DEPARTMENT OF BIOLOGY U.P. MANILA
1991

Course Module in BIO 180

GENERAL BIOSTATISTICS

Santiago, Anna Theresa A.
Sia Su, Glenn L.
Vicencio, Jeremy G.

Department of Biology
College of Arts and Sciences
University of the Philippines Manila
Padre Faura, Ermita, Manila

Table of Contents

Tables

1 Nature of Biostatistics

Biostatistics is coined from two words, *bios*, which means life, and *statistics*, which means the collection, organization, and management of data. These two words allow us to define biostatistics as the collection, organization and the management of biological data. Biostatistics is a widely taught course as it is not just limited in the study of biology. It is also used in public health and in medicine.

Biostatistics encompasses the two branches of statistics: descriptive and inferential statistics. Descriptive statistics applies to the summarization and presentation of data in a form that will make them easier to understand for the reader. It involves methods like tabulation, graphical presentation, and computation of averages to name a few. Inferential statistics on the other hand, applies to the making of generalizations and conclusions about a target population based on the results from a sample of a population. It involves the estimation of parameters and the testing of hypothesis to name a few.

Why is biostatistics needed as a science?

Biostatistics is essential because it enables each one of us to deal with the phenomenon of variation. What is the phenomenon of variation? The phenomenon of variation refers to the tendency of a measurable characteristic to change from one individual or one setting to another or from one instant of time to another instant within the same individual or setting. One example of the phenomenon of variation is when you want to take your own blood pressure reading. At one time, you may get your blood pressure reading and on a different time, the blood pressure reading that you obtained at that time may not be totally similar to your first blood pressure reading. The use of biostatistics to deal with the phenomenon of variation provides a systematic way of describing and analyzing the variability of the different phenomena that we may encounter.

How important is biostatistics?

Biostatistics provides us with numerous applications in the study of biology, public health, and medicine. Biostatistics is commonly used as a tool in the decision-making process as it allows us to properly identify the problem, assess the needs, allocate the limited resources that we have, and even evaluate programs that necessitate a way to have a systematic process of collecting data.

Biostatistics makes use of two kinds of data, namely qualitative data and quantitative data. Both kinds of data can either be a constant or a variable in the study of biostatistics. Let us distinguish between and constant and a variable. A constant is a phenomenon whose values remain the same either from person to person, or time to time, or place to place. Common examples of constants are the number of grams in a kilogram, the speed of light in a vacuum, the number of minutes in an hour, and the number of days in a week. These examples give us values that more or less remain the same. On the other hand, a variable is a phenomenon whose values or categories cannot be predicted with certainty. Common examples of variables include the color of a person's hair, the number of children in a family, attitudes towards certain issues, the weight of a person, and educational attainment. Even the certainty of whether a person is a smoker or a non-smoker is something that cannot be predicted easily.

Types of variables

There are two types of variables used in the study of biostatistics: quantitative variables and qualitative variables.

Quantitative variables are those variables that can be measured and ordered according to their quantity or amount or whose values can be expressed numerically. Examples of quantitative variables include birth weight, head circumference, and population size.

A quantitative variable can either be discrete or continuous. Discrete quantitative variables are those that have numerical values that are integers or whole numbers. They are characterized by gaps or interruptions in the values that it can assume. A few examples of which are hospital bed capacity and household size. We say that in a particular tertiary hospital, they have a hospital bed capacity of 30 beds. The 30 bed-capacity of the hospital can only serve 30 individuals. Hence, the quantity 30 is indicated as an integer or a whole number. We cannot say that the hospital has a bed capacity of 30.5 since half a bed does not exist. The next value that this variable can take on after 30 is 31. On the other hand, continuous variables have values that are potentially associated with real numbers. This means that they can attain any value including fractions or decimals. They do not possess the gaps or interruptions characteristic of a discrete random variable. Some examples include weight and height where we can indicate that a person weighs 14.5 kilos or that a person has a height of 176.25 cm.

Qualitative variables are those variables that are used as labels to distinguish one from the other. They have values that are intrinsically nonnumerical (categorical). Examples of qualitative variables include sex, urban/rural classification, and the regions of the Philippines. A person can either be a male or a female when categorized according to sex. In urban/rural classification, a person can inhabit an urban area or a rural area, while when classifying according to the region of the country, you have regions 1 up to the ARMM region.

Data Collection

In biostatistics, data collection is a major activity as the data that we collect in biostatistics is the same data that we organize and manage. There are two categories of data according to source: primary data and secondary data.

Primary data is obtained first hand whereas secondary data is obtained from existing data. One important feature why researchers want to collect primary data is because it can specifically answer the purpose of the investigator whereas in secondary data, the purpose of the original author may not necessarily be the same as those of the investigator.

There are a number of sources where we can obtain the data that we need. These sources can be from literature, expert's judgment, census, interview, direct measurement, actual observations, etc. The primary and secondary data obtained from these different sources have their advantages and disadvantages. The table on the next page shows some examples of methods that enable us to collect data together with their advantages and disadvantages.

Table 1-1. Advantages and Disadvantages of different methods of data collection

Method of Data Collection	Advantage	Disadvantage
Documented sources	Pre-collected, savings in time, money and energy	May not answer specific questions of investigator
Making observations	Answer objectives	Expensive
Interview	Subjects or variables not amenable to observations like opinions or feelings	Expensive, time consuming
Questionnaires	Less expensive and time consuming	Lower yield of respondents

The choice of method for data acquisition depend on the level of qualitative and quantitative data the researcher desires to collect, as well as time, money and manpower available for the data collection.

The data that we collect in the field also has certain qualities. The qualities of good data, or statistical data for that matter, are timeliness, completeness, accuracy, precision, relevance, and adequacy.

1. Timeliness – the interval between the date of occurrence of the different events considered and the time the data is ready to be used or disseminated by the researcher.
2. Completeness – the data is of coverage and can accomplish all the items necessarily needed by the researcher.
3. Accuracy – the data collected is close to the measurement or to its true value.

   Ways to check accuracy of data:

   a. Compare the collected data with expected trends
   b. Compare the data with totals of other record systems
   c. Compare the levels of collected data with those of other places with comparable conditions
4. Precision – the extent of the data collected refers to the consistency of information.
5. Relevance – the data collected meets the objectives of the data users. This is attained when it answers the objectives of the users and enhanced when there is communication between the data producers and the data users.
6. Adequacy – the collected data provide all the basic information needed to meet the requirements of the user.

LABORATORY EXERCISE 1
Nature of Biostatistics (25 points)

NAME:____ DATE:_____________

SECTION:___________________________ INSTRUCTOR:_______________________

1. In the following situations/objectives, identify the possible problems that you may encounter and indicate the data that you need to answer the objectives of your study. Indicate your answers on the provided table. (12 points)

Situation/Objectives	Possible Problems	Data Required	Type of Data needed	Method of Collecting Data
a. A student is interested in determining the prevalence of anemia using pallor of conjunctiva as the indicator.				
b. A doctor is interested in obtaining weekly reports on the number of cases of notifiable diseases in the neighborhood.				
c. A researcher wants to determine the mean birth weight of Filipino infants based on the data indicated from the infant's birth certificates.				

2. As a research assistant, you were assigned to collect and report the quarterly data on the ectoparasites in freely roaming animals in the community. Explain how your reports can have problems with respect to: timeliness, accuracy, precision, completeness, relevance and adequacy. (7 points)

3. What is the best method for collecting data on each of the following and why? (6 points)

Situation	Best method for collecting data	Reason for choosing the best method for collecting data
Prevalence of drug and substance abuse among the U.P. students.		
Effect of iron supplementation on the hemoglobin levels of pregnant women in Barangay X.		
Average heavy metal concentrations on the plant tissues in a garden Y.		

2 Statistics and the Biological Research Process

Biologists are in constant pursuit of new knowledge through observations, explorations and analyses conducted in the laboratory and in the field. These acts of observations, explorations and analyses constitute the process called research. The biological research process is described as systematic, objective and reproducible.

It is systematic because it is akin to a problem-solving technique that follows the scientific method of inquiry. The objectivity of biological research is based on its empiricism, whereby conclusions are founded on observed facts and logical reasoning. Statistics is not only limited to the summary, presentation and analysis of data but also in the design and execution of research to avoid haphazard data gathering and faulty data analysis.

Biological research carried out by a biology student, whether in the form of a special research problem or undergraduate thesis, must be reproducible such that it can be validated by peers. Good biological research must be disseminated in scientific gatherings so that it can serve as a stimulus for further study and contribute to the greater body of knowledge.

Basic Steps in Biological Research

1. Identification of the research problem
2. Formulation of research objective/s
3. Review of related literature (RRL)
4. Formulation of testable hypotheses
5. Construction of the research design
6. Drawing inferences or conclusions
7. Dissemination and/or utilization of results

The Research Problem

Biologists investigate research problems related to their specific expertise or field of knowledge. As biology students, a research topic may be pursued on the basis of factors such as field of interest, feasibility of investigating the research problem (e.g. limitations in time, availability of research funds), relevance (i.e. timeliness and impact), and ethical considerations (i.e. moral considerations in the use of human or animal subjects). The student researcher must avoid trivial research problems and focus on ethically sound scientific research. In student research proposals, the research problem is written succinctly, either in the form of a question or a statement.

The Research Objectives

The research objectives must reflect the questions that need to be addressed in the biological research process. It can be formulated as a single statement or in the form of general and specific objectives. When writing the research objectives, the marketing mnemonic known as "S.M.A.R.T." may be used as a guide:

- Specific: the statements are clear, unambiguous, and straight-to-the-point
- Measurable: success or achievement of the objective can be gauged or established
- Attainable: the objectives are realistic, practical and achievable based on limitations
- Relevant: the objectives are related to the research problem
- Time-bound: there is a definite end in relation to its measurability

When considering a research problem to formulate appropriate objectives, students can seek the opinions of scientists and academicians who have research experience in specific areas of biology for guidance. Coupled with personal observations and a review of similar studies that have been done on the topic of interest, students will have a clear direction toward a valid research problem that is relevant and worth pursuing.

Ideally, in student research proposals a section on "scope and limitations" is included so that assumptions, restrictions and limitation are explicitly stated with respect to the coverage of the study. Matters such as time allotted for the conduct of the study, cooperation needed from collaborating institutions or scientists, ethical considerations (e.g. method of handling live vertebrate specimens), and available funds and facilities must be specified.

Review of Related Literature (RRL)

A comprehensive review of related literature (RRL) involves the collation and integration of available information which are related to the research problem of interest. The RRL incorporates basic principles and existing studies on the topic and must not be written as a litany of previous work, but as a meaningful critique of current evidence. Attention must be given to the scientists who have contributed most to the body of knowledge, the research designs and appropriate statistical analyses executed in published research and the conclusions of these studies. The goal of the RRL is to establish a rationale or a conceptual framework for pursuing the research problem.

The Research Hypothesis

There is much confusion in the declaration of the research hypothesis. One school of thought considers the research hypothesis as the researcher's perceived answer to the research problem. In most biological research involving the use of statistical analysis, the research hypothesis is an assertion about the relationship between two or more variables which are to be observed in the research problem. The hypotheses are declared as a pair of statements known as the null and alternative hypotheses.

1. The Null Hypothesis
   - Represented symbolically as H0 or Ho
   - The hypothesis statement of equality or no difference
   - Example: On general weighted averages (GWA) of 2nd year biology students
     - Ho: The GWA of 2nd year males is equal to the GWA of 2nd year females
     - Ho: $\bar{\mathrm{x}}_{\mathrm{maleGWA}} = \bar{\mathrm{x}}_{\mathrm{femaleGWA}}$
     - Ho: (Male GWA) – (Female GWA) = 0

2. **The Alternative Hypothesis**
    - **Represented symbolically as H1 or Ha**
    - **The hypothesis statement of non-equality or difference**
    - **Possible alternative hypothesis counterparts for the example above:**
        - **Ha: The GWA of 2nd year males is <u>not</u> equal to the GWA of 2nd year females**
        - **Ha:** $\bar{x}_{\mathrm{maleGWA}} \neq \bar{x}_{\mathrm{femaleGWA}}$
        - **Ha: (Male GWA) – (Female GWA) $\neq 0$**
        - **Ha: GWA of males > GWA of females or**
        - **Ha: GWA of males < GWA of females**

The actual research hypothesis of the investigator may either be the null or alternative depending on the nature of the study. The statistical test to be used will be the basis for the rejection of one of the two hypotheses and the "non-rejection" of the other. By convention, a hypothesis may only be rejected or not rejected and never "accepted" since this poses the danger of making hasty generalizations and declaring blanket statements.

The Research Design

The research design is a careful scheme for data collection and analysis that may be considered as the "plan of attack" in order to answer the research objectives. Bulk of this plan is stipulated in the "methodology" section of student research proposals. In designing a study, the researcher must focus on the research objectives, the variables of interest, facilities and equipment required and the time frame involved. The first step is to clearly define the variables relevant to the research problem. During high school, science investigatory projects often require the identification of the research variables: the independent (control) variable, dependent (outcome) variable, and possible extraneous (confounding) variables. In statistics, the independent and dependent variables must be well-defined and the extraneous variables controlled or minimized.

- **Constant versus Variable**
    - **Constant – a phenomenon in which its value remains the same regardless of time, place or individual (e.g. speed of light, π, gravitational constant); a physical phenomenon**
    - **Variable – a phenomenon that can take different values depending on current circumstances (e.g. weight, height, exam scores); a biological phenomenon**

- **General Types of Variables**
    - **Qualitative – a categorical variable**
        - **Categories of the variable are considered as labels to distinguish one group from another**
        - **Categories cannot be used as basis for saying that one group has a higher value than another group**
        - **e.g. name, sex, nationality, religion**
        - **When using statistical software, a categorical variable can be represented numerically for simplification or coding purposes (e.g. 0-male; 1-female)**

  - Quantitative – a variable that can be measured numerically using a specific unit of measurement
    - Values of the variable specify quantity or amount and may be arranged according to magnitude
    - Quantitative variables can be further classified as discrete or continuous
      1. Discrete: only counts or whole numbers are meaningful (e.g. number of offspring)
      2. Continuous: fractions and decimals are meaningful values (e.g. tree height)

When dealing with variable data, it is imperative to classify the collected data according to the level of measurement. This allows the researcher to determine the appropriate statistical methods to summarize and analyze the data. The <u>levels of measurement</u> (or scales of measurement) include nominal, ordinal, interval and ratio.

- Nominal Level Data
  - Level of measurement where numbers or names represent a set of mutually exclusive classes to which observations of a variable may be assigned
  - Variables that yield nominal-level data are all <u>qualitative</u>
  - Examples of nominal data for qualitative variables:
    - Sex: male & female
    - Disease status: sick, not sick
    - Student number: 1998-12345, 2009-54321

- Ordinal Level Data
  - Similar to nominal in terms of being categorical in nature
  - Unique feature is that the mutually exclusive classes can be <u>*ordered or ranked*</u> (e.g. 1st, 2nd, 3rd, 4th, so on)
  - The exact distance between categories cannot be quantified
  - Both <u>quantitative and qualitative</u> variables may yield ordinal-level data
  - Examples of ordinal level data for qualitative variables:
    - Perception: strongly agree, agree, neutral, disagree, strongly disagree
    - Severity of Disease: mild, moderate, severe
    - Height: short, medium, tall
    - Weight class: lightweight, welterweight, heavyweight
      If the actual height or weight measurement based on a specific unit is used to classify the data, then the level of measurement is no longer ordinal.

