which will take the following general form (where γ is the Greek letter "gamma"):

$$\gamma = C_1\mu_1 + C_2\mu_2 + ... + C_k\mu_k$$

Check to be sure that the coefficients add up to 0 (This makes it a contrast)

$$\sum_{i-1}^{k} C_i = C_1 + C_2 + ... + C_k = 0$$

Step 2: State the hypothesis

Null hypothesis is $H_0 : \gamma = 0$

Alternative hypothesis may be:

$$H_a : \gamma \neq 0 \quad \text{or} \quad H_a : \gamma < 0 \quad \text{or} \quad H_a : \gamma > 0$$

Step 3: Calculate the estimate (sample contrast), standard error of the estimate and the t-statistic

Estimate: $\hat{\gamma} = C_1\overline{y}_1 + C_2\overline{y}_2 + ... + C_k\overline{y}_k$

$$SE(\hat{\gamma}) = s_p\sqrt{\frac{C_1^2}{n_1} + \frac{C_2^2}{n_2} + ... + \frac{C_k^2}{n_k}}$$

Where s_p = pooled (common) standard deviation, and

$$s_p = \sqrt{MSE} = \sqrt{\frac{(n_1 - 1)s_1^2 + ... + (n_k - 1)s_k^2}{n - k}}$$

$$t = \frac{\hat{\gamma} - 0}{SE(\hat{\gamma})}$$

Step 4: Decide to reject or not reject Ho by comparing the P-value at $df = n - k$ with the significance level (α) and state the strength of the evidence against Ho.

Step 5: Write the conclusion in words in terms of the research problem.

Confidence Interval for a Linear Contrast

Confidence Interval for a Linear Contrast

Step 1: For a given confidence level (1 – α), find the Critical Value ($t_{\alpha/2}$) at **df = n - k**

Step 2: Calculate (or state):

Parameter: $\gamma = C_1\mu_1 + C_2\mu_2 + \ldots + C_k\mu_k$

Estimate: $\hat{\gamma} = C_1\bar{y}_1 + C_2\bar{y}_2 + \ldots + C_k\bar{y}_k$

Standard error of the estimate:

$$SE(\hat{\gamma}) = s_p\sqrt{\frac{C_1^2}{n_1} + \frac{C_2^2}{n_2} + \ldots + \frac{C_k^2}{n_k}}$$

Where $s_p = \sqrt{MS_E} = \sqrt{\frac{(n_1-1)s_1^2 + \ldots + (n_k-1)s_k^2}{n-k}}$

Endpoints of the confidence interval:

$$\hat{\gamma} \pm Crit.Value \times SE(\hat{\gamma})$$

from the t-table

Step 3: Interpret the confidence interval in terms of the research problem

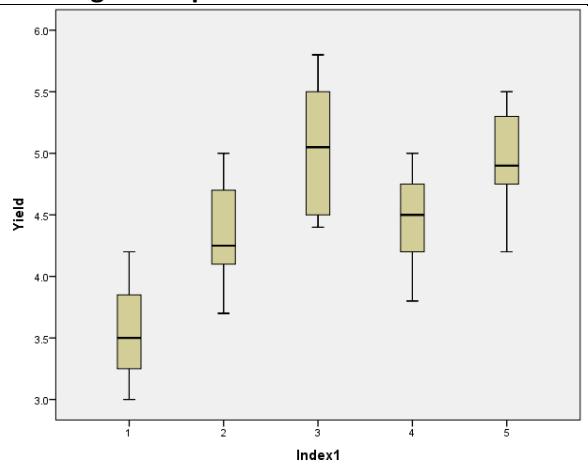
Applications of Linear Contrasts in a Rice Experiment

Azolla-Anabaena (below) is an endophytic association (symbiosis) between a water fern and a blue-green alga. *Utricularia*-Cyanophyta is an epiphytic association. Both have been used as biofertilizers ("living fertilizers") to increase rice crop yield, since both associations have been shown to fix nitrogen.

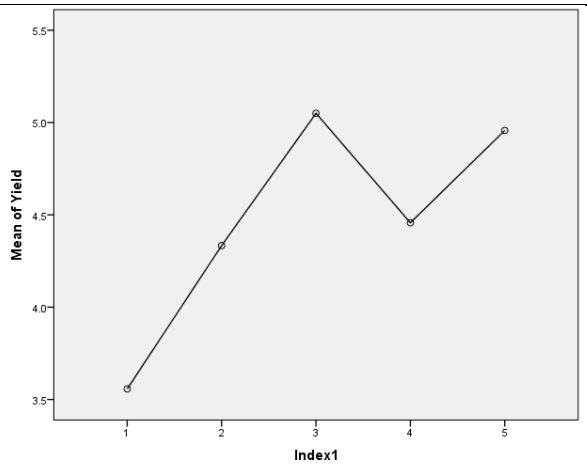
An experiment was conducted to test the effect of these biofertilizers on rice crop yield as compared to two different levels of chemical nitrogen fertilizer and a control, obtaining raw data as shown below.

Yield of rice (t ha^{-1})				
Control	***Utricularia***	***Azolla***	**N-Level 1**	**N-Level 2**
3	4.3	4.8	4.1	4.2
3.5	4.1	5.3	4.5	4.6
3.4	5	4.4	5	4.9
3.1	4.7	5.5	3.8	5.4
3.6	3.7	5.8	4.3	5.5
4.1	4.2	4.5	4.8	4.9
4.2			4.7	5.2

Checking Assumptions



Side-by-side boxplots: Standard deviations look quite similar



Means plots: Shows the relative yield of rice under the different treatments

Test of Homogeneity of Variances

Yield

Levene Statistic	df1	df2	Sig.
.410	4	28	.800

- P-value = 0.800, which is very large, so do not reject the null hypothesis. Therefore, there is insufficient evidence to conclude that there is a difference in the variances of the 5 treatments.

One-way ANOVA resulted in rejecting the null hypothesis with extremely strong evidence, as shown in the ANOVA output below.

Descriptives

Yield

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean	
					Lower Bound	Upper Bound
1 (Control)	7	3.557	.4577	.1730	3.134	3.980
2 (Utricularia)	6	4.333	.4590	.1874	3.852	4.815
3 (Azolla)	6	5.050	.5683	.2320	4.454	5.646
4 (N1)	7	4.457	.4198	.1587	4.069	4.845
5 (N2)	7	4.957	.4577	.1730	4.534	5.380
Total	33	4.458	.7040	.1226	4.208	4.707

One-Way ANOVA Hypothesis Test

ANOVA

Yield

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	9.621	4	2.405	10.793	.000
Within Groups	6.240	28	.223		
Total	15.861	32			

The exact P-value is 2.02×10^{-5}

Research Question:
At the 5% significance level, test for differences in effectiveness between the following using linear contrasts (as two-tailed tests):

1. Control versus the biofertilizers
2. The biofertilizers versus the chemical nitrogen fertilizers
3. Level 1 of both types of fertilizers (use the epiphytic association as Level 1 of the biofertilizers) versus Level 2 of both types of fertilizers (use the endophytic association as Level 2 of the biofertilizers)

Linear Contrast 1: Control (C) versus the Biofertilizers (B)

>>>>>>>>>>

Step 1: $Y = c_1\mu_1 + \cdots + c_k\mu_k$

$$Y_{C-B} = \frac{\mu_C}{1} - \frac{(\mu_V + \mu_A)}{2}$$

Parameter $Y_{C-B} = \mu_C - \frac{1}{2}\mu_V - \frac{1}{2}\mu_A = 0$

Step 2: $H_0: Y = 0 \quad H_A: Y \neq 0$

Step 3: $\hat{Y}_{C-B} = \bar{Y}_C - \frac{1}{2}\bar{Y}_V - \frac{1}{2}\bar{Y}_A$

$= 3.557 - \frac{1}{2}(4.333) - \frac{1}{2}(5.050)$

$= -1.1345$

$SE = S_p = \sqrt{MS_E} = \sqrt{0.22285} = 0.4721$

$SE(\hat{Y}) = S_p\sqrt{\frac{c_1^2}{n_1} + \cdots + \frac{c_k^2}{n_k}}$

$= (0.4721)\sqrt{\frac{(1)^2}{7} + \frac{(-\frac{1}{2})^2}{6} + \frac{(-\frac{1}{2})^2}{6}}$

$= (0.4721)(0.4756)$

$SE(\hat{Y}) = 0.2245$

$$t = \frac{\hat{Y} - 0}{SE(\hat{Y})} = \frac{-1.1345}{0.2245} = -5.033$$

Linear Contrast 2: The biofertilizers (B) versus the chemical nitrogen (N) fertilizers

$$\gamma_{B-N} = \frac{C_1(\mu_U + \mu_A)}{2} - \frac{C_2(\mu_{N_1} + \mu_{N_2})}{2}$$

$$\gamma_{B-N} = \frac{1}{2}\mu_U + \frac{1}{2}\mu_A - \frac{1}{2}\mu_{N_1} - \frac{1}{2}\mu_{N_2}$$

$H_0: \gamma = 0 \quad H_A: \gamma \neq 0$

Estimate $\bar{\gamma}_{B-N} = \frac{1}{2}\bar{Y}_U + \frac{1}{2}\bar{Y}_A - \frac{1}{2}\bar{Y}_{N_1} - \frac{1}{2}\bar{Y}_{N_2}$

$= \frac{1}{2}(4.333) + \frac{1}{2}(5.050) - \frac{1}{2}(4.457) - \frac{1}{2}(4.957)$

$= -0.0155$

$$SE(\hat{\gamma}) = (0.4721)\sqrt{\frac{(\frac{1}{2})^2}{6} + \frac{(\frac{1}{2})^2}{6} + \frac{(\frac{-1}{2})^2}{7} + \frac{(\frac{-1}{2})^2}{7}}$$

$$\sqrt{0.1547}$$

$$0.4721 \times 0.3933 = 0.1857 = \frac{-0.0155}{0.1857}$$

$= -0.083$

Linear Contrast 3: Level 1 (L1) of the both types of fertilizers versus Level 2 (L2) of both types of fertilizers

>>>>>>>>>>

$$H_0 : \gamma = 0 \qquad \text{versus } H_a : \gamma \neq 0$$

Estimate

$$\hat{\gamma}_{L1-L2} = \frac{1}{2}\bar{y}_U + \frac{1}{2}\bar{y}_{N1} - \frac{1}{2}\bar{y}_A - \frac{1}{2}\bar{y}_{N2}$$

$$= \frac{1}{2}(4.33) + \frac{1}{2}(4.46) - \frac{1}{2}(5.05) - \frac{1}{2}(4.96) = \frac{-1.22}{2} = -0.61$$

Standard error of the estimate

$$SE(\hat{\gamma}) = 0.4721\sqrt{\frac{(1/2)^2}{6} + \frac{(1/2)^2}{7} + \frac{(-1/2)^2}{6} + \frac{(-1/2)^2}{7}}$$

$$= (0.4721)(0.3934) = 0.1857$$

Observed value of the t-statistic:

$$t = \frac{\hat{\gamma} - 0}{SE(\hat{\gamma})} = \frac{-0.61 - 0}{0.1857} = -3.285$$

Combining the results of all 3 linear contrasts

Step 4: Decide to reject or not reject Ho:

$$df = n - k = 33 - 5 = 28 \qquad \alpha = 0.05$$

Linear Contrast	t-statistic	P-value	Decision	Strength of evidence
Control vs. Biofertilizers	─ 5.053	P < 0.001	Reject Ho	Extremely strong
Biofertilizers vs. N Fert.	─ 0.083	P > 0.50	Not reject Ho	Weak
Level 1 vs. Level 2	─ 3.276	0.005 < P < 0.002	Reject Ho	Very strong

Step 5: There was extremely strong evidence that the biofertilizers (both combined) resulted in a difference in (greater) crop yield in comparison with the control. There was no difference in crop yield between the biofertilizers (both combined) and the nitrogen fertilizers (both levels combined). There was very strong evidence that Level 2 (combining both the biofertilizer and the nitrogen fertilizer) resulted in a difference in (greater) crop yield than Level 1.

Calculate a 95% confidence intervals for these Linear Contrasts

At $df = n - k = 33 - 5 = 28$, $t_{\alpha/2} = t_{0.05/2} = 2.048$

$\hat{\gamma}$ ± Critical Value of t x $SE(\hat{\gamma})$

Linear Contrast	Estimate	SE(Estimate)	Endpoints	Include 0
Control vs. Biofertilizers	─ 1.1345	0.2245	(─ 1.59, ─ 0.67)	No
Biofertilizers vs. N Fert.	─ 0.0155	0.1857	(─ 0.40, 0.36)	Yes
Level 1 vs. Level 2	─ 0.6085	0.1857	(─ 0.99, ─ 0.23)	No

Research Conclusion: Biofertilizers *Azolla-Anabaena* and *Utricularia*-Cyanophyta can be applied on rice to increase crop yield with effects comparable to the application of chemical nitrogen fertilizers. At the same time, these biofertilizers save costs and are an "environmentally-friendly" alternative. Also, the biofertilizers help to control weeds.

Post Hoc Tests (Tukey's and Bonferroni's Methods) on Rice Experiment (Using SPSS Output)

Since the one-way ANOVA table above shows that there is a difference in the mean yield of rice between the 5 treatments, it is appropriate to perform multiple comparisons tests and linear contrasts to determine which means are different.

Multiple Comparisons

Dependent Variable: Yield

	(I) Index1	(J) Index1	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Tukey HSD	C	U	-.7762*	.2626	.046	-1.541	-.011
		A	-1.4929*	.2626	.000	-2.258	-.728
		N1	-.9000*	.2523	.011	-1.635	-.165
		N2	-1.4000*	.2523	.000	-2.135	-.665
	U	C	.7762*	.2626	.046	.011	1.541
		A	-.7167	.2725	.092	-1.511	.077
		N1	-.1238	.2626	.989	-.889	.641
		N2	-.6238	.2626	.152	-1.389	.141
	A	C	1.4929*	.2626	.000	.728	2.258
		A	.7167	.2725	.092	-.077	1.511
		N1	.5929	.2626	.189	-.172	1.358
		N2	.0929	.2626	.996	-.672	.858
	N1	C	.9000*	.2523	.011	.165	1.635
		U	.1238	.2626	.989	-.641	.889
		A	-.5929	.2626	.189	-1.358	.172
		N2	-.5000	.2523	.301	-1.235	.235
	N2	C	1.4000*	.2523	.000	.665	2.135
		U	.6238	.2626	.152	-.141	1.389
		A	-.0929	.2626	.996	-.858	.672
		N1	.5000	.2523	.301	-.235	1.235
Bonferroni	C	U	-.7762	.2626	.063	-1.576	.024
		A	-1.4929*	.2626	.000	-2.293	-.693
		N1	-.9000*	.2523	.013	-1.669	-.131
		N2	-1.4000*	.2523	.000	-2.169	-.631
	U	C	.7762	.2626	.063	-.024	1.576
		A	-.7167	.2725	.137	-1.547	.114
		N1	-.1238	.2626	1.000	-.924	.676
		N2	-.6238	.2626	.246	-1.424	.176

	A	C	1.4929*	.2626	.000	.693	2.293
		U	.7167	.2725	.137	-.114	1.547
		N1	.5929	.2626	.320	-.207	1.393
		N2	.0929	.2626	1.000	-.707	.893
	N1	C	.9000*	.2523	.013	.131	1.669
		U	.1238	.2626	1.000	-.676	.924
		A	-.5929	.2626	.320	-1.393	.207
		N2	-.5000	.2523	.574	-1.269	.269
	N2	C	1.4000*	.2523	.000	.631	2.169
		U	.6238	.2626	.246	-.176	1.424
		A	-.0929	.2626	1.000	-.893	.707
		N1	.5000	.2523	.574	-.269	1.269

*. The mean difference is significant at the 0.05 level.

>>>>>>>>>>

Means tukey comp. diagram

Control	U	N_1	N_2	A
3.557	4.333	4.457	4.957	5.050

We can be 95% confident that control is different from others, but no other groups in others are different for each other.

Means Bonferroni comp. diagram

Control	U	N_1	N_2	A
3.557	4.333	4.457	4.957	5.050

We can be 95% confident that the control is different from N1, N2 and A but not different from U. The others are not different from each other.

>>>>>>>>>>

Linear Contrasts from SPSS Output (Previously done by hand calculations)

Contrast Coefficients

Contrast	Index1				
	1 (Control)	2 (Utricularia)	3 (Azolla)	4 (N1)	5 (N2)
1	1	-.5	-.5	0	0
2	0	.5	.5	-.5	-.5
3	0	.5	-.5	.5	-.5

try to use proportional coefficients with int coefficients

Contrast Tests

		Contrast	Value of Contrast	Std. Error	t	df	Sig. (2-tailed)
Yield	Assume equal variances	1	-1.135	.2245	-5.053	28	.000
		2	-.015	.1857	-.083	28	.934
		3	-.608	.1857	-3.276	28	.003
	Does not assume equal variances	1	-1.135	.2284	-4.967	13.543	.000
		2	-.015	.1898	-.082	19.193	.936
		3	-.608	.1898	-3.206	19.193	.005

Perform Contrasts at α = 0.05 (Use ONLY Output for Equal Variances)

Linear Contrast 1: Control (C) versus the Biofertilizers (B)

$$\gamma_{C-B} = \mu_C - \frac{1}{2}\mu_U - \frac{1}{2}\mu_A$$

Estimate $\hat{\gamma}_{C-B}$ = -1.135 Standard error of the estimate $SE(\hat{\gamma})$ = 0.2245

t = -5.053 ➔ P = 0.000 ➔ Reject Ho ➔ With extremely strong evidence

Linear Contrast 2: The biofertilizers (B) versus the chemical nitrogen (N) fertilizers

$$\gamma_{B-N} = \frac{1}{2}\mu_U + \frac{1}{2}\mu_A - \frac{1}{2}\mu_{N1} - \frac{1}{2}\mu_{N2}$$

Estimate $\hat{\gamma}_{B-N}$ = -0.015 Standard error of the estimate $SE(\hat{\gamma})$ = 0.1857

t = -0.083 ➔ P = 0.934 ➔ Do not reject Ho ➔ Weak evidence

Contrast 3: Level 1 (L1) of the both types of fertilizers vs. Level 2 (L2) of both types of fertilizers

$$\gamma_{L1-L2} = \frac{1}{2}\mu_U + \frac{1}{2}\mu_{N1} - \frac{1}{2}\mu_A - \frac{1}{2}\mu_{N2}$$

Estimate $\hat{\gamma}_{L1-L2}$ = -0.608 Standard error of the estimate $SE(\hat{\gamma})$ = 0.1857

t = -3.276 ➔ P = 0.003 ➔ Reject Ho ➔ With very strong evidence

Comparison of Tukey's Multiple Comparisons, Bonferroni Multiple Comparisons and Linear Contrasts

1. In this study, sample sizes were nearly equal for all treatments, making Tukey's test suitable. Tukey's test showed that it is slightly more powerful than the Bonferroni Method since it showed a difference between Control and Utricularia whereas Bonferroni did not. This is mainly because the Bonferroni Method reduces individual comparison-wise error rate and makes the confidence intervals wider and less precise.

2. Linear Contrasts were more effective (and powerful) than the multiple comparisons tests in detecting differences between groups. Thus, these planned comparisons are very useful.

3.6 Reduced Models and the Extra Sum-of-Squares F-test in Single-Factor ANOVA

- Classifies two models: a reduced model and a full model
  - Null hypothesis is the reduced model, which is a special case of the full model obtained by imposing some restrictions
  - Alternative hypothesis is the full model, which is a general model that is found to adequately describe the data

Extra-Sum-of-Squares F-test

- Also called Partial F-test or Nested F-test

Extra-Sum-of-Squares F-Test

Null and alternative hypotheses:
H_0: Reduced model
H_a: Full model

Calculations for Extra-Sum-of Squares F-test:

Extra Sum of Squares = SS_E(reduced) - SS_E(full)

Extra df = df_E (reduced) – df_E (full)

$$F = \frac{(Extra\ SS)/(Extra\ df)}{SS_E(\text{Full})/df_E(Full)} = \frac{[SS_E(reduced) - SS_E(full)]/[df_E(reduced) - df_E(full)]}{SS_E(full)/df_E(full)}$$

Examine the distribution of the F-table at:

$$df = [Extra\ df, df_E(Full)] = [Extra\ df, n-k]$$

Recall that, residual (error) = observed value – estimated value
Therefore, residual sum of squares or error sum of squares is:

SS_E = $\sum$ (observed value – estimated value)^2 $= \sum (y_i - \bar{y})^2$

Research problem: An educational researcher conducted a study to determine a possible effect of the initial interest of students (Low, Medium, Super) in a statistics course (as expressed at the start of the course) on their final grades. The study was based on a random sample of 72 students (24 in each interest group). For each level of interest, there were equal numbers of females and males.

(a) At the 5% significance level, perform the most appropriate test (showing all steps) to determine whether there is a difference in mean grades between students having different levels of interest, as expressed at the start of the course (that is, determine whether at least two means are different). For this test, use only the SPSS output shown in this part (a), that is Tables 1 and 2 (with missing values).

Table 1: Summary statistics of grades for the 3 treatment groups for level of interest.

Descriptives

Grade

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean	
					Lower Bound	Upper Bound
Low	24	68.1250	6.46269	1.31919	65.3960	70.8540
Medium	24	73.8750	6.34043	1.29424	71.1977	76.5523
Super	24	78.0000	7.10786	1.45089	74.9986	81.0014
Total	72	73.3333	7.71682	.90944	71.5200	75.1467

Table 2: ANOVA table for the comparison of grades for 3 treatment groups with respect to level of interest (ignoring gender).

ANOVA

Grade

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1180.750	2	590.375	13.368	.000012
Within Groups	3047.250	69	44.163		
Total	4228.000	71			

Suppose the numbers highlighted in yellow in the table above were not given

H_0: $\mu_1 = \mu_2 = \mu_3$ (One-mean model)
There is no difference in the mean grades of students having different levels of interest.
H_a: μ_1, μ_2, μ_3 (Three-mean model)
Not all the mean grades of students having different levels of interest are equal. (Or, there is a difference in the mean grades between at least groups.)

k = number of populations being compare = 3
n = total sample size = 72

$$F = \frac{SSTR/k-1}{SSE/n-k} = \frac{MSTR}{MSE} = \frac{590.375}{3047.250/72-3} = \frac{590.375}{44.163} = 13.37$$

For df = (k – 1, n – k) = (2, 69) P < 0.001 There is extremely strong evidence against Ho.
Since P-value < α (0.05), reject Ho.

At the 1% significance level, the data provide sufficient evidence to conclude that there is a difference in the mean grades between students having different levels of interest, as expressed at the start of the course (that is, at least two means are different).

(b) The researcher then realized that he had been ignoring the possible effect of gender in the experiment. He did further analysis of the data and came up with the SPSS output in Tables 3 – 6 below. At the 5% significance level, perform the most appropriate test, showing all steps, to determine whether there is a difference in the mean grades between students having different levels of interest after accounting for the effect of gender. For this test, you may consider using any of the SPSS output shown below (Tables 3 – 6) or shown in part (a) (Tables 1 – 2).

Table 3: Summary statistics of grades for the 2 gender groups.

Descriptives

Grade

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean	
					Lower Bound	Upper Bound
Female	36	71.7500	7.32657	1.22109	69.2710	74.2290
Male	36	74.9167	7.87174	1.31196	72.2533	77.5801
Total	72	73.3333	7.71682	.90944	71.5200	75.1467

Table 4: ANOVA table for the comparison of grades for the two gender groups (ignoring level of interest).

ANOVA

Grade

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	180.500	1	180.500	3.122	.082
Within Groups	4047.500	70	57.821		
Total	4228.000	71			

Table 5: Summary statistics of grades for 6 groups (for all the combinations of levels of interest and gender).

Descriptives

Grade

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean	
					Lower Bound	Upper Bound
Low-Female	12	65.4167	5.93079	1.71207	61.6484	69.1849
Medium-Female	12	75.5833	6.11196	1.76437	71.7000	79.4667
Super-Female	12	74.2500	5.62664	1.62427	70.6750	77.8250
Low-Male	12	70.8333	6.01261	1.73569	67.0131	74.6536
Medium-Male	12	72.1667	6.35085	1.83333	68.1315	76.2018
Super-Male	12	81.7500	6.57993	1.89946	77.5693	85.9307
Total	72	73.3333	7.71682	.90944	71.5200	75.1467

Table 6: ANOVA table for the comparison of grades for 6 groups (for all the combinations of levels of interest and gender).

ANOVA

Grade

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1764.333	5	352.867	9.453	.000001
Within Groups	2463.667	66	37.328		
Total	4228.000	71			

>>>>>>>>>>

If there is no effect for females

$\mu_{L-F} = \mu_{M-F} = \mu_{S-F}$

If there is not effect for males

$\mu_{L-M} = \mu_{M-M} = \mu_{S-M}$

$H_0: \mu_{L-M} = \mu_{M-M} = \mu_{S-M}, \mu_{L-F} = \mu_{M-F} = \mu_{S-F}$

Reduced model = 2-mean model (table 4)

$H_A: \mu_{L-M}, \mu_{M-M}, \mu_{S-M}, \mu_{L-F}, \mu_{M-F}, \mu_{S-F}$

Full model = 6-mean model (table 6)

$$F = \frac{[SS_E(\text{reduced}) - SS_E(\text{full})] / [df_E(\text{reduced}) - df_E(\text{full})]}{SS_E(\text{full}) / df_E(\text{full})}$$

$$= \frac{[4047.500 - 2463.667] / [70 - 66]}{2463.667 / 66} \quad F = 10.604$$

$df = [4, 66]$ the diff between the full model and the reduced model.

$P < 0.001$, ∴ Therefore there is an extremely strong evidence against H0

$P = 0.00000107$

>>>>>>>>>>

Example on Application of One-Way ANOVA and the Extra-Sum-of-Squares F-Test
In a certain university there are ten sections of Statistics 252 being taught in the same semester. There are four instructors (A, B, C, and D) and each instructor teaches the sections shown in the table below. Each section has 50 students enrolled. For parts (a), (b), and (c) below, clearly define the best procedure to be applied, but you do not need to actually perform the test since no data is given. In particular, choose the most appropriate test, state the null and alternative hypotheses in terms of the parameters defined in the table, and state the null distribution of the test statistic (that is, name the distribution of the test statistic and indicate the degrees of freedom). Assume all the required assumptions are satisfied.

Define: μ_i =mean mark of the i^{th} section, i = 1, 2, ..., 10

Instructor	Number of lecture sections	Parameters (subscript is the lecture section number)
A	3	μ_1, μ_2, μ_3
B	3	μ_4, μ_5, μ_6
C	2	μ_7, μ_8
D	2	μ_9, μ_{10}

>>>>>>>>>>

(a) Determine whether there is any difference in mean marks between the 10 sections.

$H_0: \mu_1 = \mu_2 = \cdots = \mu_{10}$ one-mean model

$H_A: \mu_1, \mu_2, \ldots, \mu_{10}$ 10-mean model

One factor ANOVA F-test for all 10 sections $k=10$, each with 50 students $\therefore n = 500$ $df[9, 490]$

F-distribution

(b) Determine if any instructor has different mean marks between their own sections.

$H_0: \mu_1 = \mu_2 = \mu_3,\ \mu_4 = \mu_5 = \mu_6,\ \mu_7 = \mu_8,\ \mu_9 = \mu_{10}$ 4-mean model

$H_A: \mu_1, \mu_2, \ldots, \mu_{10}$ 10-mean model

Null distribution

Extra SS F-test

$df[6, 490]$

(c) Suppose lecture sections 1, 4, 7, and 9 are evening classes and all the other sections are daytime classes, determine whether there is any difference in mean marks either between the evening classes or between the daytime classes.

$H_0: \mu_1 = \mu_4 = \mu_7 = \mu_9$, $\mu_2 = \mu_3 = \mu_5 = \mu_6 = \mu_8 = \mu_{10}$ 2-mean model

$H_A: \mu_1, \mu_2, \ldots, \mu_{10}$

ESS F-test

$df(8, 490)$

>>>>>>>>>>

Example Combining Extra-Sum-of-Squares F-Test with a Review of Other Procedures Covered in This Section

Research Problem: A coral reef researcher measured the heights of randomly sampled *Acropora formosa* colonies along the reef crests of Mbudya Island and Fungu Yasin, on both the landward and seaward sides, making a total of four sites. At the four sites, the heights were normally distributed and the standard deviations were approximately equal. One-way ANOVA, performed at the 5% significance level, showed that there was a difference in the mean heights at the four sites. Use the output in Tables 1 – 5 to answer the questions below.

Table 1: Two-Sample t-test (assuming Equal Variances and independent samples) for the difference in mean height of *Acropora formosa* at Mbudya and Fungu Yasin (data from landward and seaward combined)

	Mbudya	*Fungu Yasin*
Mean	63.4375	60.10714286
Variance	131.47984	124.6917989
Observations	32	28
Pooled Variance	128.31989	
Hypothesized Mean Difference	0	
df	58	
t Stat	1.1361148	
P(T<=t) one-tail	0.1302906	
P(T<=t) two-tail	0.2605812	

Table 2: Summary Statistics for the Four Coral Reef Sites

SUMMARY						
Groups	*Count*	*Sum*	*Average*	*Variance*		
Mbudya-Landward	18	1214	67.44444	111.7908		
Mbudya-Seaward	14	816	58.28571	116.5275		
Fungu Yasin-Landward	16	1023	63.9375	107.7958		
Fungu Yasin-Seaward	12	660	55	109.2727		

Table 3: ANOVA table for comparison of all four means

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	1373.944	3	457.9814	4.113888	0.010445	2.769431
Within Groups	6234.239	56	111.3257			
Total	7608.183	59				

Table 4: ANOVA table for comparison of Mbudya versus Fungu Yasin (landward and seaward combined)

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	165.6298	1	165.6298	1.290757	0.260581	4.006873
Within Groups	7442.554	58	128.3199			
Total	7608.183	59				

Table 5: ANOVA table for comparison of Landward sites versus Seaward sites (Mbudya and Fungu Yasin combined)

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	1200.009	1	1200.009	10.86121	0.001679	4.006873
Within Groups	6408.174	58	110.4858			
Total	7608.183	59				

Define the parameters as follows:

μ_{LM} = mean Acropora height at Landward side of Mbudya

μ_{LF} = mean ... Landward side of Fungu Yasin

μ_{SM} = mean ... Seaward side of Mbudya

μ_{SF} = mean ... Seaward side of Fungu Yasin

(a) The coral reef researcher suspected that the difference between sites was mainly due to the effects of the landward environment (sheltered) versus the seaward environment (exposed to strong wave action). Perform the most appropriate test (a single **overall** test), at the 5% significance level, to determine whether there was a difference in *Acropora formosa* heights between the landward and seaward sides of these reefs, after accounting for the effects of different reefs.

Parameters: (Defined same as above)

If there is no Landward/Seaward effect at Mbudya, then: $\mu_{LM} = \mu_{SM}$

If there is no Landward/Seaward effect at Fungu Yasin, then: $\mu_{LF} = \mu_{SF}$

$H_0 : \mu_{LM} = \mu_{SM}$ and $\mu_{LF} = \mu_{SF}$

[Reduced model: Two means model for only Mbudya and Fungu Yasin]

$H_a : \mu_{LM}, \mu_{SM}, \mu_{LF}, \mu_{SF}$ [Not all four reef sites have the same mean height]

[Full model: Four means model]

From ANOVA table for comparison of Mbudya versus Fungu Yasin (reduced model) (Table 4)

SS_E(Two means model) = 7442.554 and df_E =58

From ANOVA table for comparison of all four means (full model)

SS_E(Four means model) = 6234.239 and df_E =56 (Table 3)

Hence,

Extra Sum of Squares = SS_E(reduced) - SS_E(full)

Extra SS = 7442.554 – 6234.239 = 1208.315

Extra df = df_E (reduced) – df_E (full) = 58 – 56 = 2

$$F = \frac{Extra \quad SS / Extra \quad df}{MS_E(\text{Full} \quad \text{model})}$$

$$= \frac{1208.315/2}{6234.239/56} = 5.427$$

For the Extra-Sum-of-Squares F-test, $df = (Extra\_df, n-k) = (2,56)$

Thus, 0.005 < P < 0.01, which provides very strong evidence against the null hypothesis

Since P < α (0.05), reject Ho

Conclusion: At the 5% significance level, there is sufficient evidence to conclude that there is a difference in mean height of the coral *Acropora formosa* between the landward and seaward sides of these coral reefs (Mbudya and Fungu Yasin combined).

