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

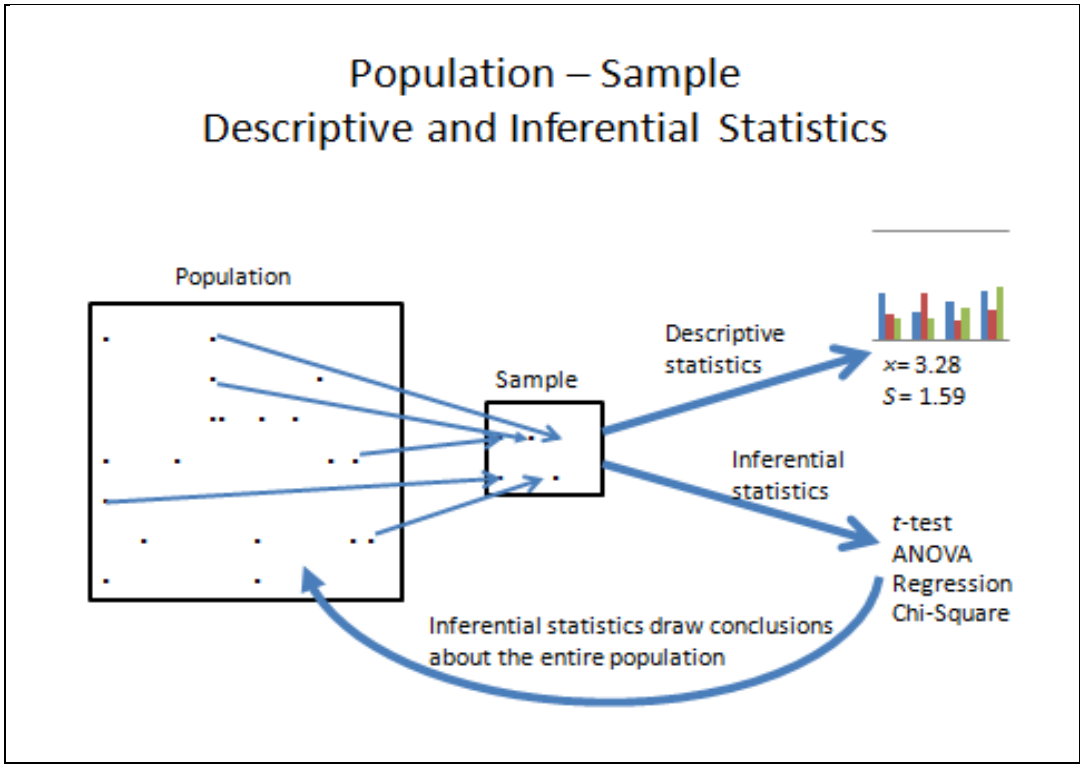
- Often too expensive or even impossible to undertake

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:

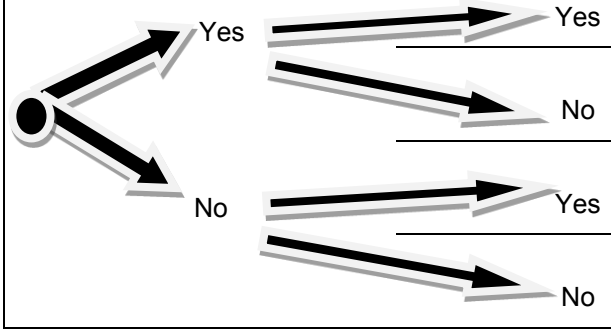
- Those who could affect the results (the subjects, treatment administrators or technicians).
- Those who evaluate the results (judges, treating physicians, researchers).
- **Single-blind experiment** – all individuals of one or the other of the above groups are blind.
- **Double blind experiment** –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

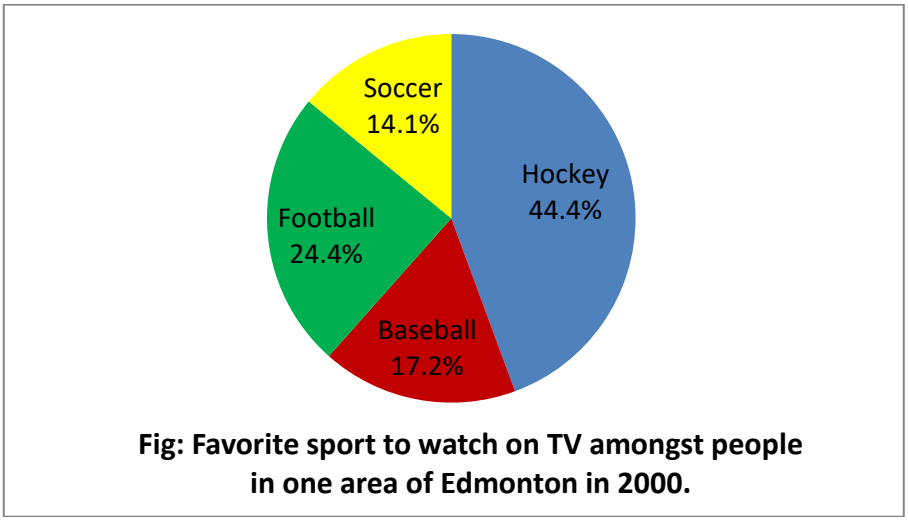


Fig: Favorite sport to watch on TV amongst people in one area of Edmonton in 2000.

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

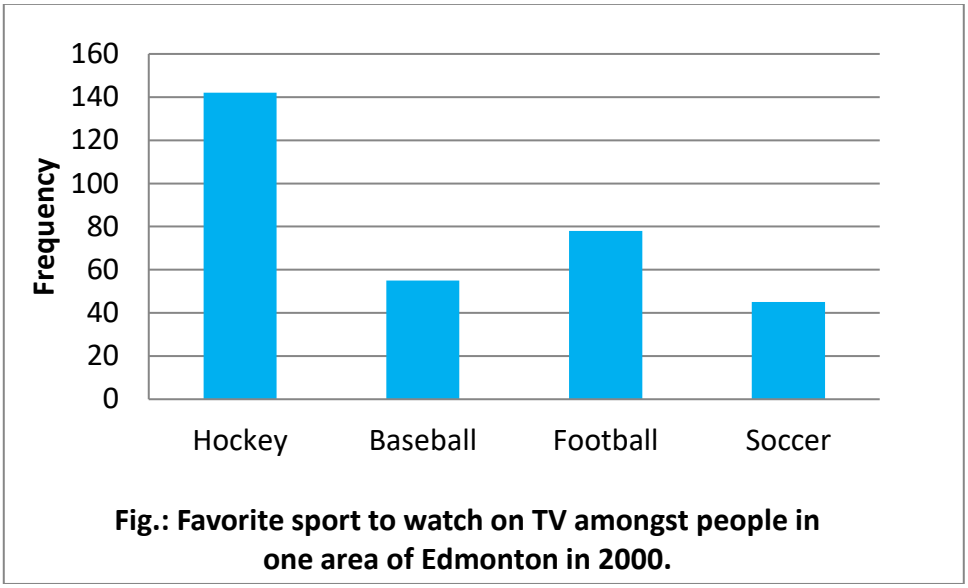


Fig.: Favorite sport to watch on TV amongst people in one area of Edmonton in 2000.