- Interval Level Data

  - Numerical observations of a <u>quantitative</u> variable, in <u>which 0 (zero) is not absolute</u> (artificial)
  - Conceptually, these scales are infinite (-∞ to + ∞)
  - Examples of variables with interval data:
    - Temperature (°C) – 0 °C doesn't mean absence of temperature (or heat)
    - Voltage – 0 volts doesn't mean absence of charged particles

- **Ratio Level Data**
    - **Numerical observations of a quantitative variable, in which 0 (zero) is absolute (fixed zero)**
    - **Ratio of two numbers is meaningful**
    - **Examples of variables with ratio data:**
        - **Weight in Kg, Income in $, Total population**
        - ***In the ratio scale, we can say that 2kg is twice as heavy as 1kg or $60 is 3x more than $20…***

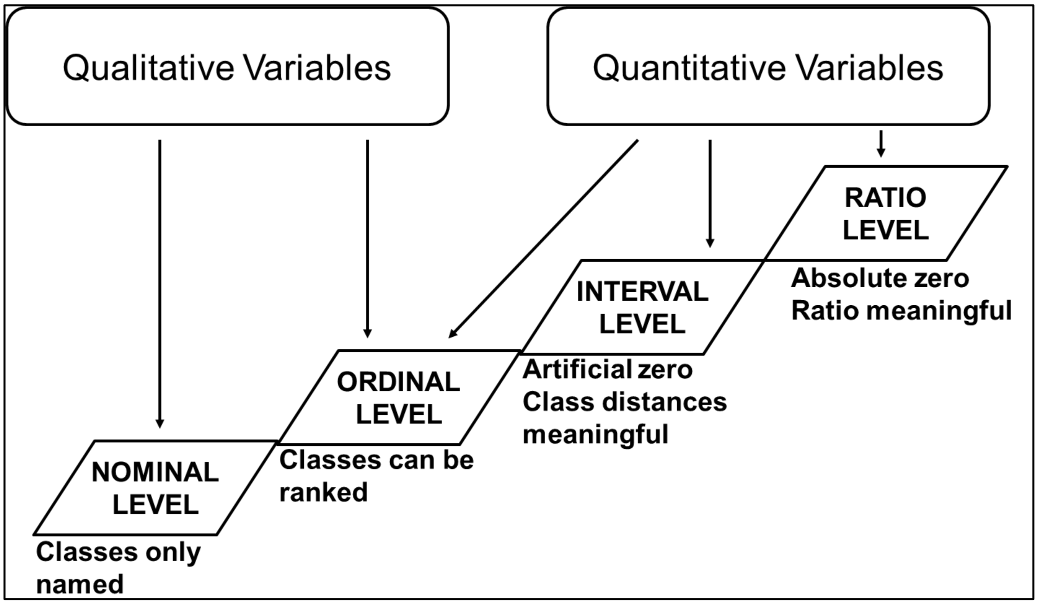


Figure 2-1. Variable Types in Relation to the Levels of Data Measurement

Once the research variables have been identified and defined, the research design can be planned. Although a course on research methodology is a more appropriate venue to delve into the nitty-gritty concepts of research design and exhaust all possible kinds of studies, it is worthy to discuss the general and common research schemes suitable for students learning the basics of biostatistics.

Research designs may be generally divided into observational and experimental.

1. **An observational study design involves the acquisition of variable data from existing phenomena, whereby the researcher serves as an observer.**
    - **A design in which the levels of all the explanatory variables are determined as part of the observational process**
    - **The investigator has not control over the explanatory variables**
    - **Example: Study on the effect of pregnant women's alcohol consumption on birth-weights of their babies**

Various observational study designs are common in medicine and public health in the form of surveys known as cross-sectional studies, cohort studies and case-control studies. Data such as patient history and demographics are collected and inferences are made with respect to a disease or condition of interest. In biological research, a common form of observational study design would be a field survey or prevalence study involving collection of specimens for identification and reporting. Although there is no manipulation of variables involved, observational study designs require a systematic method of data collection involving some form of randomization so that selection bias may be avoided. Bias can be avoided by making sure that all possible subgroups of a target population, such as humans, are represented in the data collection scheme. The chapter on sampling methods will introduce the biostatistics student to the various means of acquiring data based on probabilistic assumptions.

2. An experimental study design involves the manipulation of certain conditions (i.e. independent variable) and measures the dependent variable.
    - A stringent study design performed in a controlled environment wherein extraneous variables are minimized if not eradicated.

    - Basic components of an experiment:
        - Treatment: a combination of the levels of one or more independent variables; also known as factors
        - Experimental/Observational unit: smallest unit of the study material sharing a common treatment; e.g. an animal, a plot of land, a specimen sample, a human subject, etc.

The experimental unit may be subjected to various conditions (independent variable or treatment) in order to measure a reaction (dependent variable). The objective of an experiment is to separate the treatment effects from the uncontrolled variation among units. Sources of variation in observed responses of the experimental unit may include:

- Variation due to the effects of the independent variable/s
- Variation due to the effects of identified extraneous variable/s (e.g. genetics or health condition of subject)
- Variation due to unidentified sources (i.e. error variation)

An experimental research design must be formulated in a manner wherein variation in observed data is mainly due to the effects of the independent variable/s. In order to minimize variation in observed data due to extraneous variables or unidentified sources, data acquisition techniques may be incorporated in the experimental design:

- Randomization
    - Any experimental study requires random allocation of treatments to experimental units to avoid bias.
    - All experimental units have the same chance (equal probability) of being given any of the treatments

- **Replication**
  - **Commonly known as “replicates”**
  - **Number of experimental units in a single treatment at one time = number of replicates for that treatment (e.g. 5 mice per treatment, 3 plants per treatment, etc.)**
  - **The mean of the observations from each replicate can be subjected to statistical tests/ inferences**

- **Blocking**
  - **Groups of experimental units sharing a common level of an extraneous variable**
  - **Based on agricultural field experiments where blocks such as plots of land share the same soil conditions**
  - **e.g. In comparing the effects of soil characteristics to plant yield: Experimental units were 3 plots of land and the treatments/ independent variable is variety of a particular plant. Plant yield is the outcome of interest.**

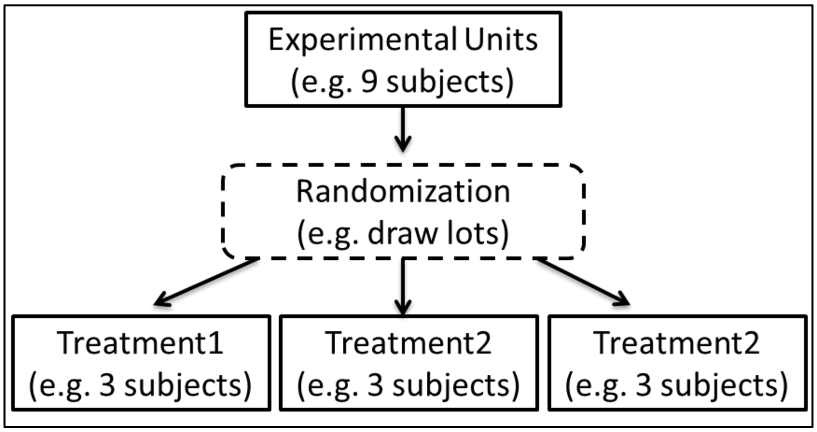


Figure 2-2. Sample Schematic of a Randomized Design with Replication

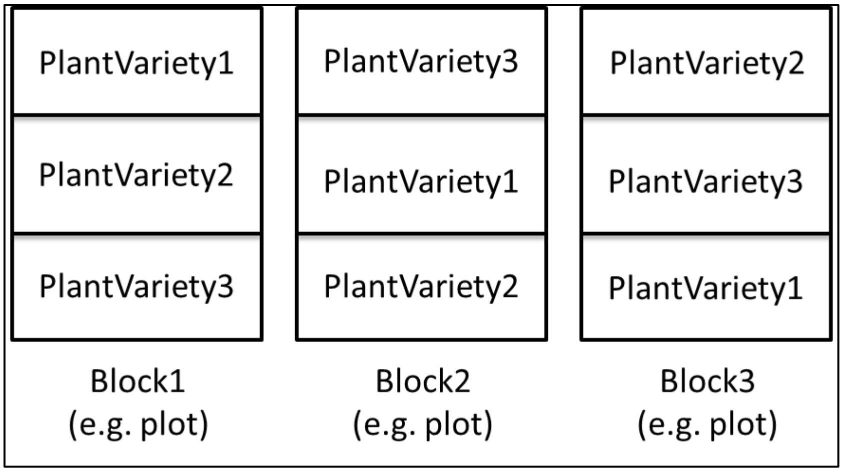


Figure 2-3. Sample Layout of a Randomized Blocks Design for Plant Studies

Drawing Inferences or Conclusions

After the complete and successful implementation of the research design, the data acquired must be summarized and analyzed appropriately. Both descriptive and inferential statistics play key roles in making sense out of the data collected and in answering the research objectives. Separate chapters on data presentation, data summary and various methods of statistical tests will be covered in this module. It is important to note that descriptive and inferential statistical methods are applicable based on the type of variables (qualitative or quantitative), level of measurement of the data acquired (nominal, ordinal, interval, ratio) and the number of variables being analyzed.

Dissemination and/or Utilization of Results

The ultimate goal of all research is to contribute to the greater body of knowledge. Dissemination of research output may be achieved through publication in reputable, peer-reviewed science journals, technical reports, oral or poster presentations in scientific meetings/conferences and even social media for real-time critique. For immediate utilization of results, technical reports may be submitted to policy-making bodies such as local government units for policy development or improvement.

LABORATORY EXERCISE 2

Statistics and Biological Research Process (30 points)

NAME:____ DATE:_______________

SECTION:______________________________ INSTRUCTOR:____________________________

1. Classify whether each listed item is a constant or variable by checking on the appropriate box. If it is a variable, specify whether qualitative, quantitative-discrete or quantitative-continuous. (10 points)

a. Blood glucose level in mg/100ml ☐constant ☐variable ________________________

b. Blood type ☐constant ☐variable ________________________

c. Avogadro's number ☐constant ☐variable ________________________

d. Dialect ☐constant ☐variable ________________________

e. Cases of pneumonia ☐constant ☐variable ________________________

f. Nutritional status ☐constant ☐variable ________________________

g. Grams in a Kilogram ☐constant ☐variable ________________________

h. Civil status ☐constant ☐variable ________________________

i. Moons of Saturn ☐constant ☐variable ________________________

j. Per capita income ☐constant ☐variable ________________________

2. Specify the level of measurement of the given variables. (10 points)

a. Patient Name ___________________________________

b. Age in years and months ___________________________________

c. Gender ___________________________________

d. Digital Systolic Blood Pressure ___________________________________

e. History of Surgery ___________________________________

f. Pain Scale ___________________________________

g. Degree Program ____

h. Students Enrolled ____

i. Distance Traveled ____

j. Calories per Serving ____

3. Answer the questions related to the given research topics.

   a. Methionine is a sulfur-amino acid that is added to the diets of turkeys to enhance growth. Three types of methionine, labeled as M1, M2 and M3, were compared. Weight gain of young turkeys over a three-week period was measured. The experimental units were 12 cages of young turkeys. The cages were stacked on top of each other, four layers high, in a room, with three cages on the floor, three cages on the second level, three cages in the third level up and three cages on the top level, near the ceiling. Because of concern over temperature as a possible confounder, the cages were grouped: Group1 - 3 floor cages, Group 2 - 3 second level cages, Group3 - 3 third level cages, Group4 - 3 ceiling cages. In each group, the cages were assigned to M1, M2 or M3 at random.

      i. What general type of research design is described above?

      ____

      ii. Independent Variable ____

      iii. Dependent Variable ____

      iv. Variable Type of the Dependent Variable and Level of Measurement:

      ____

      v. Data Acquisition Techniques Employed in the Design:

      ____

vi. **Draw a schematic diagram of the research design described above.**

b. **In a study that would like to compare the effects of two different drugs, Drug A and Drug B on the white blood cell count of human volunteers, the medical researcher formulated the following null hypothesis: There is no difference in the white blood cell counts of patients administered with Drug A and Drug B.**

i. **Formulate an appropriate alternative hypothesis** ______________________________

____

____

ii. **What general type of research design is described above?**

____

iii. **Dependent Variable** ____

iv. **Variable Type of the Dependent Variable and Level of Measurement:**

____

3 Data Presentation

After the collection of data, the researcher must make sense of raw data in a manner that can contribute to the achievement of the research objectives. Data must be presented so that the research results can be understood by peers, mentors or the general public. Data presentation methods are part of descriptive statistics and may serve as preludes to inferential statistical methods.

Basic methods of data presentation:

1. The Narrative
2. Tabular Presentation
3. Graphical Methods

The Narrative

Also known as the textual method of presentation, the narrative method is a written account of data that is more appropriate for small data sets or a few numerical figures. The narrative can be confusing for larger data sets and should only be used to introduce or supplement tables or graphs. For example:

"Out of the 7 subjects included in the investigation to determine the dependence of nutritional needs on body size, 4 subjects had fat-free mass sizes that were below 68Kg. All subjects had recorded 24-hour energy expenditures that were beyond 1,500Kcal."

Ideally, the narrative shown in the example above should be followed by other data presentation methods including tables and/or graphs so that other information gathered that may be relevant to the research problem will not be omitted or taken for granted.

Tabular Presentation

Tabular presentation involves the systematic arrangement of statistical data in columns and rows for a specific purpose. Tables are suitable for presenting large sets of numerical data in a compact and orderly fashion. A large set of nominal data may also be tabulated for presentation. A well-constructed table can clearly emphasize trends or relationships among different variables, which may not be obvious in a narrative format. Tables must be *simple*, *direct*, and *clear* to the point that is self-explanatory.

There are various table layout formats that researchers can follow such as those prescribed by the American Psychological Association (APA), Modern Language Association (MLA) and Chicago/ Turabian. Use the layout format that your mentor, school, institution or organization recommends and be consistent. Do not mix different table layout formats in a single research paper.

Regardless of the layout format, the table must possess the following parts:

- **Table number and title**
  - **Tables must be numbered chronologically within a research paper. A table must appear after it is mentioned in the body of the paper.**
  - **The title must mention the variables shown in the paper. The general location where the data was collected and the year of data collection may be included.**

- **Labels for the columns and rows including the unit (e.g. %, g/dL, m/s, years, $, etc.)**
- **Table cells containing data**
  - **String data (non-numerical) must be aligned consistently (all left or all center)**
  - **Numerical data must be aligned at the right so that the place values are observed; values may be centered as place values are aligned**
  - **Numerical data must be consistent in the number of decimal places and must use commas to delineate thousands, millions, billions and so on**
- **Footnotes or sources of data, if necessary.**

Follow the font (face and size), emphasis (bold or capitalization) and capitalization rules prescribed by the specific table layout format recommended for your paper. The tables below show the essential elements of standard tables for numerical and textual data.