(b) Suppose the researcher, then also wanted to check if there was any difference between Mbudya and Fungu Yasin, after accounting for the effect of landward versus seaward sides. Again, perform the most appropriate test (a single **overall** test) at the 5% significance level.

If there is no effect of reef (Mbudya/Fungu Yasin) on the Landward side, then: $\mu_{LM} = \mu_{LF}$

If there is no effect of reef (Mbudya/Fungu Yasin) on the Seaward side, then: $\mu_{SM} = \mu_{SF}$

$H_0 : \mu_{LM} = \mu_{LF}$ and $\mu_{SM} = \mu_{SF}$

[Reduced model: Two means model for only Landward and Seaward]

$H_a : \mu_{LM}, \mu_{LF}, \mu_{SM}, \mu_{SF}$

[Full model: Four means model]

Using the ANOVA table for Landward versus Seaward (reduced model) (=Two means model) (Table 5)
And the ANOVA table for comparison of all four means (full model) (=Four means model) (Table 3)

$$F = \frac{[SS_E(reduced) - SS_E(full)]/[df_E(reduced) - df_E(full)]}{SS_E(full)/df_E(full)}$$
$$= \frac{[6408.174 - 6234.239]/[58-56]}{6234.239/56} = \frac{173.935/2}{6234.239/56} = 0.781$$

For the Extra-Sum-of-Squares F-test, $df = (Extra\_df, n-k) = (2,56)$

Thus, P > 0.25, which provides weak evidence against the null hypothesis
Since P > α (0.05), do not reject Ho

Conclusion: At the 5% significance level, there is insufficient evidence to conclude that there is a difference in mean height of the coral *Acropora formosa* between the Mbudya and Fungu Yasin (Landward and Seaward sides combined).

(c) Compare of the pooled two-mean t-test and the single-factor ANOVA F-test with respect to purpose, assumptions, hypotheses and statistical results.

>>>>>>>>>>

Compare the purpose: The purpose of both is to test for a difference between means

Compare assumptions:

Compare hypotheses:

Statistical results of the pooled two-mean t-test:

Statistical results of the single-factor ANOVA F-test:

>>>>>>>>>>

(d) Define a linear combination (using the 4 parameters (means) defined above) to compare the overall mean height of *Acropora formosa* at Mbudya and Fungu Yasin. In addition, determine the estimate of the contrast using the output in Table 2, but you don't have to perform a complete test. How does this estimate of the difference compare to your estimate in part (c)? Whether it is the same or different, explain the reason.

$$\gamma = \frac{(\mu_{LM} + \mu_{SM})}{2} - \frac{(\mu_{LF} + \mu_{SF})}{2}$$

$$\gamma = \frac{1}{2}\mu_{LM} + \frac{1}{2}\mu_{SM} - \frac{1}{2}\mu_{LF} - \frac{1}{2}\mu_{SF}$$

$$\hat{\gamma} = \frac{1}{2}\bar{y}_{LM} + \frac{1}{2}\bar{y}_{SM} - \frac{1}{2}\bar{y}_{LF} - \frac{1}{2}\bar{y}_{SF}$$

$$= \frac{1}{2}(67.44444) + \frac{1}{2}(58.28571) - \frac{1}{2}(63.9375) - \frac{1}{2}(55.0000) = 3.3963$$

This estimate (3.3963) is slightly different from the estimate in part (c) (3.3304). This difference is due to different sample sizes. If the same sizes had been the same, the estimates would have been exactly the same.

(e) Define a linear combination to compare the overall mean height of *Acropora formosa* between landward and seaward sides (regardless of the reef). Use this linear combination to carry out a test, at the 5% significance level, whether there is a difference in mean height between landward and seaward sides.

Contrast: $\gamma = \frac{(\mu_{LM} + \mu_{LF})}{2} - \frac{(\mu_{SM} + \mu_{SF})}{2}$

$$H_0 : \gamma = 0 \qquad H_a : \gamma \neq 0$$

Estimate: $\hat{\gamma} = \frac{1}{2}\bar{y}_{LM} + \frac{1}{2}\bar{y}_{LF} - \frac{1}{2}\bar{y}_{SM} - \frac{1}{2}\bar{y}_{SF}$

$$= \frac{1}{2}(67.44444) + \frac{1}{2}(63.9375) - \frac{1}{2}(58.28571) - \frac{1}{2}(55.0000) = 9.0481$$

Standard error of the estimate:

$$SE(\hat{\gamma}) = s_p\sqrt{\frac{C_1^2}{n_1} + \frac{C_2^2}{n_2} + \ldots + \frac{C_k^2}{n_k}} \qquad s_p = \sqrt{MSE} = \sqrt{111.3257} = 10.5511$$

$$SE(\hat{\gamma}) = (10.5511)\sqrt{\frac{(1/2)^2}{18} + \frac{(1/2)^2}{16} + \frac{(-1/2)^2}{14} + \frac{(-1/2)^2}{12}} = 2.75534$$

Observed value of the test-statistic

$$t = \frac{\hat{\gamma} - 0}{SE(\hat{\gamma})} = \frac{9.0481}{2.75534} = 3.284$$

$df = n - k = 60 - 4 = 56$ P-value: (0.0005 < P < 0.001) x 2 ➔ 0.001 < P < 0.002

Since P < α (0.05), reject H_0 with very strong evidence.

Conclusion: At the 5% significance level, there is a difference in mean height of *Acropora formosa* between landward and seaward sides.

(f) Does the effect of the side of the reef (landward or seaward) depend on the reef (Mbudya or Fungu Yasin)? Define a linear combination and carry out a test (at α = 0.05) to answer this question.

The effect of the side of the reef are:

For Mbudya: $\mu_{LM} - \mu_{SM}$

For Fungu Yasin: $\mu_{LF} - \mu_{SF}$

If the effect of the side of the reef does not depend upon which reef it is, then:

$\mu_{LM} - \mu_{SM} = \mu_{LF} - \mu_{SF}$ ➔which means that: $\mu_{LM} - \mu_{SM} - \mu_{LF} + \mu_{SF} = 0$

Thus, the linear combination is:

$$\gamma = \mu_{LM} - \mu_{SM} - \mu_{LF} + \mu_{SF}$$

The estimate for the linear combination is:

$$\hat{\gamma} = \bar{y}_{LM} - \bar{y}_{SM} - \bar{y}_{LF} + \bar{y}_{SF} = 67.44444 - 58.28571 - 63.9375 + 55 = 0.2212$$

Standard error of the estimate:

$$SE(\hat{\gamma}) = (10.5511)\sqrt{\frac{(1)^2}{18} + \frac{(-1)^2}{16} + \frac{(-1)^2}{14} + \frac{(1)^2}{12}} = 5.51107$$

Observed value of the t-statistic:

$$t = \frac{\hat{\gamma} - 0}{SE(\hat{\gamma})} = \frac{0.2212}{5.51107} = 0.04014$$

$$df = n - k = 60 - 4 = 56$$

P-value: (P > 0.25) x 2 ➔ P > 0.50

Since P > α (0.05), do not reject H_0 since there is weak evidence against it.

Conclusion: At the 5% significance level, the effect of the side of the reef (landward or seaward) does not depend on the reef (Mbudya or Fungu Yasin).

(g) Use the Bonferroni method to calculate two simultaneous 96% confidence intervals for the difference in mean height of Acropora Formosa between the landward side and seawards side for each reef separately.

So, we need to find 96% familywise confidence intervals for:

i) Effect of the side of the reef for Mbudya: $\gamma_M = \mu_{LM} - \mu_{SM}$

ii) Effect of the side of the reef for Fungu Yasin: $\gamma_F = \mu_{LF} - \mu_{SF}$

For 96% confidence, α = 0.04

$$\alpha_I = \frac{\alpha_F}{m} = \frac{0.04}{2} = 0.02$$

[**Note:** Here we do not use the formula $m = \frac{k(k-1)}{2}$ because this is not multiple comparisons where we want to compare all possible means pairwise; but rather, we are calculating 2 simultaneous confidence intervals, so m = 2.]

The critical value = $t_{n-k,\alpha_I/2} = t_{60-4,0.02/2} = t_{56,0.01} = 2.403$

Using the formula: $\hat{\gamma} \pm \text{Critical value} \times SE(\hat{\gamma})$

For the effect of the side of the reef at Mbudya:

$\hat{\gamma}_M = \overline{y}_{LM} - \overline{y}_{SM} = 67.44444 - 58.28571 = 9.15873$

$s_p = \sqrt{MSE} = \sqrt{111.3257} = 10.5511$

$SE(\hat{\gamma}) = (10.5511)\sqrt{\frac{(1)^2}{18} + \frac{(-1)^2}{14}} = 3.75987$

$9.15873 \pm 2.403 \times 3.75987 \Rightarrow (9.15873 \pm 9.0350)$

$(0.124, 18.194)$

For the effect of the side of the reef at Fungu Yasin:

$\gamma_F = \overline{y}_{LF} - \overline{y}_{SF} = 63.9375 - 55.0000 = 8.9375$

$SE(\hat{\gamma}) = (10.5511)\sqrt{\frac{(1)^2}{16} + \frac{(-1)^2}{12}} = 4.02927$

$8.9375 \pm 2.403 \times 4.02927 \Rightarrow (8.9375 \pm 9.6823)$

$(-0.745, 18.620)$

3.7 The Kruskal-Wallis test (a Nonparametric Equivalent of One-Way ANOVA)

- Can be used in all situations where there are k independent samples and k > 2
- If the data fit the assumptions of ANOVA, the Kruskal-Wallis Test will be $3/\pi$ =95.5% as powerful as ANOVA
- If the data do not fit the assumptions of ANOVA, the Kruskal-Wallis Test will be more powerful than ANOVA
- The data are ranked in order from lowest to highest (across all k groups) and calculations are performed on the ranks
- Where there are tied observations, assign the average rank to the tied observations

Importance of the Kruskal-Wallis test and other Nonparametric tests

If

- one or more of the data sets being compared are not normally distributed nor are they lognormal,

And

- when the one or more sample sizes are less than 30 (Central Limit Theorem),

The Kruskal-Wallis test is the only valid option (one-factor ANOVA cannot be performed)

- Also, since the Kruskal-Wallis test converts the raw data to ranks, it is not affected by outliers or unequal standard deviations, while these would affect one-way ANOVA.

Kruskal-Wallis Test

Purpose: To test for a difference between k population (where k > 2)

Assumptions:

1. Simple Random samples
2. Independent samples
3. Same-shape populations
4. All sample sizes are 5 or greater

The null and alternative hypotheses:

H_0: The population distributions of ***k*** populations are identical.

H_a: The population distributions of ***k*** populations are not all identical, that is, at least two are different.

Calculating the test statistic:

First, rank the data from all k samples combined, from lowest to highest
Assign average ranks where there are tied observations

$$H = \frac{12}{n(n+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3(n+1)$$

Where n = total number of observations

$n_1, n_2, \ldots, n_k$ denote sample sizes of samples 1, 2,...k

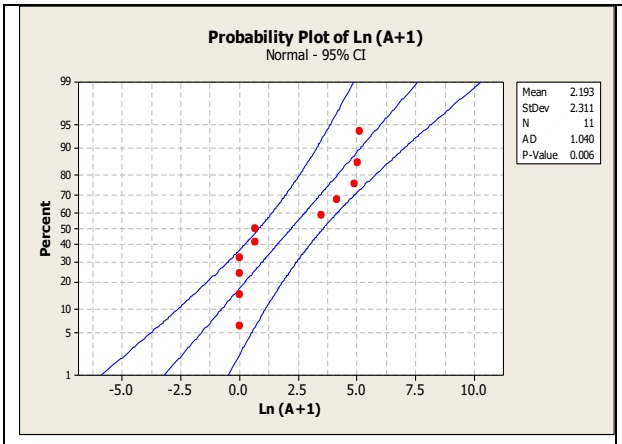
$R_1, R_2, \ldots, R_k$ denote the sums of the ranks for the sample data

Critical values of ***H*** follow the χ_α^2 distribution with **df = k – 1**

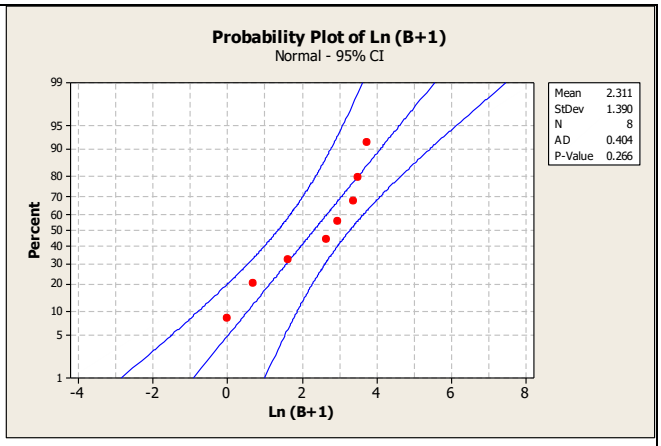
Research Problem:
Pitfall traps are inserted into the soil at ground level in three grasslands (A, B and C) in order to determine whether there is a difference in the abundance of ants in the three grasslands. Test this hypothesis at the 10% significance level.

	Number of ants per pitfall trap										
Grassland A	168	0	62	0	1	135	0	0	155	32	1
Grassland B	0	13	28	32	18	4	41	1			
Grassland C	144	1	3	135	45	3	122	4	110		

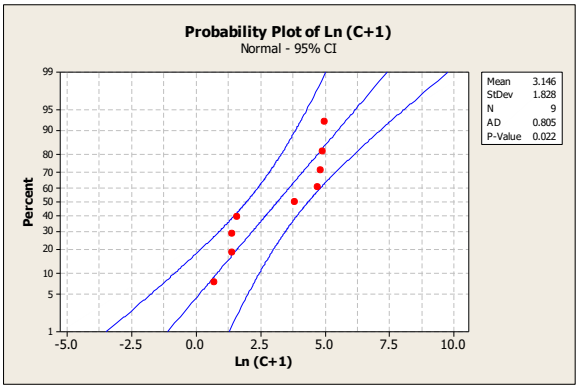
This shows an aggregated distribution.
(**Note** the difference between aggregated, random and regular (even) distributions in space or in time.)



Normal Probability Plot for logged transformed data from grassland A



Normal Probability Plot for logged transformed data from grassland B



Normal Probability Plot for logged transformed data from grassland C

Step 1: The purpose is to compare k populations

- 3 independent random samples
  However:
  - The data are neither normal nor lognormal
  - Sample size is < 30, therefore the Central Limit Theorem does not apply
  - Therefore, the Kruskal-Wallis Test must be performed
- Same shape distributions, as indicated in the NPPs above
- Sample size of all groups ≥ 5.

Step 2: H_0: There is no difference in the abundance of ants in the three grasslands.
H_a: There is a difference in the abundance of ants in the three grasslands (at least two are different).

Step 3: Calculate the test statistic H
Rank the data from lowest to highest, assigning average ranks where there are tied observations.
>>>>>>>>>>

	Grassland A		Grassland B		Grassland C	
	No. of ants	**Rank**	**No. of ants**	**Rank**	**No. of ants**	**Rank**
	168	28	0	3	144	26
	0	3	13	14	1	2.5
	62	21	28	16	3	10.5
	0	3	32	17.5	135	24.5
	1	7.5	18	15	45	20
	135	24	4	12.5	3	10.5
	0	3	41	19	122	33
	0	3	1	7.5	4	12.5
	155	27			110	22
	32	17.5				
	1	7.5				
Sum of ranks (R_j)		145		104.5		156.5
Sample size (n_j)		11		8		9

$N = 28$

$$H = \frac{12}{n(n+1)} \sum \ldots = \frac{12}{28(29)}\left[\frac{(145)^2}{11} + \frac{(104.5)^2}{8} + \frac{(156.5)^2}{9}\right] - 3(28+1)$$

$$= 1.647$$

$$df = k - 1 = 2$$

$$p > 0.200$$

>>>>>>>>>>

SECTION 4: SIMPLE LINEAR REGRESSION AND SIMPLE LINEAR CORRELATION

Linear Regression and Linear Correlation

- analyze the **relationship** between **quantitative variables**
- **"Simple"** means **only two variables** (x and y) are involved
- In the next section, we will discuss situations where there are more than two variables (multiple linear regression)
- There are other types of regression that are not linear (e.g., curvilinear), but the analysis of these are not dealt with in this course

Scatterplot or scatter diagram

- Illustrates the relationship between two quantitative variables, by placing points on the graph that correspond to the values of both variables **(x, y)** at the same time
- If a the data points in a scatterplot fall (roughly) in a straight line, this indicates a probably linear relationship between the two variables

Simple Linear Regression

- Used to analyze the relationship between two quantitative variables when one variable responds to the other
- **Explanatory variable** (or **predictor variable**) (plotted on ***x***-axis)
  = the variable that may affect the other variable or that can be used to make predictions about the other variable
- **Response variable** (plotted on ***y***-axis)
  = the variable that reacts to or is affected by the explanatory variable. It responds to changes in the predictor variable

Simple Linear Correlation

- Used to analyze the relationship between two quantitative variables when a change in one variable appears to be related to (or associated with) a change in the other, but one is not necessarily responding to the other
- So, the two variables are co-related
- Either variable may be plotted as ***x*** or ***y***
- Therefore, correlation can be applied in a wider variety of situations even when the variables cannot be identified as explanatory and response variables

Notation and Formulas for Quantities Used in Regression and Correlation

Quantity	Defining formula	Computing formula
Sum of Squares of (the deviations in) ***x***	$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$	$S_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$
Sum of Squares of (the deviations in) ***y***	$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$	$S_{yy} = \sum y_i^2 - \frac{(\sum y_i)^2}{n}$
Sum of Products of (the deviations in) ***x*** and ***y***	$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$	$S_{xy} = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}$

4.1 The Linear Regression Model

- Usually, data obtained from a sample of a population do not fall exactly along a straight line
- **Linear regression line** – the "best fit" line that passes through the points and is calculated using the **"least squares criterion"**

Simple Linear Regression Model

Model for the population regression line:

$$\mu(Y \mid X) = \beta_0 + \beta_1 X$$

$[\mu(Y \mid X)$ means "predicted mean of Y at a given $X]$

Where Y is the response variable

X is the explanatory variable

Parameters: β_0 is the y-intercept

β_1 is the slope (change in Y over change in X)

Estimated Model or sample regression line (based on a set of ***n*** data points):

$$\hat{y} = \hat{\mu}(Y \mid X) = \hat{\beta}_0 + \hat{\beta}_1 x$$ ['^' denotes an estimate]

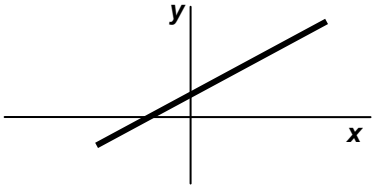
Where $\hat{y}$ is used to denote the y-value predicted by a regression equation

$\hat{\beta}_0$ and $\hat{\beta}_1$ are least squares estimates

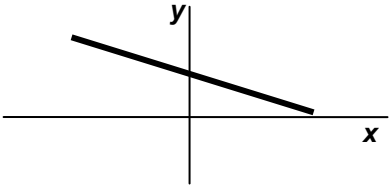
Slope = $\hat{\beta}_1 = \dfrac{S_{xy}}{S_{xx}} = \dfrac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2}$

y-intercept = $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = \dfrac{\sum y}{n} - \hat{\beta}_1 \dfrac{\sum x}{n}$

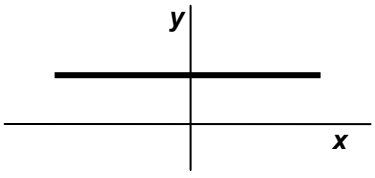
Different types of slopes



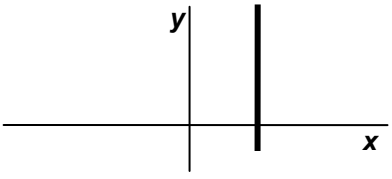
Positive slope ($\beta_1 > 0$)



Negative slope ($\beta_1 < 0$)

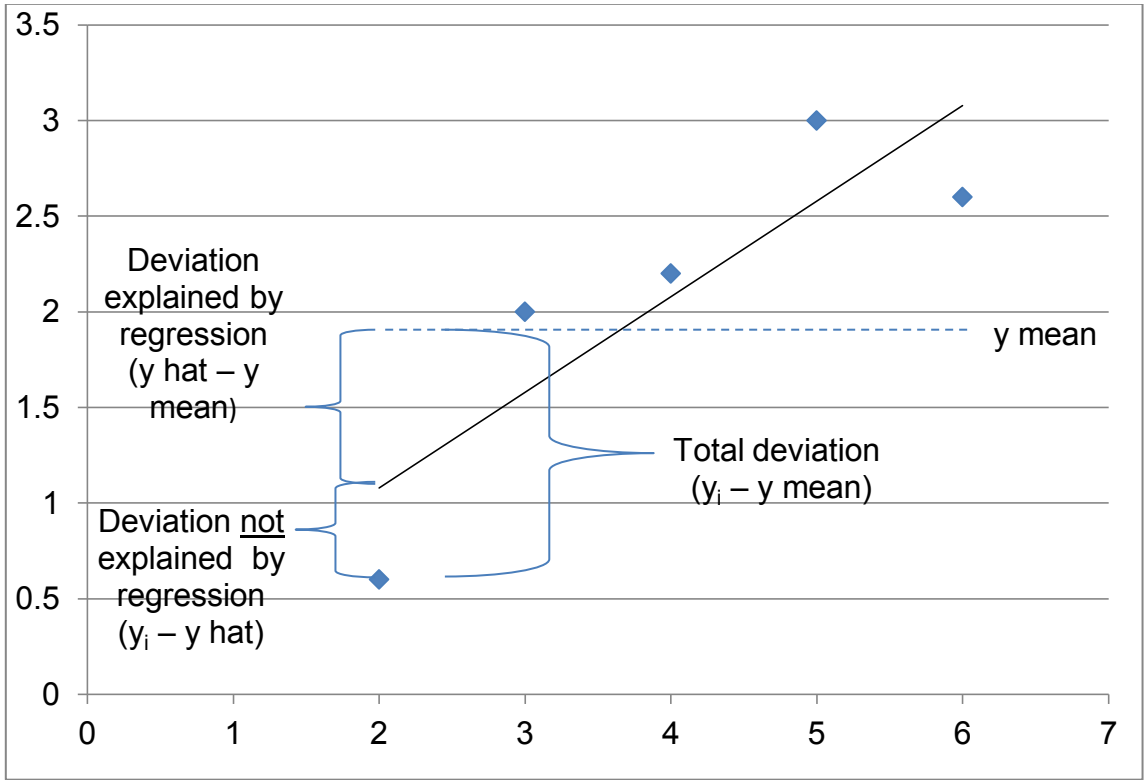


No slope ($\beta_1 = 0$)



Infinite slope ($\beta_1 = \infty$)

Three Sources of Variation (Deviation) in Regression



Three Sums of Squares in Regression
[Representing the three sources of variation or deviation]

Total Sum of Squares (SS_{TOTAL} = Syy)
= total variation in the observed values of the response variable

$$SS_{TOTAL} = S_{yy} = \sum (y_i - \bar{y})^2$$

Regression Sum of Squares (SS_{REGR})
= variation in the observed values of the response variable explained by regression model

$$SS_{REGR} = \sum (\hat{y}_i - \bar{y})^2 = \frac{(S_{xy})^2}{S_{xx}} = \frac{\left[\sum (x_i - \bar{x})(y_i - \bar{y})\right]^2}{\sum (x_i - \bar{x})^2}$$

Error (Residual) Sum of Squares (SS_{ERROR})
= variation in the observed values of the response variable that is <u>not</u> explained by the regression model

$$SS_{ERROR} = SS_{RES} = \sum (y_i - \hat{y}_i)^2$$

$$SS_{ERROR} = SS_{TOTAL} - SS_{REGR}$$

- It is SS_{ERROR} that regression tries to minimize in order to obtain the "best fit" line

Regression identity: $SS_{TOTAL} = SS_{REGR} + SS_{ERROR}$

Analysis of Residuals

Residual = error (e) = vertical distance from the regression line to a data point (may be + or –)

$$e = y_i - \hat{y}_i$$

Residual Sum of Squares = Error Sum of Squares

= the variation in the observed values of the response variable that is <u>not</u> explained by the regression

$$SS_{\text{RES}} = SS_{Error} = \sum (y_i - \hat{y}_i)^2$$

Least-squares criterion

- Tries to minimize the Residual or Error Sum of Squares (SSE) in order to get the "best fit" line
- Thus, regression tries to minimize the errors due to deviations <u>not</u> explained by the regression equation

Residual Plots and Residual Analysis

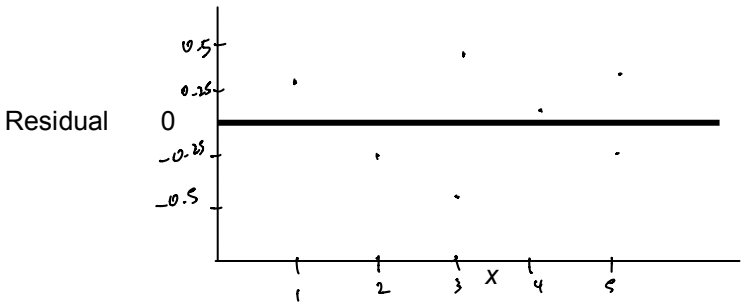
- Can be used for checking whether a certain data set fits the assumptions of regression analysis
- Each data point is plotted as the residual (error) against the corresponding the x-value
- If the assumptions for regression inferences are met, the plot of the residuals against the values of the predictor variable should:
  <u>fall roughly in a horizontal band centered and symmetric about the x-axis</u>

Example of Residual Analysis (heights of the trees of different ages)

Using the linear regression equation ($\hat{y} = -0.08087 + 0.66809x$), we can determine $\hat{y}$ and the residual for every data point and find the Residual Sum of Squares (SSE)

Age (years) x	**Height (m) y**	$\hat{y}$	**Residual (error)** $(y_i - \hat{y}_i)$	**(error)2** $(y_i - \hat{y}_i)^2$
1	0.9	0.587	0.313	0.0980
2	1.0	1.255	-0.255	0.0652
3	1.4	1.923	-0.523	0.2735
3	2.2	1.923	0.277	0.0767
4	2.6	2.591	0.009	0.0001
5	3.0	3.260	-0.260	0.0676
5	3.7	3.260	0.440	0.1936
			Sum = 0	$SS_{\text{RES}} = \sum (y_i - \hat{y}_i)^2 = 0.7747$

Sketch of a residual plot for the heights of trees [plot Error against x]



Standard error of the model (= Common standard deviation of the model)
(= Standard deviation of the residuals)

- Quantifies the amount of scatter around the regression line
- Given by: $s_e = \hat{\sigma} = \sqrt{\frac{\sum e^2}{n-2}} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{SS_{ERROR}}{n-2}} = \sqrt{MS_{ERROR}}$

Prediction: Interpolation and Extrapolation

Interpolation = using the regression equation to make predictions about the response variable, within the range of the observed values of ***x***

- can be reasonably accurate

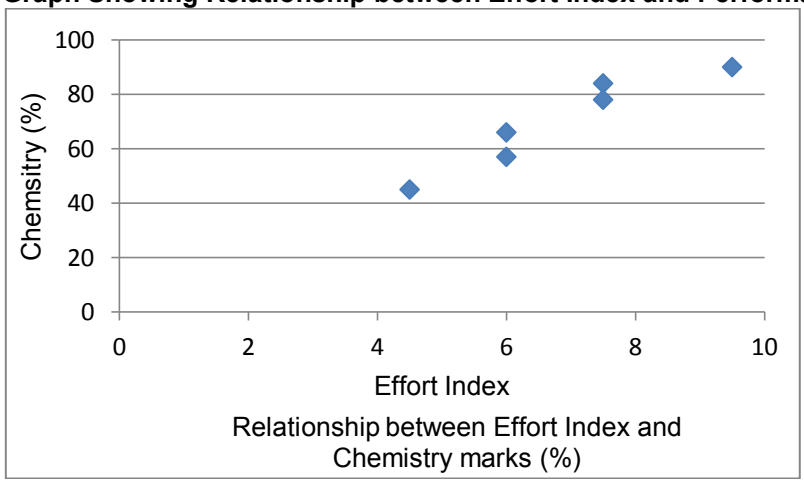
Extrapolation = using the regression equation to make predictions about the response variable, outside the range of the observed values of ***x***

- can sometimes lead to seriously incorrect predictions because the relationship may not hold beyond the observed range
- to avoid errors due to extrapolation, researchers sometimes define the range of the observed values along with a regression equation
  - e.g., $\hat{y} = 24.7 + 3.4x$, 6 ≤ ***x*** ≤ 45

Example of Calculation of the Regression Line
Effort Index (on a scale of 0 – 10) was calculated based on a combination of factors such as attendance in classes, hours per week spent studying and completion of assignments on time. The table below shows the Effort Index and Chemistry and Biology marks (in %) of a random sample of 6 students.

	Halima	John	Jing	Jasmin	Vanessa	Harry
Effort Index	9.5	4.5	7.5	6	7.5	6
Chemistry (%)	90	45	84	66	78	57
Biology (%)	90	56	96	65	81	74

Graph Showing Relationship between Effort Index and Performance in Chemistry



This above graph shows that the relationship is appropriate for linear regression analysis because: (1) linear relationship (2) no significant outliers

Calculation of the regression line:

Table showing calculation of the deviations in ***x*** and ***y*** and the product of the deviations

	Effort index ***x***	Chemistry (%) ***y***	Deviations in ***x*** $(x_i - \bar{x})$	Deviations in ***y*** $(y_i - \bar{y})$	Product of deviations in ***x*** and ***y*** $(x_i - \bar{x})(y_i - \bar{y})$
Halima	9.5	90	2.6667	20	53.334
John	4.5	45	-2.3333	-25	58.3325
Jing	7.5	84	0.6667	14	9.3338
Jasmin	6.0	66	-0.8333	-4	3.3332
Vanessa	7.5	78	0.6667	8	5.3336
Harry	6.0	57	-0.8333	-13	10.8329
Totals	**41**	**420**	**0**	**0**	$S_{xy} = \sum(x_i - \bar{x})(y_i - \bar{y}) = 140.5$
Mean	**6.833333**	**70**			
SD	**1.722401**	**17.146428**			

Squared deviations in ***x*** $(x_i - \bar{x})^2$	Squared deviations in ***y*** $(y_i - \bar{y})^2$
7.1113	400
5.4443	625
0.4445	196
0.6944	16
0.4445	64
0.6944	169
$S_{xx} = \sum(x_i - \bar{x})^2 = 14.8333$	$S_{yy} = \sum(y_i - \bar{y})^2 = 1470$

Calculate the linear regression equation describing the relationship between Effort Index and performance (%) in Chemistry. Also, interpret the meaning of the slope.