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

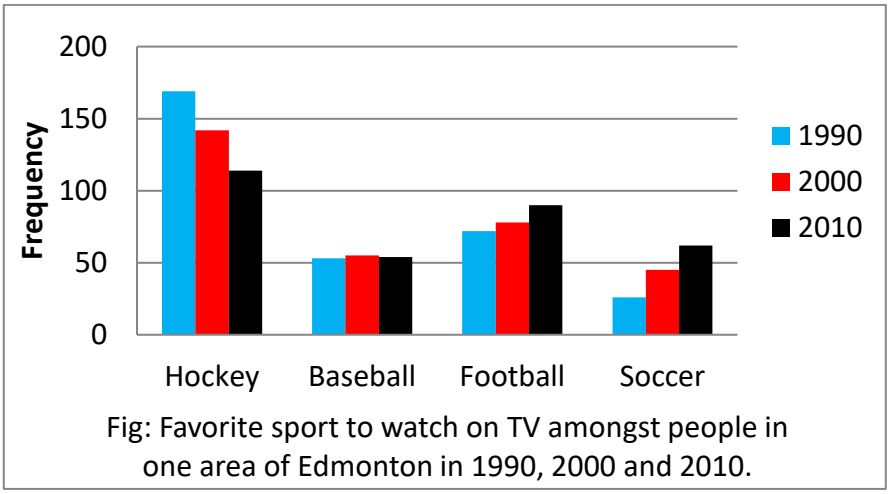


Fig: Favorite sport to watch on TV amongst people in one area of Edmonton in 1990, 2000 and 2010.

- 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

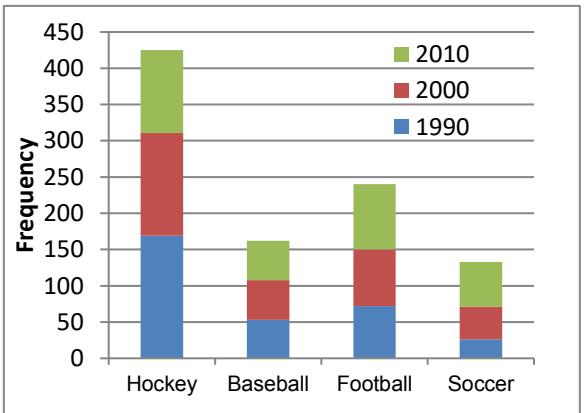


Fig.: Favorite sport to watch on TV amongst people in one area of Edmonton in 1990, 2000 and 2010.

Percent Segmented (Component) Bar Graph

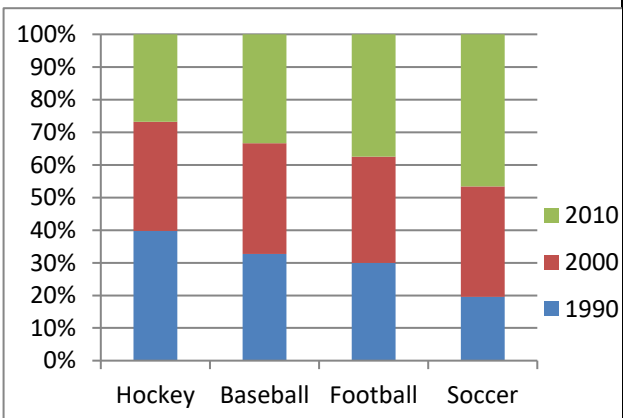


Fig.: Favorite sport to watch on TV amongst people in one area of Edmonton in 1990, 2000 and 2010.

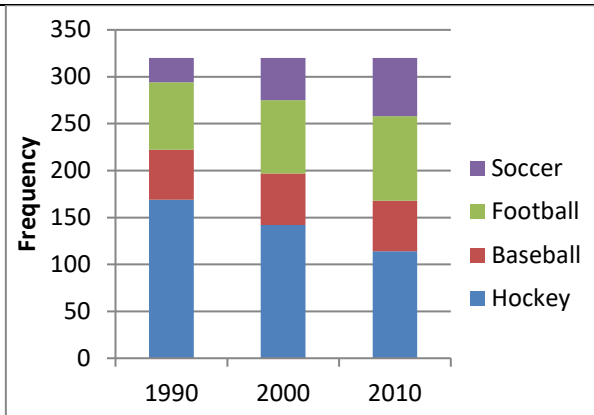


Fig.: Favorite sport to watch on TV amongst people in one area of Edmonton in 1990, 2000 and 2010.

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

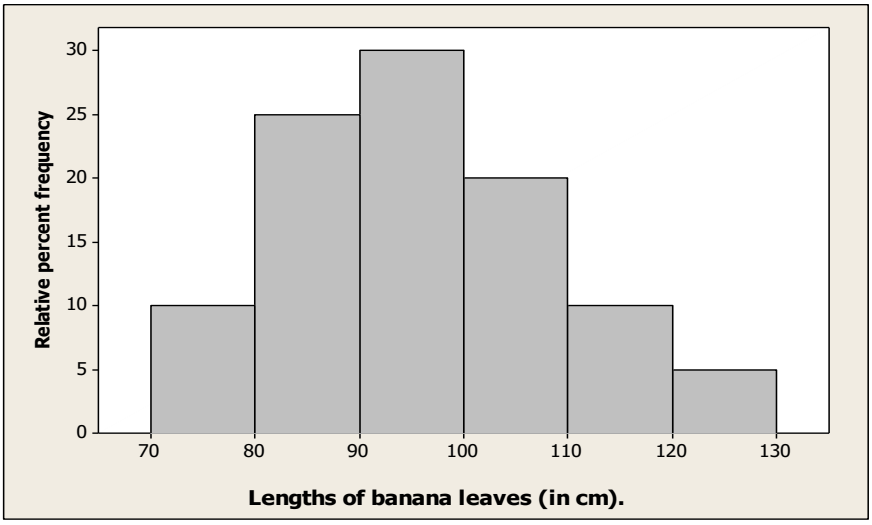
- Like a bar graph, but **no space** between bars
- Y-axis can show frequency, relative frequency or relative percent frequency