Table 3-1. Energy Expenditure Based on Student Sports Club, Manila, 1988

Student Sports Club	Average Daily Expenditure (Kcal)
Chess Varsity	1,885
Basketball Varsity	2,423
Swimming Varsity	2,667

Table 3-2. Top 3 Ballpen Preferences of B.S. Biology Students in Manila City, Mandaluyong and Quezon City, 1985

Rank	Manila City	Mandaluyong	Quezon City
1	Scribbler Footlong	Kilometrico	Kilometrico
2	Kilometrico	Scribbler Footlong	Panda
3	Panda	Funny Friends	Funny Friends

A common table that is used for presenting count data, or a tally, is the *frequency distribution table*. The frequency distribution table usually shows the number of observations (frequency) for each category and the percentage of occurrence (relative frequency in %), such as the sample table below:

Table 3-3. Color of Roses at the College of Arts and Sciences Garden, University of Blooms, 1999

Color	Frequency	Relative Frequency (%)
Red	636	86.89
White	81	11.06
Pink	15	2.05
Total	**732**	**100.00**

Frequency distribution tables may also include cumulative frequencies and relative cumulative frequencies. By definition:

- Cumulative frequency – sum of the frequency of the class under consideration and frequencies of preceding classes
- Cumulative relative frequency – same as cumulative frequency applied to the relative frequency

Modifying Table 3-3 above wherein columns for cumulative frequency and relative cumulative frequency are included, the following frequency distribution table can be constructed:

Table 3-4. Frequency Distribution of Rose Colors at the College of Arts and Sciences Garden, University of Blooms, 1999

Color	Frequency	Relative Frequency (%)	Cumulative Frequency	Relative Cumulative Frequency (%)
Red	636	86.89	636	86.89
White	81	11.06	717	97.95
Pink	15	2.05	732	100.00

In the previous examples for the frequency distribution table, the row items are straightforward categorical. There are instances wherein a large numerical data set, whether interval or ratio level data, must be reduced or condensed for better comprehension of trends. The *grouped frequency distribution table* will require the construction of *class intervals* or groups, such as the sample grouped frequency distribution below.

Table 3-5. Weights of Female College Students, Bataan Province, 1989.

Weight in kilos	Frequency	Relative Frequency (%)	Cumulative frequency	Relative Cumulative Frequency (%)
28-34	9	1.2	9	1.2
35-41	137	18.9	146	20.1
42-48	306	42.3	452	62.4
49-55	196	27.1	648	89.5
56-62	75	10.4	723	99.9

In constructing a grouped frequency distribution table, examine the raw data from the following hypothetical study: "A study on serum creatine phosphokinase (CK) enzyme in U/l and its effect on skeletal muscle activity among 36 male sprinters resulted to the serum CK values enumerated below."

58	82	151	121	100	68
163	145	201	95	64	163
94	57	60	84	139	94
203	104	113	119	110	203
110	83	93	62	67	110
42	123	48	25	70	42

First, arrange the data in ascending order from the lowest observed value to the highest.

25	60	82	95	113	151
42	62	83	100	119	163
42	64	84	104	121	163
48	67	93	110	123	201
57	68	94	110	139	203
58	70	94	110	145	203

Calculate the *range* of the data set, that is, subtract the lowest observed value from the highest = 203-25 =178. Determine the number of class intervals or groups for the data set which may be from 5 groups to 15 groups. The researcher's discretion as to the number of classes to be used must be based on how meaningful the grouped frequency distribution will be. The table must not hide (too few classes) nor be irrelevant (too many classes). For the example above, 10 groups will be used for the serum CK in U/l.

The next step is to calculate the *class width*, or magnitude of each group, by dividing the *range* by the number of classes desired = 178/10 = 17.8 or 18. The class width must have the same number of decimal places as the observed data. After determining the class width, tally the frequency of observations under each group then construct the grouped frequency distribution table.

Table 3-6. Frequency Distribution of Serum CK (U/l) for 36 Male Sprinters, Manila, 2000

Serum CK (U/l)	Frequency (no. of men)	Relative Frequency (%)	Cumulative frequency	Relative Cumulative Frequency (%)
25-43	3	8.3	3	8.3
44-62	5	13.9	8	22.2
63-81	4	11.1	12	33.3
82-100	8	22.2	20	55.5
101-119	6	16.7	26	72.2
120-138	2	5.6	28	77.8
139-157	3	8.3	31	86.1
158-176	2	5.6	33	91.7
177-195	0	0.0	33	91.7
196-214	3	8.3	36	100.0

It is apparent from the table above that the most frequently occurring serum CK range is 82-100 U/l and that around half the total serum CK levels observed (55.5%) lie from 100 U/l and lower. Grouped frequency distribution tables give meaning to numerical data sets and allow the researcher to recognize trends.

Other tables relevant to research are *dummy tables* and *master*. A dummy table contains a proposed table number, title, column and row headings but cells are empty. Dummy tables are essential in student research proposals, because these guide the student researcher as to how the data will be collected and analyzed. A master table is a single table that contains raw data on each variable used/ examined with respect to all elementary units. The master table is where the researcher encodes collected data in preparation for better tabulation, graphical presentation or analysis.

LABORATORY EXERCISE 3
Tabular Presentation (25 points)

NAME:____ DATE:______________

SECTION:_________________________________ INSTRUCTOR:___________________________

For the listed research topics, construct appropriate and complete tables for the given data using word processing software (Open Office Document or Microsoft Word) at the Biostatistics computer laboratory. Submit electronic and/or printed copies (as specified by laboratory teacher) at the end of the laboratory period.

1. **In the study by Allen and Gorski (1992) on the sexual orientation and the size of the anterior commissure in the human brain published in the *Proceedings of the National Academy of Science,* 89:7199-7202, the midsagittal area of the anterior commissure of the brain of homosexual men, heterosexual men and heterosexual women were measured in square millimeters. The average midsagittal areas of the anterior commissures for homosexual men, heterosexual men and heterosexual women were 14.20, 10.61 and 12.03, respectively (5 pts).**

2. **Tripepi and Mitchell (1984) studied the metabolic response of river birch and European birch roots to hypoxia which was published in *Plant Physiology*, 76:31-35. The researchers flooded 4 birch tree seedlings and kept 4 other seedlings as controls. All seedlings were harvested and the amounts of ATP (nmoles per milligram tissue) in the roots were analyzed. The flooded seedlings had ATP amounts of 1.45, 1.19, 1.05 and 1.07. Recorded ATP for controls were 1.70, 2.04, 1.49 and 1.91 (5 pts).**

3. **The hypocholesterolemic effects of oat-bran or bean intake for hypercholesterolemic men was investigated by Anderson et al (1984) and published in *The American Journal of Clinical Nutrition,* 40:1146-1155. The study measured the decrease in serum cholesterol level (milligrams per deciliter) of subjects given either an oat diet or bean diet. The mean fall in cholesterol level for oat diet subjects was at 53.6 with a standard deviation of 31.1, while for bean diet subjects the mean decrease was at 55.5 with a standard deviation of 29.4 (5 pts).**

4. **Paleontologists at the Museum of Natural History measured the width (in millimeters) of the terminal molar at the right side of the upper jaw of 36 specimens of the extinct mammal "auroch" (*Bos primigenius*). The measurements are listed below (10 points):**

5.9	**6.1**	**6.1**	**5.7**	**5.9**	**6.0**
6.5	**6.1**	**6.7**	**5.8**	**5.9**	**6.2**
6.3	**6.3**	**5.7**	**6.2**	**5.7**	**6.0**
6.2	**6.2**	**5.7**	**5.8**	**6.1**	**6.2**
6.1	**6.2**	**6.0**	**6.1**	**5.9**	**5.4**
6.5	**5.9**	**5.7**	**6.1**	**6.0**	**6.1**

a) **What is your calculated range?**

b) **If the number of classes prescribed is 5, what would be the class width?**

c) **Construct a grouped frequency distribution table with relative, cumulative and relative cumulative frequencies.**

Graphical Presentation

When dealing with more voluminous raw data, various graphs may be used to show trends or patterns which could be missed in tabulated data. Certain situations may require the use of both a table and graph to show different perspectives of the same data set. Usually, tables and graphs should not be redundant so it is best to use either tables or graphs. Just like tabular presentation of data, graphs must also be simple and self-explanatory.

A graph must be labeled as "Figure," numbered chronologically within a research paper, contain a title and should be position after it is mentioned in the body of the text. Various statistical software (e.g. SAS, Stata, IBM SPSS, R Statistics), technical computing software (e.g. MatLab) and spreadsheet software (e.g. Microsoft Excel) allow the construction of graphs after encoding of raw data through a spreadsheet interface or plain text(e.g. .csv). The common graphs used by student researchers are discussed in this section of the module.

- Histogram
  - Graphical form of a grouped frequency distribution table
  - The horizontal axis (abscissa) contains the classes or groups and the vertical axis (ordinate) corresponds to the frequency or relative frequency
  - Horizontal axis variables: quantitative discrete or continuous, often used in graphing the distribution across age groups; the frequencies are represented by the areas of the bars
  - Vertical axis must always start at zero (0).
  - The example below is the histogram complement of Table 3-6.

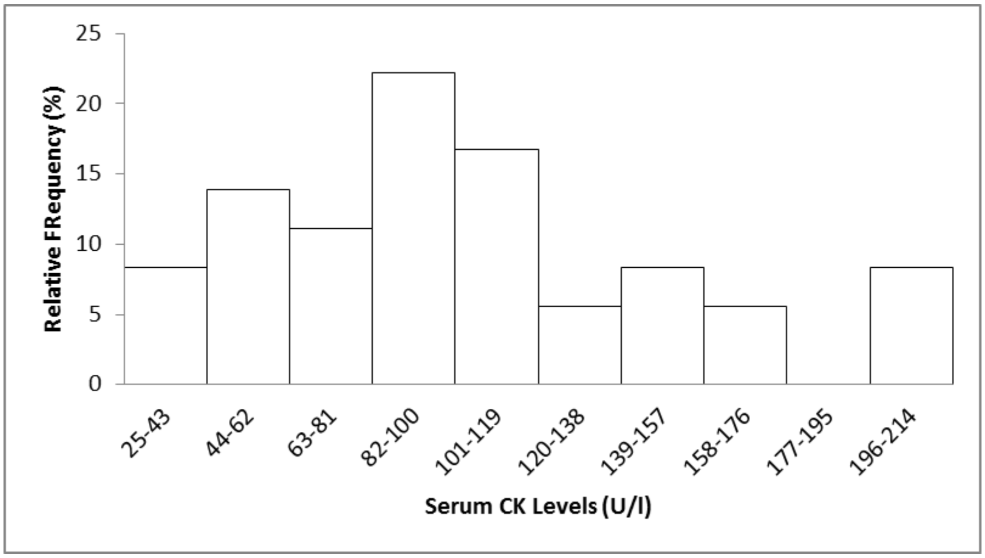


Figure 3-1. Histogram of CKU (U/l) levels of 36 male sprinters, Manila, 2000.

- Frequency Polygon
  - Also known as an area chart, this is an alternate form of the histogram wherein the class midpoints (middlemost value of each class) is plotted against the corresponding frequency or relative frequency in the vertical axis. The points may be connected to each other through a line and immediately show a trend.

- **Frequency Polygons may be used to juxtapose and compare frequency distributions of two groups.**
- **Horizontal axis variables: quantitative discrete or continuous**

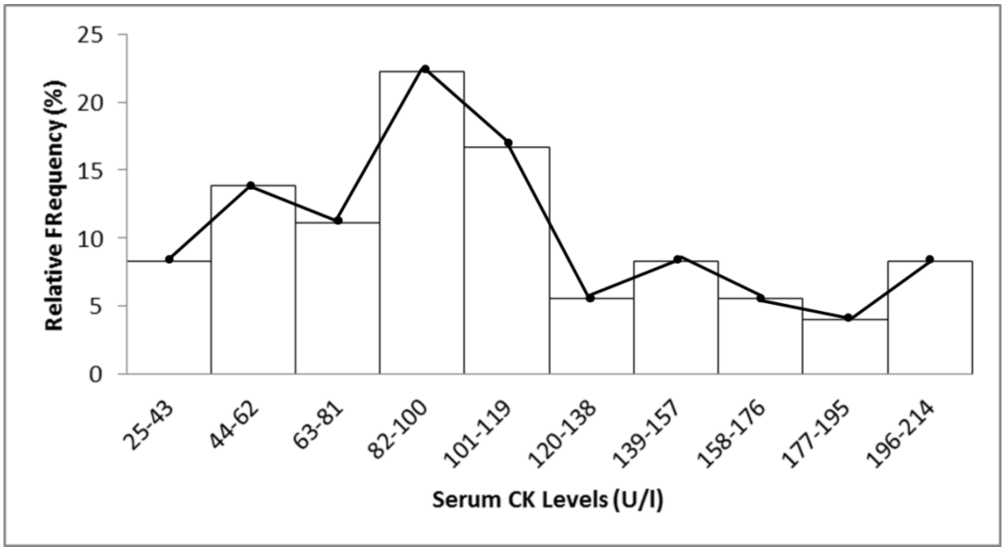


Figure 3-2. Histogram and Polygon of CKU (U/l) levels of 36 male sprinters, Manila, 2000.

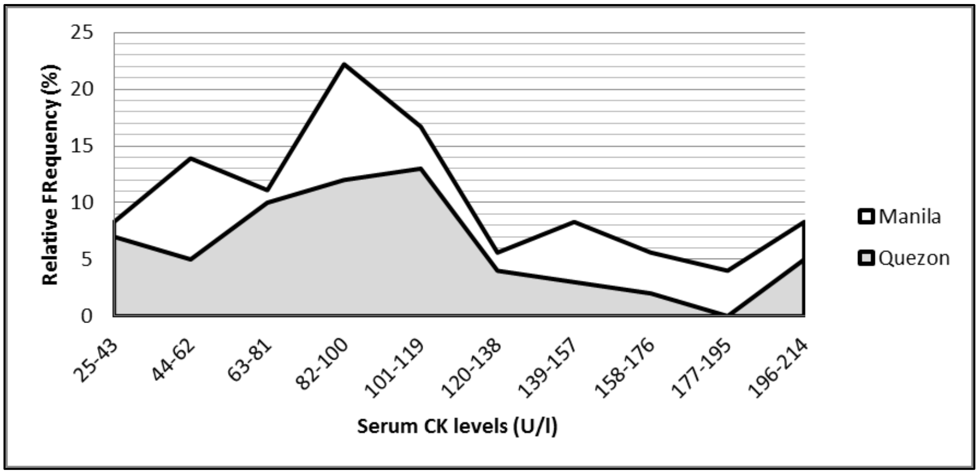


Figure 3-3. Frequency Polygon of CKU (U/l) levels of Males in Manila and Quezon Province, 2000.