>>>>>>>>>>

Slope: $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$

$= \frac{140.5}{14.8333} = 9.4719$ % patients effort index

y-intercept: $\hat{\beta} = \bar{y} - \hat{\beta}_1 \bar{x}$

$= 70.0 - 9.471968333$

$= 5.27528$

Regression Eq. $\hat{y} = 5.275 + 9.472x$

where $4.5 \leq x \leq 9.5$

>>>>>>>>>>

4.2 Linear Correlation (r) and Coefficient of Determination (R^2)

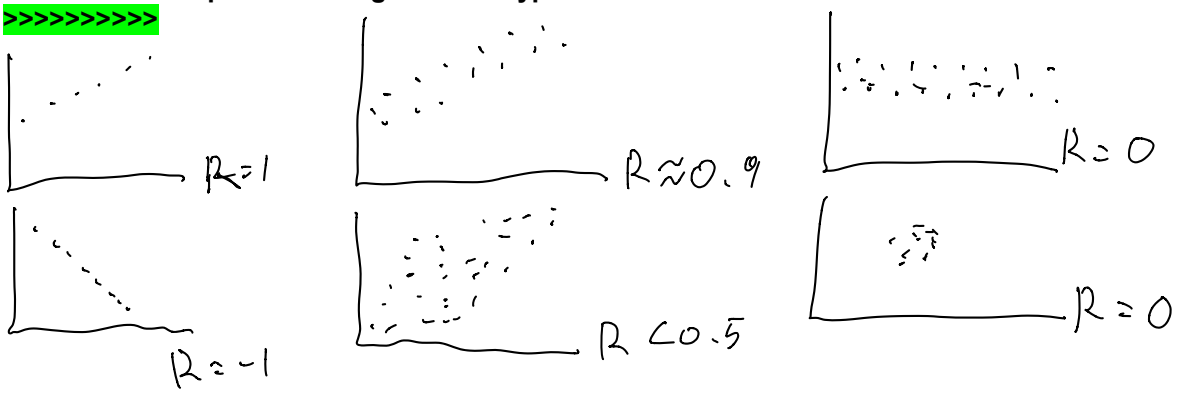
- The most common measure of correlation is the Pearson product-moment correlation coefficient.

Three Aspects of a Relationship between Variables (Correlation and Regression)

- **Direction:**
  - Positive correlation: if ***r*** has + sign, then the slope must have + sign
  - Negative correlation: if ***r*** has — sign, then the slope must have — sign
- **Form:**
  - May be a straight line relationship (linear) or curved (Here we only deal with linear)
- **Strength:** The magnitude of ***r*** indicates the strength of the linear relationship between the two variables.
  - ***r*** close to -1 or 1 indicates a strong linear relationship and the regression equation is very useful for making predictions
  - ***r*** close to 0 indicates no relationship or a weak linear relation and the regression equation is either useless or not very useful for making predictions

Sketch 6 scatterplots showing different types of linear correlation

>>>>>>>>>>



>>>>>>>>>>

Warnings on the Use of the Linear Correlation Coefficient (and linear regression analysis)

- The linear correlation coefficient should only be used when a scatterplot indicates that the data points are scattered roughly about a straight line

Outliers, Leverage and Influential Points

Outlier

- Any data point that does not follow the general pattern of the rest of the data or that stands away from other data points
- May be either due to having a large residual (on the y-scale) or having high leverage (being far away from other points on the x-scale)

Leverage

- Data points whose x-values are far from the mean of x have leverage
- Data points having leverage can have a great affect on the linear regression line
- They can completely change the slope and y-intercept

Influential Points

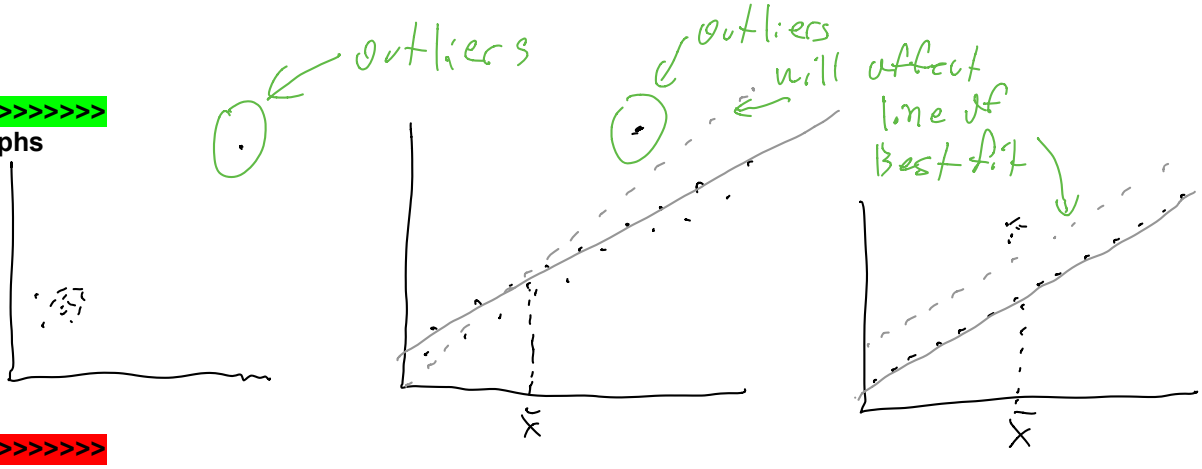
- A data point which, if omitted, results in a very different regression model

Serious outliers and influential observations

- They can make a weak correlation appear to be a strong correlation
- They can change the slope considerably
- They can even make a positive correlation to be calculate as a negative correlation

>>>>>>>>>>

Graphs



>>>>>>>>>>

Correlation versus Prediction/Response

- Correlation between variables does not necessarily mean that one variable affects or can be used to predict the other variable

Calculation of the Correlation Coefficient

The Linear Correlation Coefficient, *r*

For a set of ***n*** data points,

$$r = \frac{\text{covariance of } x \text{ and } y}{(\text{standard deviation of } x) \times (\text{standard deviation of } y)}$$

Where: Covariance of ***x*** and ***y*** = $s_{xy} = \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(n-1)}$

Thus, $r = \dfrac{s_{xy}}{s_x \times s_y} = \dfrac{\dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(n-1)}}{\sqrt{\dfrac{\sum (x_i - \bar{x})^2}{n-1}} \times \sqrt{\dfrac{\sum (y_i - \bar{y})^2}{n-1}}}$

Where s_x and s_y are the sample standard deviations of the x-values and y-values, respectively.

Also, $r = \dfrac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left[\sum (x_i - \bar{x})^2\right]\left[\sum (y_i - \bar{y})^2\right]}}$

Also, $r = \sqrt{\dfrac{SS_{REGR}}{SS_{TOTAL}}}$ (then **add correct sign**, + or –)

Always: $-1 \le r \le 1$

r has no units because they cancel out during calculations

Coefficient of Determination (R^2) and Adjusted R^2

- Adjusted R^2 takes into account (adjusts for) samples size, though it makes little difference in SLR

Coefficient of determination (R^2) = [correlation coefficient]^2

***R^2* =** the fraction or percentage of variation in the observed values of the response variable that is accounted for by the regression analysis

$$R^2 = r^2 = \frac{\text{Explained variability}}{\text{Total variability}}$$

$$R^2 = \frac{SS_{REGR}}{SS_{TOTAL}} = 1 - \frac{SS_{Error}}{SS_{TOTAL}} = \frac{SS_{TOTAL} - SS_{Error}}{SS_{TOTAL}}$$

Adjusted Coefficient of Determination

$$R^2_{adj} = 1 - \frac{MSE}{MST}$$

Always: $0 \le R^2 \le 1$
OR: $0\% \le R^2 \le 100\%$

This implies that $1 - R^2$ of the variation in the observed values of the response variable are accounted for by other factors, not the explanatory variable used in the regression analysis

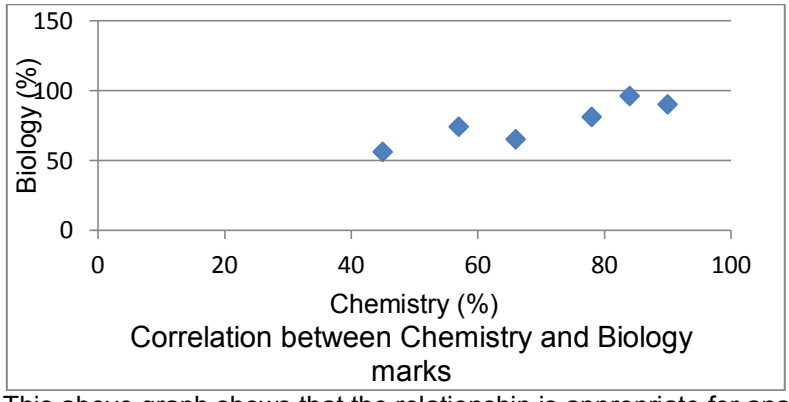
- If $\mathbf{r^2}$ is close to 0, this suggests that the regression equation is not very useful for accounting for the response variable or for making predictions
- If $\mathbf{r^2}$ is close to 1, this suggests that the regression equation is very useful for accounting for the response variable or for making predictions

Relationship Between the Correlation Coefficient and the Coefficient of Determination

- Both are measures for indicating the strength of a linear relationship or the usefulness of the regression equation for making predictions.
- If you have already calculated the coefficient of determination and we know the direction of the relationship (positive or negative), we can calculate the correlation coefficient as:

$$r = \sqrt{R^2}$$ [And then add the appropriate sign, **+** or **-**]

Example: Correlation between Chemistry and Biology marks
Note: Since this is correlation, either variable could be considered as ***x*** or ***y***



This above graph shows that the relationship is appropriate for analysis with linear correlation because: (1) linear relationship (2) no significant outliers

Table showing calculation of the deviations in ***x*** and ***y*** and the product of the deviations

	Chem (%) ***x***	Biol (%) ***y***	Deviations in ***x*** $(x_i - \bar{x})$	Deviations in ***y*** $(y_i - \bar{y})$	Product of deviations in ***x*** and ***y*** $(x_i - \bar{x})(y_i - \bar{y})$
Halima	90	90	20	13	20 x 13 = 260
John	45	56	-25	-21	525
Jing	84	96	14	19	266
Jasmin	66	65	-4	-12	48
Vanessa	78	81	8	4	32
Harry	57	74	-13	-3	39
Totals	**420**	**462**	**0**	**0**	$S_{xy} = \sum(x_i - \bar{x})(y_i - \bar{y}) = 1{,}170$
Mean	**70**	**77**			
SD	**17.14643**	**15.09967**			

Squared deviations in ***x*** $(x_i - \bar{x})^2$	Squared deviations in ***y*** $(y_i - \bar{y})^2$
400	169
625	441
196	361
16	144
64	16
169	9
$S_{xx} = \sum(x_i - \bar{x})^2 = 1{,}470$	$S_{yy} = \sum(y_i - \bar{y})^2 = 1{,}150$

Covariance of x and y is:

$$s_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{(n-1)} = \frac{1{,}170}{6-1} = 234$$

Correlation coefficient (***r***) is:

$$r = \frac{s_{xy}}{s_x \times s_y} = \frac{234}{17.14643 \times 15.09967} = 0.9038$$

Correlation coefficient can also be calculated as:

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left[\sum(x_i - \bar{x})^2\right]\left[\sum(y_i - \bar{y})^2\right]}} = \frac{1{,}170}{\sqrt{(1{,}470)(1{,}150)}} = 0.900$$

Does the strong positive correlation indicate that the performance of the students in Chemistry <u>explains</u> their performance in Biology?

4.3 Assumptions for Regression Inferences and Analysis of Residuals

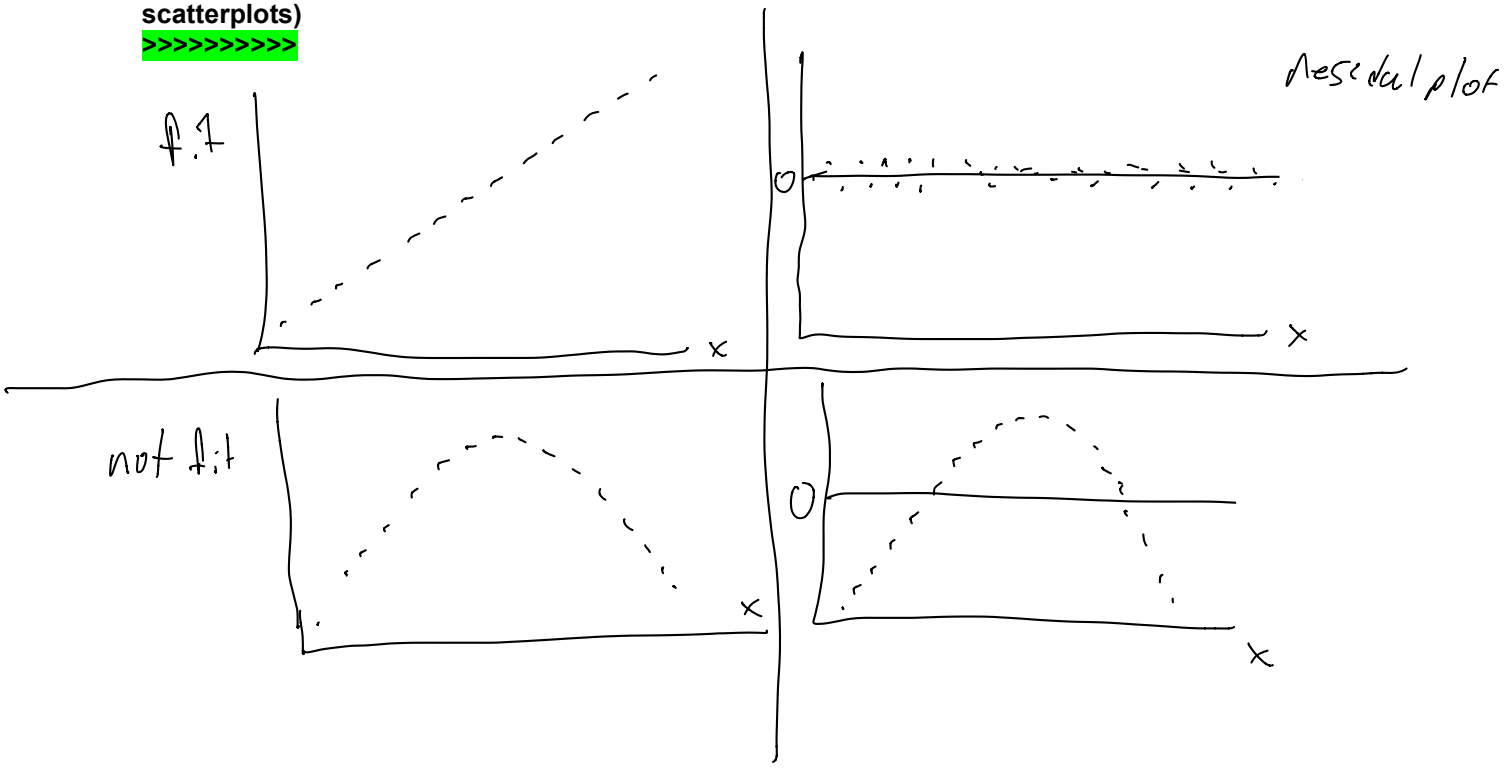
Assumptions (Conditions) for Regression Inferences

1. **Population regression line (<u>linearity</u>):** The relationship between the two variables must be approximately linear. In other words, there are constants β_0 and β_1 such that, for each value x of the predictor variable, the conditional mean of the response variable is $\beta_0 + \beta_1 x$.
2. **Equal standard deviations (homoscedasticity):** The standard deviations of ***y***-values must be approximately the same for all values of ***x***
3. **Normal populations:** For each value of ***x***, the corresponding ***y***-values must be normally distributed)
4. **No Serious Outliers:** Significant outliers can drastically change the regression model (just as for correlation)
5. **Independent observations:** The observations of the response variable are independent of one another. This implies that the observations of the predictor variable not need to be independent.

Note: All assumptions (except independence) can be <u>checked</u> by examining a scatterplot and/or residual plot. The assumption for normality is best checked with a NPP.

Illustrations for the first 2 assumptions of regression inferences each with xy scatterplot and a residual plot, for data that fit each assumption and data that do not (for normality, just use xy scatterplots)

>>>>>>>>>>



>>>>>>>>>

4.4 Testing the Significance of the Model using the Regression ANOVA Test

- Regression ANOVA tests the overall significance of the regression model
- In SLR, regression ANOVA can also be used to test for the significance of the slope, since there is only one slope

Mean Squares and F-Statistic in Simple Linear Regression ANOVA

Regression mean square (MS_{Regr}) = regression sum of squares divided by regression df

$$MS_{REGR} = SS_{REGR} / 1$$

Error mean square (MS_{ERROR}) = error sum of squares divided by error df

$$MS_{ERROR} = SS_{Error} / (n-2)$$

F-Statistic (F) $$F = \frac{SS_{REGR}/1}{SS_{Error}/(n-2)} = \frac{MS_{REGR}}{MS_{ERROR}}$$ Where n = number of x,y observations

ANOVA Test for Significance of the Model and the Slope of a Population Regression line

Purpose: To determine whether there is a relationship between two quantitative variables OR to decide whether the slope of the line is significantly different from zero.
Assumptions: see textbox on assumptions above

Step 1: Select appropriate test by checking purpose and assumptions

Step 2: H_0: $\beta_1 = 0$ (There is no relationship between the two variables)
H_a: $\beta_1 \neq 0$ (There is a relationship between the two variables)

Step 3: Calculate the three sums of squares (see page 3) and construct an ANOVA table
ANOVA Table for Simple Linear Regression

Source of variation	SS	df	MS = SS/df	F-statistic
Regression	SS_{REGR}	2 - 1	$MS_{REGR} = SS_{REGR}/1$	$F = MS_{REGR} / MS_{ERROR}$
Error	SS_{Error}	n – 2	$\hat{\sigma}^2 = MS_{ERROR} = SS_{Error}/(n-2)$	
Total	SS_{TOTAL}	n - 1		

$$F = \frac{SS_{REGR}/1}{SS_{Error}/(n-2)} = \frac{MS_{REGR}}{MS_{ERROR}}$$

Step 4: Decide to reject or not reject Ho
***df* =** (numerator degrees of freedom, denominator degrees of freedom)
$df = (1, n-2)$ (Where ***n*** = no. of ***xy*** observations)
If P-value ≤ α, reject H_0 (otherwise, do not reject H_0)
Step 5: Conclusion in terms of the research problem

Note:

1. $MS_{ERROR} = \hat{\sigma}^2 = (\text{standard error of the model})^2 = \text{variance of the model}$
2. $t^2 = F$
3. **The P-value will be the same for both the F-test and the t-test (two-tailed)**

Note: In general, the regression ***df*** is the number of parameters being estimated minus 1. Since here we are estimating two parameters, the y-intercept and slope (β_0 and β_1), the regression ***df*** = 2 – 1 = 1

Computer Output for the Example on Effort Index and Chemistry Performance

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.951477
R Square	0.905308
Adjusted R Square	0.881636
Standard Error	5.899081
Observations	6

model squared

$\sqrt{F} = t$

ANOVA table for Simple Linear Regression					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1330.8034	1330.8034	38.2424	0.003474604
Residual	4	139.19663	34.799157		
Total	5	1470			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	5.275281	10.739903	0.4911851	0.649026	-24.54347	35.09403
Effort index	9.47191	1.5316692	6.1840442	0.003475	5.21931	13.72451

Example: At the 5% significance level, test for a relationship between Effort Index and performance in Chemistry. [In other words, test whether the slope of the regression line is significant.] The regression line has been shown to be linear. Assume that all other assumptions are met.

>>>>>>>>>>

$H_0: \beta_1 = 0$ (there is no relationship)

$H_A: \beta_1 \neq 0$ (there is a relationship)

$SS_T = SS_{YY} = 1,470$

$SS_R = \frac{(SS_{XY})^2}{S_{XX}} = \frac{(140.5)^2}{14.8333} = 1330.8064$

$SS_E = SS_T - SS_R$

$= 1470 - 1330.8064$

$= 139.1936$

$df(1, 4)$

$0.001 < p < 0.005$

$F = \frac{SS_R/1}{SS_E/n-2} = \frac{MS_R}{MS_E} = \frac{1330.8064}{34.7984} = 38.242$

>>>>>>>>>>

Comparison between Regression t-test and ANOVA F-test (Applies <u>ONLY</u> to SLR)

F-statistic = (t-statistic)^2 = $(6.1840442)^2$ = 38.24

Two-tailed P-values are always equal for the F-test and t-test

For both the t-test and the F-test, the exact P-value = 0.003475

4.5 Inferences for the Slope of the Population Regression Line

- In SLR, either a Regression t-test OR Regression ANOVA can be used to test the slope
- However, the Regression t-test is more flexible because it is suitable for doing two-tailed tests or one-tailed tests

Regression t-Test for Significance of the Slope of a Population Regression line

Purpose: To determine whether there is a relationship between two quantitative variables OR to decide whether the slope of the line is significantly different from zero.

Assumptions: see textbox on assumptions above

Step 1: Select appropriate test by checking purpose and assumptions

Step 2:

Null hypothesis

H_0: $\beta_1 = 0$ (T**here is no relationship between the two variables**)

Alternative hypotheses

Two-tailed test: H_a: $\beta_1 \neq 0$ **(There is a relationship between the two variables**)

Left-tailed test: H_a: $\beta_1 < 0$ **(There is a negative relationship between the two variables**)

Right-tailed test: H_a: $\beta_1 > 0$ **(There is a positive relationship between the two variables**)

Step 3: Compute the calculated value of the test statistic

$$SS_{TOTAL} = S_{yy} = \sum (y_i - \bar{y})^2$$

$$SS_{REGR} = \sum (\hat{y}_i - \bar{y})^2 = \frac{(S_{xy})^2}{S_{xx}} = \frac{\left[\sum (x_i - \bar{x})(y_i - \bar{y})\right]^2}{\sum (x_i - \bar{x})^2}$$

$$SS_{ERROR} = SS_{TOTAL} - SS_{REGR}$$

Standard error of the model (residual standard deviation) (s_e):

$$\hat{\sigma} = \sqrt{\frac{SS_{ERROR}}{n-2}} = \sqrt{MS_{ERROR}}$$

Standard Error of the Estimate of the Slope:

$$SE(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{S_{xx}}}$$

The Regression t-statistic:

$$t = \frac{\hat{\beta}_1}{\hat{\sigma} / \sqrt{S_{xx}}} = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$$

Step 4: Decide to reject or not reject Ho and state the strength of the evidence against Ho.

$df = n - 2$ (Where ***n*** = no. of ***xy*** observations)

If P-value ≤ α, reject H_0; otherwise, do not reject H_0

Step 5: Conclusion in terms of the research problem

Testing the Significance of the Slope using the Regression t-test
At the 5% significance level, test for a relationship between Effort Index and performance in Chemistry. [In other words, test whether the slope of the regression line is significant.] The regression line has been shown to be linear. Assume that all other assumptions are met.

Step 1: Regression t-test is selected because the purpose is to test if the slope is significantly different from 0.

Step 2: H_0: $\beta_1 = 0$ (There is no relationship between performance in Chemistry and Effort index.)
H_a: $\beta_1 \neq 0$ (There is a relationship between performance in Chemistry and Effort index.)

Step 3: Compute the three sums of squares

$$SS_{TOTAL} = S_{yy} = \sum (y_i - \bar{y})^2 = 1,470$$

$$SS_{REGR} = \frac{(S_{xy})^2}{S_{xx}} = \frac{(140.5)^2}{14.8333} = 1330.8064$$

$$SS_{ERROR} = SS_{TOTAL} - SS_{REGR} = 1470 - 1330.8064 = 139.1936$$

>>>>>>>>>>

SE of model

$\hat{\sigma} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{139.1936}{6-2}} = 5.8990$

SE of slope

$SE(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{S_{xx}}} = \frac{5.8990}{14.8333} = 1.5316$

$t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} = \frac{9.4719}{1.5316} = 6.184$

$df = 6-2 = 4$

$(0.001 < p < 0.0025) \times 2 = (0.002 < p < 0.005)$

There is a very strong evidence against H_0

Since p-value $\leq \alpha$, reject H_0.

>>>>>>>>>>

4.6 Confidence Interval for the Slope of the Population Regression Line

- The confidence interval for the slope follows the same general formula as for other t-procedures:

Confidence Interval: $Estimate \pm Critical\ Value \times SE(Estimate)$

Confidence Interval for the Slope of the Population Regression Line
[Regression t-Interval Procedure]

Purpose: To find a confidence interval for the slope, $\boldsymbol{\beta_1}$, of the population regression line

Assumptions: The four assumptions for regression inferences

Step 1: For a given confidence level (1 – α), use the t-table to find $t_{\alpha/2}$ in the row for the appropriate ***df,*** where ***df = n – 2***.

Step 2: The endpoints of the confidence interval for $\boldsymbol{\beta_1}$ are:

$$\hat{\beta}_1 \pm t_{\alpha/2} \times \frac{\hat{\sigma}}{\sqrt{S_{xx}}} \quad \Rightarrow \quad \hat{\beta}_1 \pm t_{\alpha/2} \times SE(\hat{\beta}_1)$$

Step 3: Interpret the confidence interval.

Example of Finding a Confidence Interval For the Slope
Calculate a 95% confidence interval for the slope of the regression line for the relationship between Effort Index and Chemistry Performance.

[Previously calculated: $Slope(\hat{\beta}_1) = 9.4719, S_{xx} = 14.8333, \hat{\sigma} = 5.8990$ (Standard error of the model)]

>>>>>>>>>>

For a 95% C.I. $\alpha = 0.05$

① $df = n-2 = 6-2 = 4$, $t_{\alpha/2} = t_{0.025} = 2.776$

$\hat{\beta}_1 \pm t_{\alpha/2} \times SE(\hat{\beta}_1)$

$9.4719 \pm 2.776 \times 1.5316$

$= (5.22, 13.72)$

∴ you get the idea for the conclusion.

>>>>>>>>>>

See Computer Output on page 13

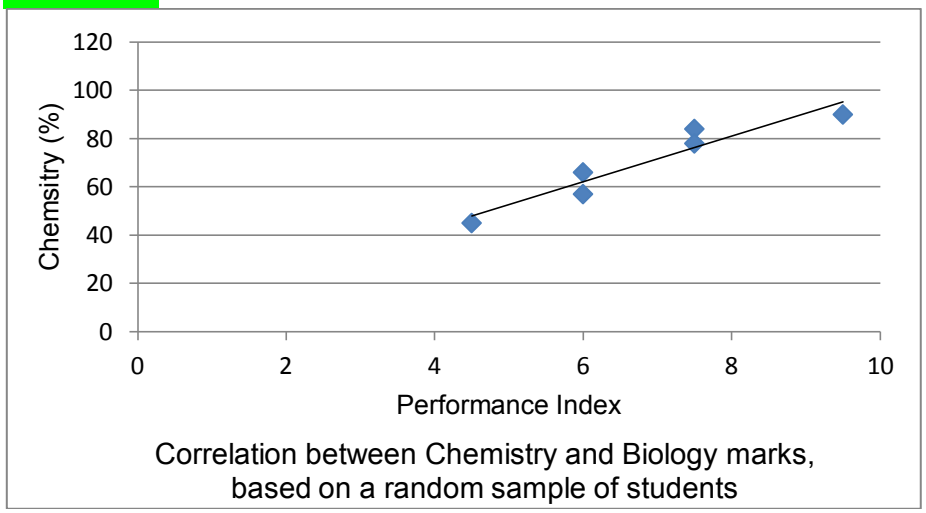
$$\hat{\sigma} = \sqrt{\frac{SS_{ERROR}}{n-2}} = \sqrt{\frac{139.19663}{6-2}} = \sqrt{MS_{ERROR}} = 5.8991$$

$$t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} = \frac{9.47191}{1.53167} = 6.184$$

Confidence interval (β_1) = (5.21931, 13.7245)

Sketch the confidence interval for the slope:

>>>>>>>>>>



Correlation between Chemistry and Biology marks, based on a random sample of students

>>>>>>>>>>

4.7 Confidence Intervals for Estimation of Mean Response and Predicted Response

Confidence Interval for Mean Response (or Conditional Mean)

- Used to estimate the confidence interval for the mean response of the response variable for a given value of the explanatory variable
- The predicted value of the response variable (y) for a given value of the predictor variable (x) can be determined by simply substituting that value of the given x into the regression equation.

Confidence Interval for the Mean response of *y* for a Given *x*
[Conditional Mean t-interval Procedure]

Purpose: To find a confidence interval for the mean response of the response variable for any given value (***x*_{*p*}**) of the predictor or explanatory variable

Assumptions: The four assumptions for regression inferences.

Step 1: For a given confidence level (1 – α), use the t-table to find $t_{\alpha/2}$ at ***df = n – 2***

Step 2: Compute the point estimate (the predicted value of response variable for a given value of the predictor variable) by using the linear regression equation:

$$\hat{y}_p = \hat{\beta}_0 + \hat{\beta}_1 x_p$$

Then, calculate the endpoints of the confidence interval by:

$$\hat{y}_p \pm t_{\alpha/2,n-2} \times \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}} \qquad \text{Where } S_{xx} = (n-1)s_x^2$$

Step 3: Interpret the confidence interval.

Prediction Interval

- A confidence interval for predicting all single observations of the response variable at a given value of the explanatory or predictor variable

Prediction Interval (OR Confidence Interval for the prediction of <u>all single observations</u> of the response of *y* for a Given *x*]

Purpose: To find a prediction interval for <u>all single observation responses</u> of the response variable for any given value (x_p) of the predictor or response variable

Assumptions: The four assumptions for regression inferences.

Step 1: For a given confidence level (1 – α), use the t-table to find $t_{\alpha/2}$ at ***df = n – 2***.

Step 2: Compute the point estimate (the predicted value of response variable for a given value of the predictor variable) by using the linear regression equation:

$$\hat{y}_p = \hat{\beta}_0 + \hat{\beta}_1 x_p$$

Then, calculate the endpoints of the confidence interval by:

$$\hat{y}_p \pm t_{\alpha/2,n-2} \times \hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_p-\bar{x})^2}{S_{xx}}}$$

Note also: $S_{xx} = (n-1)s_x^2$

Step 4: Interpret the confidence interval in terms of the research problem.

Example of Finding the Mean Response and Predicted Response

Age (years) x	Height (m) y
1	0.9
2	1.0
3	1.4
3	2.2
4	2.6
5	3.0
5	3.7

Given: $\hat{y} = -0.08087 + 0.66809x$

$S_{xx} = 13.42857$

$\hat{\sigma} = 0.39367$

$\bar{x} = 3.28571$

$\sum x = 23$

Find a 95% confidence interval for the mean height of all 3-year-old trees.
Also, find a 95% prediction interval for the height of a 3-year-old tree.

>>>>>>>>>>

for a 95% C.I, $\alpha = 0.05$ $df = n-2 = 7-2 = 5$ $t_{\alpha/2} = t_{0.025} = 2.571$

$\hat{y}_p = \hat{\beta}_0 + \hat{\beta}_1 x_p = 1.923$

$\hat{y}_p \pm t_{\alpha/2,n-2} \times \hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_p-\bar{x})^2}{S_{xx}}}$

$1.923 \pm 2.571 \times 0.39367 \times \sqrt{1+\frac{1}{7}+\frac{(3-3.28571)^2}{13.42857}}$

(0.838, 3.008) m

$S_x = 1.496026$

$S_{xx} = (n-1)S_x^2$

$= (7-1)(1.496026)$

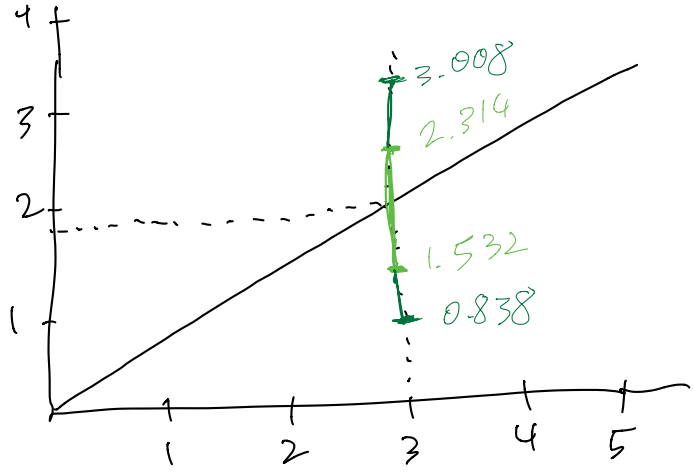
$= 13.42857$

1. we can be 95% confident that the mean height of 3-year old trees is somewhere between 1.532 and 2.314 m

2. any 3-year old tree be somewhere between 0.838 and 3.008 m

Comparison Of Confidence Intervals For the Mean Response and Predicted Response

Graph



>>>>>>>>>>

Note: The prediction interval is always wider than the confidence interval because:

- The estimate of a mean is always close to the population mean, whereas variation in all observed values is more disperse

4.8 Hypothesis Test for Linear Correlation

- A hypothesis test for the significance of the correlation between two quantitative variables
- This hypothesis test can be performed either:
  - When one variable can be identified as the explanatory variable (and regression can also be performed).
  - Or, when neither variable can be considered as the explanatory variable (and regression would not be performed)

Linear Correlation Hypothesis Test

Purpose: To perform a hypothesis test to decide whether two quantitative variables are significantly correlated.