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**	



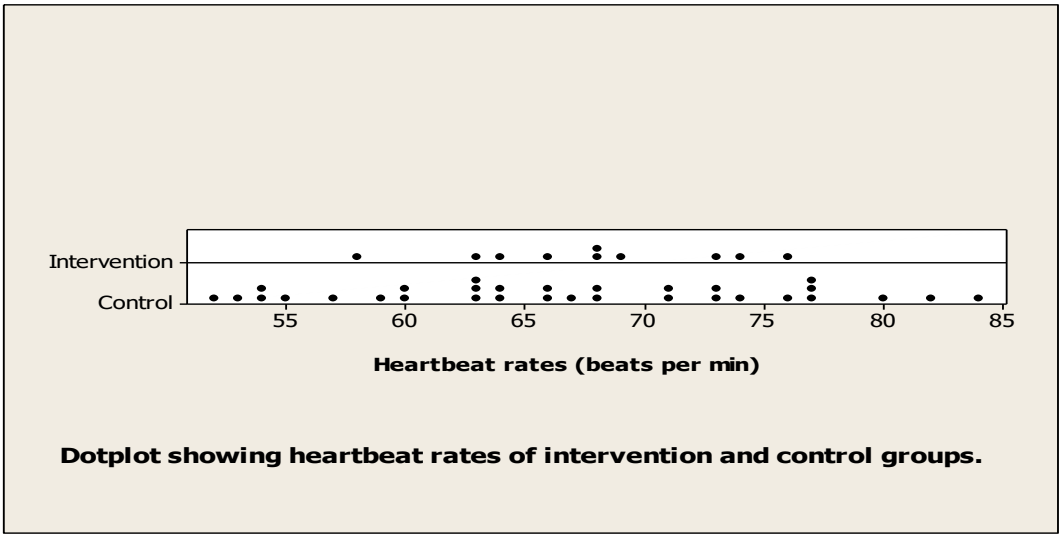
1.8.2 Dotplots

- 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

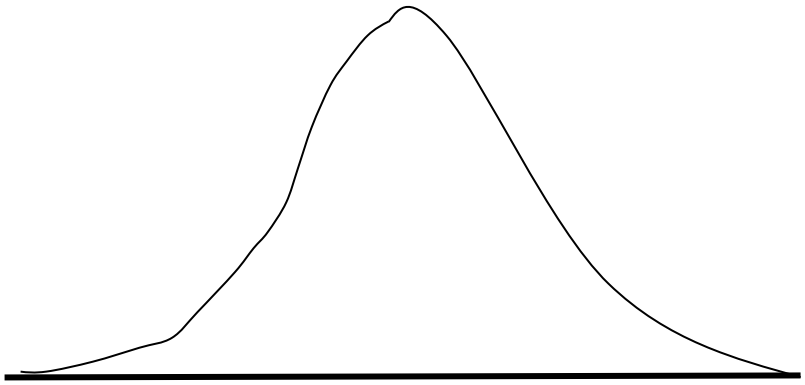
Sample distribution = the distribution of sample data

- 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

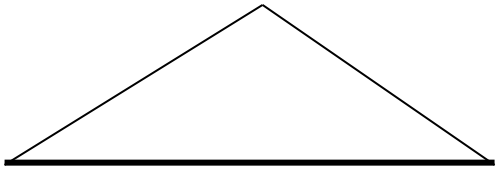
Symmetrical versus Skewed

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



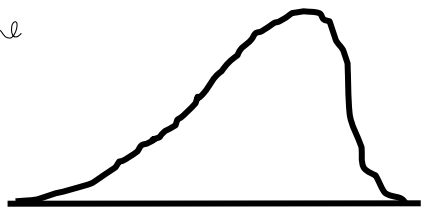
Triangular

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.



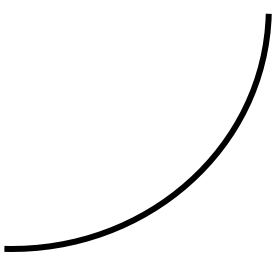
Left skewed

Right skewed

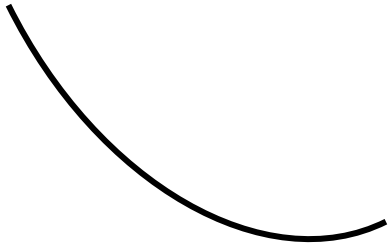
J-shaped = special type of negatively skewed distribution that has no right tail

- 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

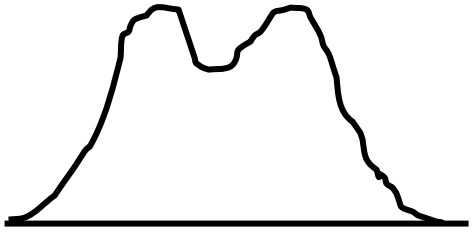


Reverse J-Shaped

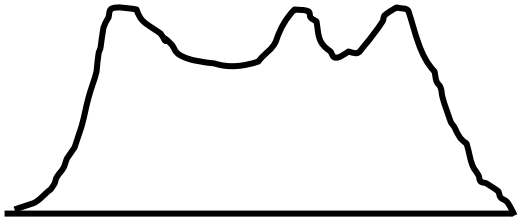
Modality

- 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

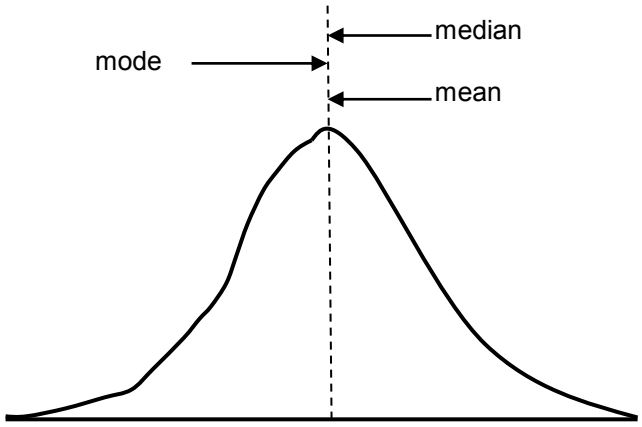
- The sample mean is considered as the best estimate of the population mean
- There can be only one population mean, but every sample from that population will have a different sample mean, though they may be close to each other
- Increasing sample size will increase the closeness of the sample means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

Median = the middle observation in a distribution

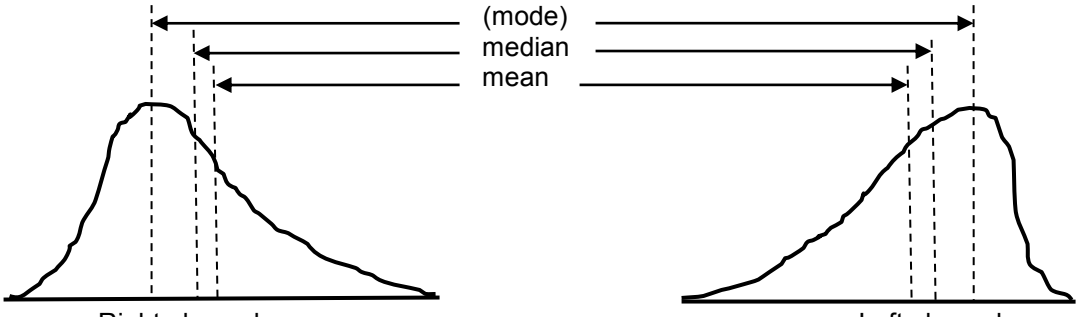
Median class = the class in a frequency distribution in which the median is found

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

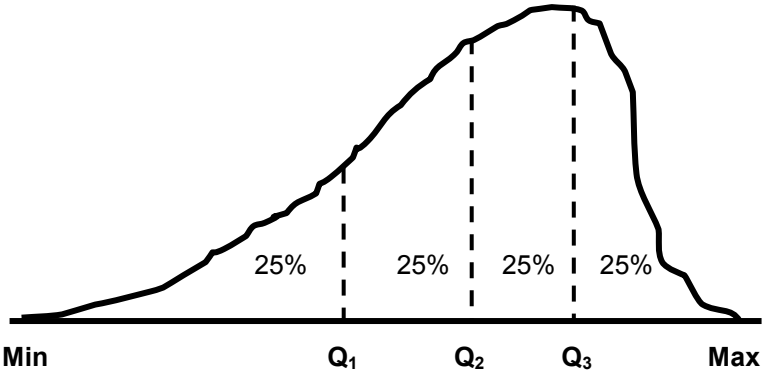
Population standard deviation (σ) (pronounced sigma)

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

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).