- **Bar Graph**
  - **The bar graph is commonly used to compare the frequencies or relative frequencies among different qualitative variables (nominal or ordinal). It may be oriented vertically or horizontally.**
  - **Vertical bar graph: the categories are situated along the horizontal axis; compared to the histogram, the bars are separated from each other. Subcategories of a group can be represented by adjacent bars.**
  - **Horizontal bar graph: the categories are situated along the vertical axis**

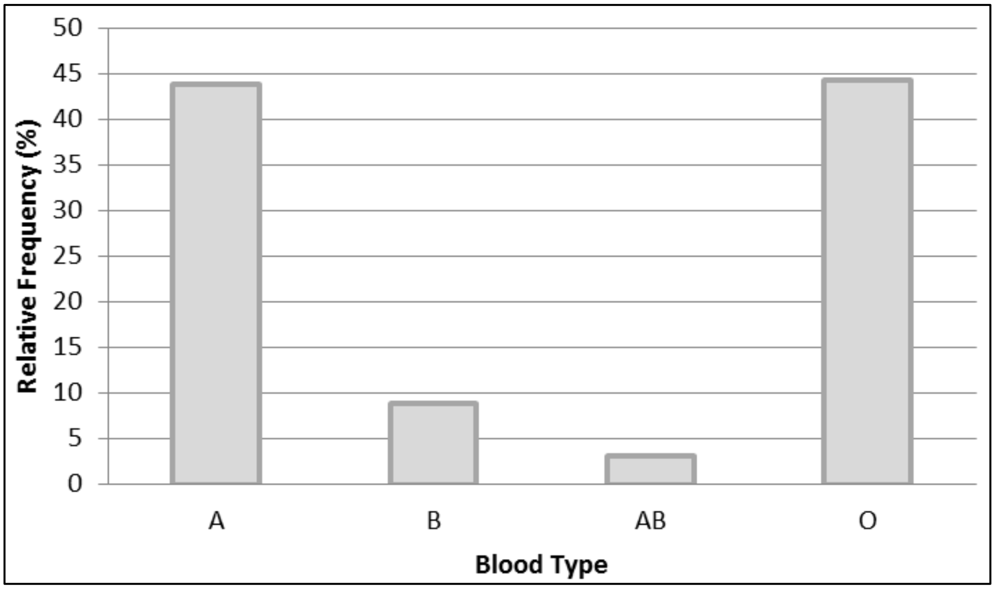


Figure 3-4. Blood Type Frequencies among Biology 22 Students, U.P. Manila, 2002.

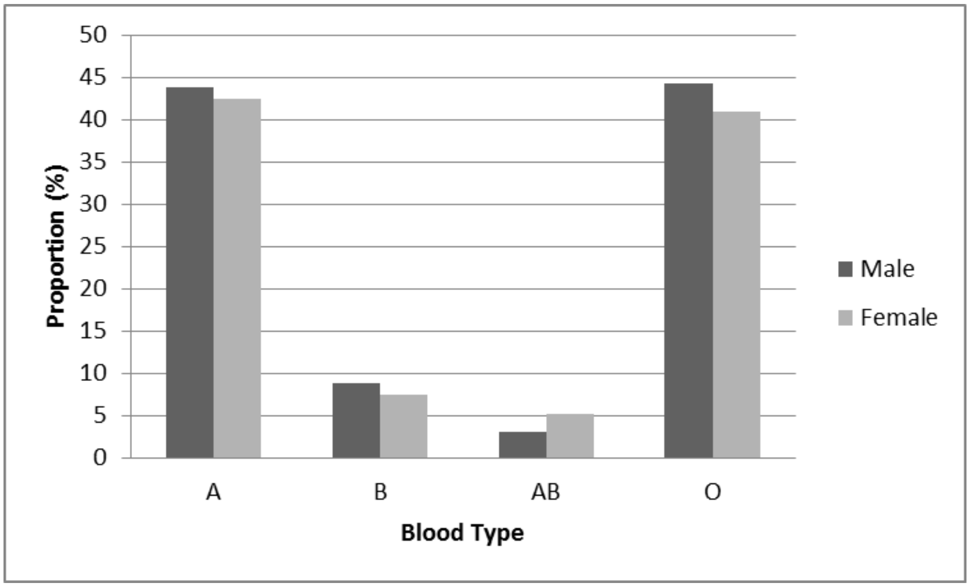


Figure 3-5. Blood type according to gender of BIO22 students, U.P. Manila, 2002.

- **Pie Chart**
  - **Similar to the bar graph, pie charts are useful in graphing relative frequencies of qualitative variables (nominal or ordinal) with few categories. There is no clear-cut rule regarding the number of categories appropriate for a pie chart or a bar graph. What would be more relevant is that the graph is clear, self-explanatory and concise.**
  - **The categories are represented by the slices and the magnitude (in percent) is represented by the size of the slice.**

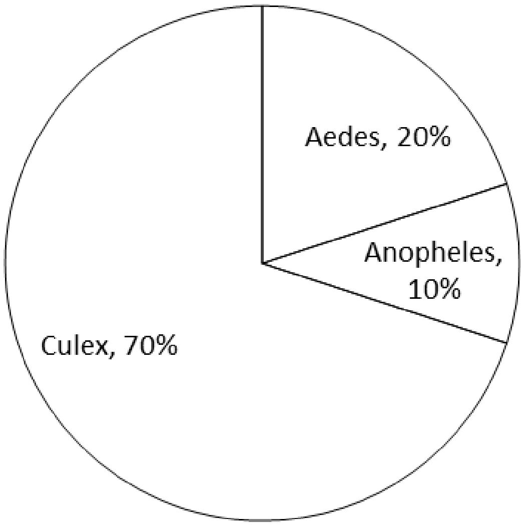


Figure 3-6. Identified mosquito genera in Barangay X, Batangas, 2007.

- Component Bar Graph
  - Instead of visualizing several pie charts of different groups, a component bar graph (a.k.a stacked bar graph) can condense the information into a single graph.
  - The component bar graph is also appropriate for qualitative variables (nominal or ordinal) with few categories.
  - The graph may be oriented vertically or horizontally, where each bar corresponds to a group and the components of the bar represent the percent distribution of the categories of interest.

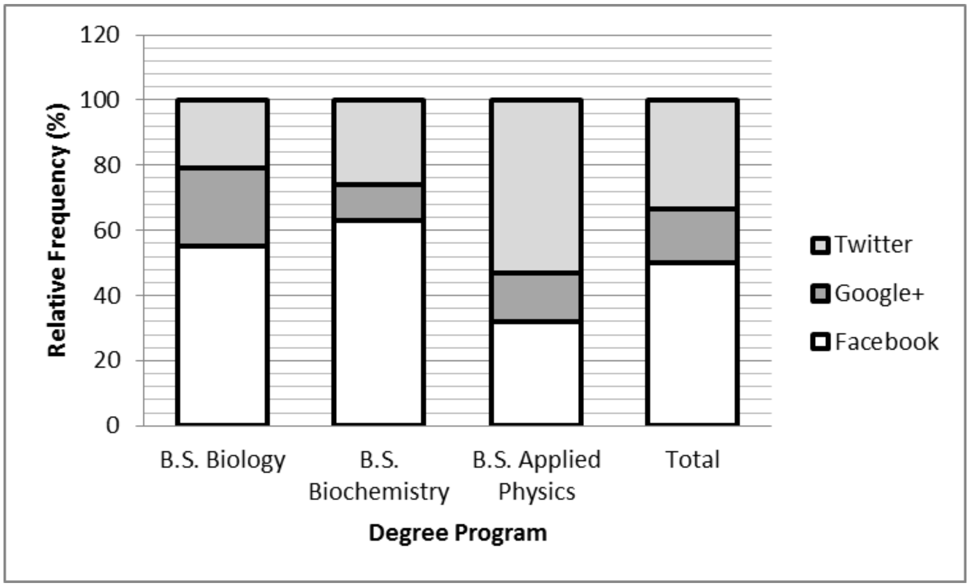


Figure 3-7. Social Network Preferences of Science Majors, Manila, 2011.

- **Line Graph**
  - **The line graph is suitable for graphing time series of frequencies of qualitative variables or quantitative variables**
  - **The horizontal axis is for time and the vertical axis is for the frequency of the qualitative variable or value of the quantitative variable. The aim is to show the trend of the frequencies or values across time.**
  - **Trends for two or more groups can be juxtaposed in a single line graph for comparison**

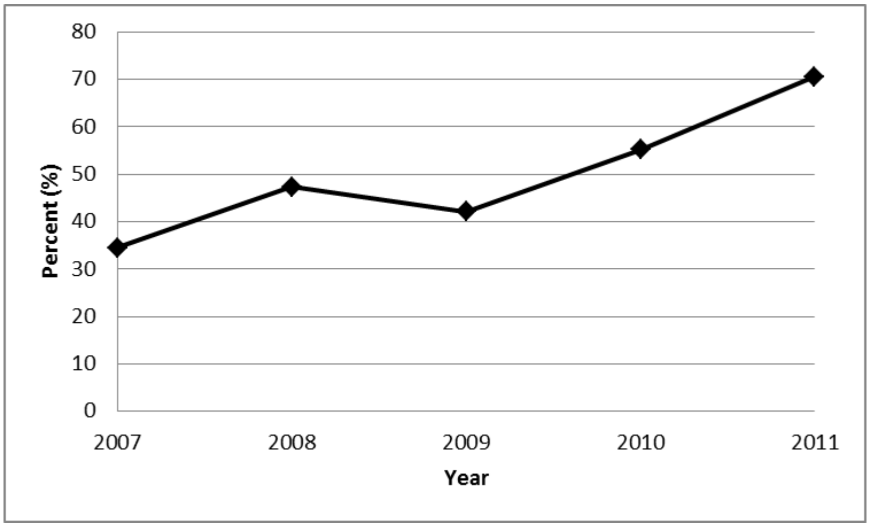


Figure 3-8. Proportion of faculty members with published journal articles, Department of Biology, University of the Philippines Manila.

- **Scatterplot**
  - **A scatterplot is a graphical representation of the relationship between two quantitative variables (interval or ratio data) and usually supplement correlation or linear regression analysis.**
  - **Typically, the independent quantitative variable is plotted in the x-axis (abscissa) and the dependent quantitative variable is plotted along the y-axis (ordinate).**

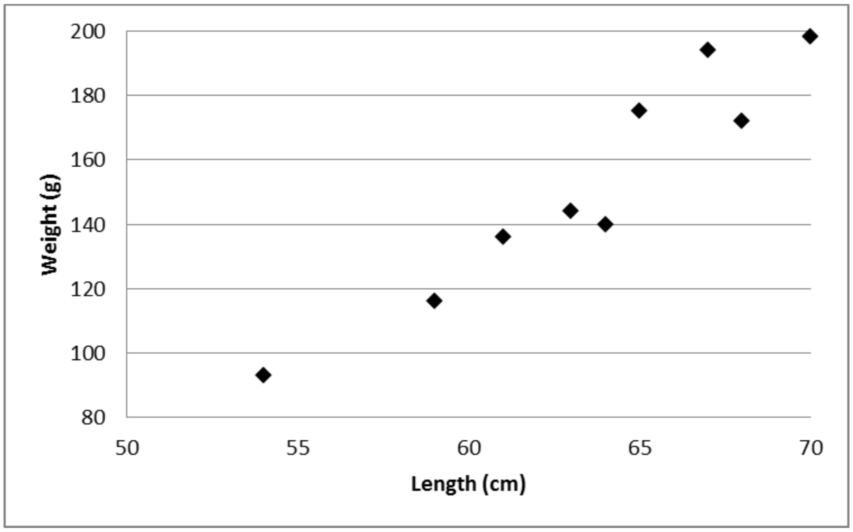


Figure 3-9. Scatterplot of snake length and weight, Manila Zoo, 1999.

- **Stem and Leaf Diagram**
  - **A simple graphical presentation of the distribution of a small set of observed quantitative variables (discrete or continuous variables that have been rounded off). This is can be considered as an oversimplified histogram.**
  - **All values of the observed data are preserved (more or less).**
  - **Values of data are arranged in ascending order. Each value us split into a "stem" (e.g. tens) and a "leaf" (e.g. ones). In the diagram, the stem is adjacent to all its leaves.**
  - **Example: radish growth in millimeters after 5 days of total darkness**

15	**11**	**20**	**29**	**8**	**30**	**33**
20	**35**	**10**	**22**	**15**	**37**	**25**

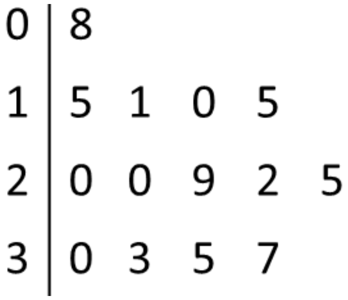


Figure 3-10. Stem-and-leaf diagram for radish Growth (mm) after 5 days in darkness

- **Boxplot**
  - **To graphically compare the distribution and measures of central tendency, dispersion and location of quantitative variables (interval or ratio data; continuous or discrete) across different groups, boxplots or box-and-whisker plots are useful.**
  - **A separate chapter in this module will discuss how to determine and calculate various measures of central tendency, dispersion and location. Typically, statistical software can generate boxplots from raw data. A single data set boxplot may be oriented horizontally.**

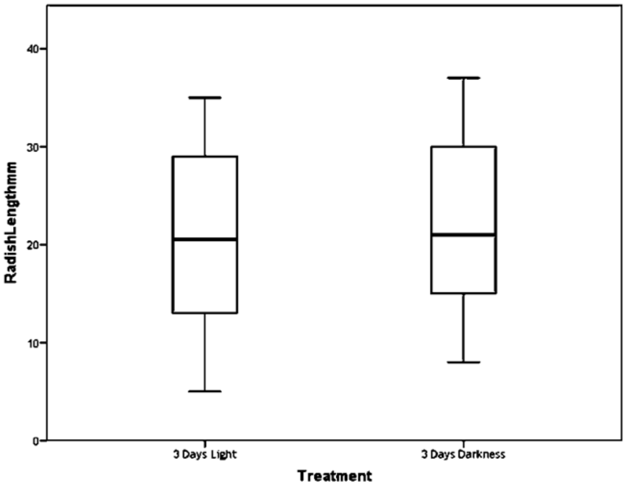


Figure 3-11. Radish length (mm) based on light conditions.

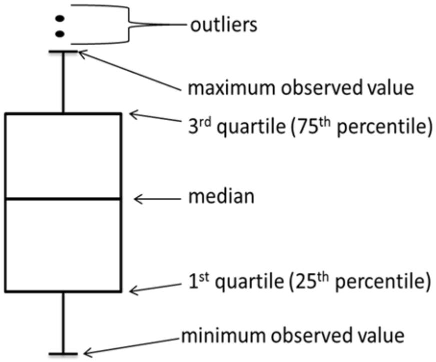

outliers
maximum observed value
3rd quartile (75th percentile)
median
1st quartile (25th percentile)
minimum observed value

Figure 3-12. Parts of a boxplot

LABORATORY EXERCISE 4
Graphical Presentation (35 points)

NAME:____ DATE:_______________

SECTION:______________________________________ INSTRUCTOR:_______________________________

Generate the required graph for the given raw data using spreadsheet or statistical software (Microsoft Excel or trial version of SPSS, STATA or open source software such as R or PSPP) at the Biostatistics computer laboratory. Combine all complete and numbered graphs in a document (e.g. M.S. Word) file, and submit electronic and/or printed copies (as specified by laboratory teacher) at the end of the laboratory period.