Step 1: Select the appropriate test by checking purpose and assumptions

Step 2: State the null and alternative hypotheses

H_0: $\rho = 0$ (There is no correlation between the two variables.)

Just like regression, correlation can be a two-tailed, left-tailed or right-tailed test, so the alternative hypothesis may be:

Two-tailed: Ha: $\rho \neq 0$; Left-tailed: Ha: $\rho < 0$ (negative); Right-tailed: Ha: $\rho > 0$ (positive)

Note: Rho (ρ) is the Greek letter for population correlation coefficient

Step 3: Compute the correlation coefficient (***r***)

Using $r = \dfrac{s_{xy}}{s_x \times s_y}$ Where $s_{xy} = \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(n-1)}$

OR $r = \dfrac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left[\sum (x_i - \bar{x})^2\right]\left[\sum (y_i - \bar{y})^2\right]}}$

OR $r = \sqrt{\dfrac{SS_{REGR}}{SS_{TOTAL}}}$ **(then add correct sign, + or −)**

Step 4: Decide to <u>reject H_0 or not reject H_0</u> and state the <u>strength of the evidence</u> against H_0
df = n – 2 (where n = no. of ***xy*** observations)

Utilize the Table for the correlation coefficient, ***r***
If the P-value ≤ α, we <u>reject H_0</u> (otherwise do not reject H_0)

Step 5: Interpretation (conclusion) in words in terms of the research problem being investigated.

Correlation p-value (two-tailed)
p-value of the t-test (two-tailed)
p-value of the F-test

Example of Linear Correlation Hypothesis Test
At the 5% significance level, test whether there was a correlation between performance in Chemistry and Effort index (based on a random sample of 6 students).

Recall: We previously calculated the following (page 5)

Sum of Products of deviations of ***x*** and ***y***: $S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = 140.5$

Sum of Squares of deviations in ***x***: $S_{xx} = \sum (x_i - \bar{x})^2 = 14.8333$

Sum of Squares of deviations in ***y***: $S_{yy} = \sum (y_i - \bar{y})^2 = 1470$

>>>>>>>>>>

$H_0: \rho = 0$ i.e. no correlation

$H_A: \rho \neq 0$ i.e correlation

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{140.5}{\sqrt{14.833 \cdot 1470}} = 0.9515$$

$df = n - 2 = 6 - 2 = 4$

There is a strong evidence against H_0 and $\alpha < 0.05$

$\therefore$ reject

Note: The p-value for the correlation test is exactly the same as the p-value of the t-test for the slope and Regression ANOVA F-test.

>>>>>>>>>>

Comparative Examples Correlation Coefficients

1. An education researcher wanted to test (at α = 0.01) whether there is significant correlation between the amount of time per week that high school students spend watching TV and their academic performance. Upon analyzing data obtained from a random sample of 50 students, she found the correlation coefficient ***r*** = – 0.374. What P-value and conclusion did she obtain?

   (a) P < 0.001; significant correlation
   (b) 0.005 < P < 0.01; no significant negative correlation
   (c) P > 0.50; no significant correlation
   (d) 0.005 < P < 0.01; significant negative correlation
   (e) 0.02 < P > 0.01; significant correlation

2. The correlation coefficient between two variables was r = 0.903, sample size = 6. State the P-value and conclude whether the correlation is significant at the 1% significance level.

>>>>>>>>>>

$df = 6 - 2 = 4$, $0.01 < p < 0.02$

>>>>>>>>>>

Example on Biotechnology: Using the water fern Azolla to produce Hydrogen Fuel
[Example Combining All Concepts]

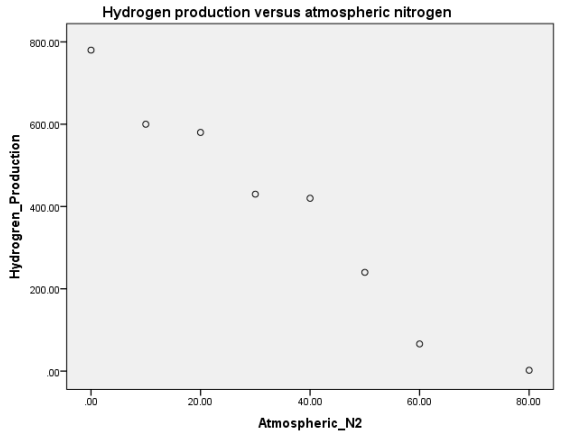
Use information in the Excel output and Residual plot below to answer questions 8 – 12:

Azolla is a water fern that fixes nitrogen and can be used as a biofertilizer on rice. However, experiments in biotechnology have shown that when *Azolla* is grown at reduced levels of atmospheric nitrogen, it produces hydrogen gas, a high energy, non-polluting fuel. Below is incomplete SPSS output showing regression analysis of data from an experiment in which *Azolla* was exposed to atmospheric N_2 levels ranging from 80% to 0% and the production of H_2 (in nmol H_2 g^{-1} fresh weight $hour^{-1}$) was measured.

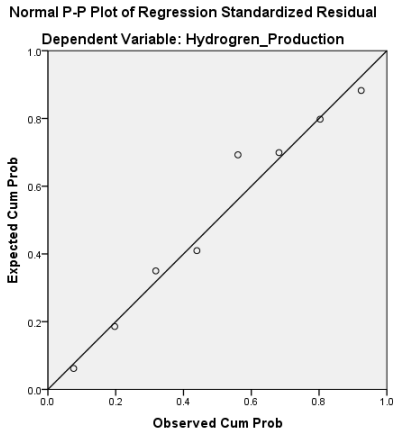
Consider also that n = 8, $\bar{x} = 36.25$ and $S_{XX} = 4{,}987.5$.

Atmospheric nitrogen (N_2)	Hydrogen Production (nmol H_2 g^{-1} fresh weight $hour^{-1}$)
80.00	2.00
60.00	66.00
50.00	240.00
40.00	420.00
30.00	430.00
20.00	580.00
10.00	600.00
0.00	780.00

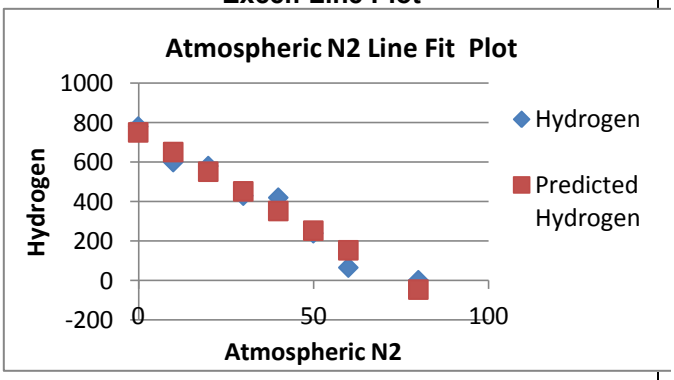
SPSS: xy graph



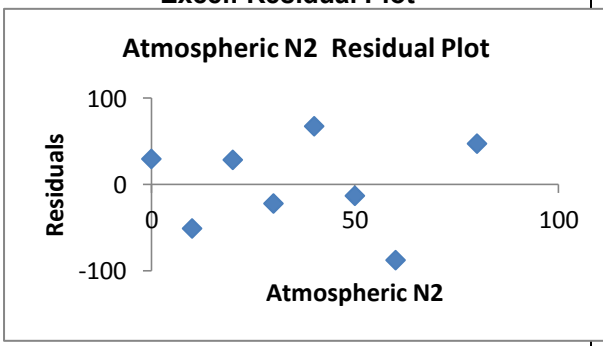
SPSS: Normal Probability Plot



Excel: Line Plot



Excel: Residual Plot



Multiple R

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.981[a]	.962	.956	56.86996

a. Predictors: (Constant), Atmospheric_N2

b. Dependent Variable: Hydrogren_Production

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	493414.347	1	493414.347	152.562	.000[b]
	Residual	19405.153	6	3234.192		
	Total	512819.500	7			

a. Dependent Variable: Hydrogren_Production

b. Predictors: (Constant), Atmospheric_N2

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	750.306	35.446		21.168	7.25E-07	663.574	837.038
	Atmospheric_N2	-9.946	.805	-.981	-12.352	1.72E-05	-11.917	-7.976

a. Dependent Variable: Hydrogren_Production

Suppose the numbers highlighted in yellow in the table above were not given

>>>>>>>>>>

(a) According to the regression model what would you predict to be the rate of the hydrogen production at 25% atmospheric nitrogen? Would this be a reliable estimate?

$\hat{Y} = 750.306 - 9.946(x)$

$= 750.306 - 4.946(25)$

$= 501.656$

The predicted rate of H_2 production @ 25% atmos. N_2 is 501.656...

This would be a reliable estimate since 25% is with the Obs. range of x (interpolation)

(b) What was the residual (error) of this regression model at an atmospheric N_2 level of 60%?

$\hat{Y} = 750.306 - 9.946(60)$

$\varepsilon = Y_i - \hat{Y}$

$66 - \hat{Y}$

$66 - 153.546 = -87.546$ nmol $H_2\ y^{-1}$ fresh wt/h

(c) Calculate the linear correlation coefficient for the relationship between atmospheric N_2 and hydrogen production. What is the exact P-value?

$SS_T = SS_R + SS_E$

$= 493,414,397 + 19,405,153 = 512,819,500$

$R^2 = \frac{SS_R}{SS_E} = \frac{493,414,397}{19,405,153} = 0.96216$

$r = -\sqrt{R^2} = -\sqrt{0.96216} = -0.981$

Exact p-value is 1.72×10^{-5}

note: this is a two-tailed test if doing a one-tailed test you would divide the p-value.

(d) What is the standard error of the model?

$\hat{\sigma} = \sqrt{MS_E} = \sqrt{\frac{SS_E}{n-2}} = \sqrt{\frac{19405.153}{8-2}} = 56.86996$

(e) What percentage of variability in hydrogen production is explained by the level of atmospheric N_2?

$R^2 = 0.96216$ calculated in part c)

96.22% of the variability in h_2 production is explained by the level of the atmos N_2 or by the regression model.

(f) At the 1% significance level, test the hypothesis that there is a negative relationship between atmospheric N_2 and hydrogen production. Carry out the most appropriate, showing all steps (give both the exact P-value and the value from the table).

$H_0: \beta_1 = 0$ there is no relationship between N_2 and H_2 production

$H_A: \beta_1 \neq 0$ there is a relationship " " " " "

$t = \frac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{S_{xx}}} = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} = \frac{-9.946}{0.805} = -12.355$

$df = n-2 = 8-2 = 6$

$P < 0.0005$ exact p-value is $\frac{1.72 \times 10^{-5}}{2} = 0.0000086$

(g) What are the value of the *F*-statistic and the P-value (both the exact value and the value from the table?

$F = t^2 = (-12.3516)^2 = 152.562$

$df(1, n-2) = (1, 8-2) = (1, 6)$

$p < 0.001$

$p = 0.0000172$

(h) Find the margin of error for a 99% confidence interval for the expected value of hydrogen production at 20% atmospheric nitrogen.

for a 99% C.I, $\alpha = 0.01$

$df = n-2, 8-2 = 6$ $t_{\alpha/2} = t_{0.005} = 3.707$

$ME = t_{\alpha/2} \times \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}}$

$= 3.707 \times 56.86996 \sqrt{\frac{1}{8} + \frac{(20 - 36.25)^2}{4987.5}}$

$ME = 88.93$

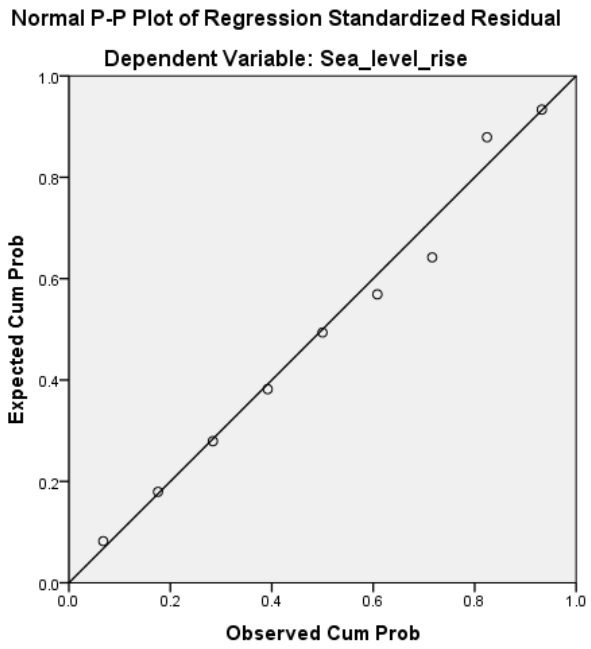
The ME for a 99% C.I. for the expected value of H_2 production @ 20% atmos n_2 is 88.93 nmol H_2/hour.

>>>>>>>>>>

Example on Sea Level Rise in Seychelles (Combing all concepts in the section)

A marine biologist used a tidal gauge to monitor mean annual sea level rise (in *mm*) in Seychelles for a period of 9 years from 2001 to 2009, using the year 2000 as baseline. The data fit the required assumptions. The table below shows incomplete SPSS output from regression analysis. Perform all calculations assuming that the numbers highlighted in yellow are not given.

Year	Sea level rise (mm) above 2000 baseline
2001	2
2002	5
2003	5
2004	9
2005	14
2006	16
2007	16
2008	18
2009	22



Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.982[a]	.964	.958	1.405

a. Predictors: (Constant), Year

b. Dependent Variable: Sea_level_rise

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	365.067	1	365.067	184.881	.000[b]
	Residual	13.822	7	1.975		
	Total	378.889	8			

a. Dependent Variable: Sea_level_rise

b. Predictors: (Constant), Year

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-4933.778	363.730		-13.564	.000	-5793.862	-4073.693
	Year	2.467	.181	.982	13.597	.000	2.038	2.896

a. Dependent Variable: Sea_level_rise

(a) At the 1% significance level, test whether there is a relationship between time (in years) and mean annual sea level rise in Seychelles. In other words, test for the significance of the slope of the regression line.

H_0: $\beta_1 = 0$ (There is no relationship between time (in years) and mean annual sea level rise)
H_a: $\beta_1 \neq 0$ (There is a relationship between time (in years) and mean annual sea level rise)

$$t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} = \frac{2.467}{0.181} = 13.630$$

df = n – 2 = 9 – 2 = 7
P-value = 2 X (P < 0.0005) ==> P < 0.001
Since P ≤ α (0.01), we reject Ho with extremely strong evidence

Interpretation: At the 1% significance level, the data provide sufficient evidence to conclude that the slope of the population regression line is not 0 and therefore there is a relationship between time (in years) and mean annual sea level rise.

(b) Find a 95% confidence interval for the slope of the regression line that relates mean annual sea level rise to time (in years). Use this confidence interval to determine whether the slope is significant.

>>>>>>>>>> $\alpha = 0.05$ at $df = n-2 = 7-2 = df = 5$ $t_{0.025,7} = 2.365$

$\hat{\beta}_1 \pm t^* \times SE(\hat{\beta}_1)$

$2.467 \pm 2.365 \times 0.181$

$(2.039, 2.895)$

We can be 95% ... slope of the regression line is somewhere between 2.039 and 2.895

Since 0 isn't in the C.I, there is significance.

(c) Calculate a 95% confidence interval for the effect of an additional 5 years on the average sea level.

The addition of 5 years is $5 \cdot \beta_1$

$5\hat{\beta}_1 \pm t_{0.025,7} \times 5SE(\hat{\beta}_1)$

$\therefore$

$(10.20, 14.48)$

>>>>>>>>>>

(d) What is the standard error of the model?

$$SS_{Error} = SS_{TOTAL} - SS_{REGR} = 378.889 - 365.067 = 13.822$$

$$\hat{\sigma} = \sqrt{MS_{ERROR}} = \sqrt{\frac{SS_{Error}}{n-2}} = \sqrt{\frac{13.822}{9-2}} = \sqrt{1.9746} = 1.405$$

Rise in Global Mean Sea Level

The change in global mean sea level per decade (10 years) from 1930 to 2010 (considering 1930 as baseline year 0, thus covering 8 decades), can be described by the following linear equation: $\hat{y} = 20.07 + 18.35x$. The correlation coefficient relating time and sea level is: r = 0.988703. [Years were coded as decades.]

>>>>>>>>>>

(a) What percentage of variability in global mean sea level is explained by time?

(oeff of determination is R^2

$r^2 = 0.98703^2 = 0.97753$

Thus 97.75% of variability ...

(b) What is the value of the F-statistic for determining whether the relationship between time and sea level is significant?

$$F = \frac{SS_R \times (n-2)}{(SS_T - SS_R)/SS_T} = \frac{(n-2)SS_R/SS_T}{(SS_T - SS_R)/SS_T}$$

$$= \frac{(n-2)R^2}{1-R^2}$$

$$= \frac{(8-2)(0.97753)}{1-0.97753}$$

$$= 261.022697$$

(c) What is the standard error of the slope of the regression line?

$$t = \sqrt{F} = \sqrt{261.022697} = 16.1562$$

$$t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} \Rightarrow SE(\hat{\beta}_1) = \frac{\hat{\beta}_1}{16.1562} = 1.1361$$

>>>>>>>>>>

4.9 More on Assumptions and Transformations of Data

Variations in notation for the linear regression equation

$$\hat{\mu}\{Y \mid X) = 3.707 - 0.012X$$

Same as: $\hat{y} = 3.707 - 0.012x$

Assumptions (Conditions) for Regression Inferences and Robustness

1. **Linearity:**
   Checking: Scatterplot or residual plot

   Robustness: If no linear relationship exists in the data, don't use linear regression!

   Solution: Consider transformation. If the problem is only with linearity (not with equal standard deviations), try transforming ***x***. If both problems exist, transform ***y*** first.

2. **Equal standard deviations:**

   Checking: Scatterplot or residual plot (but residual plot is best)

   Robustness: The consequences for violating this assumption are critical. When there is lack of equal variability, the resulting standard errors inaccurately estimate their respective parameters, thus confidence intervals and hypothesis tests can be misleading.

   Solution: Consider transformation of ***y***.

3. **Normal populations:**

   Checking: Normal Probability Plot (Look for an increasing linear pattern)

   Robustness: Standard errors are robust to non-normal distributions (*t*-tools). The consequences of violating this assumption are usually minor. The only situation of large concern is when the distributions have long tails, outliers are present, and/or sample sizes are small. In terms of prediction, normality is critical.

   Solution: Consider transformation of ***y***, only if the problem is very serious.

4. **No Serious Outliers:**

   Checking: Scatterplot or Normal Probability Plot

   Robustness: Several Significant outliers can drastically change the regression model (just as for correlation)

   Solution: Consider transformation of ***y***, if there are several serious outliers and sample size is small.

5. **Independent observations:** The observations of the response variable are independent of one another. This implies that the observations of the predictor variable not need to be independent.

   Checking: Evaluate the research design.

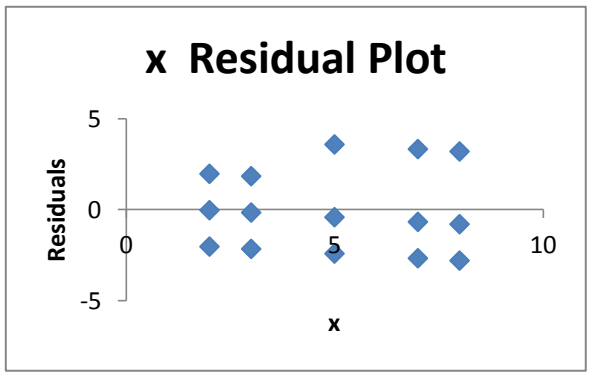
   Robustness: Lack of independence causes inaccurate standard errors.

   Solution: Difficult to solve unless you revise the research design.

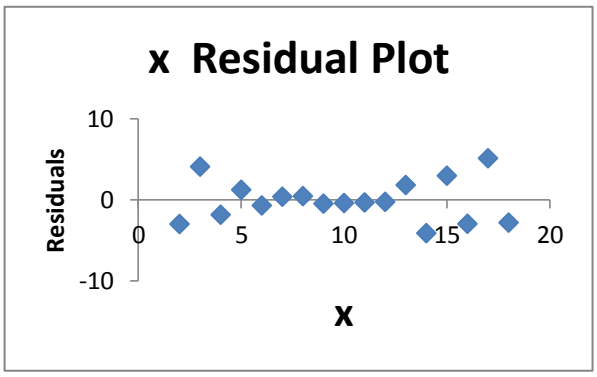
Checking Assumptions with Scatterplots, Residual Plots and Normal Probability Plots

Which assumption or assumptions are violated in the data sets below?

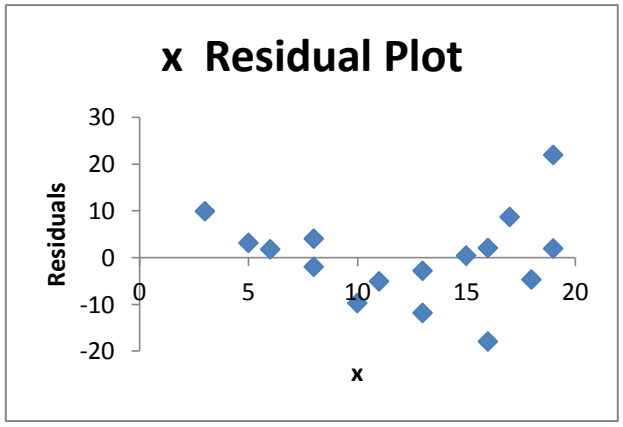
Example 1



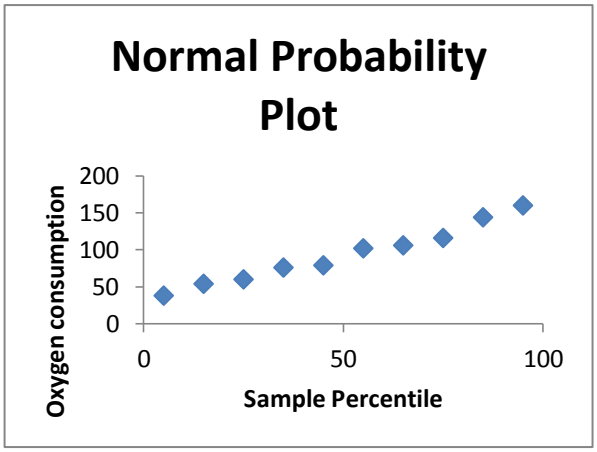
Example 2



Example 3



Example 4



Data for Example 3 above:

x	3	5	6	8	8	10	11	13	13	15	16	16	17	18	19	19
y	4	4	6	9	15	8	16	16	25	35	20	40	50	40	50	70

>>>>>>>>>>

Ex 1: no serious violations

Ex 2: only equal std dev violated

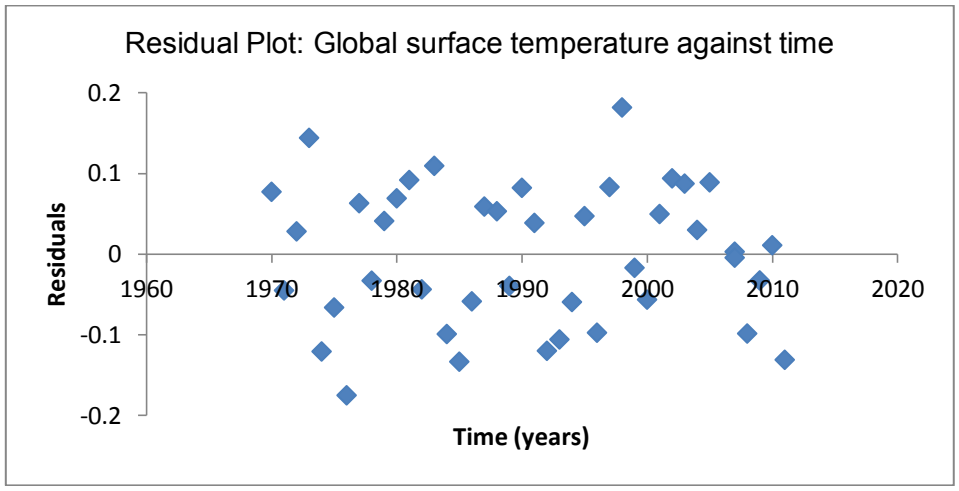
Ex 3: linearity and equal std dev and outliers are violated

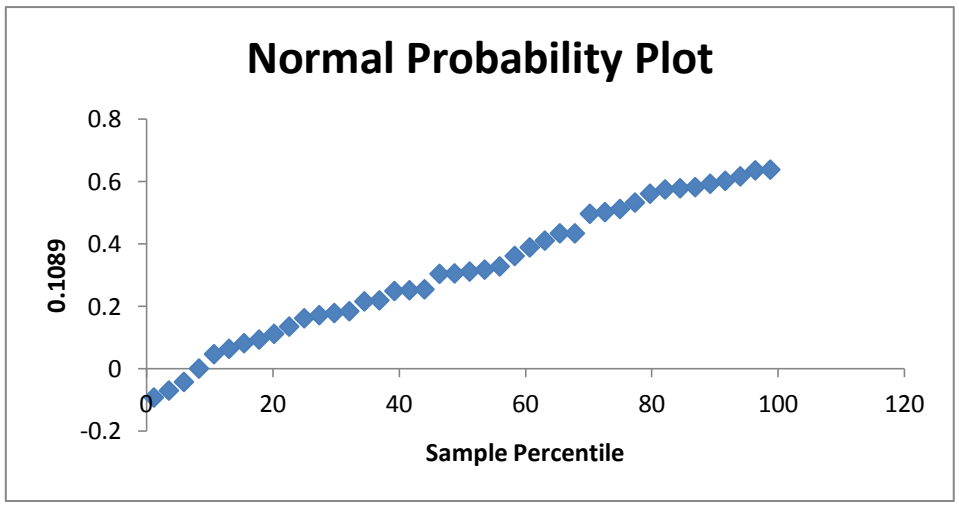
Ex 4: data is normal but normality plot cannot be used to test other violations.

>>>>>>>>>>

Checking the Assumptions of Regression for the Data on Global Surface Temperature Against Time (1969 – 2011)

[Examine the NOAA Graph: 1880 – 2011]





Conclusion:

- The data points in the residual plot fall roughly in a horizontal band centered and symmetric about the x-axis
- The normal probability plot is approximately linear
- Therefore, these data fit all the assumptions of regression analysis

Interpretation of Model Effects in SLR after Log Transformation

***SLR model of Blood Pressure on Age*:** The effect of age on mean blood pressure is measured as the change in the mean blood pressure associated with a 1-unit (one year) increase in age. This effect is measured as β_1 in the model below.

$\mu(bp \mid age) = \beta_0 + \beta_1 age$

This is known as an additive effect.

IN GENERAL, for *Y* vs. *X*:

Additive change of *k* units in *X* → *Additive* change of $k\beta_1$ in the mean of *Y*.
(*X* + *k*)

Suppose (in separate circumstances) that the following natural log transformations were required. Interpretation of the model effect on the original scale will follow.

most common →

i) a natural log transformation was used on the response variable only. (ln *Y* vs. *X*)
ii) a natural log transformation was used on the predictor variable only. (*Y* vs. ln *X*)
iii) a natural log transformation was used on both variables. (ln *Y* vs. ln *X*)

***Case i)*: (ln *Y* vs. *X*)**

IN GENERAL, for ln *Y* vs. *X*:

Additive change of *k* units in *X* → *Multiplicative* change of $e^{k\beta_1}$ in the median of *Y*. (Take antilog of the slope)

k = final − initial

Suppose we had this model:

Model: $\mu(\ln(bp) \mid age) = \beta_0 + \beta_1 age$ → $\hat{\mu}(\ln(bp) \mid age) = 4.481 + 0.010 age$

The additive effect of *age* on $\mu_{\ln(bp)}$ is β_1.

Back Transforming to the Original Scale

Example 1: the additive effect of 1-year (*k* = 1) in *age* is associated with a multiplicative effect of $e^{1(\beta_1)} = e^{\beta_1}$ on *Median*(*bp*). It is estimated that a 1-year increase in *age* is associated with a multiplicative change of $e^{0.010}$ = 1.01 in *Median*(*bp*). In other words, the median *bp* at *age* + 1 is estimated to be 1.01 times (1.01 – 1 = 0.01 ==> 1% higher than) the median *bp* at the given *age*.

Example 2: It is estimated that a 5-year increase in *age* is associated with a multiplicative change of $e^{5(0.010)}$ = 1.051 in *Median*(*bp*). In other words, the median *bp* at *age* + 5 is estimated to be 1.051 times (1.051 – 1 = 0.051 ==> 5.1% higher than) the median *bp* at the given *age*.

Example 3: The median blood pressure of 55-year-olds will be 1.22 times (1.22 – 1 = 0.22 ==> 22% higher than) the median blood pressure of 35-year-olds ($e^{20(0.010)}$ = 1.22).

[Likewise for confidence intervals, take the antilog of the two endpoints.]

***Case ii)*: (*Y* vs. ln *X*)**

IN GENERAL, for *Y* vs. ln *X*:

Multiplicative change by a factor of *k* in *X* (*X* × *k*) → *Additive* change of $\beta_1\ln(k)$ in the mean of *Y*.

$k = \frac{\text{final}}{\text{initial}}$

Suppose we had this model:

Model: $\mu(bp \mid \ln(age)) = \beta_0 + \beta_1\ln(age)$ → $\hat{\mu}(bp \mid \ln(age))$ = -81.784 + 58.967ln(*age*)

The additive effect of ln(*age*) on μ_{bp} is β_1.

Example 1: A multiplicative change in *age* by a factor of *k* is associated with an additive change of ($\beta_1\ln(k)$) in μ_{bp}. It is estimated that a multiplicative change in *age* by a factor of *k* is associated with an additive change of 58.967ln(*k*) in μ_{bp}.

Example 2: It is estimated that aging from 40 to 50 (*k* = 50/40 = 1.25) is associated with an additive increase in μ_{bp} of 58.967(ln(1.25)) = (58.967)(0.22314) = 13.16.

Example 3: It is estimated that aging from 28 to 35 (*k* = 35/28 = 1.25) is associated with an additive increase in μ_{bp} of 58.967(ln(1.25)) = (58.967)(0.22314) = 13.16.

Example 4: It is estimated that aging from 30 to 50 (*k* = 50/30 = 1.67) is associated with an additive increase in μ_{bp} of 58.967(ln(5/3)) = (58.967)(0.5108) = 30.12.

***Case iii)*: (ln *Y* vs. ln *X*)**

IN GENERAL, for ln *Y* vs. ln *X*:

Multiplicative change by a factor of *k* in *X* (*X* × *k*) → *Multiplicative* change of k^{β_1} in the median of *Y*.