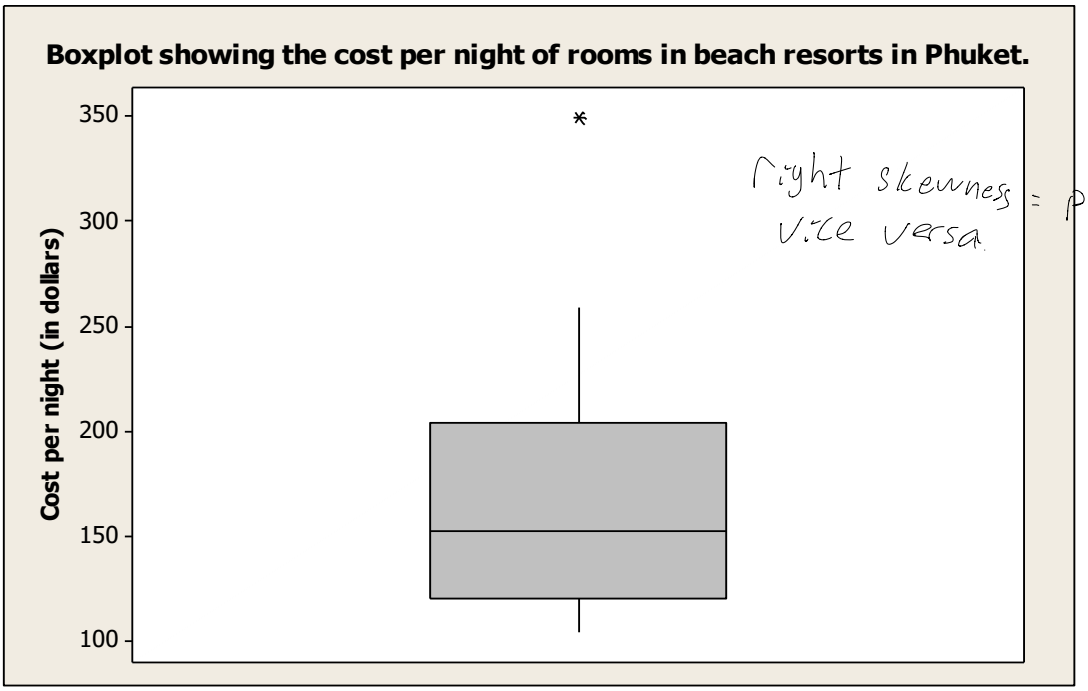
Lower limit = Q1 – 1.5 x IQR = 122.5 – 1.5(200 – 122.5) = 122.5 – 116.25 = 6.25 US dollars
Upper limit = Q3 + 1.5 x IQR = 200 + 1.5(77.5) = 200 + 116.25 = 316.25 US dollars

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.

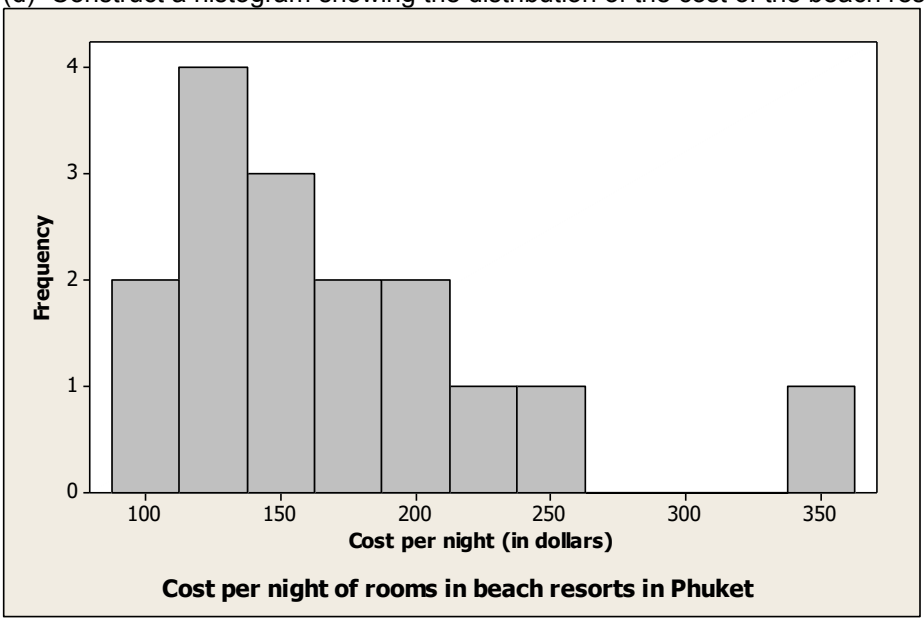
Adjacent values = 105 (on the lower end of the distribution) (also is the min)
= 259 (on the upper end of the distribution)

(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

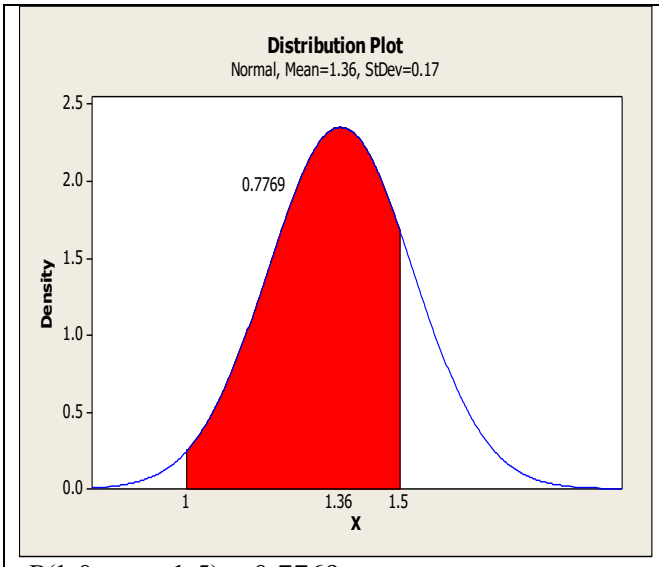
Area between is 0.7769

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.