1. Stem-and-Leaf Diagram and Single Boxplot. A study by Connolly (1968) on fruit fly preening behavior published in *Animal Behaviour*, 16:385-391, the total preening time spent (in seconds) by each fruit fly specimen (n=20) during a 6-minute observation period were recorded as follows: (10 points)

48	22	48	29	19
18	26	57	32	25
76	33	31	46	24
34	24	10	16	52

2. Bar Graph. Based on the CIA World Fact Book, the estimated life expectancies at birth (2011) of ASEAN countries are as follows: (5 points)

Country	Life Expectancy (years)	Country	Life Expectancy (years)
Brunei	75.74	Malaysia	73.29
Burma	63.39	Philippines	71.09
Cambodia	62.1	Singapore	82.14
Indonesia	70.76	Thailand	73.10
Laos	56.68	Vietnam	71.58

3. Pie Chart. The Manila Zoological and Botanical Garden has the following groups of animal specimens: (5 points)

Group	Number of Species
Mammals	30
Reptiles	63
Birds	13
TOTAL	106

4. **Component Bar Graph. A marketing student conducted a survey on the preferential rice varieties of consumers in the 4th district of Metro Manila (CAMANAVA) and discovered that the top three preferred rice varieties were Sinandomeng, Dinorado and Thai Jasmine. The survey results are as follows: (5 points)**

City	Dinorado (%)	Sinandomeng (%)	Thai Jasmine (%)
Caloocan	32	40	28
Malabon	28	56	16
Navotas	56	13	31
Valenzuela	41	32	27

5. **Line Graph. The average monthly precipitation data from DOST-PAGASA are summarized below: (5 points)**

Month	Ave. Rainfall (cm)
Jan	1.5
Feb	0.8
Mar	1.5
Apr	2.5
May	12.1
Jun	29.2
Jul	35.3
Aug	47.5
Sep	40.5
Oct	18.5
Nov	12.3
Dec	6.2

6. **Scatterplot. A resident pediatrician measured the weights and heights of 13 newborn babies delivered for the month of August 2012 at the Philippine General Hospital for routine census. Data are tabulated below. (5 points)**

Baby	1	2	3	4	5	6	7	8	9	10	11	12	13
Weight (oz)	81	125	83	106	88	118	86	101	86	102	98	95	88
Height (in)	17.0	19.7	17.3	19.0	17.2	20.0	17.5	18.7	17.0	18.8	19.4	18.0	18.1

4 Descriptive Measures

In this chapter, we will learn several techniques for organizing and summarizing data so that we may easily determine what information they contain. The ultimate in data summarization is the calculation of a single value that in some ways conveys important information about the data from which it was calculated. Such single values that are used to describe data are called *descriptive measures*. After studying this chapter, you will be able to calculate several descriptive measures for both populations and samples of data.

First, we will learn how to calculate measures of central tendency, location, dispersion when data are ungrouped.

MEASURES OF CENTRAL TENDENCY

- Convey information regarding the *average* value of a set of values.
- They indicate where the majority of values in the distribution are located.
- These measures can be considered as the center of the probability distribution from which the data were sampled.

Arithmetic Mean

- Sum of the individual values in a data set divided by the number of values in the data set.

Table 4-1. Formulae for arithmetic mean

General formula for a finite population mean	General formula for the sample mean
$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$	$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$
where: ▪ μ = population mean ▪ N = population size ▪ x_i = individual observation ▪ i = position of observation ▪ Σ = summation sign	where: ▪ $\bar{x}$ = sample mean ▪ n = no. of observations ▪ x_i = individual observation ▪ i = position of observation ▪ Σ = summation sign

Example:

In an outbreak of dengue fever, 10 people became ill with clinical symptoms 3 to 14 days after exposure to the virus. In this example, we will illustrate how to calculate the sample mean period for the dengue outbreak. The incubation periods of the affected people (x_i) were 7, 4, 3, 12, 8, 9, 6, 5, 5, and 14 days.

1. **To calculate the numerator, sum the individual observations**

$$\sum_{i=1}^{10} x_i = 7 + 4 + 3 + 12 + 8 + 9 + 6 + 5 + 5 + 14 = 73$$

2. **For the denominator, count the number of observations: *n* = 10**
3. **To calculate the mean, divide the numerator (sum of observations) by the denominator (number of observations):**

$$\bar{x} = 74/10 = \mathbf{7.3}\ \boldsymbol{days}*$$

***A reasonable rule is to express the mean with one more significant digit than the observations**

Therefore, the mean incubation period for this outbreak was 7.3 days.

Geometric Mean

- **The mean or average of a set of data measured on a logarithmic scale.**
- **Consider the value of 100 and a base of 10 and recall that the logarithm is the power to which a base is raised. For example, the logarithm of 100 at base 10 is 2 since 10^2 is equal to 100.**

$$\log_{10} 100 = 2$$

- **An antilog raises the base to the power (logarithm). For example, the antilog of 4 at base 2 is 16.**

$$\text{antilog}_2\, 4 = 16$$

- **Is calculated as the n^{th} root of the product of *n* observations.**
- **The geometric mean is a good summary measure in situations when data follow an exponential or logarithmic pattern typical in dilution assays such as serum antibodies, as well as environmental sampling data.**
  Note: to calculate the geometric mean, you will need a scientific calculator with *log* and y^x keys.

Formula for calculating the geometric mean from individual data:

$$\bar{x}_{geo} = \sqrt{x_1 \times x_2 \times \ldots x_n}$$

In practice, the geometric mean is calculated as:

$$\bar{x}_{geo} = antilog\left(\frac{1}{n}\sum \log x_i\right)$$

where:

- $\bar{x}_{geo}$ = **geometric mean**
- x_1 = **lowest value in the set of observations**
- x_n = **highest value in the set of observations**
- n = **number of observations**
- Σ = **summation sign**

Example:

Using the titers given on the next page, calculate the geometric mean titer of antibodies against human parainfluenza virus among the seven patients.

Table 4-2. Antibody titers of patients against human parainfluenza virus (HPIV)

ID#	Dilution	Titer
1	1:256	256
2	1:512	512
3	1:4	4
4	1:2	2
5	1:16	16
6	1:32	32
7	1:64	64

Using the second formula, we get

$$\bar{x}_{geo} = antilog_2\left[{}^{1}/_{7} \times (\log_2 256 + \log_2 512 + \log_2 4 + \log_2 2 + \log_2 16 + \log_2 32 + \log_2 64)\right]$$

$$\bar{x}_{geo} = antilog_2\left[{}^{1}/_{7} \times (8 + 9 + 2 + 1 + 4 + 5 + 6)\right]$$

$$\bar{x}_{geo} = antilog_2\left({}^{1}/_{7} \times 35\right)$$

$$\bar{x}_{geo} = antilog_2(5)$$

$$\bar{x}_{geo} = 32$$

Therefore, the geometric mean titer is 32, and the geometric mean dilution is 1:32

Properties of the Mean

- **The mean is the "center of mass"**
- **It uses all the observed values in the calculation**
- **It may or may not be an actual observed value in the data set**
- **It is algebraically tractable, which means that mean values can be computed directly**
- **Its value is affected by outliers**
- **The mean of a finite data set always exists and is unique**
- **Data values should be measured using at least an interval scale**

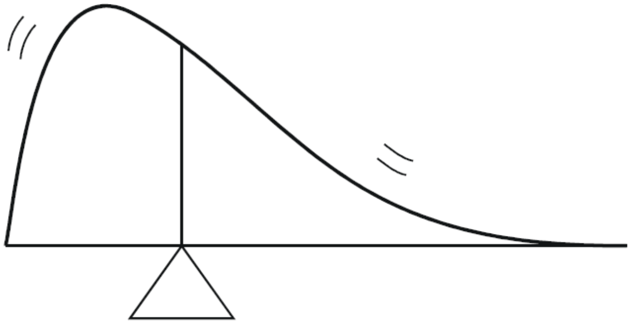
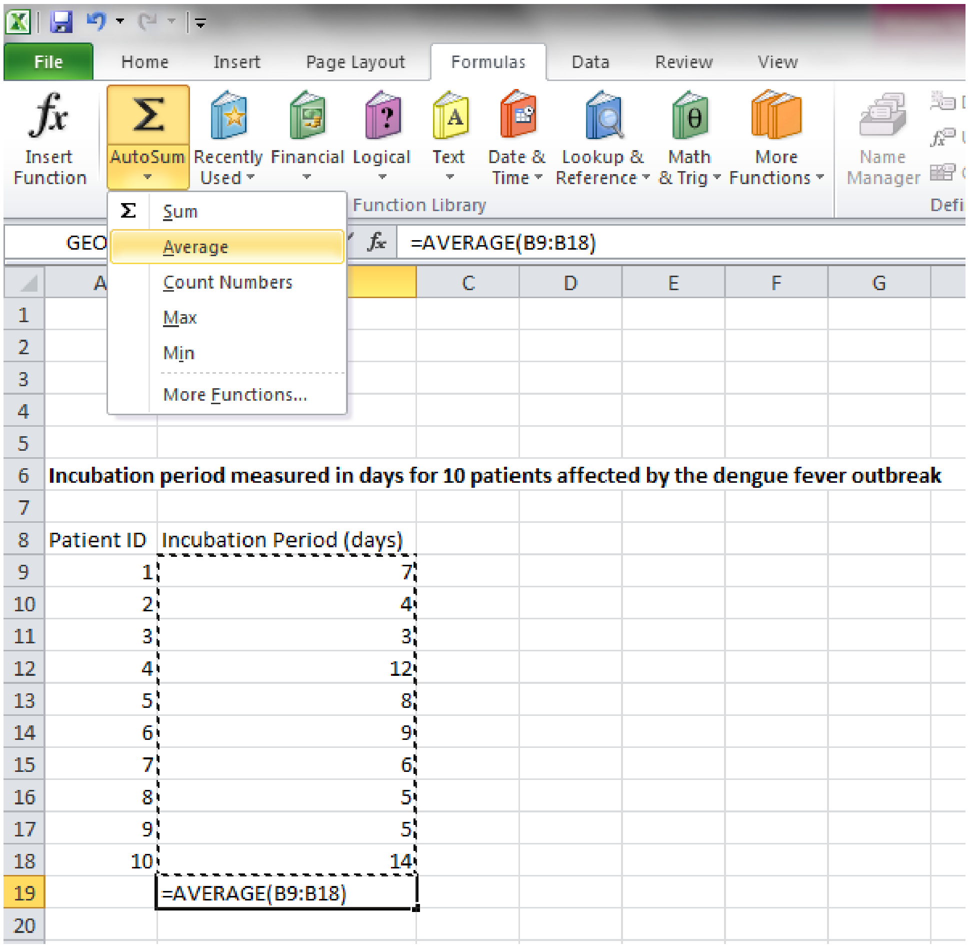


Figure 4-1. The mean is the center of gravity of the distribution

Calculating the arithmetic mean using Microsoft Excel:



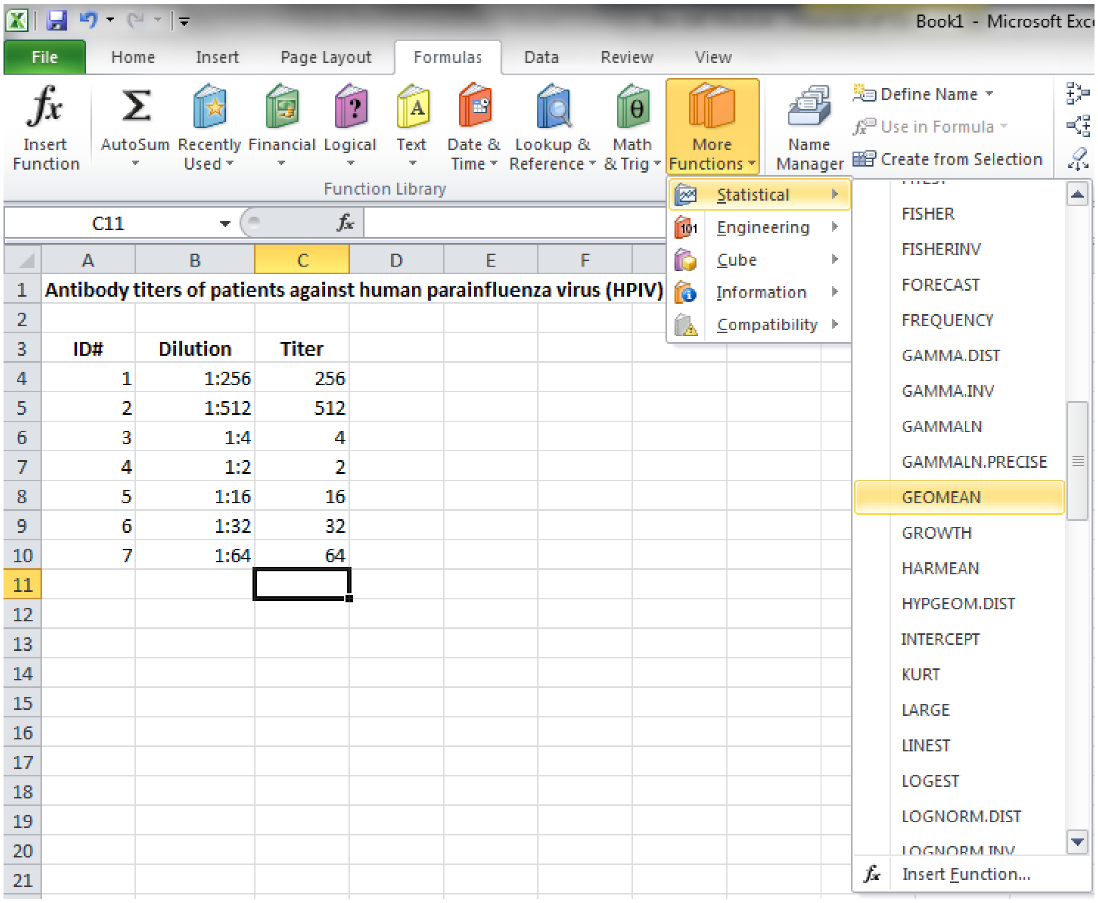
=AVERAGE(B9:B18)

Incubation period measured in days for 10 patients affected by the dengue fever outbreak

Patient ID	Incubation Period (days)
1	7
2	4
3	3
4	12
5	8
6	9
7	6
8	5
9	5
10	14
	=AVERAGE(B9:B18)

1. Input your data into the spreadsheet in an organized manner.
2. Select the cell directly underneath the individual observations. In this case, cell B19.
3. Under the *Formulas* tab, click *Auto Sum* and wait for the drop down menu to appear.
4. Click on *Average*.
5. Press Enter.