$k = \frac{\text{final}}{\text{initial}}$

Suppose we had this model:

Model: $\mu(\ln(bp) \mid \ln(age)) = \beta_0 + \beta_1\ln(age)$ → $\hat{\mu}(\ln(bp) \mid \ln(age))$ = 3.332 + 0.426ln(*age*)

The additive effect of ln(*age*) on $\mu_{\ln(bp)}$ is β_1.

Example 1: A multiplicative change in *age* by a factor of *k* is associated with a multiplicative change of k^{β_1} in *Median*(*bp*). It is estimated that a multiplicative change in *age* by a factor of *k* is associated with a multiplicative change of $k^{0.426}$ in *Median*(*bp*).

Example 2: It is estimated that aging from 40 to 50 (*k* = 50/40 = 1.25) is associated with a multiplicative change of $1.25^{0.426}$ = 1.0997 in *Median*(*bp*). That is, the Median blood pressure will be 1.0997 times (1.0997 – 1 = 0.0997 ==> 9.97% higher) what it was at the age of 40.

Example 3: It is estimated that aging from 35 to 55 (*k* = 55/35 = 1.57) is associated with a multiplicative change of $1.57^{0.426}$ = 1.212 in *Median*(*bp*), that is there is a 21.2% increase (1.212 – 1 = 0.212).

Example on Log Transformed Data [Like Case (i) described above]
Suppose the relationship between the annual rate of hip fractures (per 100,000 people) and age follows the following model: $\hat{\mu}(\ln(fractures)\,|\,age) = -2.09 + 0.0912\ age$. Suppose further that a 98% confidence interval for the slope based on the logged data is (0.0723, 0.1101)

(a) For an increase in age from 40 to 50 years old, what would be your interpretation regarding the rate of hip fractures on the original scale?

>>>>>>>>>>

$e^{10\beta_1} = e^{10(\beta_1)} = e^{10(0.0912)} = 2.489$

The median rate of hip fractures / 100,000 ppl. at 50 years of age will be 2.489 times the median rate of hip fractures at 40 years.

(b) What is the estimated rate of hip fractures on the original scale for people that are 80 years old?

$\hat{\mu}(\ln(\text{fractures}) \mid \text{age}) = -2.09 + 0.0912(80) = 5.206$

$e^{5.206} = 182.363$

It is estimated that the rate of hip fractures for ppl. that are 80 years old is on average 18

(c) Interpret the 98% confidence interval for the slope on the original scale. Also, based on this confidence interval back transformed to the original scale, would you conclude that there is a relationship between the annual rate of hip fractures and age? Explain your answer.

Take the antilog

$(e^{0.703}, e^{0.1101}) = (1.075, 1.116)$

It is estimated with 98% confidence that a 1 year increase in age is associated with a multiplicative change of 1.075 to 1.116 in the median rate of hip fractures per 100,000.

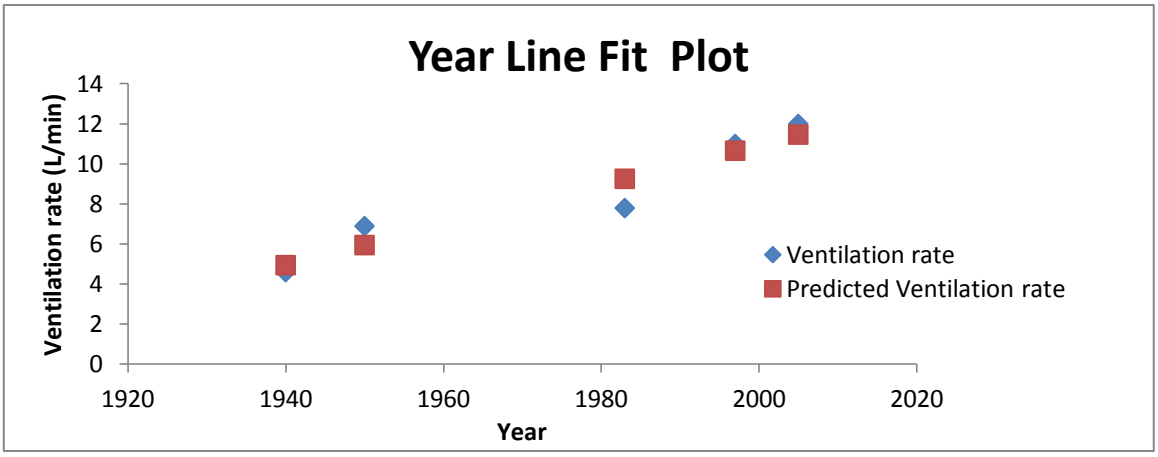
The slope is sig. as 1 isn't in the C.I.

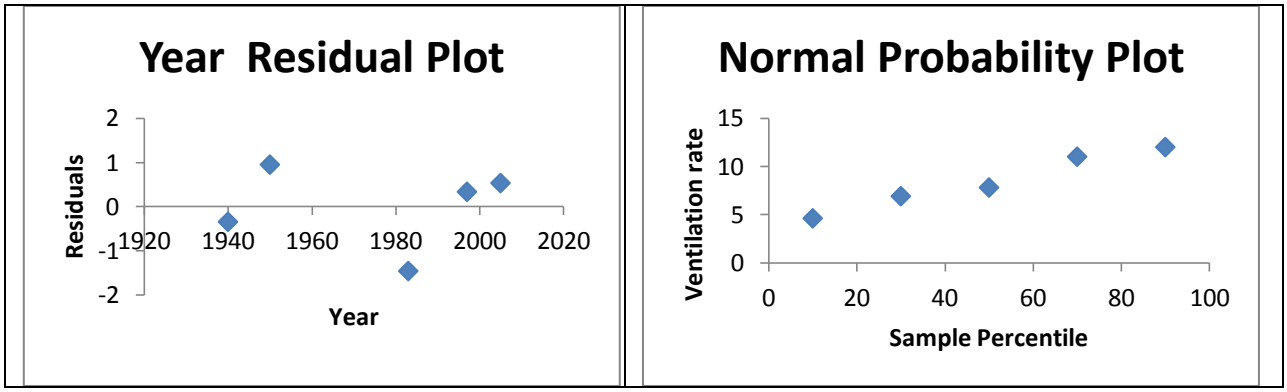
>>>>>>>>>>

Practice Question on Inferences for Regression: Changes in Human Ventilation Rate over Time
A group of medical researchers suspected that various growing health issues such as lack of fitness, obesity and increasing prevalence of diseases such as cancer, diabetes and heart disease may have caused ventilation rates in humans to change over the past few decades. Examination of past records provided data as shown below, using 1940 as a baseline. At the 5% significance level, test whether the data provide sufficient evidence to conclude that there has been significant change in human ventilation rates over time.

(a) Check whether the data fit the assumptions of regression analysis.

Year	Average human ventilation rate (L/min)
1940	4.6
1950	6.9
1983	7.8
1997	11
2005	12
$\bar{x} = 1975$	
$s_x = 28.714108$	





Conclusion regarding the assumptions: These graphs show that the data approximately fit the linear regression model. Though there are few data points, the residual plot (right) shows no noticeable violation of equal standard deviations and linearity, while the normal probability plot shows that the data are approximately normally distributed.

Regression Statistics	
Multiple R	0.950454
R Square	0.903362
Adjusted R Square	0.87115
Standard Error	1.088061
Observations	5

ANOVA table for Regression analysis					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	33.20037	33.20037	28.04375	0.01314
Residual	3	3.55163	1.183878		
Total	4	36.752			

***t*-test for the Significance of the Slope**						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-189.699	37.42242	-5.06912	0.014823	-308.794	-70.6039
Year	0.100334	0.018946	5.295635	0.01314	0.040037	0.16063

Note: Suppose the numbers highlighted in yellow are missing values in the table.

Linear Correlation Coefficient (r) = 0.950

(a) Give the regression equation and interpret the meaning of the slope and the y-intercept in terms of the research problem.

The regression equation is: $\hat{y} = -189.699 + 0.1003x$

The meaning of the slope is that ventilation rates in humans have increased, on average, by 0.1003 L/min per year from 1940 to 2005 (or 1.003 liters per decade).
The meaning of the y-intercept is WOW!!!!

(b) At the 5% significance level, test for the significance of the slope. In other words, determine if there has been a significant change in human ventilation rates over time.

Step 1: Regression t-test is selected because the purpose is to test of the slope is significantly different from 0.

Step 2:
H_0: β1 = 0 (There has been no significant change in human ventilation rates over time or there is no relationship between time and human ventilation rates.)
H_a: β1 ≠ 0 (There has been significant change in human ventilation rates over time or there is a relationship between time and human ventilation rates.)

Step 3:

Standard Error of the Slope:

Given in the computer output as: $SE(\hat{\beta}_1) = 0.018946$

The Regression t-statistic:

$$t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} = \frac{0.100334}{0.018946} = 5.296$$

Step 4: Decide to reject or not reject Ho

$df = n - 2 = 5 - 2 = 3$

So, P-value is between (0.005 and 0.01) x 2: 0.01 < P < 0.02
Since P-value ≤ α, reject H_0 with strong evidence.

Step 5: At the 5% significance level, the data provide sufficient evidence to conclude that there has been significant change in human ventilation rates over time.

Alternate Method

The significance of the slope could likewise be tested with ANOVA as follows:

$$F = \frac{SS_{REGR}/1}{SS_{Error}/(n-2)} = \frac{33.20037/1}{3.55163/3} = \frac{33.20037}{1.183878} = 28.04$$

At $df = (1, n-2) = (1,3)$, P-value: 0.01 < P < 0.025

[**Note:** Exact P-value = 0.01314 for both tests]

(c) Calculate a 95% confidence interval for the slope of the regression line.

At the 95% confidence level and df = 5 – 2 = 3, $t_{\alpha/2} = 3.182$

$$\hat{\beta}_1 \pm t_{\alpha/2} \times SE(\hat{\beta}_1)$$

$$0.100334 \pm 3.182 \times 0.018946$$

0.100334 ± 0.060286

Or 0.040 to 0.161 L/min per year

Interpretation: We can be 95% confident that the increase in human ventilation rates over time is somewhere between 0.040 and 0.161 L/min per year.

(d) Calculate a 95% confidence interval for the mean response of variable y (human ventilation rate) in 1990.

At the 95% confidence level and df = 5 – 2 = 3, t_{Crit} = 3.182

The point estimate of variable y in 1990 was:

$$\hat{y}_p = \hat{\beta}_0 + \hat{\beta}_1 x_p$$

$$\hat{y} = -189.699 + 0.1003(1990) = 9.966 \text{ L/min}$$

The endpoints are given by:

$$\hat{y}_p \pm t_{\alpha/2,n-2} \times \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}}$$

Three Methods of Finding S_{xx}

Method 1:

$$SE(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{S_{xx}}} \Rightarrow \sqrt{S_{xx}} = \frac{\hat{\sigma}}{SE(\hat{\beta}_1)}$$

$$\Rightarrow S_{xx} = \left(\frac{\hat{\sigma}}{SE(\hat{\beta}_1)}\right)^2 = \left(\frac{1.08806}{0.018946}\right)^2 = 3298.15 = 3298$$

Method 2:

$$S_{xx} = (n-1)s_x^2$$
$$= (5-1)(28.714108)^2 = 3298.0000$$

Method 3:

$$S_{xx} = \sum (x_i - \bar{x})^2$$

[Calculate from raw data]

$$9.966 \pm 3.182 \times 1.08806\sqrt{\frac{1}{5} + \frac{(1990-1975)^2}{3298}}$$

9.966 ± 1.793

Or 8.173 to 11.759 L/min

Interpretation: We can be 95% confident that the mean ventilation rate of humans at 30 years after baseline was somewhere between 8.173 and 11.759 L/min.

SECTION 5: MULTIPLE REGRESSION ANALYSIS

5.1 The Multiple Linear Regression Model

- Multiple Linear Regression develops a model where there is only one response variable (*y*), but more than one explanatory or predictor variables ($x_1, x_2, ..., x_k$)

- The general model for multiple linear regression is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon$$

Where,

- *y* is the response variable
- $x_1, x_2, ..., x_k$ are the explanatory variables
- $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$ is the deterministic part of the model
- β_i determines the contribution of the explanatory variable x_i to the model
- ε is the random error, which is assumed to be normally distributed with mean 0 and standard deviation σ

- When the least squares criterion is applied this leads to the general model for the <u>population</u> multiple linear regression equation as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$

Or

$$\mu(y \mid x_1, x_2, ..., x_k) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$

- The general formula for the <u>sample</u> multiple linear regression equation is:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_k x_k$$

Or

$$\hat{\mu}\left(y \mid x_1, x_2, ..., x_k\right) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_k x_k$$

- The y-intercept ($\hat{\beta}_0$) is the value of *y* when all explanatory variables have a value of 0 ($x_1=0, x_2=0, ..., x_k=0$).
- The values $\hat{\beta}_1, +\hat{\beta}_2, +... + \hat{\beta}_k$ are referred to as **partial slopes** or **partial regression coefficients**
- Each $\hat{\beta}_i$ tells us the change in *y* per unit increase in *x*, holding <u>all other explanatory variables constant</u>

5.2 Inferences Concerning the Overall Usefulness of the Multiple Regression Model

Assumptions for Multiple Regression Inference

Assumptions (Conditions) for Regression Inferences

1. **Linearity of the population regression line:** The relationship between the variables as described by the population regression equation $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$ must be approximately linear.
2. **Equal standard deviations (homoscedasticity):** The standard deviations of ***y***-values must be approximately the same for all sets of values of $x_1, x_2, ..., x_k$
3. **Normal populations:** For each set of values of $x_1, x_2, ..., x_k$, the corresponding ***y***-values must be normally distributed
4. **No Serious Outliers:** Significant outliers can drastically change the regression model
5. **Independent observations:** The observations of the response variable are independent of one another. This implies that the observations of the predictor variable not need to be independent.

Note: All assumptions (except independence) can be checked graphically.

Regression Identity for Multiple Linear Regression

Regression Identity:

$$SS_{TOTAL} = SS_{REGR} + SS_{ERROR}$$

Regression Identity for Degrees of Freedom:

$$df(SS_{TOTAL}) = df(SS_{REGR}) + df(SS_{ERROR})$$

Or $$n - 1 = k + (n - (k + 1))$$

Where *n* is sample size and *k* is the number of predictor variables

- If the sample multiple linear regression equation fits the data well, then the observed values and predicted values of the response variable (based on the regression model) will be "close" together
- AND thus, SS_{ERROR} will be small relative to SS_{TOTAL} and SS_{REGR} will be large relative to SS_{TOTAL}

Overall usefulness or significance of the multiple regression model can be determined by:

1. Multiple regression ANOVA F-test
2. Multiple R (Multiple correlation coefficient)
3. Coefficient of multiple determination

Multiple Regression ANOVA Test (F-Test)

Multiple Regression ANOVA Test (F-Test)

Purpose: To test whether a multiple linear regression model is useful for making predictions

Assumptions: The assumptions shown above

Step 1: Selection of the test based on the purpose and assumptions

Step 2: The null and alternative hypotheses are:

$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$

H_a : At least one of the slopes β_i s is not zero

Step 3: Obtain the three sums of squares (SS_{TOTAL}, SS_{REGR} and SS_{ERROR}) and Compute the calculated value of the F-statistic

ANOVA Table for Multiple Linear Regression

Source of variation	SS	*df*	MS = SS/df	F-statistic
Regression	SS_{REGR}	*k*	$MS_{REGR} = \frac{SS_{REGR}}{k}$	$F = \frac{MS_{REGR}}{MS_{ERROR}}$
Error (Residual)	SS_{ERROR}	*n – (k+1)*	$MS_{ERROR} = \frac{SS_{ERROR}}{n-(k+1)}$	
Total	SS_{TOTAL}	*n - 1*		

$$F = \frac{SS_{REGR} / k}{SS_{ERROR} / (n-(k+1))} = \frac{MS_{REGR}}{MS_{ERROR}}$$

Step 4: Decide to reject or not reject Ho

***df* =** (numerator degrees of freedom, denominator degrees of freedom)

$df = (k, n-(k+1))$

(Where ***n*** = no. of ***xy*** observations and ***k*** = the number of predictor variables)

If P-value ≤ α, reject H_0

Step 5: Conclusion in terms of the research problem

Note: Recall that, in general, in simple linear regression, the Regression ***df*** is the number of coefficients (y-intercept + slope) being estimated minus 1, that is 2 – 1 = 1. For multiple linear regression, the coefficients are the y-intercept plus the slopes of k predictor variables, that is, there are *1 + k* coefficients. Thus, Regression ***df*** *= (1 + k) – 1 = k*

Multiple R (Multiple Correlation Coefficient)

- Measures the overall correlation between the all the variables involved in the model
- Multiple $R = +\sqrt{R^2}$ (see below)

Coefficient of Multiple Determination

Coefficient of multiple determination (R^2) = [multiple correlation coefficient]^2
[Also called **Multiple R^2**]

= the fraction or percentage of variation in the observed values of the response variable that is accounted for by the regression analysis involving more than one explanatory variable

$$R^2 = \frac{\text{Explained variability}}{\text{Total variability}}$$

$$R^2 = \frac{SS_{REGR}}{SS_{TOTAL}} = 1 - \frac{SS_{Error}}{SS_{TOTAL}} = \frac{SS_{TOTAL} - SS_{Error}}{SS_{TOTAL}}$$

$0 \le R^2 \le 1$ OR $0\% \le R^2 \le 100\%$

This implies that $1 - R^2$ of the variation in the observed values of the response variable are accounted for by other factors, not the explanatory variable used in the regression analysis

Adjusted Coefficient of Determination

- If the sample size equals the number of parameters (regression coefficients), then $R^2 = 1$, which can give the impression that the estimated model is a good fit of the population regression model, even when the estimated model may actually may not give an accurate representation of the real population model.
- Therefore, the adjusted R^2 is a more accurate measure of the fit of the model

Adjusted Coefficient of Determination

$$R^2_{adj} = 1 - \frac{MS_{ERROR}}{MS_{TOTAL}}$$

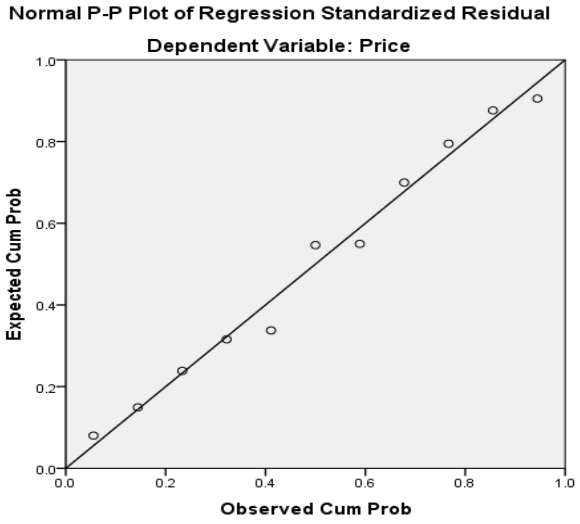
$$R^2_{adj} = 1 - \frac{\frac{SS_{ERROR}}{(n-(k+1))}}{\frac{SS_{TOTAL}}{(n-1)}} = 1 - \frac{(n-1)SS_{ERROR}}{(n-(k+1))SS_{TOTAL}}$$

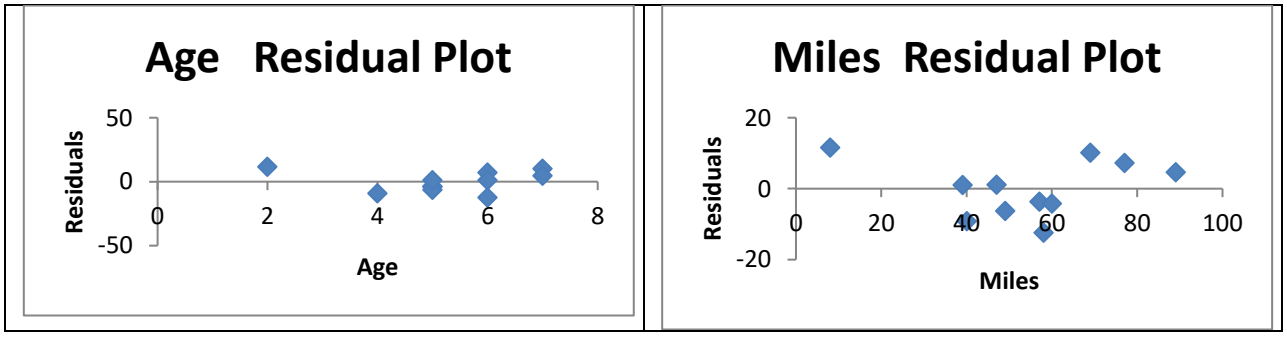
$$R^2_{adj} = 1 - \frac{(n-1)}{[n-(k+1)]}(1-R^2)$$

Example: Effect of age and miles driven on the price of Orion cars
The age, miles driven and price of a random sample of 11 Orion cars along with SPSS output are shown below.

Car	Age (yrs)	Miles (1000)	Price ($100s)
1	5	57	85
2	4	40	103
3	6	77	70
4	5	60	82
5	5	49	89
6	5	47	98
7	6	58	66
8	6	39	95
9	2	8	169
10	7	69	70
11	7	89	48

Checking Assumptions for the Orion Price regression model (SPSS ouput)





SPSS Output

Descriptive Statistics

	Mean	Std. Deviation	N
Price	88.6364	31.15854	11
Age	5.2727	1.42063	11
Miles	53.9091	21.56597	11

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.968[a]	.936	.920	8.80505	.936	58.612	2	8	.000

a. Predictors: (Constant), Miles, Age

b. Dependent Variable: Price

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	9088.314	2	4544.157	58.612	.000[b]
	Residual	620.232	8	77.529		
	Total	9708.545	10			

a. Dependent Variable: Price

b. Predictors: (Constant), Miles, Age

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	183.035	11.348		16.130	.000	156.868	209.203
	Age	-9.504	3.874	-.433	-2.453	.040	-18.438	-.570
	Miles	-.821	.255	-.569	-3.219	.012	-1.410	-.233

a. Dependent Variable: Price

*Suppose that the numbers highlighted in yellow were not given

Research Problem: Overall Assessment of the Model

>>>>>>>>>>

(a) At the 5% significance level, perform a hypothesis test to determine whether the overall multiple linear regression model is useful for making predictions, that is, whether the variables age and miles driven, taken together, are useful for predicting the price of the Orions.

Step 1: The purpose is to perform a hypothesis test for the usefulness of the overall regression model where there is more than 1 predictor var.

Step 2: $H_0: \beta_1 = \beta_2 = 0$

$H_A: \exists i \in \mathbb{N}$ s.t. $\beta_i \neq 0$

Step 3: $n = 11 \quad k = \max(\text{predictor var}) = 2$

$$F = \frac{SS_R/k}{SS_E/(n-(k+1))} = \frac{MS_R}{MS_E} = SS_E = SS_T - SS_R$$

$$= 9708.545 - 9088.314 = 620.232$$

$$= \frac{9088.314/2}{620.232/(11-(2+1))} = \frac{4544.157}{77.524} = 58.612$$

Step 4: $p < 0.001$ there is strong evidence against H_0

Step 5: Conclusion, you get the idea.

(b) What percentage of the variation in Orion price is explained by the regression model? Determine the unadjusted percentage.

$$R^2 = \frac{SS_R}{SS_T} = \frac{9088.314}{9708.545} = 0.9361$$

93.61% of the variation in orion price is explained by the model

(c) What percentage of the variation in Orion price is explained by the regression model? Determine the adjusted percentage and compare it with the unadjusted percentage calculated in part (b).

$$R^2_{Adj} = 1 - \frac{MS_E}{MS_T}$$

$MS_E = SS_E/(n-(k+1))$

$MS_T = SS_T/(n-1)$

$$= 1 - \frac{77.529}{970.8545} = 0.920$$

The diff is not very large. However, R^2_{Adj} is more accurate.

>>>>>>>>>>

5.3 Inferences Concerning the Usefulness of Particular Predictor Variables: The Multiple Regression *t*-test and Confidence Interval for Particular Slopes

- The ANOVA F-test determines whether the overall model is useful in explaining the relationship between all the variables involved.
- However, the Multiple Regression t-test is required to determine if particular predictor variables are useful in making predictions.

Multiple Regression t-test for the Usefulness of Particular Predictor Variables

State the hypotheses

Ho: $\beta_i = 0$

(Predictor variable x_i is not useful in making predictions about the response variable)

Ha: $\beta_i \neq 0$ (two-tailed) or $\beta_i < 0$ (left tailed) or $\beta_i > 0$ (right-tailed)

(Predictor variable x_i is useful in making predictions about the response variable)

Calculate the test statistic for each particular predictor variable using computer output

$$t = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)}$$

Decide to reject or not reject Ho by looking in the t-table at $df = n - (k+1)$

Interpretation in words in terms of the research problem

Note: $t^2 \neq F$ in Multiple Linear Regression, though it did in Simple Linear Regression

Confidence Interval for a Slope, β_i in Multiple Regression

1. For a confidence level of *1 – α*, use the table of the t-distribution to find $t_{\alpha/2}$ with *df = n – (k + 1)*

2. The endpoints of the confidence interval for β_i are:

$$\hat{\beta}_i \pm t_{\alpha/2} \times SE(\hat{\beta}_i)$$

3. Interpret the confidence interval in terms of the research problem

Example (Orion Prices): Refer to the data set and full SPSS output on previous pages

SPSS Output

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	183.035	11.348		16.130	.000	156.868	209.203
	Age	-9.504	3.874	-.433	-2.453	.040	-18.438	-.570
	Miles	-.821	.255	-.569	-3.219	.012	-1.410	-.233

a. Dependent Variable: Price

>>>>>>>>>>

(a) At the 5% significance level, test whether the data provide sufficient evidence to conclude that the number of miles driven, in conjunction with age, is useful for predicting price.

The regression eq. is $\hat{y} = 183.035 - 9.504x_1 - 0.821x_2$

$H_0: \beta_2 = 0$ $H_A: \beta_2 \neq 0$ Miles driven is not useful / vs. useful

$t = \frac{\hat{\beta}_2}{SE(\hat{\beta}_2)} = \frac{-0.821}{0.255} = -3.219$

$df = n - (k+1) = 11 - (2+1) = 8$

$(0.005 < p < 0.01) \times 2 = (0.01 < p < 0.02)$ there is strong evidence against H_0.

Since $p < \alpha$ of 0.05 we reject H_0.

you get the idea for the conclusion.

(b) Calculate a 95% confidence interval for the partial slope for miles driven.

for 95% C.I, $\alpha = 0.05$ @ $df = 8$ $t_{\alpha/2, df} = t_{0.025, 8} = 2.306$

$= \hat{\beta}_2 \pm t_{0.025,8} \times SE(\hat{\beta}_2)$

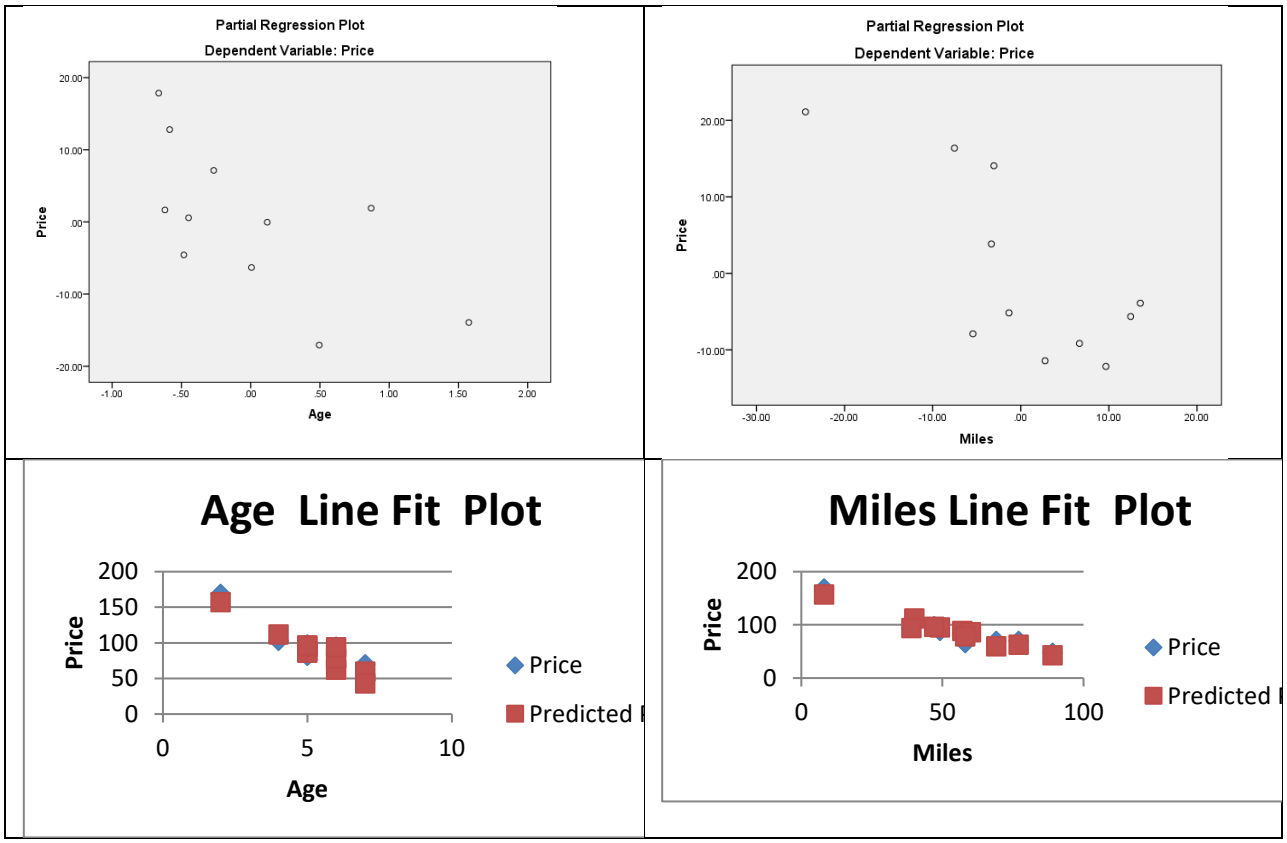
$= -0.821 \pm 2.306 \times 0.255$

$= -0.821 \pm 0.588$

$(-1.409, -0.233)$ we can be 95% confident that the partial slope for miles driven is in between -1.409 and -0.233.