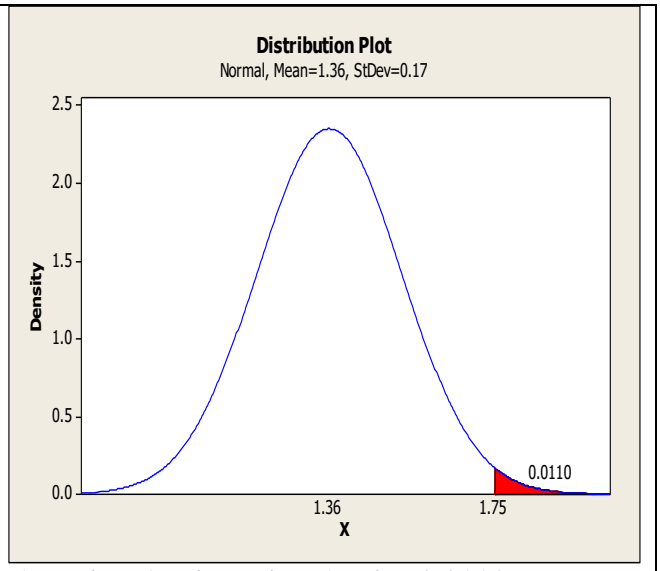
$$z = \frac{y-\mu}{\sigma} = \frac{1.75-1.36}{0.17} = 2.29$$

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%.



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

Graph for part (a)



$\text{Area } (y > 1.75) = P(y > 1.75) = 0.0110$

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)

$y = 1.36 + (1.28)(0.17) = 1.58 \text{ kg}$

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):

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

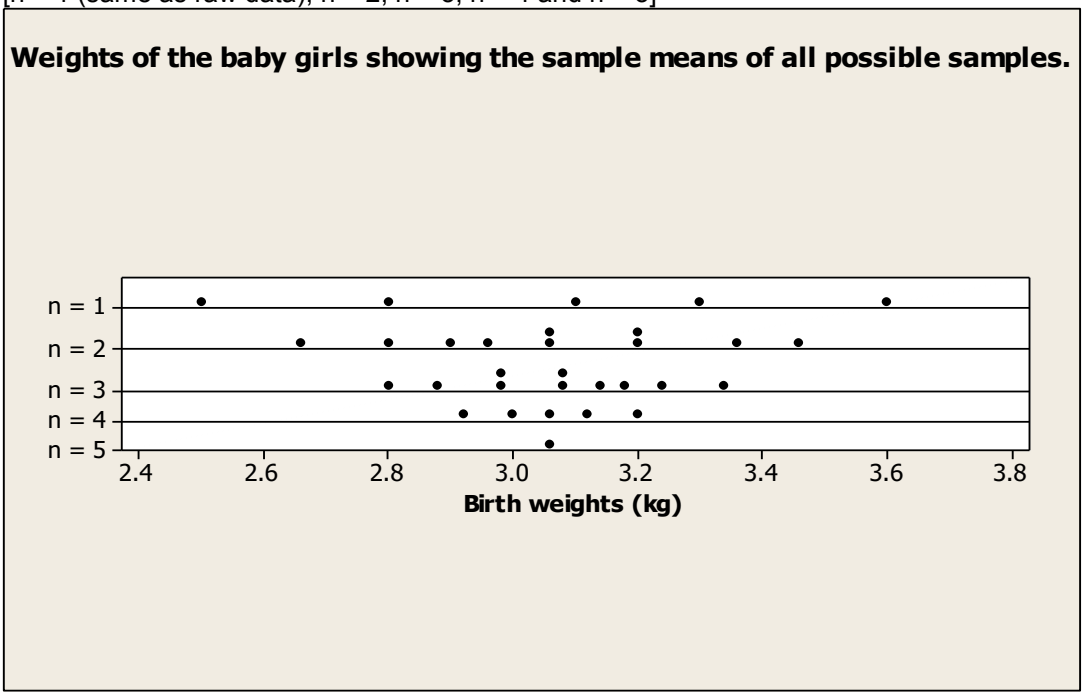
Table: All possible samples (5) and sample means for samples of size 4.

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

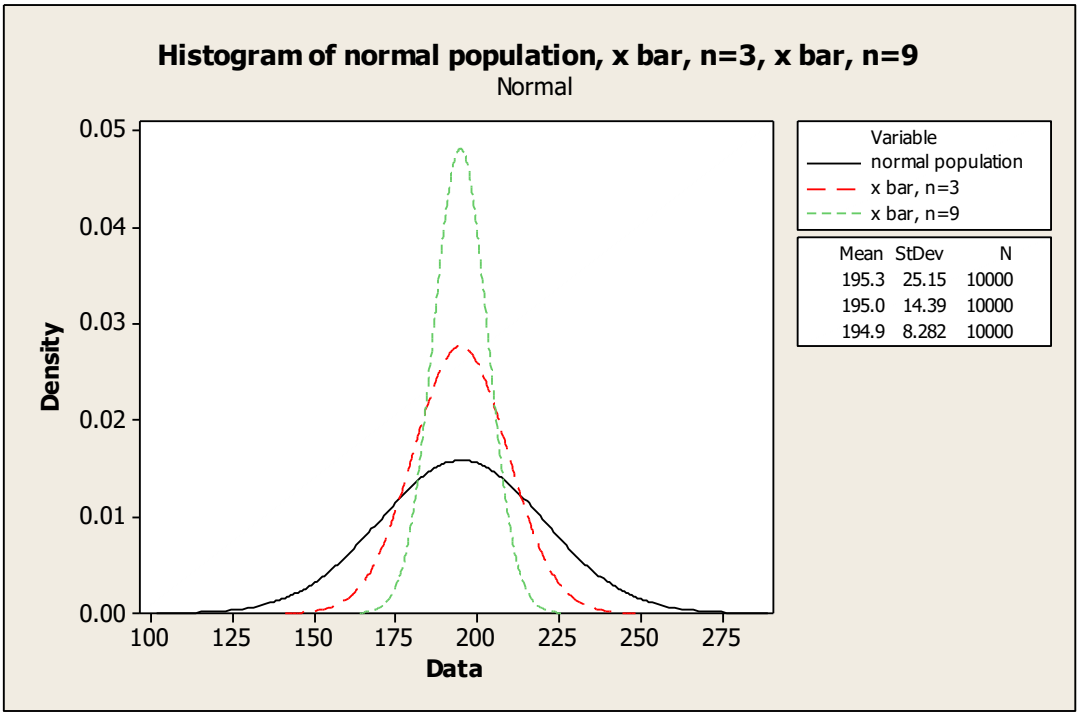
Sampling Distribution of the Sample Mean for a Normally Distributed Variable

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 ANY 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 large sample size, 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.