Calculating the geometric mean using Microsoft Excel:

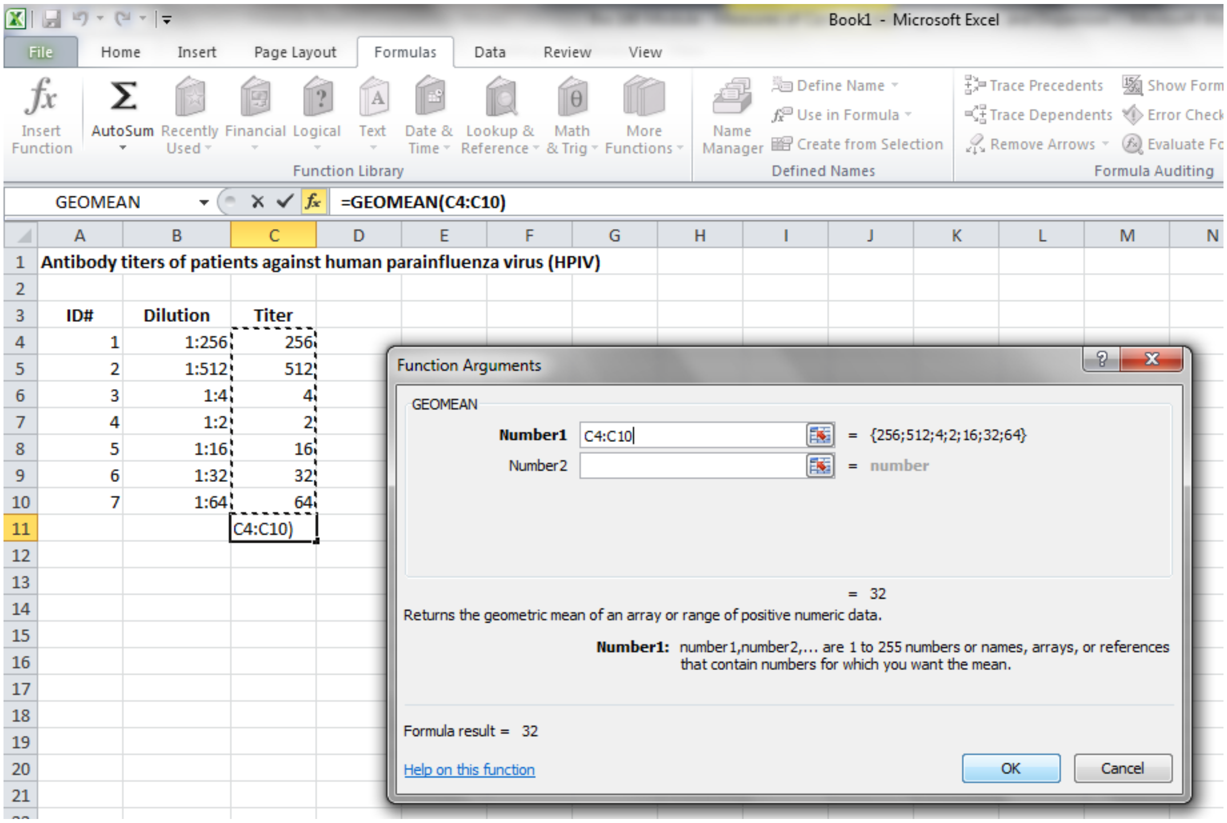


C11

Antibody titers of patients against human parainfluenza virus (HPIV)

ID#	Dilution	Titer
1	1:256	256
2	1:512	512
3	1:4	4
4	1:2	2
5	1:16	16
6	1:32	32
7	1:64	64

1. Input your data into the spreadsheet in an organized manner.
2. Select the cell directly underneath the individual observations. In this case, cell C11.
3. Under the *Formulas* tab, click *More Functions* > *Statistical* and wait for the drop down menu to appear.
4. Click on *GEOMEAN*.



GEOMEAN =GEOMEAN(C4:C10)

Antibody titers of patients against human parainfluenza virus (HPIV)

ID#	Dilution	Titer
1	1:256	256
2	1:512	512
3	1:4	4
4	1:2	2
5	1:16	16
6	1:32	32
7	1:64	64
		C4:C10)

Function Arguments

GEOMEAN

Number1 C4:C10 = {256;512;4;2;16;32;64}

Number2 = number

= 32

Returns the geometric mean of an array or range of positive numeric data.

Number1: number1,number2,... are 1 to 255 numbers or names, arrays, or references that contain numbers for which you want the mean.

Formula result = 32

Help on this function

OK Cancel

5. A menu called *Function Arguments* will appear.
6. Select the cells comprising the individual observations. In this case, select C4 through C10.
7. Click on *OK*.

Median

- From the previous example on the average incubation period during a dengue fever outbreak, only four observations were higher than the mean of 7.3 days.

$$x_i: 7, 4, 3, 12, 8, 9, 6, 5, 5, 14$$

- Thus, the mean is not very representative of the data set as a whole. The median might be more appropriate

- The median is defined as the value which divides a finite set of values into two equal parts such that the number of values equal to or greater than the median is equal to the number of values equal or less than the median.

- In other words, the median is the middlemost observation in a set of observations arranged in numerical order or in an array.

Identifying the median from individual data

1. **Arrange the observations in increasing or decreasing order. Microsoft Excel has a sort function for arranging data into an array.**
2. **Find the middle rank with the following formula:**

   $$Middle\ rank = \frac{n+1}{2}$$

   a. **If the number of observations (*n*) is odd, the middle rank falls on an observation.**

   b. **If *n* is even, the middle rank falls between two observations**
   NOTE: The formula only computes for the position of the median and the median itself!
3. **Identify the value of the median:**

   a. **If the middle rank falls on a specific observation (that is, if n is odd), the median is equal to the value of that observation.**

   b. **If the middle rank falls between two observations (that is, if n is even), the median is equal to the average (i.e., the arithmetic mean) of the values of those observations.**

***Example*:**

From the previous example, the incubation period of 10 patients affected by the dengue fever outbreak, arranged from lowest to highest, are:

3	4	5	5	6	7	8	9	12	14
1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th

$Middle\ rank = \frac{10+1}{2} = 5.5$ **[The median is between the 5th and 6th observations]**

Median = (6+7)/2 = 6.5 days

Interpretation: Half of the patients who had dengue fever had an incubation period less than 6.5 days and half had an incubation period greater than 6.5 days.

Suppose there were 11 patients, and the 11th had an incubation period of 14 days.

3	4	5	5	6	7	8	9	12	14	14
1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	11th

$Middle\ rank = \frac{11+1}{2} = 6$ **[The median falls exactly on the 6th observation]**

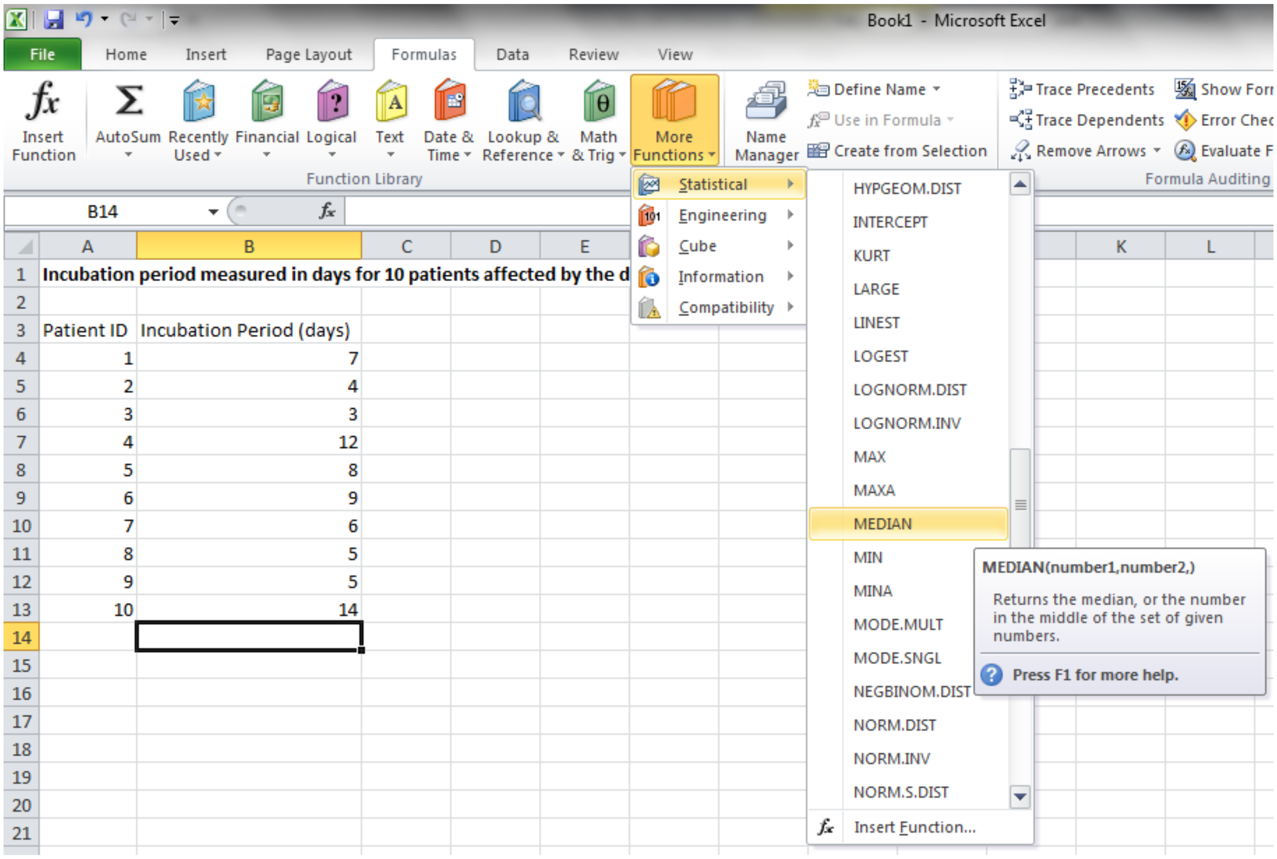
Thus, the median is 7 days.

Interpretation: Half of the patients who had dengue fever had an incubation period less than or equal to 7 days and half had an incubation period greater than 7 days.

Properties of the Median

- The median is the “center” of the array
- Unlike the mean, it uses only the middle value(s) in the array for its computation
- The median is not affected by outliers
- The median itself is not algebraically tractable. Only the position of the median is.
- The median is still interpretable when the scale of measurement used is as low as ordinal
- The median will always exist and is unique

Calculating the median using Microsoft Excel:



Incubation period measured in days for 10 patients affected by the d

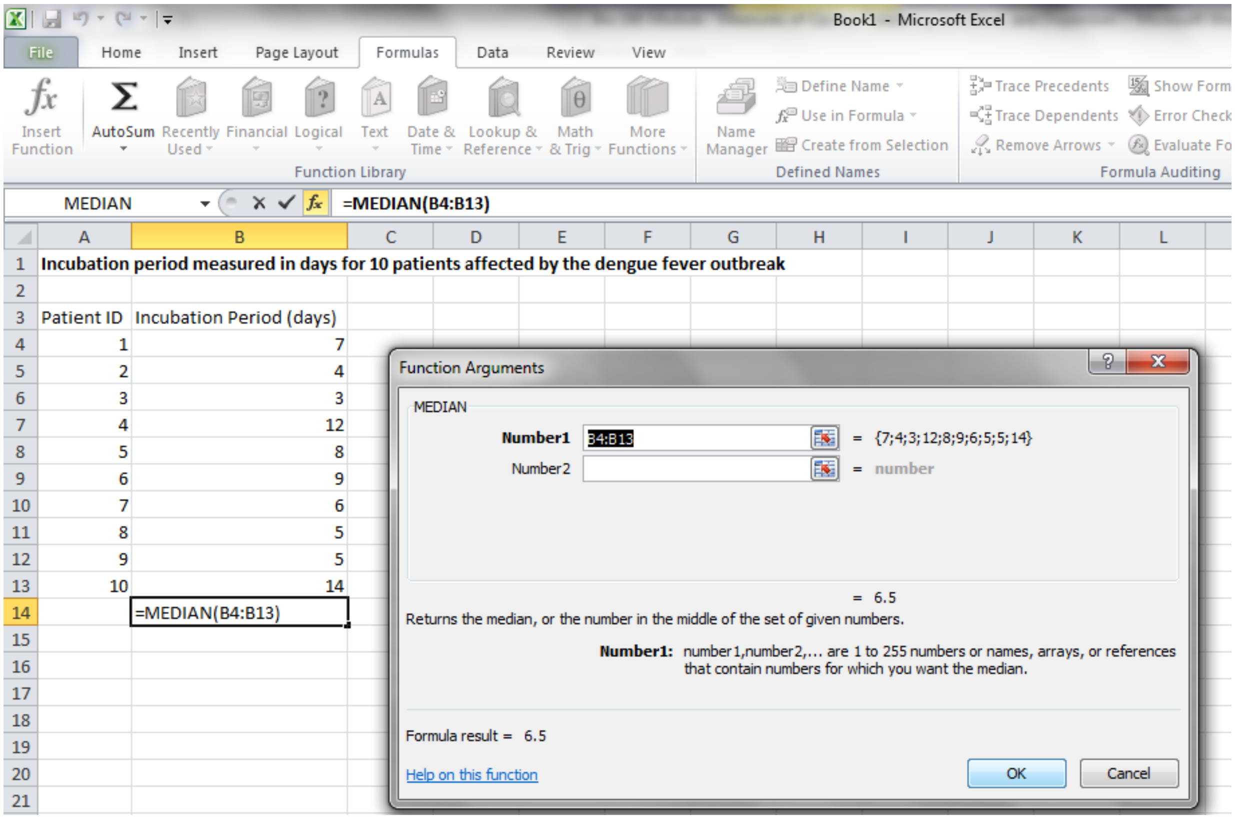
Patient ID	Incubation Period (days)
1	7
2	4
3	3
4	12
5	8
6	9
7	6
8	5
9	5
10	14

MEDIAN(number1,number2,)

Returns the median, or the number in the middle of the set of given numbers.

Press F1 for more help.

1. Input your data into the spreadsheet in an organized manner.
2. Select the cell directly underneath the individual observations. In this case, cell B14.
3. Under the *Formulas* tab, click *More Functions* > *Statistical* and wait for the drop down menu to appear.
4. Click on *MEDIAN*.



=MEDIAN(B4:B13)

Incubation period measured in days for 10 patients affected by the dengue fever outbreak

Patient ID	Incubation Period (days)
1	7
2	4
3	3
4	12
5	8
6	9
7	6
8	5
9	5
10	14
	=MEDIAN(B4:B13)

MEDIAN

Number1 B4:B13 = {7;4;3;12;8;9;6;5;5;14}

Number2 = number

= 6.5

Returns the median, or the number in the middle of the set of given numbers.

Number1: number1,number2,... are 1 to 255 numbers or names, arrays, or references that contain numbers for which you want the median.

Formula result = 6.5

5. A menu called *Function Arguments* will appear.
6. The cells of your individual observations will be automatically selected.
7. Click on *OK*.

Mode

- The value which occurs most frequently in a set of observations.
- The mode is usually found by creating a frequency distribution in which we tally how often each value occurs.
- It is possible to have no mode, one mode (unimodal distribution), two modes (bimodal distribution), or more than two modes (multimodal distribution)

Example:

Using the previous example where the incubation period of 10 patients during a dengue fever outbreak was recorded, the mode would be 5 days since the value occurred twice while all other values occurred only once.