>>>>>>>>>>

Compare Age and Miles Driven with respect to Usefulness in making predictions



Correlation Matrix: For all variables in the data set for Orion prices

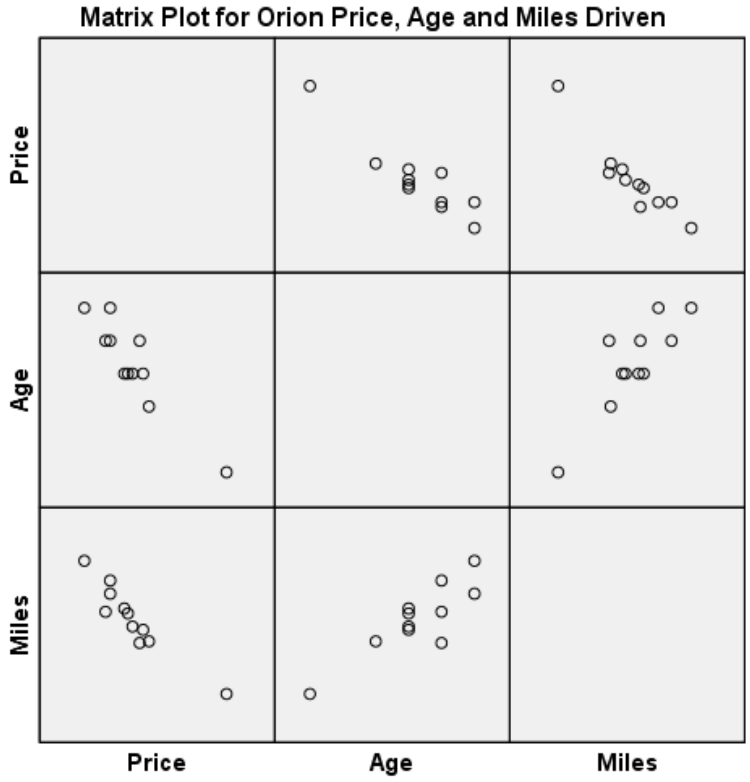
Correlations

		Price	Age	Miles
Pearson Correlation	Price	1.000	-.924	-.942
	Age	-.924	1.000	.863
	Miles	-.942	.863	1.000
Sig. (1-tailed)	Price	.	.000	.000
	Age	.000	.	.000
	Miles	.000	.000	.
N	Price	11	11	11
	Age	11	11	11
	Miles	11	11	11

R

p-value 1-tailed

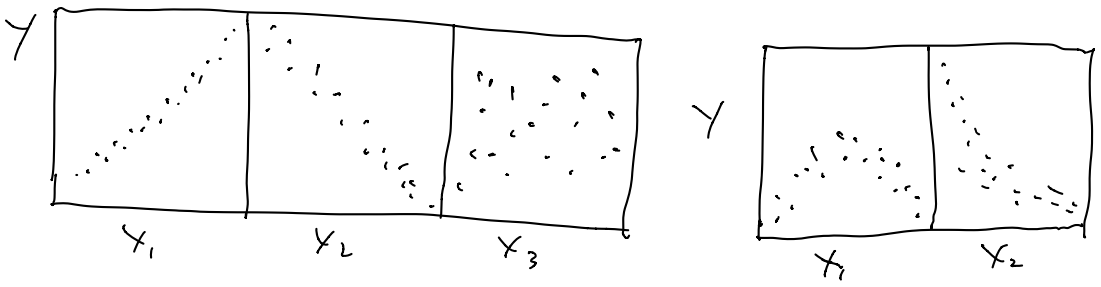
n



age vs miles
transposed with mile vs age

Note the following:

1. Miles driven has a higher t-statistic than age
2. Miles driven has a slightly lower P-value than age
3. Miles driven have a "tighter" confidence interval for the slope than age
4. Miles driven is more highly correlated with price (r = -0.942) than is age (r = -0.924), at $df = n - (k+1) = 11 - (2+1) = 8$



these are curvelinear and cannot be used.

5.4 Confidence Interval and Prediction Interval for the Response Variable

Confidence Interval for Mean Response (or Conditional Mean) in Multiple Regression

1. For a confidence level of $1 - \alpha$, use the t-distribution table to find $t_{\alpha/2}$ with $df = n - (k + 1)$
2. Compute the point estimate by using the multiple regression equation. At particular values of the predictor variables: $x_1, x_2, ..., x_k$, the point estimate $\hat{y}_p$ of the mean response of the response variable is found as follows:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_k x_k$$

   The endpoints of the confidence interval are:

   Point estimate or "Fit" ± Critical value x SE(Fit)

   OR $\hat{y}_p \pm t_{\alpha/2} \times SE(Fit)$

   [Note: SE(Fit) = standard deviation of the predicted y-value = $s_{\hat{y}_p}$]

3. Interpret the confidence interval in terms of the research problem

Prediction Interval (for all Single Observations) for the Response Variable in Multiple Regression

1. For a confidence level of $1 - \alpha$, use the t-distribution table to find $t_{\alpha/2}$ with $df = n - (k + 1)$
2. Compute the point estimate by using the multiple regression equation. At particular values of the predictor variables: $x_1, x_2, ..., x_k$, the point estimate $\hat{y}_p$ of the mean response of the response variable is found as follows:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_k x_k$$

   The endpoints of the prediction interval are:

   Point estimate or "Fit" ± Critical value x $\sqrt{MSE + [SE(Fit)]^2}$

   OR $\hat{y}_p \pm t_{\alpha/2} \times \sqrt{\hat{\sigma}^2 + [SE(Fit)]^2}$

   [Note: SE(Fit) = standard deviation of the predicted y-value = $s_{\hat{y}_p}$]

3. Interpret the confidence interval in terms of the research problem

[**Note:** Since exact calculations of the standard deviation of the predicted y-value ($s_{\hat{y}_p}$) is rather complicated, we usually use computer output to obtain SE(Fit).]

Example (Price of Orions against age and miles driven)
Find:

1. A 95% confidence interval for the mean price of Orions that are 5 years old and have been driven 52,000 miles
2. A 95% prediction interval for the price of an Orion (any single observation) that is 5 years old and has been driven 52,000 miles

MINITAB Output
[See Weiss, Module A, page A-55]
Regression Analysis: Price versus Age, Miles

The regression equation is
Price = 183 - 9.50 Age - 0.821 Miles

Predictor	Coef	SE Coef	T	P
Constant	183.04	11.35	16.13	0.000
Age	-9.504	3.874	-2.45	0.040
Miles	-0.8215	0.2552	-3.22	0.012

s_e = 8.80505 R-Sq = 93.6% R-Sq(adj) = 92.0%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	9088.3	4544.2	58.61	0.000
Residual Error	8	620.2	77.5		
Total	10	9708.5			

Predicted Values for New Observations

New Obs	Fit	SE Fit	95% CI	95% PI
1	92.80	2.74	(86.48, 99.12)	(71.53, 114.06)

Values of Predictors for New Observations

New Obs	Age	Miles
1	5.00	52.0

Find a 95% confidence for the mean price of all Orions that are 5 years old and have been driven 52,000 miles

>>>>>>>>>>

1. $df = n-(k+1) = 11-(2+1) = 8$ @ $df = 8$, $t_{\alpha/2} = t_{0.025,8} = 2.306$
2. $\hat{Y}_p = 183.04 - 9.504(5) - 0.821(52) = \$92.80$ (x100)
3. $\hat{Y}_p \pm t_{\alpha/2} \times SE(Fit)$

   $92.80 \pm 2.306 \times 2.74$

   92.80 ± 6.32

   $(86.48, 99.12)$

We can be 95% confident that the mean price of all orions that are 5 years old and have been driven 52,000 miles is in between 86.48 and 99.12 (x100).

>>>>>>>>>>

Calculate a 95% prediction interval for the price of an Orion (any single observation) that is 5 years old and has been driven 52,000 miles

1. At df = 8, $t_{\alpha/2} = t_{0.05/2} = t_{0.025} = 2.306$
2. The point estimate for the price of 5-year-old Orions that has been driven 52,000 miles is:

   $\hat{y}_p$ = 183 - 9.50 (5) - 0.821 (52) = 92.80 (in hundreds of dollars)

>>>>>>>>>>

$\hat{Y}_p \pm t_{\alpha/2} \times \sqrt{\sigma^2 + (SE(Fit))^2}$

$92.80 \pm 2.306 \times \sqrt{(8.805)^2 + (2.74)^2}$

92.80 ± 21.26

$(71.54, 114.06)$

3. We can be 95% confident that the price of an Orion (any single observation) that is 5 years old and have been driven 52K miles is in between 71.54 and 114.06 (x100).

>>>>>>>>>>

5.5 Multiple Regression Models Involving Indicator Variables (= Dummy Variables)

- These are categorical variables that are used as one of the predictor variables
- It is coded as 0 or 1

Example involving an Indicator Variable

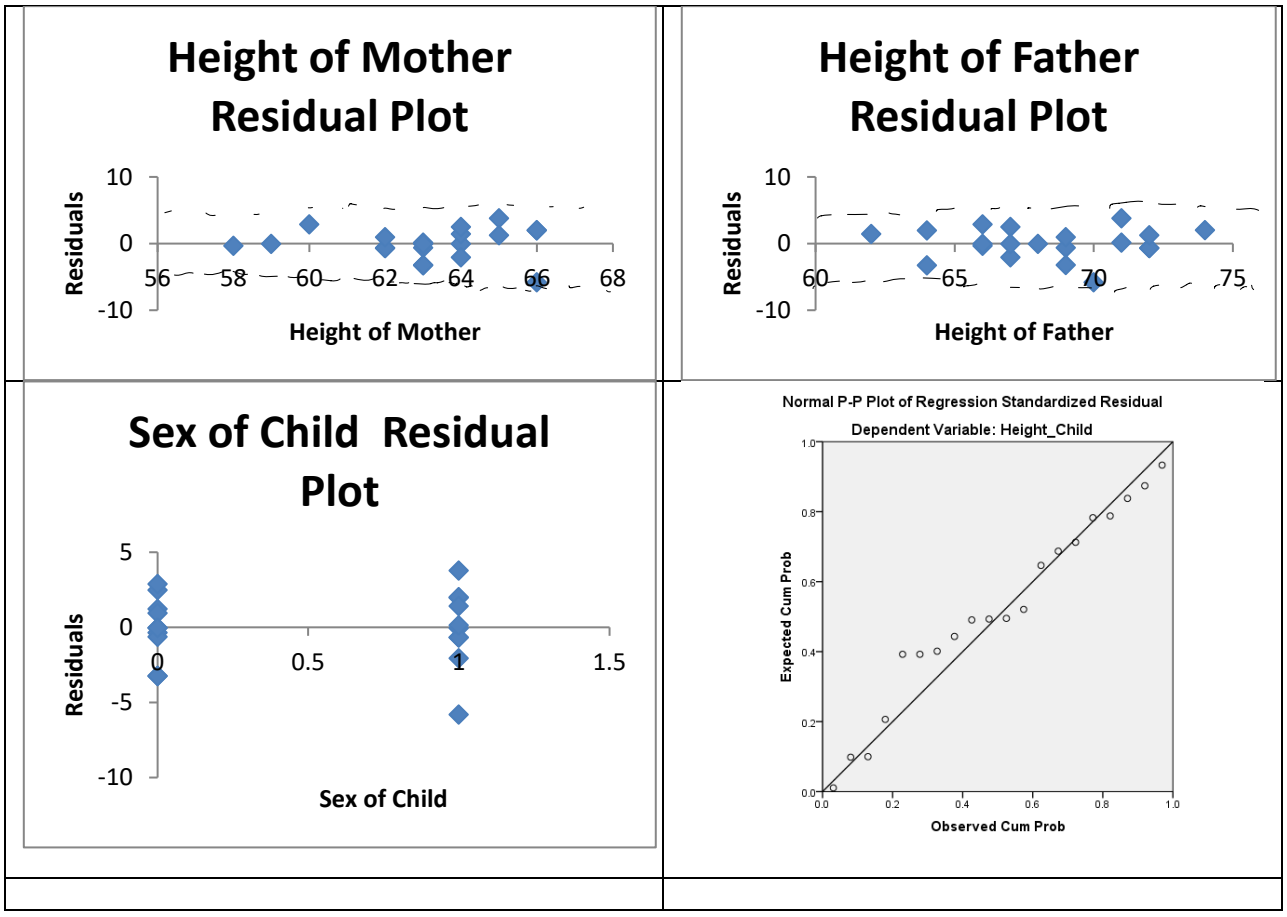
Indicator variable = sex of the child (Coded as 0 for female and 1 for male)

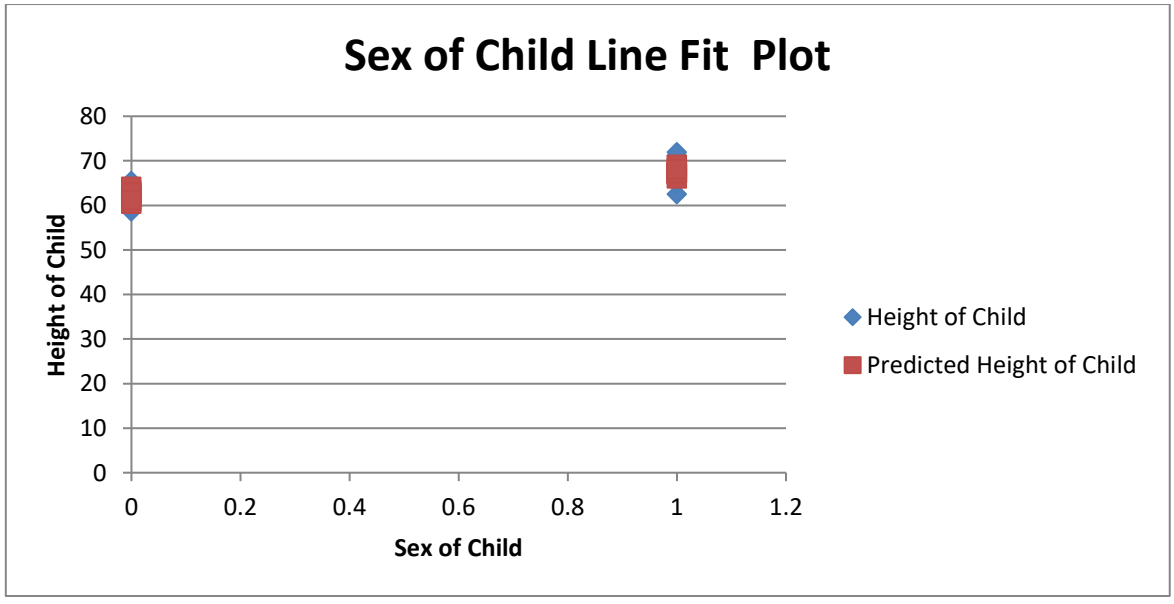
Height of Mother	Height of Father	Sex of Child	Height of Child
66	70	1	62.5
66	64	1	69.1
64	68	1	67.1
66	74	1	71.1
64	62	1	67.4
64	67	1	64.9
62	72	1	66.5
62	72	1	66.5
63	71	1	67.5
65	71	1	71.9
63	64	0	58.6
64	67	0	65.3
65	72	0	65.4
59	67	0	60.9
58	66	0	60
63	69	0	62.2
62	69	0	63.4
63	66	0	62.2
63	69	0	59.6
60	66	0	64

Descriptive Statistics

	Mean	Std. Deviation	N
Height_Child	64.805	3.6954	20
Height_Mother	63.10	2.198	20
Height_Father	68.30	3.164	20
Sex_of_Child	.50	.513	20

Checking Assumptions





The R squares are a bit far from 1, but close enough to use the model

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.780[a]	.609	.535	2.5195

a. Predictors: (Constant), Sex_of_Child, Height_Father, Height_Mother

b. Dependent Variable: Height_Child

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	157.902	3	52.634	8.291	.001[b]
	Residual	101.568	16	6.348		
	Total	259.470	19			

a. Dependent Variable: Height_Child

b. Predictors: (Constant), Sex_of_Child, Height_Father, Height_Mother

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	25.588	21.942		1.166	.261	-20.928	72.104
	Height_Mother	.377	.308	.224	1.224	.239	-.276	1.030
	Height_Father	.195	.190	.167	1.028	.319	-.207	.598
	Sex_of_Child	4.148	1.334	.576	3.108	.007	1.319	6.976

a. Dependent Variable: Height_Child

The sex of the child overrides the effect of the parents.

Regression equation:

Height of child = 25.588 + 0.377(Height of Mother) + 0.195(Height of Father) + 4.148(Sex)

Prediction:

Suppose a mother is 63 inches and a father is 69 inches

Predicted height of a daughter is:

Height of a daughter = 25.588 + 0.377(63) + 0.195(69) + 4.148(0) = 62.8 inches

Predicted height of a son is:

Height of a son = 25.588 + 0.377(63) + 0.195(69) + 4.148(1) = 67.0 inches

The coefficient 4.148 means that for given heights of mothers and fathers, a son will have a predicted height that is 4.148 inches more than the height of a daughter.

Adjusted Coefficient of Determination:

$$R^2_{adj} = 1 - \frac{\frac{SS_{ERROR}}{(n-(k+1))}}{\frac{SS_{TOTAL}}{(n-1)}} = 1 - \frac{\frac{101.568}{(20-(3+1))}}{\frac{259.470}{(20-1)}} = 1 - \frac{6.348}{13.6563} = 0.535$$

Note: This is fairly different from the coefficient of determination (unadjusted), which is 0.609. This is because there are 4 regression coefficients (intercept and 3 slopes)

Calculate 95% confidence intervals for the partial slopes of the regression equation that relate:

1. Heights of children to the heights of mothers
2. Heights of children to their sex

df = n – (k + 1) = 20 – (3+1) = 16

At df = 16, $t_{\alpha/2} = t_{0.05/2} = t_{0.025} = 2.120$

Heights of children to the heights of mothers

$$\hat{\beta}_i \pm t_{\alpha/2} \times SE(\hat{\beta}_i)$$
$$0.377 \pm 2.120 \times 0.308$$
$$0.377 \pm 0.6530$$
$$(-0.276, 1.030)$$

Heights of children to their sex

$$\hat{\beta}_i \pm t_{\alpha/2} \times SE(\hat{\beta}_i)$$
$$4.148 \pm 2.120 \times 1.334$$
$$4.148 \pm 2.8288$$
(1.319, 6.976)

Note: The slope that relates heights of children to their sex does not have a negative value as one of the endpoints. This is in agreement with the greater significance of that slope when the multiple regression t-test was performed.

Does this mean that the heights of children are not related to the heights of their parents?

5.6 Interaction Models in Multiple Regression

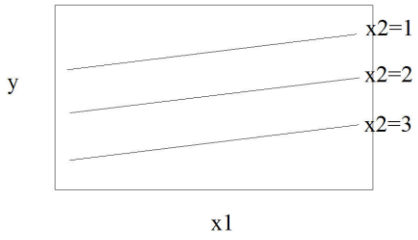
- Without interaction, the general model for multiple linear regression was:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

each term have their own predictor var

The predicted response of y with changes in x1 has the same slope for all values of x2 (and the same holds true for all xi variables involved)

This results in a parallel-lines model as shown below:



- When interaction between variables occurs, the interaction model for multiple linear regression (for two interacting predictor variables) is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

each term may have multiple predictor vars

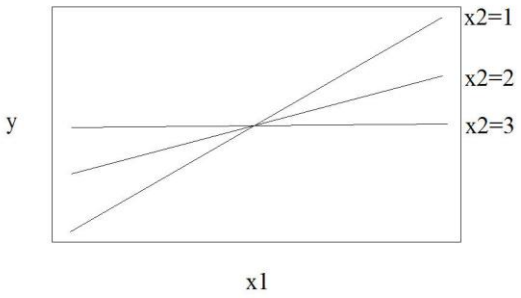
Where,

- y is the response variable
- x_1, x_2 are the explanatory (predictor) variables
- $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$ is the deterministic part of the model
- $\beta_1 + \beta_3 x_2$ represents the change in y for a 1-unit increase in x_1

  [Since $\beta_1 x_1 + \beta_3 x_1 x_2 \Rightarrow x_1(\beta_1 + \beta_3 x_2)$]
- $\beta_2 + \beta_3 x_1$ represents the change in y for a 1-unit increase in x_2

  [Since $\beta_2 x_2 + \beta_3 x_1 x_2 \Rightarrow x_2(\beta_2 + \beta_3 x_1)$]
- ε is the random error, which is assumed to be normally distributed with mean 0 and standard deviation σ

This results in non-parallel lines (often intersecting lines) as shown below:



Research Problem Involving an Interaction Term (and Combining all Previous MLR Concepts):

Effect of BMI and Salt Intake (and their Interaction) on Systolic Blood Pressure
It has been hypothesized that increased salt intake associated with greater food intake by obese people may be the mechanism for the relationship between obesity and high blood pressure. A random sample of 14 people with high blood pressure was selected and their body mass index (BMI) (body weight/(height)^2), as a measure of obesity, was measured along with their sodium intake (in 100s of mg/day). These two variables were used to calculate the interaction term (BMI x sodium intake). Their systolic blood pressure (SBP) was measured in mm Hg as the response variable. The raw data are shown below along with incomplete SPSS output.

BMI (kg/m^2)	Sodium intake (100 mg/day)	Interaction	SBP (mm Hg)
30	30	900	143
30	31	930	144
33	32	1056	146
34	35	1190	150
36	36	1296	152
37	37	1369	154
38	38	1444	156
39	39	1521	158
40	41	1640	161
40	42	1680	163
41	43	1763	165
43	44	1892	168
44	45	1980	170
47	49	2303	176

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.999[a]	.997	.997	.586

a. Predictors: (Constant), Interaction, BMI, Salt_intake

b. Dependent Variable: SBP

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1330.000	3	443.333	1293.138	.000[b]
	Residual	3.428	10	.343		
	Total	1333.429	13			

a. Dependent Variable: SBP

b. Predictors: (Constant), Interaction, BMI, Salt_intake

Coefficients[a]						
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	108.726	8.168		13.312	.000
	BMI	-.218	.285	-.109	-.765	.462
	Salt_intake	.892	.350	.496	2.546	.029
	Interaction	.015	.006	.612	2.640	.025
a. Dependent Variable: SBP						

>>>>>>>>>>

(a) At the 5% significance level, perform a hypothesis test to determine whether the overall multiple regression model is significant or useful for making predictions about systolic blood pressure (SBP). Perform ALL steps of the hypothesis test.

$H_0: \beta_1 = \beta_2 = \beta_3 = 0$

$H_A: \exists \beta_i \in \beta$ s.t $\beta_i \neq 0$

The overall regression model is not useful in making predictions about systolic blood pressure.

The overall model is useful is making predictions about systolic blood pressure.

$k=3 \quad n=14$

$F = \frac{SS_R/k}{SS_E/(n-(k+1))} = \frac{1330.000/3}{3.428/(14-(3+1))} = \frac{443.333}{0.3428}$

$F = 1293.271$

it's slightly different due to precision.

$df(3, 10)$ # $p < 0.001$

It is extremely strong evidence against H0.

Since $p < \alpha$ of 0.05

We reject the null hypothesis

At 5% significant level, the data provided sufficient evidence to conclude that at least one of the population regression coefficient is not zero OR that the overall regression model is useful for making predictions about the response variable (systolic blood pressure).

(b) At the 5% significance level, perform the most appropriate test to determine whether there is a positive relationship between salt intake and systolic blood pressure.

Has to be a t-test as we are checking for a positive relationship.

$H_0: \beta_2 = 0 \quad H_A: \beta_2 > 0$

$t = \frac{\hat{\beta}_2 - 0}{SE(\hat{\beta}_2)} = \frac{0.892}{0.350} = 2.5486$

$df = (10)$ $\quad p\text{-value} = 0.01 < p < 0.05$

Strong evidence against H0

Since $p < \alpha$ of 0.05

We reject H0

At the 5% sig level, the data provide sufficient evidence to conclude that there is a significant positive relationship between salt intake and systolic blood pressure.

(c) Calculate a 95% confidence interval for the slope of the interaction term (representing interaction between BMI and sodium intake). Using this confidence interval, what conclusion can you make about the possible interaction between body mass index and sodium intake in their effect on systolic blood pressure? Explain your answer. $Cr.t.val = t_{0.05/2,10} = t_{0.025,10} = 2.228$

$\beta_3 \pm Cr.t.val \times SE(\hat{\beta}_3)$

$0.0157 \pm 2.228 \times 0.006$

0.0157 ± 0.013368

$(0.001632, 0.02837)$

Since 0 is not in the CI we can be 95% confident that the slope of the interaction term is significant, that is, there is significant interaction between body mass index and salt intake in their effect on systolic blood pressure.

(d) What does this model tell us about effect of BMI and the relative effect of the 3 predictor variables?

That some of the variables overrides other variable BMI, thus giving a negative slope

(e) Find the standard error of the model (standard error of the estimate of the model)?

$MS_E = SS_E/n-(k+1) = 3.428/10 = 0.3428$

$\hat{\sigma} = \sqrt{0.3428} = 0.585$

(f) What percentage of the variation in systolic blood pressure is explained by (or accounted for by) the regression model? (Note: Determine the adjusted percentage.)

$R^2_{ADJ} \quad 1 - \frac{MS_E}{MS_T} = \frac{0.343}{SS_T/(n-1)} = 1 - \frac{0.343}{102.57138} = 0.9967$

The adjusted coefficient of determination shows that 99.7% of the variation in systolic blood pressure is explained by the regression model.

(g) Suppose that a person with a body mass index of 40 kg/m^2 and daily sodium intake of 42 (in 100s of mg/day) had an observed systolic blood pressure reading of 163 mm Hg. What was the residual or error of this observation?

$\hat{y} = 108.726 - 0.218(40) + 0.892(42) + 0.015(40)(42)$

$= 162.67$ mm Hg

$Residual = e = observed - predicted$

$163 - 162.67 = +0.33$ mmHg

(h) Based on the values of the predictor variables given in part (g) (BMI = 40 kg/m2, sodium intake = 42 (100) mg/day)), what is the 95% prediction interval for all single observation responses of systolic blood pressure at those values of the predictor variables? [Note: SE(Fit) = 0.337]

At $df = n-(k+1) = 10$ $t_{0.025,10} = 2.228$

based on part (f) $\hat{y} = 162.67$

$\hat{y}_p \pm t_{\alpha/2} \times \sqrt{\hat{\sigma}^2 + SE(Fit)^2}$

$162.67 \pm 2.228 \times \sqrt{(0.585)^2 + (0.337)^2}$

$162.67 \pm 2.228 \times 0.675125$

162.67 ± 1.5042

$(161.166, 164.174)$

We can be 95% confident that systolic blood pressure at the values of the predictor variable given in part (g) is in between 161.166 and 164.174 mmHg.

(i) Based on the values of the predictor variables given in part (g) (BMI = 40 kg/m2, sodium intake = 42 (100) mg/day)), what is the 95% confidence interval for mean systolic blood pressure at those values of the predictor variables? [Note again: SE(Fit) = 0.337]

At $df = 10$, $t_{\alpha/2} = 2.228$

$\hat{y}_p \pm t_{\alpha/2} \times SE(Fit)$

$162.67 \pm 2.228 \times 0.337$

162.67 ± 0.7508

$(161.919, 163.421)$

We are 95% confident that the mean systolic blood pressure at the value of the predictor variable given in part (g) is in between 161.919 and 163.421 mm Hg.

>>>>>>>>>>

(j) Compare the length of the prediction interval in part (h) with the confidence interval in part (i). Explain the difference between these two confidence intervals and explain any possible difference in their lengths.

Based on the prediction interval in part (h), if we take random samples of people having the given values of the predictor variables, we can be 95% confident that an individual would have systolic blood pressure between 161.67 and 164.174 mm Hg; whereas, based on the confidence interval in part (i), we can be 95% confident that the means of those samples will be between 161.919 and 163.421 mm Hg. This is because the confidence interval for the mean response is shorter than the prediction interval for all single observation responses.

5.7 Reduced Models and the Extra Sum-of-Squares F-test in Multiple Linear Regression

Full Model = model which includes all the parameters or predictor variables involved in the research

Reduced Model = model which hypotheses that some of the slopes of the predictor variables equal zero and, thus they are taken out of the full model to make a reduced model

Extra-Sum-of-Squares F-test in Multiple Linear Regression

- Also called Partial F-test or Nested F-test

Extra-Sum-of-Squares F-Test in MLR

Null and alternative hypotheses:

H_0: All selected beta's (slopes) equal 0. (Reduced model)

H_a: Not all selected beta's (slopes) equal 0. (Full model)

Calculations for Extra-Sum-of Squares F-test:

Extra Sum of Squares = $SSE(reduced) - SSE(full)$

Extra df = $df_{ERROR}(reduced) - df_{ERROR}(full)$

$$F = \frac{(Extra\ SS)/(Extra\ df)}{SSE(\text{Full})/df_{ERROR}(Full)}$$

OR $$F = \frac{[SS_E(reduced) - SS_E(full)]/[df_E(reduced) - df_E(full)]}{SS_E(full)/df_E(full)}$$

Examine the distribution of the F-table at:

$$df = [Extra\ df, df_{ERROR}(Full)] = [\text{Number of selected } \beta_i\text{'s}, n-(k+1)]$$

Recall that, residual (error) = observed value – estimated value

Therefore, residual sum of squares or error sum of squares is:

SSE = $\sum$ (observed value – estimated value)^2 $= \sum (x_i - \bar{x})^2$

Extra df = df_E(reduced) − df_E(full)

= [n − (k(reduced) + 1) − (n − (k(full) + 1))]

= [n − k(reduced) − 1 − n + k(full) + 1]

= k(full) − k(reduced)

Example with Interaction and Indicator Variables & Involving Extra Sum-of-Squares F-test
The table below shows the prices of a random sample of 30 homes, along with the living area, number of bedrooms, number of rooms, age, and location.

- Indicator variables z_1 and z_2 are defined as:

  $z_1 = z_2 = 0$ for downtown; $z_1 = 1$, $z_2 = 0$ for inner suburbs; $z_1 = 0$, $z_2 = 1$ for outer suburbs

- $x_1 z_1$ = interaction $x_1 \times z_1$
- $x_1 z_2$ = interaction $x_1 \times z_2$

Price ($1000) ($y$)	Living area (100s of sq. Ft.) (x_1)	No. of bedrooms (x_2)	No. of room (x_3)	Age (years) (x_4)	Location (z_1)	Location (z_2)	$x_1 z_1$	$x_1 z_2$
84	13.8	3	7	10	1	0	13.8	0
93	19	2	7	22	0	1	0	19
83.1	10	2	7	15	0	1	0	10
85.2	15	3	7	12	0	1	0	15
85.2	12	3	7	8	0	1	0	12
85.2	15	3	7	12	0	1	0	15
85.2	12	3	7	8	0	1	0	12
63.3	9.1	3	6	2	0	1	0	9.1
84.3	12.5	3	7	11	0	1	0	12.5
84.3	12.5	3	7	11	0	1	0	12.5
77.4	12	3	7	5	1	0	12	0
92.4	17.9	3	7	18	0	0	0	0
92.4	17.9	3	7	18	0	0	0	0
61.5	9.5	2	5	8	0	0	0	0
88.5	16	3	7	11	0	0	0	0
88.5	16	3	7	11	0	0	0	0
40.6	8	2	5	5	0	0	0	0
81.6	11.8	3	7	8	0	1	0	11.8
86.7	16	3	7	9	1	0	16	0
89.7	16.8	2	7	12	0	0	0	0
86.7	16	3	7	9	1	0	16	0
89.7	16.8	2	7	12	0	0	0	0
75.9	9.5	3	6	6	0	1	0	9.5
78.9	10	3	6	11	1	0	10	0
87.9	16.5	3	7	15	1	0	16.5	0
91	15.1	3	7	8	0	1	0	15.1
92	17.9	3	8	13	0	1	0	17.9
87.9	16.5	3	7	15	1	0	16.5	0
90.9	15	3	7	8	0	1	0	15
91.9	17.8	3	8	13	0	1	0	17.8

Overall multiple regression model

Selecting some of the above predictor variables, the overall model describing the effect of living area, location and the interaction between living area and location (leaving out the number of bedrooms, number of rooms and age) is as follows:

Overall (Full) model: $y = \beta_0 + \beta_1 x_1 + \beta_2 z_1 + \beta_3 z_2 + \beta_4 x_1 z_1 + \beta_5 x_1 z_2 + \varepsilon$

We can determine the fitted straight line for each location by finding 3 simple linear regression equations based on simplification of the overall model

Downtown: ($z_1 = z_2 = 0$)

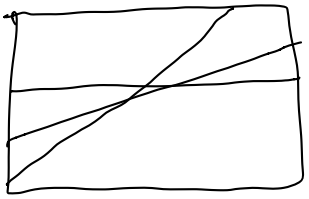
$$y = \beta_0 + \beta_1 x_1 + \beta_2(0) + \beta_3(0) + \beta_4 x_1(0) + \beta_5 x_1(0) + \varepsilon$$
$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

Inner suburbs: ($z_1 = 1, z_2 = 0$)

$$y = \beta_0 + \beta_1 x_1 + \beta_2(1) + \beta_3(0) + \beta_4 x_1(1) + \beta_5 x_1(0) + \varepsilon$$
$$y = \beta_0 + \beta_2 + (\beta_1 + \beta_4)x_1 + \varepsilon$$

Outer suburbs: ($z_1 = 0, z_2 = 1$)

$$y = \beta_0 + \beta_1 x_1 + \beta_2(0) + \beta_3(1) + \beta_4 x_1(0) + \beta_5 x_1(1) + \varepsilon$$
$$y = \beta_0 + \beta_3 + (\beta_1 + \beta_5)x_1 + \varepsilon$$



From this we write 3 models:

Model 1 (Separate Lines Model = Full Model, which includes all predictor variables):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 z_1 + \beta_3 z_2 + \beta_4 x_1 z_1 + \beta_5 x_1 z_2$$

OR $\mu(price \mid area, location, \text{interaction}) = \beta_0 + \beta_1 area + \beta_2 z_1 + \beta_3 z_2 + \beta_4 x_1 z_1 + \beta_5 x_1 z_2$

Model 2 (Parallel Lines Model = Reduced model assuming there is no interaction effect):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 z_1 + \beta_3 z_2$$

OR $\mu(price \mid area, location) = \beta_0 + \beta_1 area + \beta_2 z_1 + \beta_3 z_2$

Explanation: If no interaction effect, then $\beta_4 = \beta_5 = 0$ so $\beta_1 = \beta_1 + \beta_4 = \beta_1 + \beta_5$ (slopes are equal)

And thus the 3 SLR lines are parallel.