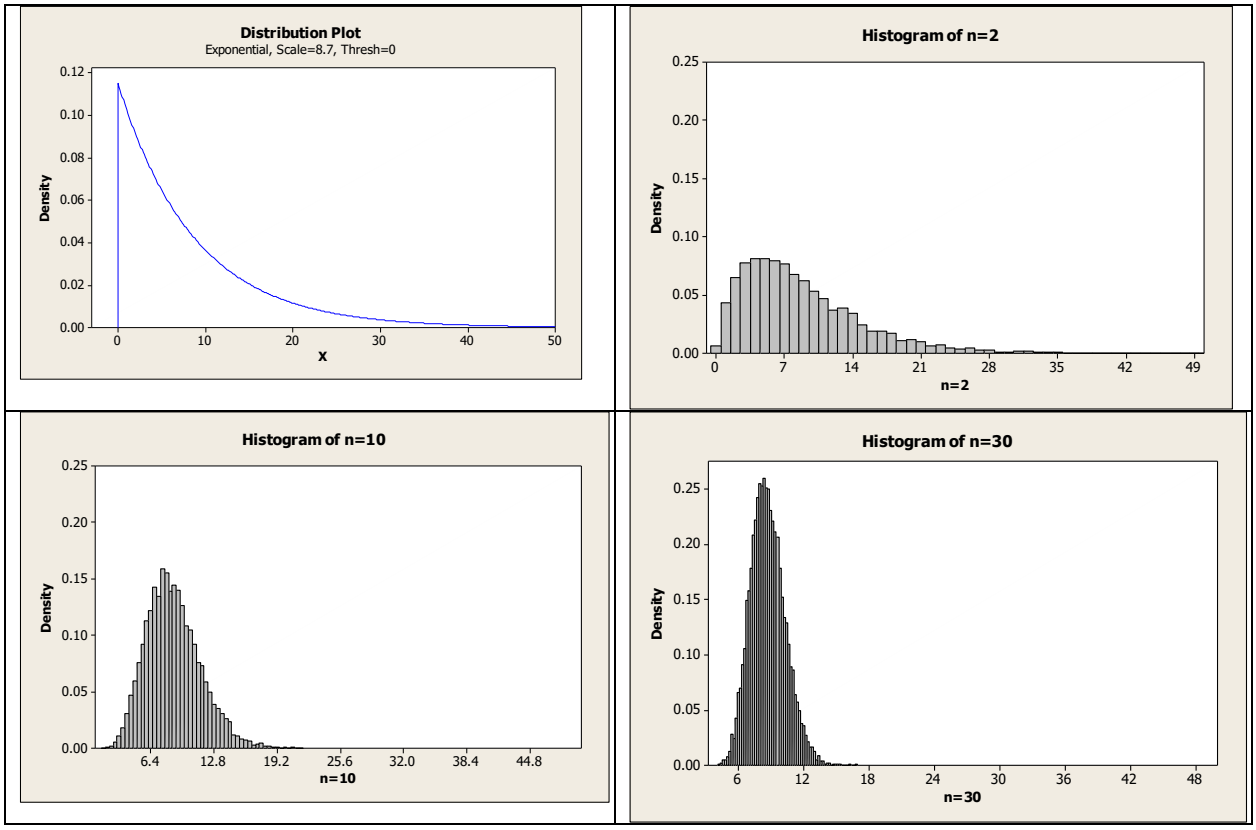
2. **Center:** The mean of the sampling distribution is: $\mu_{\bar{x}} = \mu$

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

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.

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

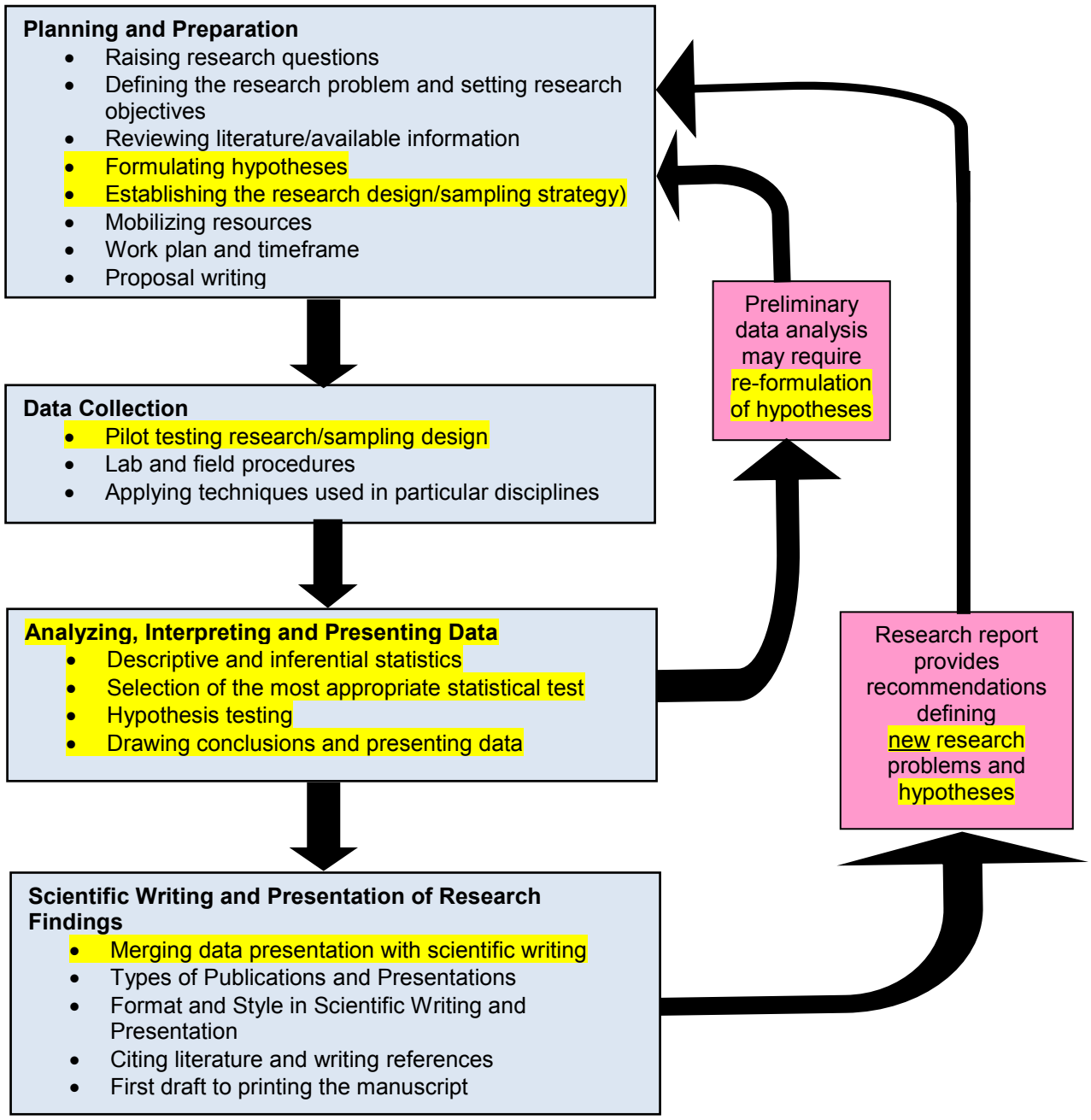
For z = 0.90, area to the left = 0.8159
So, area to the right = 1 – 0.8159 = 0.1841