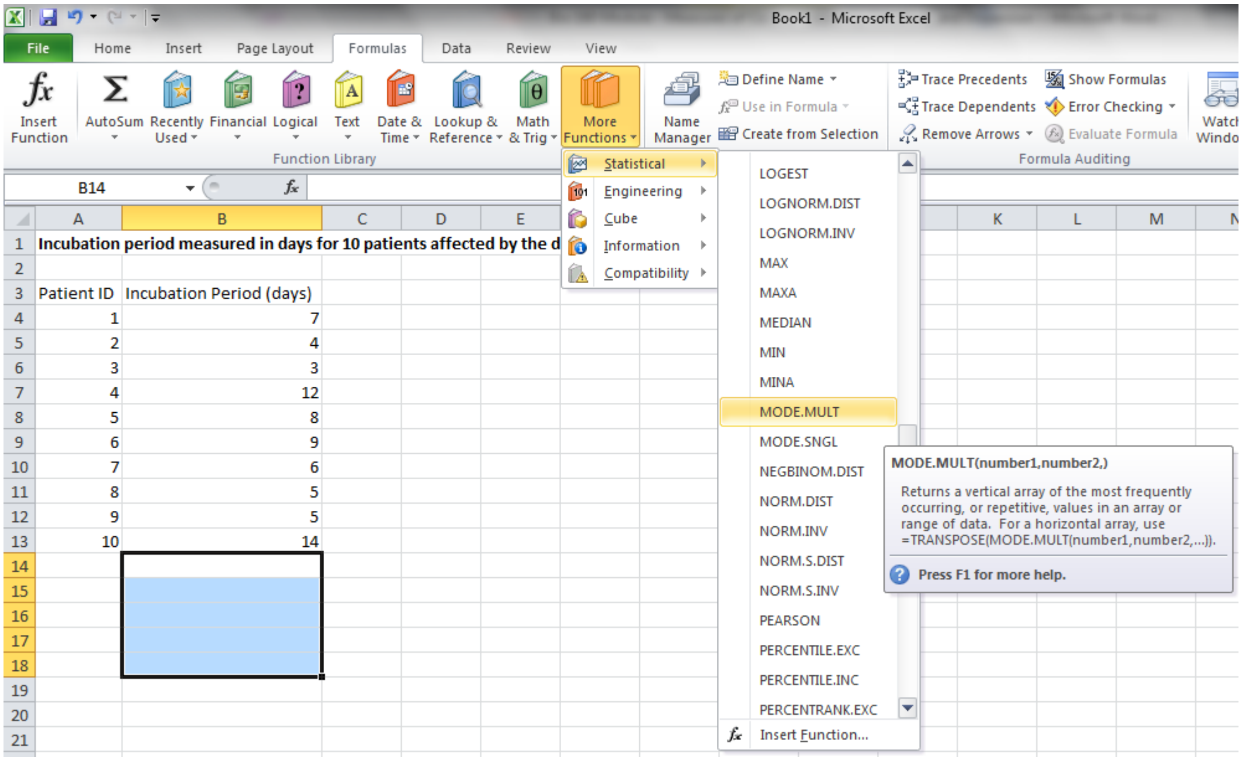
7, 4, 3, 12, 8, 9, 6, 5, 5, 14

Interpretation: The usual incubation period of a dengue fever patient is 5 days.

Properties of the Mode

- The mode is the "center" in the sense that it is the most typical value in a set of observations
- Outliers do not affect the mode
- The mode is not algebraically tractable for ungrouped data
- The mode will not always exist; and if it does, it may not be unique
- The value of the mode is always one of the observed values in the data set
- The mode can be obtained for both quantitative and qualitative types of data; that is, the mode is interpretable even if the scale of measurement is as low as nominal
- Generally, the mode is not as useful as a measure of central tendency as the mean and the median when the data consists of only a few observations. For example, for the values 5, 14, 20, 26, 37, 37, the mode is 37 since it occurred twice and all other numbers only once. However, 37 cannot be considered a good measure of central tendency for this set of data since it is in fact, at the extreme high end of the values and its frequency exceeds the frequency of the other values by only 1.

Calculating the mode using Microsoft Excel:



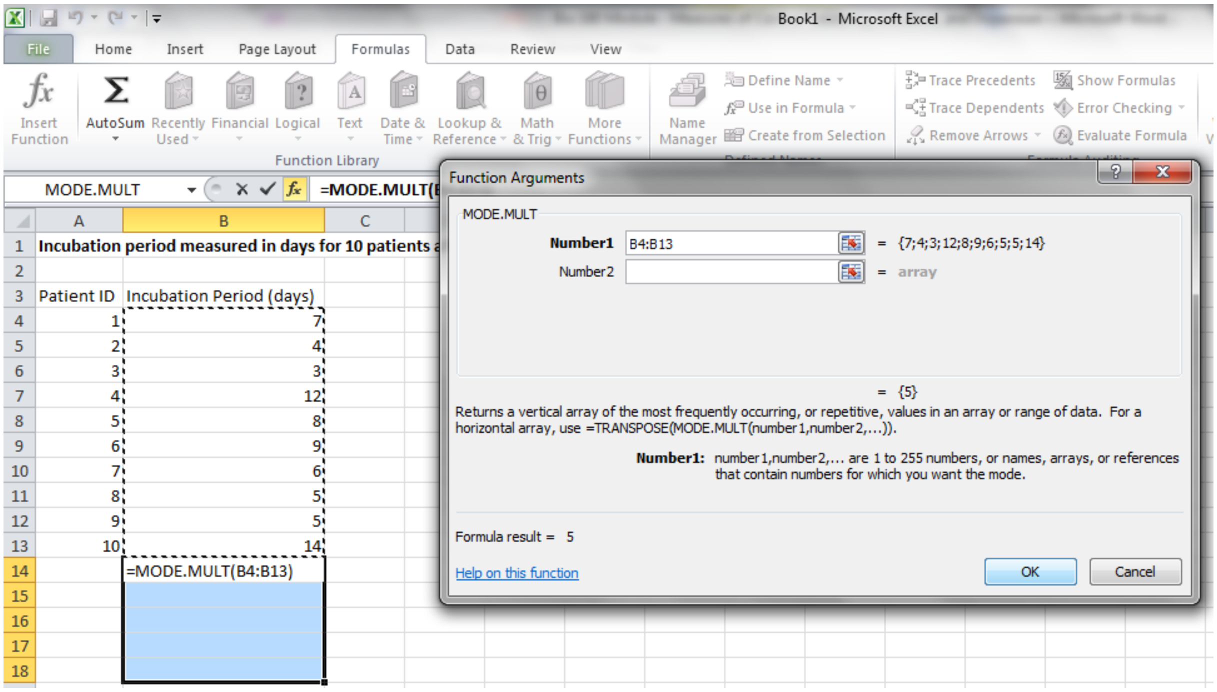
	A	B
1	Incubation period measured in days for 10 patients affected by the d	
2		
3	Patient ID	Incubation Period (days)
4	1	7
5	2	4
6	3	3
7	4	12
8	5	8
9	6	9
10	7	6
11	8	5
12	9	5
13	10	14

MODE.MULT(number1,number2,)

Returns a vertical array of the most frequently occurring, or repetitive, values in an array or range of data. For a horizontal array, use =TRANSPOSE(MODE.MULT(number1,number2,...)).

1. Input your data into the spreadsheet in an organized manner.
2. Select a few cells directly underneath the individual observations. In this case, cell B14. This is important in case the distribution is multimodal. Selecting only a single cell will give only one of the modes if the distribution is multimodal.
3. Under the *Formulas* tab, click *More Functions* > *Statistical* and wait for the drop down menu to appear.

4. **Click on *MODE.MULT*. [This will ensure that the values of all modes are returned in case the distribution is multimodal in contrast to selecting *MODE.SNGL* which will only return one mode]**



	A	B	C
1	Incubation period measured in days for 10 patients a		
2			
3	Patient ID	Incubation Period (days)	
4	1	7	
5	2	4	
6	3	3	
7	4	12	
8	5	8	
9	6	9	
10	7	6	
11	8	5	
12	9	5	
13	10	14	
14		=MODE.MULT(B4:B13)	
15			
16			
17			
18			

Function Arguments

MODE.MULT

Number1 B4:B13 = {7;4;3;12;8;9;6;5;5;14}

Number2 = array

= {5}

Returns a vertical array of the most frequently occurring, or repetitive, values in an array or range of data. For a horizontal array, use =TRANSPOSE(MODE.MULT(number1,number2,...)).

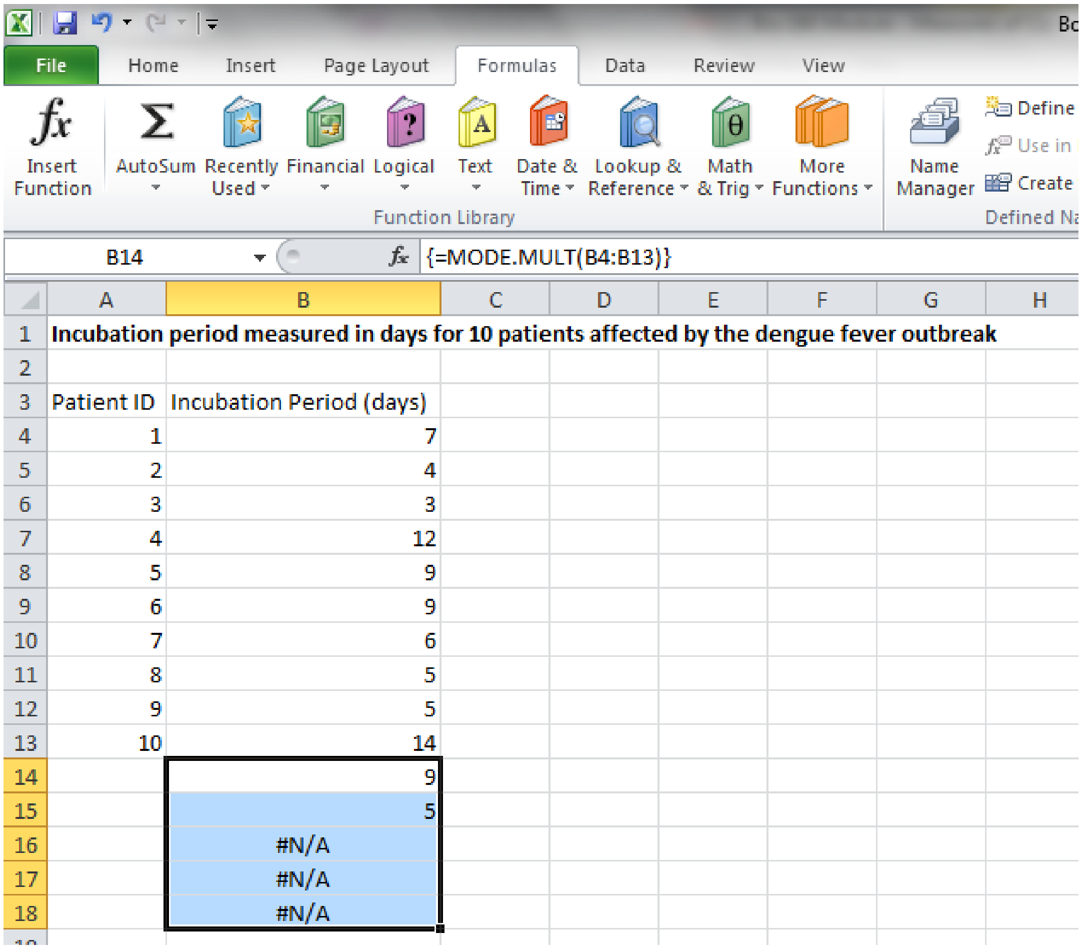
Number1: number1,number2,... are 1 to 255 numbers, or names, arrays, or references that contain numbers for which you want the mode.

Formula result = 5

Help on this function

OK Cancel

5. **A menu called *Function Arguments* will appear.**
6. **Select the cells comprising the individual observations. In this case, select B4 through B14.**
7. **Press CTRL + SHIFT + ENTER. [Clicking on OK will return only a single mode even if the distribution is multimodal which is similar to using the *MODE.SNGL* function].**
8. **In this case, all the cells selected in the vertical array will contain the value “5”. If the data were changed to have two modes, the output will look like this:**



B14 {=MODE.MULT(B4:B13)}

	A	B
1	Incubation period measured in days for 10 patients affected by the dengue fever outbreak	
2		
3	Patient ID	Incubation Period (days)
4	1	7
5	2	4
6	3	3
7	4	12
8	5	9
9	6	9
10	7	6
11	8	5
12	9	5
13	10	14
14		9
15		5
16		#N/A
17		#N/A
18		#N/A

***The value in B8 was changed from 8 to 9, making the distribution multimodal.**

Note that both modes (9 and 5) were given in the vertical array. The remainder of the cells in the vertical array show #N/A which means that there are no other modes to be displayed within those cells.

MEASURES OF DISPERSION

When we observe the graph of a frequency distribution, we generally notice two main features: 1) The graph has a peak, commonly near the center; and 2) it spreads out on either side of the peak. Just as a measure of central tendency is used to describe where the peak is located, a measure of dispersion is used to describe how much spread there is in the distribution. Several measures of dispersion are available.

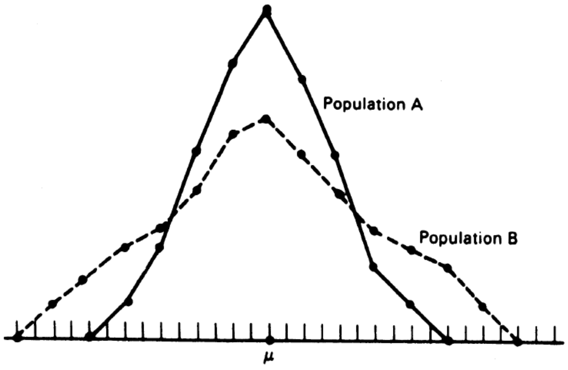


Figure 4-2. Two frequency distributions with equal means but different amounts of dispersion.

Range

- The difference between the smallest and largest value in a set of observations.
- In the statistical world, the range is reported as a single number, the difference between the maximum and minimum. In the epidemiologic community, the range is often reported as "from (the minimum) to (the maximum)", i.e., two numbers.

$$R = x_L - x_S$$

where:

- R = range
- x_L = largest value
- x_S = smallest value

Example:

Still using the previous example where the incubation period of 10 patients during a dengue fever outbreak was recorded (x_i : 7, 4, 3, 12, 8, 9, 6, 5, 5, 14), the range is equal to:

$$R = 14 - 3 = \mathbf{11\ days}$$

Interpretation: This means that the patients differed in the time when they began presenting symptoms of dengue fever by as much as 11 days.