Model 3 (Equal Lines Model = Reduced model assuming location and their interaction have no effect):

$$y = \beta_0 + \beta_1 x_1$$

OR $\mu(price \mid area) = \beta_0 + \beta_1 area$

Explanation: If no effect of location and interaction, then $\beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ so

$\beta_0 = \beta_0 + \beta_2 = \beta_0 + \beta_3$ (y-intercepts are equal) and $\beta_1 = \beta_1 + \beta_4 = \beta_1 + \beta_5$ (slopes are equal)

And thus the 3 SLR lines are equal.

SPSS output:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.943[a]	.889	.866	4.05994

a. Predictors: (Constant), x1z2, x1, z1, z2, x1z1

Model 1 (Full Model or Separate Lines Model)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3158.414	5	631.683	38.323	.000[b]
	Residual	395.595	24	16.483		
	Total	3554.010	29			

a. Dependent Variable: y

b. Predictors: (Constant), x1z2, x1, z1, z2, x1z1

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	8.969	6.078		1.476	.153
	x1	4.807	.397	1.366	12.098	.000
	z1	52.122	11.225	2.025	4.643	.000
	z2	48.558	7.797	2.231	6.228	.000
	x1z1	-3.201	.759	-1.823	-4.218	.000
	x1z2	-2.803	.530	-1.836	-5.291	.000

a. Dependent Variable: y

Model 2 (Parallel Lines Model): Effect of area and location (Reduced model assuming there is no interaction effect, i.e., assuming slopes for interaction = 0)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2607.733	3	869.244	23.883	.000[b]
	Residual	946.277	26	36.395		
	Total	3554.010	29			

a. Dependent Variable: y

b. Predictors: (Constant), z2, x1, z1

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	35.825	5.785		6.193	.000
	x1	3.000	.362	.852	8.292	.000
	z1	5.189	3.127	.202	1.660	.109
	z2	8.142	2.680	.374	3.038	.005

a. Dependent Variable: y

Model 3 (Equal Lines Model): Effect of Area only (Reduced model assuming location and interaction have no effect, i.e., assuming all slopes for location and interaction = 0)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2271.714	1	2271.714	49.605	.000[b]
	Residual	1282.296	28	45.796		
	Total	3554.010	29			

a. Dependent Variable: y
b. Predictors: (Constant), x1

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	43.732	5.780		7.567	.000
	x1	2.814	.400	.799	7.043	.000

a. Dependent Variable: y

→ F = 57.259

(a) At the 5% significance level, perform a hypothesis test to determine whether the overall multiple regression model is significant or useful for making predictions about house price. Perform ALL steps of the hypothesis test.

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$

[The overall multiple regression model is not useful for making predictions about house price.]

H_a : At least one β_i is not zero

[The overall multiple regression model is useful for making predictions about house price.]

k = number of predictor variables = 5, n = 30 (random sample of 30 homes)

$$F = \frac{SS_{REGR} / k}{SS_{ERROR} / (n-(k+1))} = \frac{3158.414 / 5}{395.595 / (30-(5+1))} = \frac{631.683}{16.483} = 38.323$$

df (regression) = k = 5 df (error) = n – (k + 1) = 30 – (5 + 1) = 24
At df = (5, 24), P < 0.001 There is extremely strong evidence against Ho.
Since P < α (0.05), reject Ho.

Conclusion: At the 5% significance level, the data provide sufficient evidence to conclude that at least one of the population regression coefficients is not zero OR that the overall regression model is useful for making predictions about the response variable (house price).

these are involved in our model.

>>>>>>>>>>

(b) At the 5% significance level, perform an Extra Sum-of-Squares F-test to determine if there is interaction between location and living area in the way that they affect house price, after accounting for area and location. In other words, test whether the 3 simple regression lines are parallel, that is, whether the slopes are the same for all 3 lines.

$H_0: \beta_4 = \beta_5 = 0$ (interaction term = 0) (reduced model) (model 2) (additive model)

$H_A: \beta_i \neq 0,\ i = 4, 5$ (full model) (model 1)

$Y = \beta_1 area + \beta_2 z_1 + \beta_3 z_2$

$y = \beta_0 + \beta_1 area + \beta_2 z_1 + \beta_3 z_2 + \beta_4 x_1 z_1 + \beta_5 x_1 z_2$

$$F = \frac{[SS_E(reduced) - SS_E(full)]}{[df_E\ reduced - df_E(full)]} \Big/ SS_E(full) / df_E(full)$$

$$\frac{946277 - 395.595 / 26 - 24}{395.595 / 24} = 16.7045$$

$df = [2, 24]$ $\quad p < 0.001$ $\quad$ Since $p < \alpha$ of 0.05 we reject H_0.

At the 5% sig level, we can conclude that there is interaction between location and living area in the way that they affect house price, after accounting for area and location, in other words the 3 SDR are not parallel.

Finding the Residual Sum-of-Squares

Suppose you are given that the F-statistic for the Parallel Lines Model is F = 16.7045, but you are not given the ANOVA table on the previous page for this model. What is the Residual Sum-of-Squares (SS_{ERROR}) for this Parallel Lines Model?

$$16.7045 = \frac{[SS_E(reduced) - 395.595]/[2]}{395.595 / 24}$$

$$(16.7045)(16.483125) = (SS_E(reduced) / 2) - 197.7975$$

$$275.34236 + 197.7975 = SS_E(reduced) / 2$$

$$SS_E(reduced) \approx 946.28$$

(c) At the 5% significance level, perform an Extra Sum-of-Squares F-test to determine if there is an effect of location and/or the interaction between location and living area on house price, after accounting for area. In other words, test whether the 3 simple regression lines are equal, that is, whether the y-intercepts and slopes are the same for all 3 lines.

$H_0: \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ (model 3)

$H_A: \beta_i \neq 0 : i \in \{2, \ldots, 5\}$ (model 1)

$$F = \frac{[SS_E(\text{reduced}) - SS_E(\text{full})]}{[df_E(\text{reduced}) - df_E(\text{full})]} \Big/ SS_E(\text{full}) / df_E(\text{full})$$

$$\frac{1282.298 - 395.595 / 28 - 24}{395.595 / 24} = 13.4486$$

$df = [4, 24]$ $p < 0.001$ Since $p < \alpha$ of 0.05 we reject H_0.

At the 5% the sig level, we can conclude that there is an affect of location and/or the interaction between location and living area on house prices, after accounting for area. In other words the three SDR lines are not equal.

>>>>>>>>>>

Comparing the 3 SLR Equations for Downtown, Inner Suburbs, and Outer Suburbs

Using the output to get the overall regression model, we get the following:

$$\hat{y} = 8.969 + 4.807x_1 + 52.122z_1 + 48.558z_2 + (-3.201)x_1z_1 + (-2.803)x_1z_2$$

Note: all partial slopes, including those for the interaction terms, are significant.

We can determine the fitted straight line for each location by finding 3 simple linear regression equations by simplifying the overall model

Overall model: $y = \beta_0 + \beta_1x_1 + \beta_2z_1 + \beta_3z_2 + \beta_4x_1z_1 + \beta_5x_1z_2 + \varepsilon$

Downtown: $y = \beta_0 + \beta_1x_1 + \varepsilon$

$\hat{y} = 8.969 + 4.807x_1$

Inner suburbs: $y = \beta_0 + \beta_2 + (\beta_1 + \beta_4)x_1 + \varepsilon$

$\hat{y} = 8.969 + 52.122 + (4.807 - 3.201)x_1$

$\hat{y} = 61.091 + 1.606x_1$

Outer suburbs: $y = \beta_0 + \beta_3 + (\beta_1 + \beta_5)x_1 + \varepsilon$

$\hat{y} = 8.969 + 48.558 + (4.807 - 2.803)x_1$

$\hat{y} = 57.527 + 2.004x_1$

Overall Conclusion:

1. Downtown houses have a much lower baseline price relative to the suburbs, judging by the lower end of the simple linear regression line (indicated by the low y-intercept).
2. At least some of the slopes are significantly different, so they contribute differently to the model.
3. Downtown prices increase faster than the suburbs as the house size increases. (Based on the slopes of the simple linear regression equations.)
4. Both types of suburbs (inner and outer) are similar in baseline prices as well as the increase in price with increasing house size.

5.8 Building Models in Multiple Linear Regression

Example on Refractive Surgery
Radial keratotomy is a type of refractive surgery in which radial incisions are made in a myopic (nearsighted) patient's cornea to reduce the person's myopia. The incisions extend radially from the periphery toward the centre of the cornea. A circular central portion of the cornea, known as the clear zone, remains uncut. A researcher examined the variables associated with the five-year post-surgical change in refractive error. She selected 413 patients for the study who met strict entry criteria. In fact, four clear zone sizes were used: 2.5 mm, 3.0 mm, 3.5 mm, and 4.0 mm. The following is the description of variables under study.

Variable	**Description of Variables**
Gender	Gender (Male, Female),
Diameter	Diameter of the clear zone (remains uncut) (2.5 mm, 3.0 mm, 3.5 mm, and 4.0 mm),
Age	Age of patients (in years),
Depth	Depth of incision (in mm),
CRE	Change in refractive error.

Define the gender and diameter of the clear zone variables using the following indicator variables:

Male = 1 for a male and Male = 0 for a female,
D1 = 1 if diameter of the clear zone is 2.5 mm and D1 = 0 otherwise,
D2 = 1 if diameter of the clear zone is 3.0 mm and D2 = 0 otherwise,
D3 = 1 if diameter of the clear zone is 3.5 mm and D3 = 0 otherwise,
D4 = 0 (no incision)

Consider the following as the ORIGINAL regression model with change in refractive error (CRE) as the response:

$$\begin{aligned}\mu\{CRE \mid Age, Gender, Diameter) &= \beta_0 + \beta_1 Age + \beta_2 Male + \beta_3 D1 + \beta_4 D2 + \beta_5 D3 \\ &+ \beta_6 (Age \times Male) + \beta_7 (Age \times D1) + \beta_8 (Age \times D2) + \beta_9 (Age \times D3) \\ &+ \beta_{10} (Age \times Male \times D1) + \beta_{11} (Age \times Male \times D2) + \beta_{12} (Age \times Male \times D3)\end{aligned}$$

a) Referring to the original model, in terms of the regression coefficients, what is the effect of age on mean change in refractive error (CRE), after accounting for gender and diameter? Define this effect in general, then summarize the effect for each combination of gender and diameter of the clear zone? Summarize your results in the chart below.

Solution:
Logic: For the general effect of age, consider only terms that include age, thus all terms without age are excluded, that is,

$\beta_0, \beta_2, \beta_3, \beta_4, \beta_5$ are excluded.

The general effect of age on mean CRE is:

$$\begin{aligned}&\mu\{CRE \mid Age + 1, Gender, Diameter\} - \mu\{CRE \mid Age, Gender, Diameter\} \\ &= \beta_1 + \beta_6 male + \beta_7 D1 + \beta_8 D2 + \beta_9 D3 + \beta_{10} (male \times D1) + \beta_{11} (male \times D2) + \beta_{12} (male \times D3)\end{aligned}$$

Logic: For the effect of age on each combination below, include only slopes for age by itself or for age in combination with either gender and/or diameter of the clear zone.

Therefore, for each combination of gender and diameter, we have:

Gender	Diameter of the clear zone	Logic	Effect of age on mean CRE
Male	2.5		$\beta_1 + \beta_6 + \beta_7 + \beta_{10}$
Male	3.0		$\beta_1 + \beta_6 + \beta_8 + \beta_{11}$
Male	3.5		$\beta_1 + \beta_6 + \beta_9 + \beta_{12}$
Male	4.0		$\beta_1 + \beta_6$
Female	2.5		$\beta_1 + \beta_7$
Female	3.0		$\beta_1 + \beta_8$
Female	3.5		$\beta_1 + \beta_9$
Female	4.0		β_1

b) Modify the original model to specify that the effect of age on the mean of CRE is the same for males and females with the same diameter of the clear zone; otherwise, the effect of age on the mean of CRE is possibly different for males and females without having the same diameter of the clear zone. Just state the constraint(s) needed. You do not have to rewrite the model.

$$\left.\begin{array}{l} Diameter = 2.5 : \beta_1 + \beta_6 + \beta_7 + \beta_{10} = \beta_1 + \beta_7 \Rightarrow \beta_6 + \beta_{10} = 0 \\ Diameter = 3.0 : \beta_1 + \beta_6 + \beta_8 + \beta_{11} = \beta_1 + \beta_8 \Rightarrow \beta_6 + \beta_{11} = 0 \\ Diameter = 3.5 : \beta_1 + \beta_6 + \beta_9 + \beta_{12} = \beta_1 + \beta_9 \Rightarrow \beta_6 + \beta_{12} = 0 \\ Diameter = 4.0 : \beta_1 + \beta_6 = \beta_1 \qquad\qquad \Rightarrow \beta_6 = 0 \end{array}\right\} \Rightarrow \beta_6 = \beta_{10} = \beta_{11} = \beta_{12} = 0$$

Explanation:

c) Referring to the original model, write the null and alternative hypotheses, in terms of the coefficients, to test whether the effect of age is the same for all diameters of the clear zone for females. What is the distribution of the test statistic under the null hypothesis?

Solution: The effect of age on the mean of CRE is the same for all diameters of the clear zone for females if $\beta_1 + \beta_7 = \beta_1 + \beta_8 = \beta_1 + \beta_9 = \beta_1$.

Therefore, H_0: $\beta_7 = \beta_8 = \beta_9 = 0$,
H_A: at least one $\beta_i \neq 0$ $\quad i = 7, 8, 9$

If H_0 is true, the test statistic has an *F*-distribution with degrees of freedom of:

$$df = [Extra\ df, df_{ERROR}(Full)] = [\text{Number of selected } \beta_i\text{'s}, n-(k+1)] = (3, 413-(12+1)) = (3, 400)$$

d) Referring to the original model, in terms of the regression coefficients, what is the effect of gender (male vs. female) on the mean CRE, after accounting for age and diameter? Define this effect in general, then summarize the effect for each diameter of the clear zone in the table below.

Logic: For the general effect of gender, consider only terms that include male.

Solution: The effect of gender (male vs. female) on the mean of CRE is:

$$\begin{aligned}
&\mu\{CRE \mid Age, Male, Diameter\} - \mu\{CRE \mid Age, Female, Diameter\} \\
&= \mu\{CRE \mid Age, Male = 1, Diameter\} - \mu\{CRE \mid Age, Male = 0, Diameter\} \\
&= \beta_2 + \beta_6 Age + \beta_{10}(Age \times D1) + \beta_{11}(Age \times D2) + \beta_{12}(Age \times D3)
\end{aligned}$$

Diameter of the clear zone	Logic	Effect of gender (male vs. female) on the mean CRE
2.5 D1	includes slopes for male by itself	$\beta_2 + (\beta_6 + \beta_{10})Age$
3.0 D2	of male with age and by diameter	$\beta_2 + (\beta_6 + \beta_{11})Age$
3.5 D3		$\beta_2 + (\beta_6 + \beta_{12})Age$
4.0 0		$\beta_2 + \beta_6 Age$

e) Re-write the original model indicating that gender has no effect on mean CRE.

Solution: Gender has no effect on mean CRE if there is no gender in the model. Therefore,

$$\begin{aligned}
\mu\{CRE \mid Age, Diameter\} &= \beta_0 + \beta_1 Age + \beta_3 D1 + \beta_4 D2 + \beta_5 D3 \\
&+ \beta_7(Age \times D1) + \beta_8(Age \times D2) + \beta_9(Age \times D3)
\end{aligned}$$

SECTION 6: TWO-FACTOR ANOVA

- Factorial design involves two or more factors affecting the response variable and each factor has two or more levels
- When there are two factors affecting the response variable, it is known as two-factor ANOVA
- Just like one-factor ANOVA, there is only one variable being measured, that is, the response variable
- There are two factors being tested at the same time to determine whether they have an effect on the variable being measured
- For each of the two factors, there are two or more levels or treatments being applied to the individuals or experimental units being test
- Often the two factors are categorical
- Sometimes one or both of the two factors are quantitative, having several measureable levels or treatments
    - In this case, two-factor ANOVA is very similar to multiple linear regression
- Some of the advantages setting up one experiment to test the effects of two-factors on the response variable at the same time and analyzing the data with two-factor ANOVA, as opposed to doing two separate experiments for the effects of these two factors and analyzing them separately using one-factor ANOVA are as follows:
    1. Can test the effect of two factors at the same time, thus saving time and expenses
    2. Can compare the significance of the effects of the two factors
    3. Allows testing for interaction between the two factors
- **Main effects** = the effect of each factor considered separately
- **Interaction effect** = interaction between the two factors, that is, the effect of one factor on the response variable depends on the level of the other factor
    - The interaction effect may or may not occur

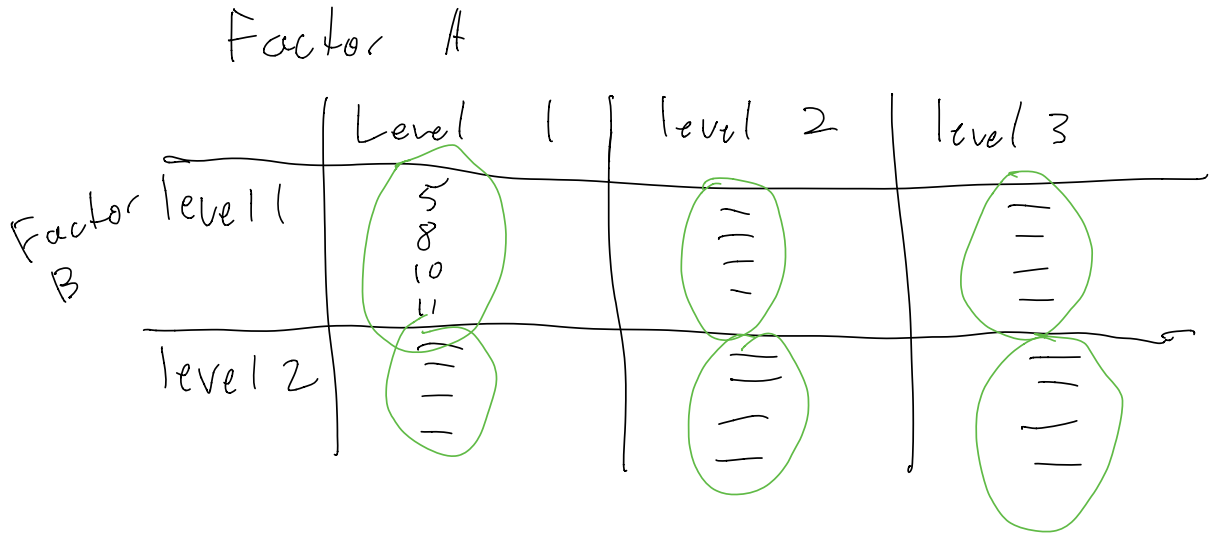
Fixed Effect Factors and Random Effect Factors

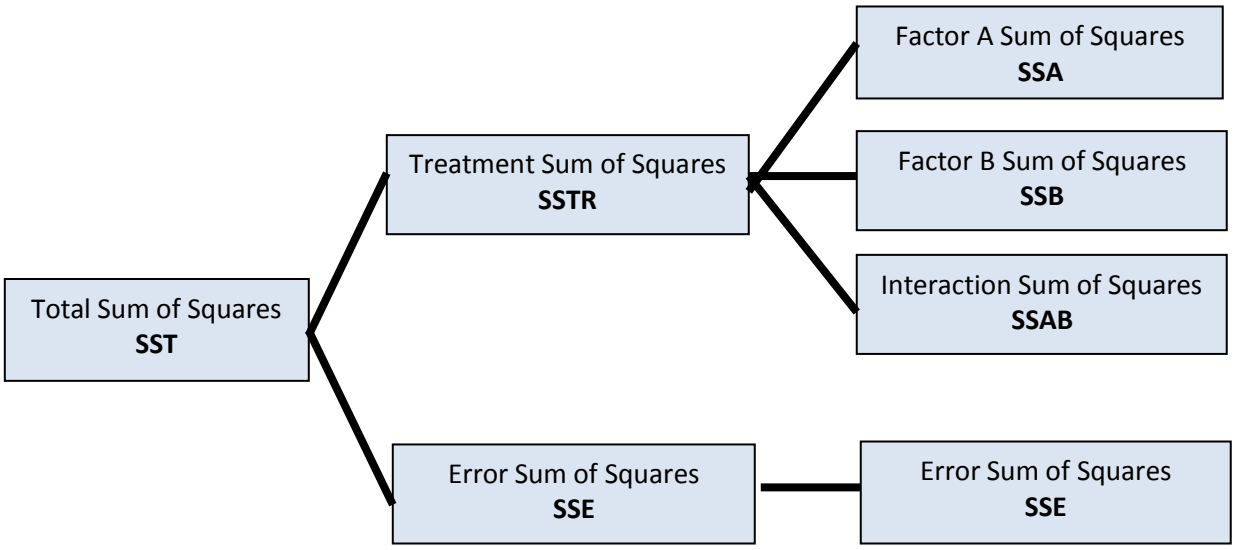
- **Fixed Effect Factors** are factors whereby the researcher deliberately fixes the levels of the factor because those are the levels of interest to him/her
- **Random Effect Factors** are factors for which the levels are selected or occur at random from a collection of possible levels

Assumptions of Two-Factor ANOVA

1. Random sampling
2. Independent observations: The observations of the response variable are independent of one another, though the levels of the factors do not need to be independent
3. Normal Distributions: For each combination of treatments, the response variable is normally distributed
4. Equal Standard deviations: The standard deviations of the response variable are the same for all combinations of treatments

Example


Factor A
Level 1
level 2
level 3
Factor B
level 1
level 2
5
8
10
11



Response = Overall mean + A Main Effect + B Main Effect + AB Interaction Effect + Error

6.1 Two-Factor ANOVA with Replication, Balanced Data

- Sample size (i.e., number of observations) at least 2 for each combination of treatments
- **Balanced data** means that the sample size is the same for all combinations of treatments

Two-Factor ANOVA Identity for Sums of Squares (balanced data):

Total Sum of Squares = Factor A Sum of Squares + Factor B Sum of Squares
+ AB Interaction Sum of Squares + Error Sum of Squares

$$SST = SSA + SSB + SSAB + SSE$$

Two-Factor ANOVA Identity for Degrees of Freedom (balanced data):

$$df(SST) = df(SSA) + df(SSB) + df(SSAB) + df(SSE)$$

Or

$$n - 1 = (a - 1) + (b - 1) + (a - 1)(b - 1) + (n - ab)$$

Two-Factor ANOVA Hypothesis Test (With Interaction) (Non-Additive Model)

Purpose: To perform hypothesis tests for the main effects and interaction effects of two factors
Assumptions: Given above

Null and Alternative Hypotheses

Overall Model: H_0: The overall model is not useful for making predictions
H_a: The overall model is useful for making predictions
Factor A main effect: H_0: There is no main effect due to Factor A
Ha: There is a main effect due to Factor A
Factor B main effect: H_0: There is no main effect due to Factor B
Ha: There is a main effect due to Factor B
AB interaction effect: H_0: The two factors do not interact
Ha: The two factors interact

ANOVA table for Two-Factor Analysis of Variance

Source of variation	**SS**	**df**	**MS = SS/df**	**F-statistic**
Overall model Corrected model	SSA +SSB +SSAB	df(A) +df(B) +dfAB) = ab – 1	$Corr\ MS = \frac{Corr\ \mathrm{SS}}{Corr\ df}$	$F_{Overall} = \frac{Corr\ MS}{MSE}$
Factor *A*	*SSA*	*a* – 1	$MSA = \frac{SSA}{a-1}$	$F_A = \frac{MSA}{MSE}$
Factor *B*	*SSB*	*b* – 1	$MSB = \frac{SSB}{b-1}$	$F_B = \frac{MSB}{MSE}$
AB Interaction	*SSAB*	(*a* – 1)(*b* – 1)	$MSAB = \frac{SSAB}{(a-1)(b-1)}$	$F_{AB} = \frac{MSAB}{MSE}$
Error (within)	*SSE*	*n – ab*	$MSE = \frac{SSE}{n-ab}$	
Total	***SST***	***n* – 1**		

F (Overall model) = F (Corrected model) $= \frac{\text{Corrected } SS/(ab-1)}{\text{Error } SS/(n-ab)} = \frac{\text{Corrected } MS}{\mathrm{M}SE}$

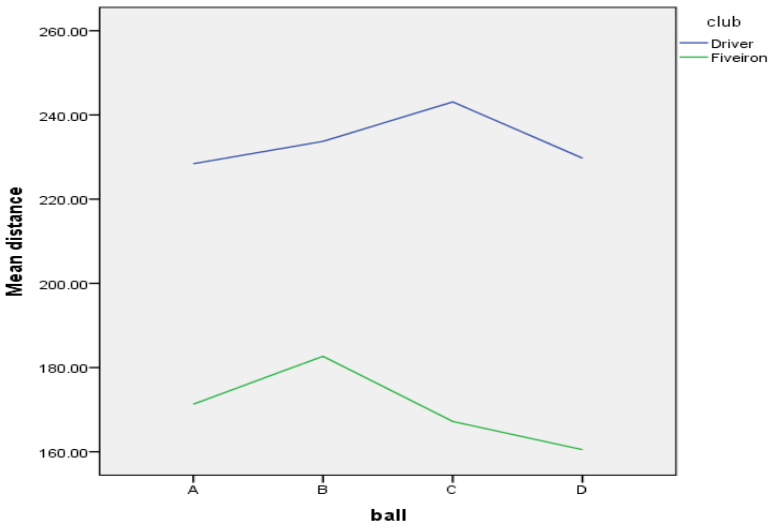
$$F_A = \frac{SSA/(a-1)}{SSE/(n-ab)} = \frac{MSA}{MSE} \qquad F_B = \frac{SSB/(b-1)}{SSE/(n-ab)} = \frac{MSB}{MSE} \qquad F_{AB} = \frac{SSAB/(a-1)(b-1)}{SSE/(n-ab)} = \frac{MSAB}{MSE}$$

Where: a = number of levels of Factor A
b = number of levels of Factor B
n = total number of observations
= a x b x (no. of replicates per combination of treatments)

Research Problem Involving Two-Factor ANOVA, Followed by Multiple Comparisons Tests
Example on Golf Balls

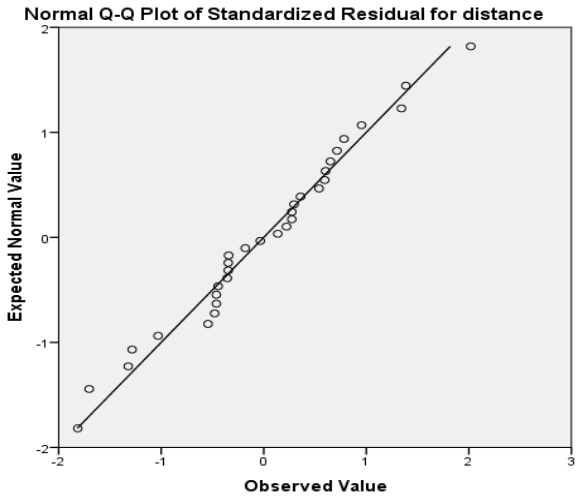
A golfer investigated the distances that golf balls are hit and how this is affected by the brand of ball used and the club used. This is a balanced design, with 4 replicates per combination of the two factors. Thus he was testing the following effects:

1. Effect of the four brands of balls (A, B, C and D)
2. Effect of two types of clubs (driver and five iron)
3. Effect of interaction between club and ball



Observe the following in this Line Chart:

1. The huge separation of the two lines indicates a virtually certain effect of the club factor, that is, a difference between the average of the means for the driver and the average of the means for the five iron.
2. The fact that the two lines go considerably up and down from A to B to C to D indicates a likely effect of the brand of the balls.
3. The non-parallel lines indicate an interaction effect between club and ball.
    - The segments from B to C run in completely different directions
    - They don't have to cross; if they are not parallel that indicates interaction



Checking for Normality
The above Q-Q Plot shows that the points are reasonably close to a straight line. Thus the distance variable is approximately normally distributed.

Tests of Between-Subjects Effects

Dependent Variable: distance

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	33652.830[a]	7	4807.547	140.689	.000
Intercept	1306819.028	1	1306819.028	38243.115	.000
club	32086.778	1	32086.778	938.996	.000
ball	801.348	3	267.116	7.817	.001
club * ball	764.703	3	254.901	7.459	.001
Error	820.113	24	34.171		
Total	1341291.970	32			
Corrected Total	34472.942	31			

a. R Squared = .976 (Adjusted R Squared = .969)

- Suppose the numbers highlighted in yellow are not given

>>>>>>>>>>

(a) At the 5% significance level, perform the most appropriate test to determine whether the overall model is significant.

H_0: All treatment combinations have equal means

H_A: At least 2 treatments have different means

Overall SS = $SS_A + SS_B + SS_{AB}$

$= 32086.778 + 801.348 + 764.703$

$= 33652.829$

or Correct total SS − Error SS

df => overall model df = $df(A) + df(B) + df(AB) = ab - 1 = 3 + 3 + 1 = 7$

$F(\text{overall}) = F(\text{corrected}) = \dfrac{\text{corrected SS / corrected df}}{\text{Error SS / Error df}}$

$\dfrac{33652.830/7}{820.113/24} = 140.689$

$df = (7, 24)$ $p < 0.001$ Extremely strong evidence $p < \alpha$ of 0.05 ∴ we reject H_0.

Conclusion: you get the idea.

(b) At the 5% significance level, perform the most appropriate test to determine whether there is a main effect of club type on mean distance.

H_0: There is no main effect on club type
H_A: There is a main effect on club type
F (main effect of club type)

$$F_A = \frac{SSA/(a-1)}{SSE/(n-ab)} = \frac{MSA}{MSE} = \frac{32086.778/(2-1)}{820.113/(32-(2)(4))} = 938.996$$

$df = (1, 24)$ $p < 0.001$ since $p < \alpha$ of 0.001

At the 5% sig level, the data provide sufficient evidence that there is a significant main effect of club type, that is, the mean distance are not all the same for the different club types, averaging over all brands.

(c) At the 5% significance level, perform the most appropriate test to determine whether there is a main effect of ball brand on mean distance.

H_0: There is no effect on ball brand
H_A: There is an effect on ball brand

$$F_B = \frac{SSB/(b-1)}{SSE/(n-ab)} = \frac{MSB}{MSE} = \frac{801.348/(4-1)}{820.113/(32-(2)(4))} = 7.817$$

$df = (3, 24)$ $p < 0.001$ Extremely strong evidence
Since $p < \alpha$ of 0.05 we reject H_0.
you get the idea at this point for the conclusion.

(d) At the 5% significance level, perform the most appropriate test to determine whether the effect of club type on mean distance depends on ball brand. (In other words, test whether there is an interaction effect between club type and ball brand.)