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 <u>two</u> 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 <u>no</u> difference (among groups, etc.) or <u>no</u> 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 <u>is</u> a difference or there <u>is</u> 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, <u>based on knowledge of the field</u> 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 **<u>sample</u>** 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 **<u>population</u>** 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

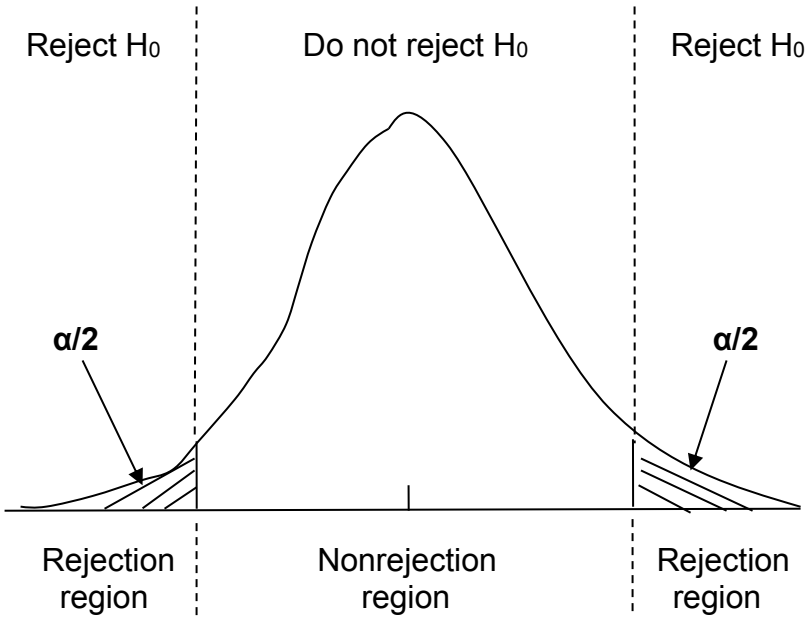
Test statistic = the statistic used as a basis for deciding whether the null hypothesis should be rejected, e.g., t statistic, z statistic or correlation coefficient, r.

- **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)

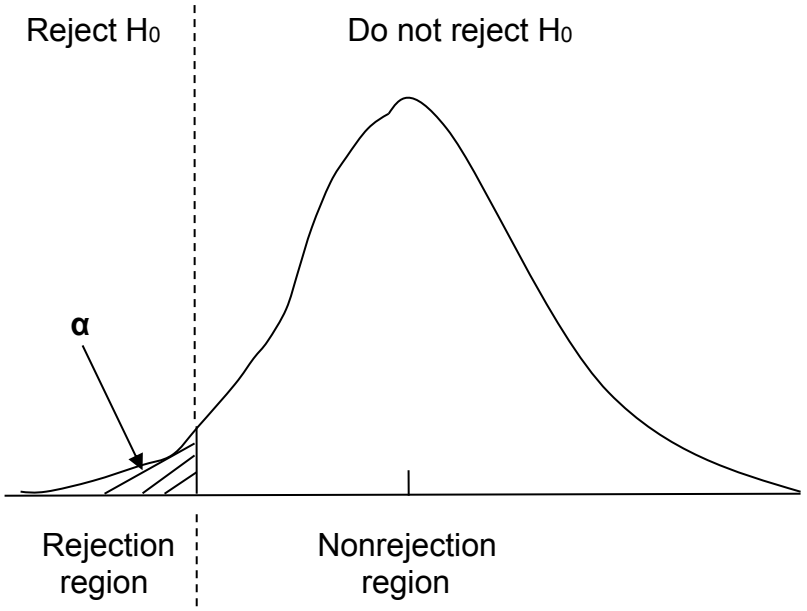
Rejection region = the set of values for the test statistic that leads to rejection of the null hypothesis

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

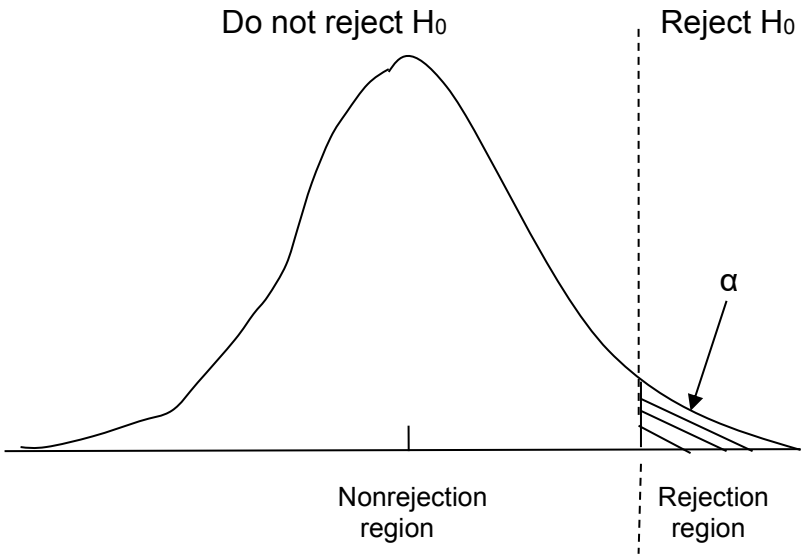
Alpha (α) = Type I error (defined below)



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

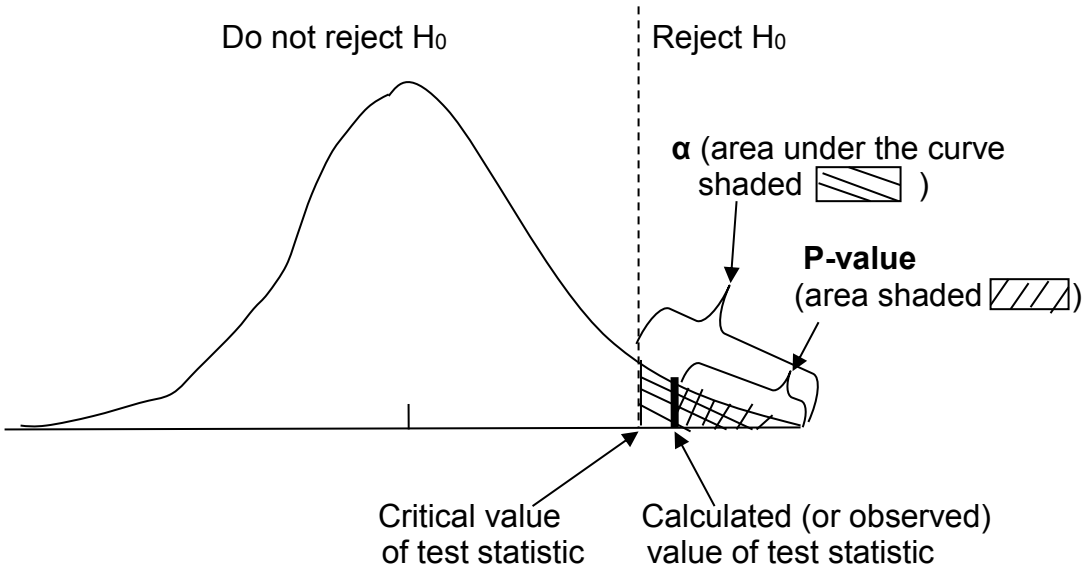
Difference between alpha (α) and P-value