H_0: There is no interaction between club type and ball brand
H_A There is an interaction between club type and ball brand
F (interactions) $F_{AB} = \frac{SSAB/(a-1)(b-1)}{SSE/(n-ab)} = \frac{MSAB}{MSE} = \frac{769.703/(2-1)(4-1)}{820.113/(32-(2)(4))} = 7.459$
$df(3, 24)$ $0.001 < p < 0.005$ very strong evidence against H_0
Since $p < 0.05$ we reject H_0.
you get the idea at this point for the conclusion.

>>>>>>>>>>

Multiple Comparisons

Point Estimates + Confidence Intervals for each Combination of Club Type and Ball Brand

Estimates					
Dependent Variable: distance					
club	Ball	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Driver	A	228.425	2.923	222.393	234.457
	B	233.725	2.923	227.693	239.757
	C	243.100	2.923	237.068	249.132
	D	229.750	2.923	223.718	235.782
Fiveiron	A	171.300	2.923	165.268	177.332
	B	182.675	2.923	176.643	188.707
	C	167.200	2.923	161.168	173.232
	D	160.500	2.923	154.468	166.532

Pairwise Comparisons: For each ball brand, comparisons between club types

Pairwise Comparisons							
Dependent Variable: distance							
ball	(I) club	(J) club	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b]	
						Lower Bound	Upper Bound
A	Driver	Fiveiron	57.125*	4.133	.000	48.594	65.656
	Fiveiron	Driver	-57.125*	4.133	.000	-65.656	-48.594
B	Driver	Fiveiron	51.050*	4.133	.000	42.519	59.581
	Fiveiron	Driver	-51.050*	4.133	.000	-59.581	-42.519
C	Driver	Fiveiron	75.900*	4.133	.000	67.369	84.431
	Fiveiron	Driver	-75.900*	4.133	.000	-84.431	-67.369
D	Driver	Fiveiron	69.250*	4.133	.000	60.719	77.781
	Fiveiron	Driver	-69.250*	4.133	.000	-77.781	-60.719
Based on estimated marginal means							
*. The mean difference is significant at the .05 level.							
b. Adjustment for multiple comparisons: Bonferroni.							

- The point estimate for the multiple comparisons
  = Difference between each pair of means (SPSS calls this Mean Difference)
  = $(x_i - x_j)$
- For example, difference between the means for Ball A, Driver and Ball A, Fiveiron
  $= (x_i - x_j) = (228.425 - 171.300) = 57.125$ (shown in the table above)

Pairwise Comparisons for Interactions (Club*Ball)
(Shown in this table for each club type, comparisons between ball brands)

- For example, the first comparison is: (Driver*Ball A compared with Driver*Ball B)

Pairwise Comparisons							
Dependent Variable: distance							
club	(I) ball	(J) ball	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b]	
						Lower Bound	Upper Bound
Driver	A	B	-5.300	4.133	1.000	-17.184	6.584
		C	-14.675*	4.133	.010	-26.559	-2.791
		D	-1.325	4.133	1.000	-13.209	10.559
	B	A	5.300	4.133	1.000	-6.584	17.184
		C	-9.375	4.133	.196	-21.259	2.509
		D	3.975	4.133	1.000	-7.909	15.859
	C	A	14.675*	4.133	.010	2.791	26.559
		B	9.375	4.133	.196	-2.509	21.259
		D	13.350*	4.133	.021	1.466	25.234
	D	A	1.325	4.133	1.000	-10.559	13.209
		B	-3.975	4.133	1.000	-15.859	7.909
		C	-13.350*	4.133	.021	-25.234	-1.466
Fiveiron	A	B	-11.375	4.133	.067	-23.259	.509
		C	4.100	4.133	1.000	-7.784	15.984
		D	10.800	4.133	.092	-1.084	22.684
	B	A	11.375	4.133	.067	-.509	23.259
		C	15.475*	4.133	.006	3.591	27.359
		D	22.175*	4.133	.000	10.291	34.059
	C	A	-4.100	4.133	1.000	-15.984	7.784
		B	-15.475*	4.133	.006	-27.359	-3.591
		D	6.700	4.133	.709	-5.184	18.584
	D	A	-10.800	4.133	.092	-22.684	1.084
		B	-22.175*	4.133	.000	-34.059	-10.291
		C	-6.700	4.133	.709	-18.584	5.184
Based on estimated marginal means							
*. The mean difference is significant at the .05 level.							
b. Adjustment for multiple comparisons: Bonferroni.							

>>>>>>>>>> → for Driver

Means Comparisons Diagrams (based on the Bonferroni Method)

A 228.425 | D 229.250 | B 233.725 | C 243.100

you get the idea for conclusion

We can be 95% confident that, when using the Driver, the mean distance for ball C is different from A and D but no other difference between the rest.

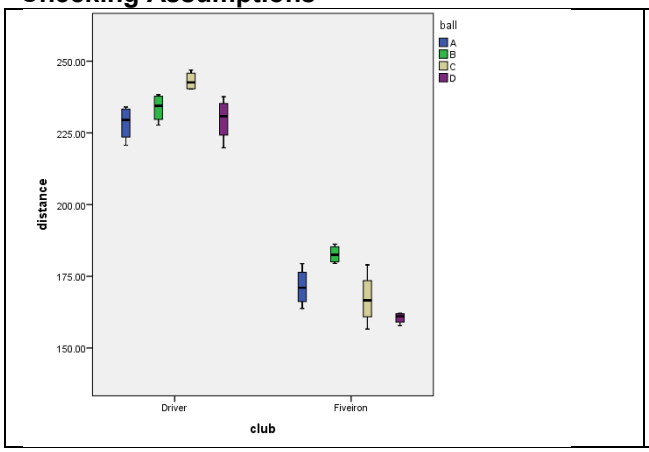
>>>>>>>>>>

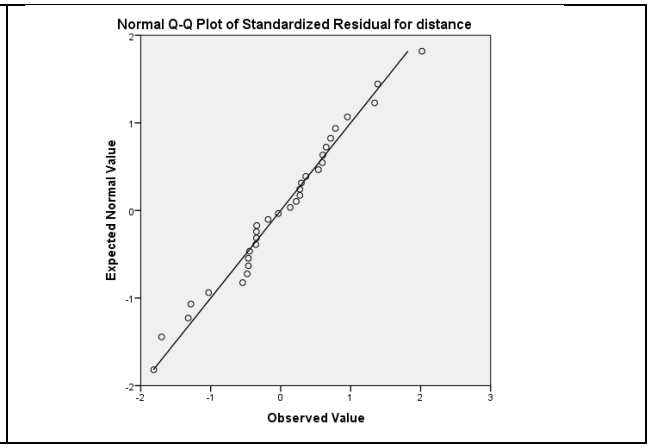
- The very different results of the multiple comparisons for Driver as opposed to Five Iron (since the ascending order of the means are different) also indicates interaction between club type and ball type
- These could be joined together to compare all 8 means in one diagram, but there would be complete separation between the means for Driver and means for Five Iron.

Means Comparisons for Interactions

- The multiple comparisons for interactions on the previous page does not show all of them because it just gives them separately for the two club types.
- Combinations = 4 x 2 = 8, so there are actually [8(8-1)]/2 = 28 multiple comparisons for interactions

Checking Assumptions





Levene's Test of Equality of Error Variances[a]

Dependent Variable: distance

F	df1	df2	Sig.
1.269	7	24	.307

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept + club + ball + club * ball

Non-Additive and Additive Models in Two-Factor ANOVA

- Non-additive Model (includes interaction) (Full Model)
- Additive Model (does not include interaction) (Reduced Model)
- Often compared using the Extra Sum-of-Squares F-test

Previous Example on Golf Balls (Comparing Non-Additive and Additive Model)

A golfer investigated the distances that golf balls are hit and how this is affected by the brand of ball used and the club used and the interaction between ball brand and club type.

Consider the following two models

$$\text{Model 1}: \mu(\text{Distance} \mid \text{Club, Ball}) = \beta_0 + \text{Club} + \text{Ball}$$

[No interaction –Additive model >>> Reduced model]

$$\text{Model 2}: \mu(\text{Distance} \mid \text{Club, Ball}) = \beta_0 + \text{Club} + \text{Ball} + (\text{Club*Ball})$$

[With Interaction – Non-additive model >>> Full model]

Additive Model (No Interaction – Reduced Model)

Tests of Between-Subjects Effects

Dependent Variable: distance

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	32888.127	4	8222.03175	140.076	.000
club	32086.778	1	32086.778	546.652	.000
ball	801.348	3	267.116	4.551	.010
Error	1584.816	27	58.6969		
Corrected Total	34472.942	31			

a. R Squared = .976 (Adjusted R Squared = .969)

Non-Additive Model (With Interaction – Full Model)

Tests of Between-Subjects Effects

Dependent Variable: distance

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	33652.830[a]	7	4807.547	140.689	.000
Intercept	1306819.028	1	1306819.028	38243.115	.000
club	32086.778	1	32086.778	938.996	.000
ball	801.348	3	267.116	7.817	.001
club * ball	764.703	3	254.901	7.459	.001
Error	820.113	24	34.171		
Total	1341291.970	32			
Corrected Total	34472.942	31			

a. R Squared = .976 (Adjusted R Squared = .969)

Testing for An Interaction Effect by comparing an Additive and Non-additive Model

(a) At the 5% significance level, perform an Extra Sum-of-Squares F-test, using the additive and non-additive models presented in the tables above, in order to determine whether the effect of club type depends on ball brand (in other words, whether there is an interaction effect between club type and ball brand), after accounting for club type and ball brand.

>>>>>>>>>>

H_0: μ Distance | club, ball) = β_0 + club + Ball

H_A: μ Distance | club, ball) = β_0 + club + Ball + (club * ball)

$$F = \frac{(SS_E \text{ reduced}) - SS_E(\text{full})) / (df_E \text{ reduced} - df_E(\text{full}))}{SS_E(\text{full}) / df_E(\text{full})}$$

$$= \frac{(1584.816 - 820.113) / (27 - 24)}{820.113/24} = \frac{254.901}{34.1714} = 7.459$$

$df = (3, 24)$ $0.001 < p < 0.005$ very strong evidence against H_0

Since $p < 0.05$ we reject H_0.

Conclusion: you get the idea.

Compared with the approach on page 6.

(b) Now suppose the Interaction is ignored, use the additive model to determine whether either club type or ball brand have an effect on mean distance. Perform the test at the 5% significance level.

H_0: Neither factor has an effect on mean distance

H_A: At least one factor has an effect on mean distance

$$F = \frac{[32086.778 + 801.348] / [(2-1) + (4-1)]}{1584.816/27}$$

$$= \frac{32888.126/4}{1584.816/27} = 140.076$$

$df = (4, 27)$ $p < 0.001$ Since $p < \alpha$ of 0.05 we reject H_0.

Extremely strong evidence against H_0.

At the 5% sig level, the data provide sufficient evidence to conclude that either club type or ball brand or both have an effect on mean distance.

>>>>>>>>>>

Compare the Results of the non-additive and additive models

	Full Model, OR Non-additive Model (Interaction)	**Reduced Model, OR Additive Model (No interaction)**
Overall Model	F = 140.689 Df = (7, 24) P-value = 0.00000000 $F_{0.001}$ = 5.23 (may be more significant than additive)	F = 140.076 Df = (4, 27) P-value = 0.00000000 $F_{0.001}$ = 6.33
Effect of Club type	F = 938.996 Df = (1, 24) P-value = 0.00000000	F = 546.652 Df = (1, 27) P-value = 0.00000000
Effect of Ball brand	F = 7.817 Df = (3, 24) P-value = 0.000824	F = 4.551 Df = (3, 27) P-value = 0.010478
Effect of interaction	F = 7.459 Df = (3, 24) P-value = 0.001074	**None**

Note: The overall model and the effect of club type and ball brand were all more significant with the non-additive model (with interaction) than the additive model. This is because, when the interaction term is significant (such as in this example), the interaction model is more accurate and more effective in showing significant effects.

6.2 Randomized Block Design

- Analyzed with Randomized Block Analysis of Variance (ANOVA)
- Considered as an extension of the paired design
- A special type of Two-Factor ANOVA where the "block" factor is not of interest to the researcher.
- One advantage is that it allows the researcher to eliminate the effect of the block factor so that it does not affect the real factor of interest.
- Thus it is much more powerful in testing the effect of the factor of interest than ordinary ANOVA

Blocks in space

- Suppose an experiment was conducted to determine whether there is a difference in the effectiveness of four new types of fertilizers (A, B, C, and D)
- However, there is a gradient of conditions in the test area due to a gradual slope towards a river
- Experimental area is divided into blocks such that it can be assumed that the conditions are homogenous within each block, even though conditions vary among blocks.
    - Thus it eliminates the effect of extraneous variables in space.
- In each block, each treatment is represented once.

Gradient in moisture &nutrients ↓

Blocks (in space)				
1	C	A	D	B
2	B	D	A	C
3	B	C	D	A
4	D	A	B	C
5	A	C	D	B
	RIVER			

Blocks in time

- Eliminate the effect of time on the observations
    - E.g., If a researcher wants to compare 4 sites and he cannot take several measurements in all sites at the same time, he can take 1 measurement in each site every month
- This will eliminate the effect of temporal variation (e.g., seasonal variation or day-to-day variation) on the results.

Gradient in Time ↓

	Abundance of birds			
Blocks (in time)	**Site A**	**Site B**	**Site C**	**Site D**
Oct	11	8	9	15
Nov	13	10	12	16
Jan	4	2	3	7
Feb	9	6	7	14
March	20	16	17	25

Applications in Medical Sciences, Education, Psychology, Etc.

- The before-and-after "treatment" can be extended to monitoring patients every few hours, once a month, etc., or subjects during some intervention program in education or psychology
- In this case, the blocks are the subjects or patients
- Greatly increases the power of the test in detecting responses of subjects/patients to treatments or programs

Randomized Block ANOVA

Extra Assumption (in addition to the other assumptions of two-way ANOVA:
There is no significant interaction between the main factor of interest (treatment) and the blocks factor (which is not of interest). This can be checked graphically with a line graph.

Null and alternative hypotheses for the factor of interest:
H_0: There is no significant difference between population means of the main factor of interest (treatment)
H_a: There is a significant difference between population means of the main factor of interest (treatment)

Null and alternative hypotheses for blocks (not always tested because not main interest):
H_0: There is no effect of the block factor.
Ha: There is an effect of the block factor

Calculations for Randomized Block ANOVA:

Sum of Squares	Defining formula
Treatment (SSTR)	$SSTR = \sum_{i=1}^{k} b(\bar{x}_{Ti} - \bar{x})^2$
Block (SSBL)	$SSBL = \sum_{i=1}^{b} k(\bar{x}_{Bj} - \bar{x})^2$
Total (SST)	$SST = \sum_{i,j} (x_{ij} - \bar{x})^2$
Error (SSE)	$SSE = SST - SSTR - SSBL$

Source of variation	SS	df	MS = SS/df	F-statistic
Treatment	SSTR	k – 1	MSTR = SSTR / (k – 1)	F = MSTR/MSE
Block	SSBL	b - 1	MSBL = SSBL / (b – 1)	F = MSBL/MSE
Error	SSE	(k – 1)(b – 1)	MSE = SSE / (k – 1)(b – 1)	
Total	SST	n - 1		

Note: Error df = (k – 1)(b – 1) = n – k – b + 1 = Total df – (Treatment df + Blocks df)

$$F_{Treatment} = \frac{SSTR/(k-1)}{SSE/(k-1)(b-1)} = \frac{MSTR}{MSE}$$

$$df = [(k-1),(k-1)(b-1)]$$

$$F_{Blocks} = \frac{SSBL/(b-1)}{SSE/(k-1)(b-1)} = \frac{MSBL}{MSE}$$

$$df = [(b-1),(k-1)(b-1)]$$

Where k = number of treatments, b = number of blocks, n = kb = total number of observations

Research Problem: An experiment was conducted to test the effectiveness of four types of fertilizers on eggplants (*Solanum melongena*). The test area was a low-lying area near a river. Eggplants are particularly sensitive to soil fertility and structure as well as soil moisture. Therefore the gradients in these factors towards the river would have affected the experiment. Thus, a randomized block design was used to eliminate the effect of these gradients.

Table of raw data, including treatment means, block means and grand mean (Units = kg/m^2):

Block	Fertilizer Treatment				Block Means
	A	B	C	D	
1	2.7	2.8	2.9	2.8	$\bar{x}_{B1} = 2.8$
2	2.9	3.0	3.2	2.9	$\bar{x}_{B2} = 3.0$
3	3.0	3.1	3.3	3.0	$\bar{x}_{B3} = 3.1$
4	3.3	3.3	3.6	3.4	$\bar{x}_{B4} = 3.4$
5	3.3	3.4	3.5	3.4	$\bar{x}_{B5} = 3.4$
Treatment Means	$\bar{x}_{T1} = 3.04$	$\bar{x}_{T2} = 3.12$	$\bar{x}_{T3} = 3.30$	$\bar{x}_{T4} = 3.10$	$\bar{x} = 3.14$

Calculate Four Sums of Squares

Treatment SS between treatments

$$SSTR = \sum_{i=1}^{k} b(\bar{x}_{Ti} - \bar{x})^2 = 5(3.04-3.14)^2 + 5(3.12-3.14)^2 + 5(3.30-3.14)^2 + 5(3.10-3.14)^2 = 0.188$$

Block SS between blocks

$$SSBL = \sum_{i=1}^{b} k(\bar{x}_{Bj} - \bar{x})^2$$

$$= 4(2.8-3.14)^2 + 4(3.0-3.14)^2 + 4(3.1-3.14)^2 + 4(3.4-3.14)^2 + 4(3.4-3.14)^2 = 1.088$$

Total SS

$$SST = \sum_{i,j} (x_{ij} - \bar{x})^2 = (2.7-3.14)^2 + (2.8-3.14)^2 + (2.9-3.14)^2 + ... + (3.4-3.14)^2 = 1.308$$

Error SS

$$SSE = SST - SSTR - SSBL = 1.308 - 0.188 - 1.088 = 0.032$$

>>>>>>>>>>

Source of variation	SS	df	MS	F
Treatment	0.188	$k-1$ $4-1=3$	$0.188/3$ $=0.06267$	23.50
block	1.088	$b-1$ $5-1=4$	$1.088/4$ $=0.27200$	101.87
Error	0.032	$(k-1)(b-1)$ $3 \cdot 4 = 12$	$0.032/12$ $=0.00267$	
Total	1.308	$n = kb-1$ $(4 \cdot 5)-1=19$		

(a) At the 5% significance level, test whether there is a difference in mean eggplant yield between the four fertilizer treatments (the factor of interest).

H_0: There is no difference in mean eggplant yield between the 4 fertilizer

H_A: There is a difference in mean eggplant yield between the 4 fertilizer

$F_{Treatment} = \frac{SSTR/k-1}{SSE/(k-1)(6-1)} = \frac{0.1881/(4-1)}{0.0321/3.4} = \frac{0.06267}{0.00267} = 23.50$

$df(3,12)$ $p < 0.001$ extremely strong evidence against H_0

Since $p < \alpha$ of 0.05 we reject H_0.

Conclusion you get the idea. only compared between fertilizer

(b) At the 5% significance level, test whether there is a difference in mean eggplant yield between blocks (in space), which is not the factor of interest.

H_0: There is no effect on the block factor on mean eggplant yield

H_A: There is a effect on the block factor on mean eggplant yield

$F_{block} = \frac{SSBL/b-1}{SSE/(k-1)(6-1)} = \frac{1.0881/5-1}{0.0321/3.4} = \frac{0.27200}{0.00267} = 101.87$

$df(4,12)$ $p < 0.001$ extremely strong evidence against H_0.

Since $p < \alpha$ of 0.05 we reject H_0. P-value $= 3.71 \times 10^{-9}$

Conclusion you get the idea. only compared between blocks

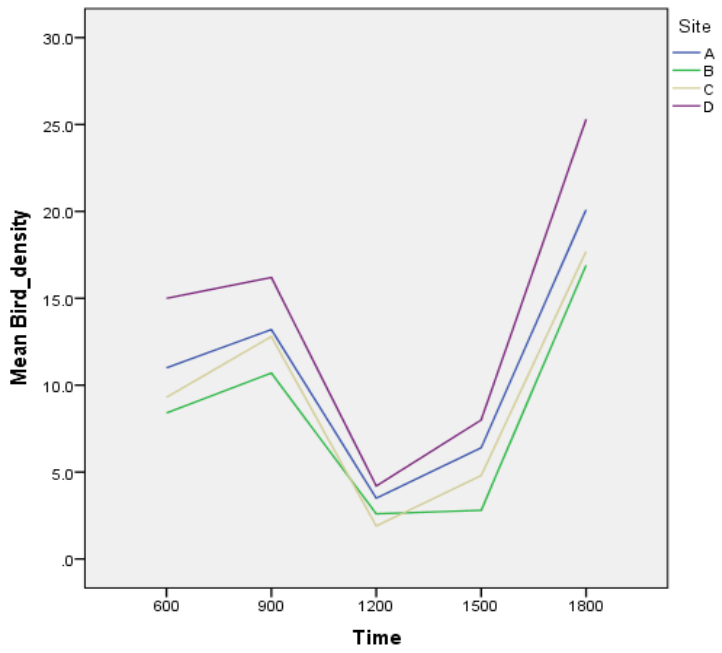
>>>>>>>>>>

Demonstrate what would happen if we IGNORE the Blocks Factor and analyze this data set with Single-Factor ANOVA (Use Excel)

Result: F = 0.895 **, df =** 3, 16 **, P =** 0.46501

Example of Applying Randomized Block ANOVA to Blocks in Time

As part of planning wildlife conservation strategies, a group of ecologists wanted to determine whether there was a significant difference in bird density between four sites (A, B, C, and D). They chose a randomized block design where blocks were times of the day and they recorded bird density simultaneously at the four sites at 5 times of the day (600 hrs, 900 hrs, 1200 hrs, 1500 hrs, and 1800 hrs). Make use of the line graph and ANOVA table with missing values to answer the questions below.



Note:

1. If a block design was not used, the great variation in time of day would hide any differences between sites.
2. The lines are almost parallel, indicating little or no interaction between subject and test.

Tests of Between-Subjects Effects

Dependent Variable: Bird_density

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	798.279[a]	7	114.040	80.267	.000
Intercept	2221.832	1	2221.832	1563.844	.000
Site	84.876	3	28.292	19.913	.000
Time	713.403	4	178.351	125.533	.000
Error	17.049	12	1.421		
Total	3037.160	20			
Corrected Total	815.328	19			

a. R Squared = .979 (Adjusted R Squared = .967)

(a) Does the line graph above indicate interaction? Explain your answer.

The lines are more or less parallel, which means that there is no interaction between the factor of interest (site) and blocks (time of the day), which is not of interest, in their effect on bird density.

(b) At the 5% significance level, test whether there is a difference in mean bird density between the four sites (the factor of interest).

H_0: There is no difference in mean bird density between the four sites
Ha: There is a difference in mean bird density between sites (means of at least two sites are different).

$$F_{Treatment} = \frac{SSTR/(k-1)}{SSE/(k-1)(b-1)} = \frac{84.876/(4-1)}{17.049/(4-1)(5-1)} = \frac{28.292}{1.421} = 19.913$$

$$df = [(k-1),(k-1)(b-1)] = [(4-1),(4-1)(5-1)] = (3,12)$$

P < 0.001 Since P < α (0.05), reject Ho since there is very strong evidence.

Conclusion: At the 5% significance level, the data provide sufficient evidence to conclude that there is a difference in mean bird density between sites (the means of at least two sites are different).

(c) At the 5% significance level, test whether there is a difference in mean bird density between blocks (time of day), which is not the factor of interest.

H_0: There is no effect of the block factor (time of day).
Ha: There is an effect of the block factor

$$F_{Blocks} = \frac{SSBL/(b-1)}{SSE/(k-1)(b-1)} = \frac{MSBL}{MSE} = \frac{713.403/(5-1)}{17.049/(4-1)(5-1)} = \frac{178.351}{1.421} = 125.533$$

$$df = [(b-1),(k-1)(b-1)] = [(5-1),(4-1)(5-1)] = (4,12)$$

P < 0.001, Since P < α (0.05), reject Ho.

Conclusion: At the 5% significance level, the data provide sufficient evidence to conclude that there is a significant effect of blocks (time of day) on bird density.

Note: The block effect is probably much greater than the effect of site which can be seen in the Line Graph.

(d) Judging by the line graph and the ANOVA output do you think the same conclusion would have been reached if a completely randomized design (analyzed with one-way ANOVA) had been applied by the researchers instead of the randomized block design. Explain the logic of your answer.

The line graph shows a huge variation in bird density from one time of the day to another. In fact, this variation is much greater than the variation in bird density between sites. Therefore, if the data were analyzed with completely randomized one-way ANOVA it would likely have shown no difference between sites because this difference would have been overshadowed by the difference over time. With such a research design, randomized block ANOVA is much more powerful than One-Way ANOVA.

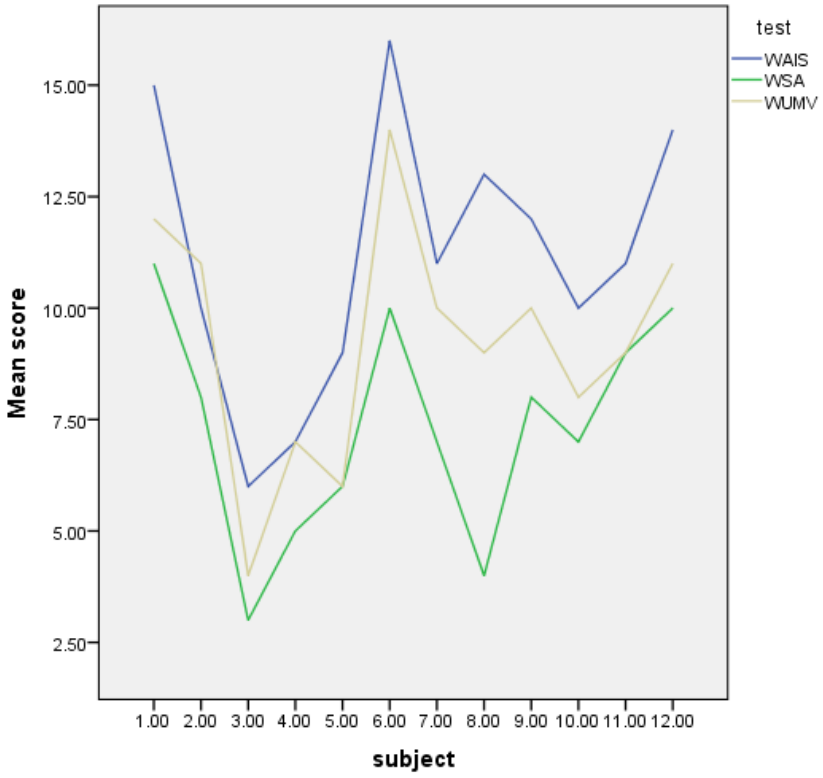
Example on Comparing Language Test Scores

A random sample of 12 subjects were given three different types of language tests as follows:

1. WAIS vocabulary (linguistic)
2. Willner Unusual Meanings Vocabulary (WUMV) pragmatic
3. Willner-Sheerer Analyogy Test (WSA) pragmatic

Since people have very different linguistic ability, in order to account for, or eliminate, the subjects' abilities, the randomized block design was selected. Therefore the same 12 subjects took all three language tests. At the 5% significance level, test whether there is any difference in mean scores attained on the three language tests.

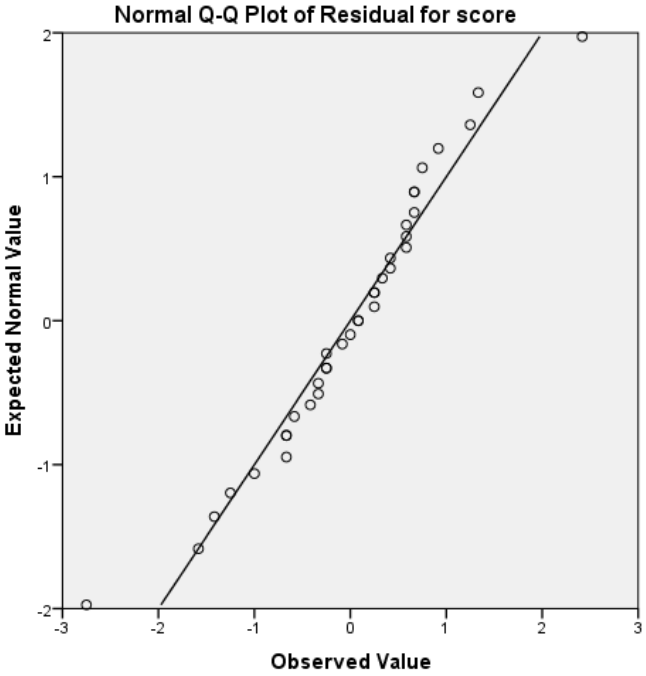
Line Graph



Note:

1. If a block design was not used, the great variation from one subject to the other would hide any differences between language tests.
2. The lines are almost parallel, indicating little or no interaction between subject and test.

Q-Q Plot to examine the assumption of normality



Note: the data points fall roughly along a straight line, indicating that the data are approximately normally distributed.

Randomized Block ANOVA (SPSS Output)

Tests of Between-Subjects Effects

Dependent Variable: score

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	310.250[a]	13	23.865	17.214	.000
Intercept	3080.250	1	3080.250	2221.820	.000
test	88.167	2	44.083	31.798	.000
subject	222.083	11	20.189	14.563	.000
Error	30.500	22	1.386		
Total	3421.000	36			
Corrected Total	340.750	35			

a. R Squared = .910 (Adjusted R Squared = .858)

Test for a Difference in Mean Score Between the Three Language Tests

H_0: There is no difference in mean score between the three language tests

Ha: There is a difference in mean score between the three language tests

F (Treatment)

$$F_{Treatment} = \frac{SSTR/(k-1)}{SSE/(k-1)(b-1)} = \frac{MSTR}{MSE} = \frac{88.167/(3-1)}{30.500/(3-1)(12-1)} = \frac{44.0835}{1.3864} = 31.798$$

$$df = [(k-1),(k-1)(b-1)] = [(3-1),(3-1)(12-1)] = (2,22)$$

P < 0.005 Since P < α (0.05), reject Ho with very strong evidence

Conclusion: The data provide sufficient evidence that there is a significant difference in mean score between the three language tests

Test for Effect of Blocks

H_0: There is no effect of the block factor.

Ha: There is an effect of the block factor

F (Blocks)

$$F_{Blocks} = \frac{SSBL/(b-1)}{SSE/(k-1)(b-1)} = \frac{MSBL}{MSE} = \frac{222.083/(12-1)}{30.500/(3-1)(12-1)} = \frac{20.189}{1.3864} = 14.563$$

$$df = [(b-1),(k-1)(b-1)] = [(12-1),(3-1)(12-1)] = (11,22)$$

P < 0.005 Since P < α (0.05), reject Ho with very strong evidence

Conclusion: The data provides sufficient evidence that there is a significant effect of blocks.

Note: The block effect is probably much greater than the effect of language score (even though the F-statistic is lower) due to the higher numerator df. Also that can be seen in the Line Graph.

Multiple Comparisons

Dependent Variable: score

Tukey HSD

(I) test	(J) test	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
WAIS	WSA	3.8333*	.48069	.000	2.6258	5.0409
	WUMV	1.9167*	.48069	.002	.7091	3.1242
WSA	WAIS	-3.8333*	.48069	.000	-5.0409	-2.6258
	WUMV	-1.9167*	.48069	.002	-3.1242	-.7091
WUMV	WAIS	-1.9167*	.48069	.002	-3.1242	-.7091
	WSA	1.9167*	.48069	.002	.7091	3.1242

Based on observed means.

The error term is Mean Square(Error) = 1.386.

*. The mean difference is significant at the 0.05 level.

Conclusion: All pairwise comparisons show significant differences.