- **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

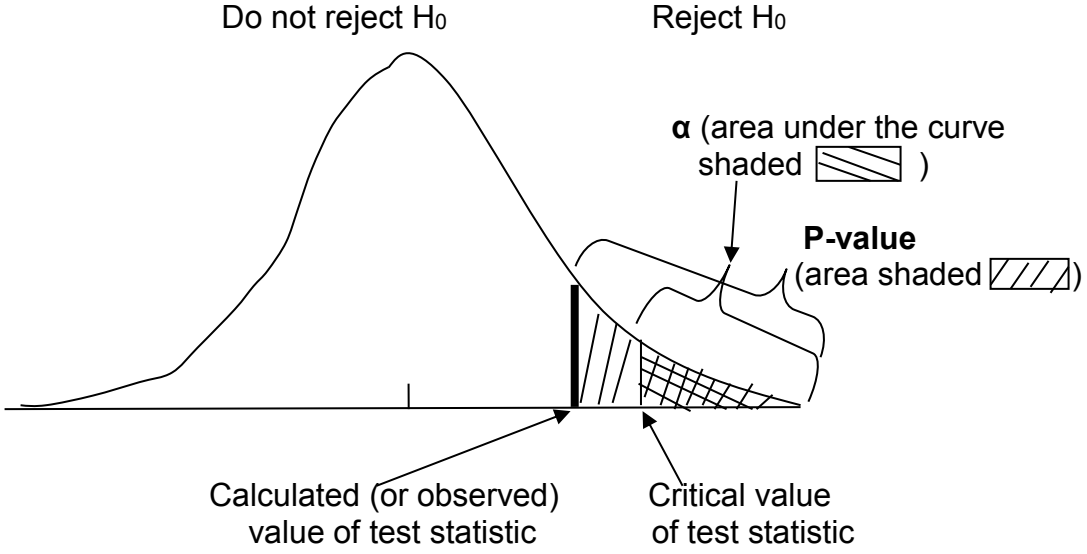
Case of Rejecting H_0



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

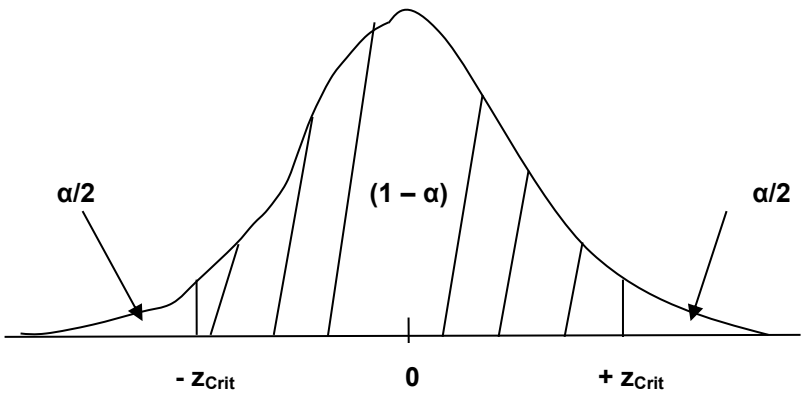
Confidence interval (CI) = an interval or range of numbers derived from a point estimate of a parameter

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]

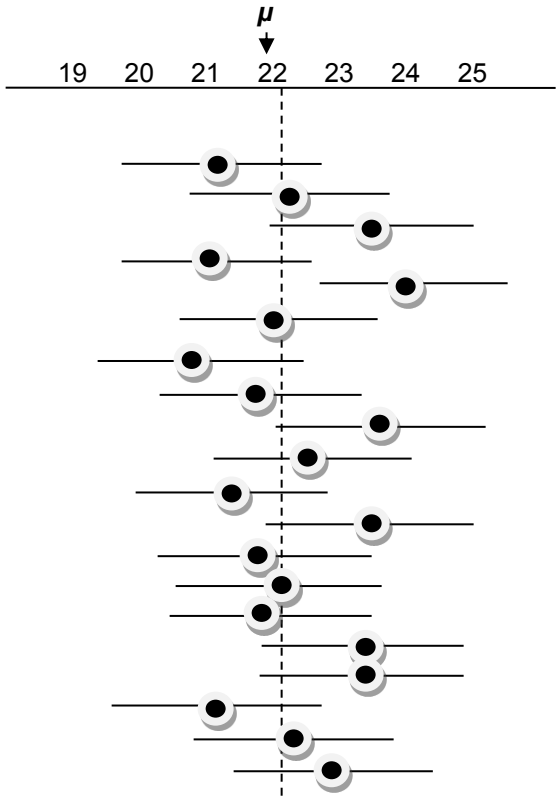
So for each sample, we calculate the confidence interval (CI) as follows:

$\bar{x} \pm 2\sigma_{\bar{x}}$ 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.

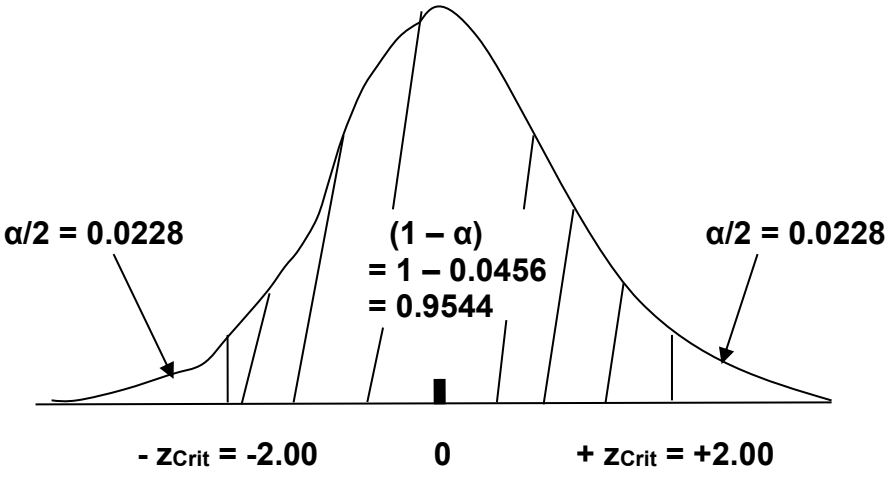
Sample	$\bar{x}$	**95.44% CI**	***μ*** in CI
1	21.04	19.58-22.50	Yes
2	22.15	20.69-23.61	Yes
3	23.34	21.88-24.80	Yes
4	20.92	19.46-22.38	Yes
5	23.86	22.40-25.32	No
6	21.86	20.40-23.32	Yes
7	20.75	19.29-22.21	Yes
8	21.67	20.21-23.13	Yes
9	23.40	21.94-24.86	Yes
10	22.38	20.92-23.84	Yes
11	21.31	19.85-22.77	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