Variance

- Measure of a variable's spread or distribution around its mean.
- Takes into account the squared deviations of individual observations from the mean

Table 4-3. Formulae for variance

General formula for the finite population variance	General formula for the sample variance
$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$	$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$
Where • σ^2 = population variance • μ = population mean • N = population size • x_i = individual observation • i = position of observation • Σ = summation sign	Where • s^2 = sample variance • $\bar{x}$ = sample mean • n = sample size • x_i = individual observation • i = position of observation • Σ = summation sign

Example:

Still from our previous example: 7, 4, 3, 12, 8, 9, 6, 5, 5, 14

$$s^2 = \frac{(7 - 7.3)^2 + (4 - 7.3)^2 + \ldots\ (3 - 7.3)^2}{9}$$

s^2 = 12.5 days^2

Alternative formula for the sample variance:

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

$\sum_{i=1}^{n} x_i^2$ means square the individual observations then take the sum of the squared observations

$\left(\sum_{i=1}^{n} x_i\right)^2$ means take the sum of the individual observations, then square the sum

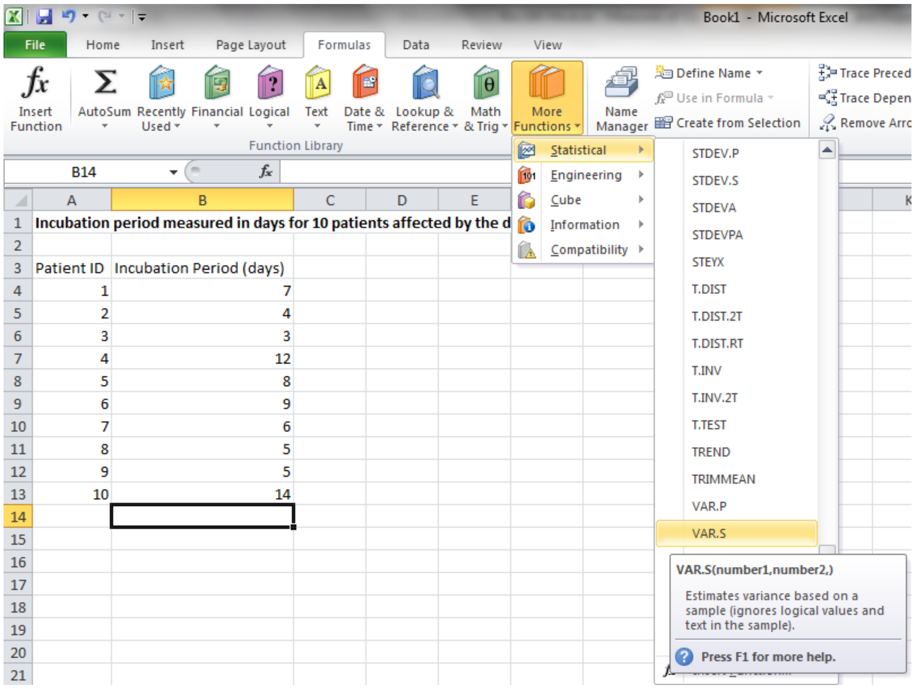
Patient No.	Incubation Period in days (x_i)	x_i^2
1	7	49
2	4	16
3	3	9
4	12	144
5	8	64
6	9	81
7	6	36
8	5	25
9	5	25
10	14	196
Total	73	645

$$\sum_{i=1}^{10} x_i \qquad \sum_{i=1}^{10} {x_i}^2$$

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

$$s^2 = \frac{645 - \frac{73^2}{10}}{9} = \frac{645 - 532.9}{9} = \frac{112.1}{9} = \mathbf{12.5\ days^2}$$

Calculating the variance using Microsoft Excel:



B14

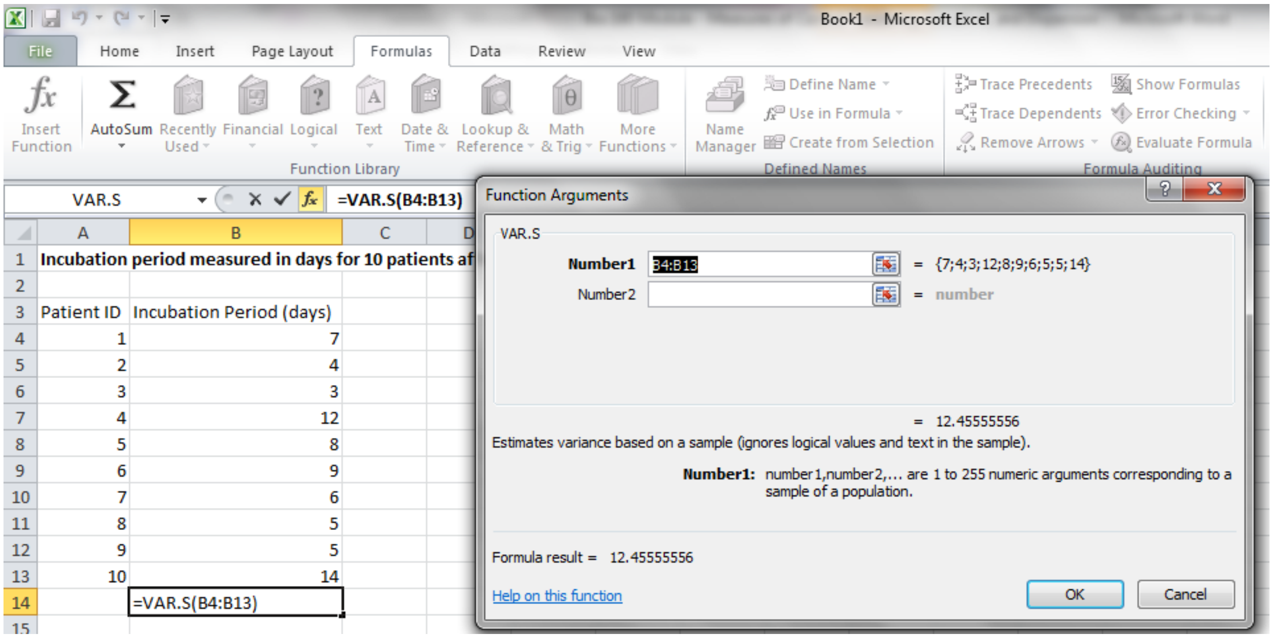
	A	B
1	Incubation period measured in days for 10 patients affected by the d	
2		
3	Patient ID	Incubation Period (days)
4	1	7
5	2	4
6	3	3
7	4	12
8	5	8
9	6	9
10	7	6
11	8	5
12	9	5
13	10	14
14		

VAR.S(number1,number2,)

Estimates variance based on a sample (ignores logical values and text in the sample).

Press F1 for more help.

1. **Input your data into the spreadsheet in an organized manner.**
2. **Select the cell directly underneath the individual observations. In this case, cell B14.**
3. **Under the *Formulas* tab, click *More Functions* > *Statistical* and wait for the drop down menu to appear.**
4. **Click on *VAR.P* if you wish to calculate the population variance or *VAR.S* if you wish to calculate the sample variance.**



VAR.S =VAR.S(B4:B13)

	A	B	C
1	Incubation period measured in days for 10 patients af		
2			
3	Patient ID	Incubation Period (days)	
4	1	7	
5	2	4	
6	3	3	
7	4	12	
8	5	8	
9	6	9	
10	7	6	
11	8	5	
12	9	5	
13	10	14	
14		=VAR.S(B4:B13)	
15			

Function Arguments

VAR.S

Number1 B4:B13 = {7;4;3;12;8;9;6;5;5;14}

Number2 = number

= 12.45555556

Estimates variance based on a sample (ignores logical values and text in the sample).

Number1: number1,number2,... are 1 to 255 numeric arguments corresponding to a sample of a population.

Formula result = 12.45555556

5. **A menu called *Function Arguments* will appear.**
6. **The cells of your individual observations will be automatically selected.**
7. **Click on *OK*.**

Standard Deviation

- **Simply the square root of the variance**
- **Note that the units of measurement for the variance are in units square (e.g. grams^2, meters^2). This is a major drawback for the variance's use as a measure of dispersion – it is difficult for most people to think in terms of squared units.**
- **The formula for the sample standard deviation is:**

$$s = \left[\sum_{i=1}^{n} (x_i - \overline{x})^2 \Big/ (n-1) \right]^{1/2}$$

$$= \left[\left(\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2 \Big/ n \right) \Big/ (n-1) \right]^{1/2}$$

From our previous example, the standard deviation is $\sqrt{12.5\ days^2} = 3.5\ days$

Interpretation: This means that on average, the incubation period of patients affected by the dengue fever outbreak is ± 3.5 units away from the mean (7.3 ± 3.5).

Properties of the Standard Deviation

- It uses every observation in its computation.
- It may be distorted by outliers. This is because squaring large deviations from the mean will give more weight to these outliers.
- It is algebraically tractable.
- It is always nonnegative. A value of 0 implies the absence of variation.
- The scale of measurement must at least be interval for the standard deviation to be interpretable.

Calculating the standard deviation using Microsoft Excel:

Follow the same steps as if you were calculating variance except that you must change the function to *STDEV.P* if you wish to calculate the population standard deviation or *STDEV.S* if you wish to calculate the sample standard deviation.

Coefficient of Variation

- A measure of relative dispersion which expresses the standard deviation as a percentage of the mean
- May be used to compare SDs of two variables measured in different units or used when two means, although measured in the same unit, differ appreciably.
- The formula for the coefficient of variation is:

  $$CV = \frac{s}{\bar{x}}(100)$$

  where *s* is the sample standard deviation, and $\bar{x}$ is the sample mean

Example:

Suppose two samples of female chimpanzees yield the following results:

	Sample A	Sample B
Age	5 years	15 years
Mean Height	45 cm	70 cm
Standard Deviation	8 cm	8 cm

We wish to know which is more variable, the heights of the 5-year old female chimpanzees or the heights of the 15-year old female chimpanzees. Comparing the standard deviation of the two samples may mislead us that both are equally variable. However, by using the coefficient of variation, we get a different impression.

- **CV for Sample A**

  $CV = \frac{8}{45}(100) = 17.8\%$

- **CV for Sample B**

  $CV = \frac{8}{70}(100) = 11.4\%$

Interpretation: The heights of the 5-year old chimpanzees are more dispersed than the heights of the 15-year old chimpanzees.

The coefficient of variation (CV) is a sample statistic. Its corresponding population parameter is called the coefficient of dispersion (CD) and is given by the following formula:

$$CD = \frac{\sigma}{\mu}(100)$$

Choosing the measures of central tendency and dispersion

Table 5-4. Preferred measures of central tendency and dispersion by type of data

Type of Distribution	Measure	
	Central Tendency	Dispersion
normal	arithmetic mean	standard deviation
skewed	median	interquartile range
exponential or logarithmic	geometric mean	consult statistician

SOURCE: Principles of Epidemiology, Centers for Disease Control and Prevention (1992)

MEASURES OF LOCATION

- **Percentile (P_i) – one of the 99 values of a variable which divides the distribution into 100 equal parts.**
    - **There are 99 percentiles, denoted by $P1, P2, \ldots, P99$. The k^{th} percentile, denoted by P_k, is a value such that at least $k\%$ of the observations are less than or equal to it and at least $(100 - k)\%$ are greater than it, where $k = 1,2,3, \ldots, 99$**
- **Decile (D_i) – one of the 9 values of a variable which divides the distribution into 10 equal parts.**
- **Quartile (Q_i) – one of the 3 values of a variable which divides the distribution into 4 equal parts.**

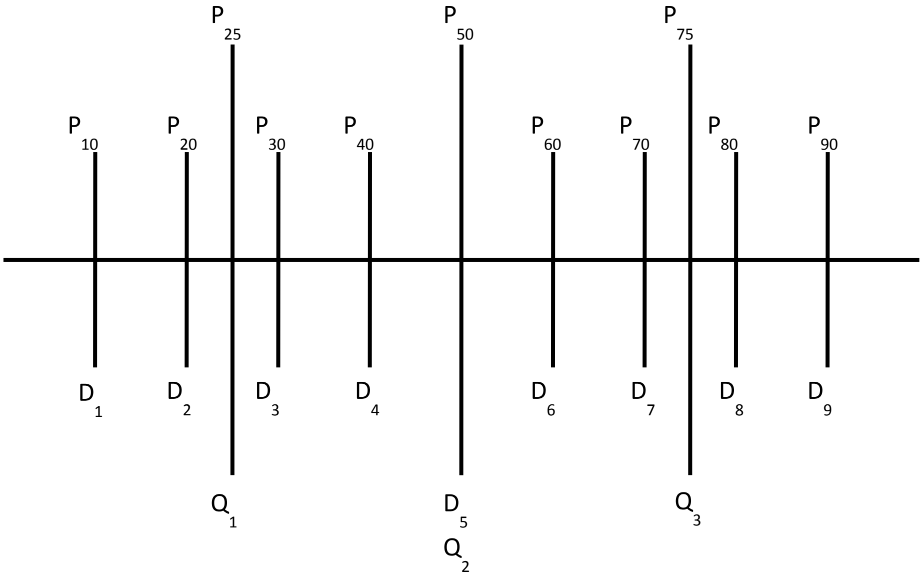


Figure 4-3. Equivalence of percentiles, deciles, and quartiles and their position along the number line.

Finding the k^{th} percentile

1. **Order the data in increasing order of magnitude.**
2. **Compute $nk/100$ where *n* is the sample size while *k* is the percentile of interest.**
3. **If $nk/100$ is not an integer, the k^{th} percentile is the $j + 1^{th}$ largest measurement where *j* is the largest integer less than $nk/100$.**

   $$P_k = x_{\frac{nk}{100}+1'}$$

   [1′ denotes adding the necessary value to $nk/100$ to *raise* it to the nearest integer]
4. **If $nk/100$ is an integer, the k^{th} percentile is the average of the $(nk/100)^{th}$ and $[(nk/100) + 1]^{th}$ largest observations.**

   $$P_k = \frac{x_{\frac{nk}{100}} + x_{\frac{nk}{100}+1}}{2}$$

Additional Notes on the Interpretation of P_k

- P_k will be an interpolated value if $nk/100$ is not an integer. If the values used in the interpolation are not tied values then P_k will not be one of the observations. In such a case, the interpretation of P_k will simplify as follows:
  "$k\%$ of observations are less thanP_k. Likewise, $(100-k)\%$ are greater than P_k".
- That is, $nk/100$ observations are less than P_k so that the remaining $n-(nk/100)$ = $n\ (1-k/100)$ are greater than P_k.

A note on the calculation of percentiles

There is no standard definition of a percentile. The method used here relies on calculating $nk/100$ to determine the position of the percentile. Another method to estimate the position of the percentile is $k(n+1)$ while Excel uses $1+k(n-1)$. In addition, the latter two methods uses linear interpolation to obtain a more accurate estimate of the percentile when it does not fall on a single observation. When calculating percentiles using Excel, we suggest the use of the PERCENTILE.EXC function. For more information about the proper use of this function, please visit the following URL: http://office.microsoft.com/en-us/excel-help/percentile-exc-function-HA010345439.aspx

Note that the PERCENTILE.EXC function is included only in the 2010 version of Microsoft Excel. The PERCENTILE function present in earlier versions of Excel is similar to the PERCENTILE.INC function currently present in Excel 2010. These two functions, including the method we discussed here for calculating percentiles by hand, yield different results. However, when the number of observations is very large, all these methods will yield similar results.
Example:

Using once again the data on the incubation period of the 10 patients affected by the dengue fever outbreak, find the 90th percentile (x_i: 3, 4, 5, 5, 6, 7, 8, 9, 12, 14).

$$\frac{nk}{100}=\frac{10(90)}{100}=\frac{900}{100}=9$$

Since 9 is an integer, the k^{th} percentile is the average of the 9th and 10th largest observations.

$$P_{90}=\frac{x_{\frac{10(90)}{100}}+x_{\frac{(10)(90)}{100}+1}}{2}=\frac{x_9+x_{10}}{2}=\frac{12+14}{2}=\frac{26}{2}=13\ days$$

Interpretation: 90% of the patients had an incubation period <u>less than</u> 13 days while 50% of the patients had an incubation period greater than 13 days.

Using the same data, this time, find the 23rd percentile.

$$\frac{nk}{100}=\frac{10(23)}{100}=\frac{230}{100}=2.3$$

Since 2.3 is not an integer, the 23rd percentile is the $j+1^{th}$ largest measurement where *j* is the largest integer less than 2.3. Thus, *j* = 2 and *j* + 1 = 3. The 23rd percentile falls on the 3rd observation.

$$P_k = x_{\frac{(10)(23)}{100}+1'} = x_{2.3+1'} = x_3 = 5\ days$$

Interpretation: 23% of the patients had an incubation period <u>less than or equal</u> to 5 days while 77% of the patients had an incubation period greater than 5 days.

Note the difference in the interpretation when $nk/100$ is an integer and when it is not.

Calculating Deciles and Quartiles

- **To compute for deciles and quartiles, first determine the equivalent percentile, then use the equation for the percentile.**
- **Using the previous data, compute for D_8.**
  1. **Determine the corresponding percentile**
     $D_8 = P_{80}$
  2. **Calculate as you would a percentile**

Interquartile Range

- **Represents the central portion of the distribution, and is calculated as the difference between the third or upper quartile (Q_3) and the first or lower quartile (Q_1).**
- **This range includes about one-half of the middlemost observations in the set, leaving one-quarter of the observations on each side.**

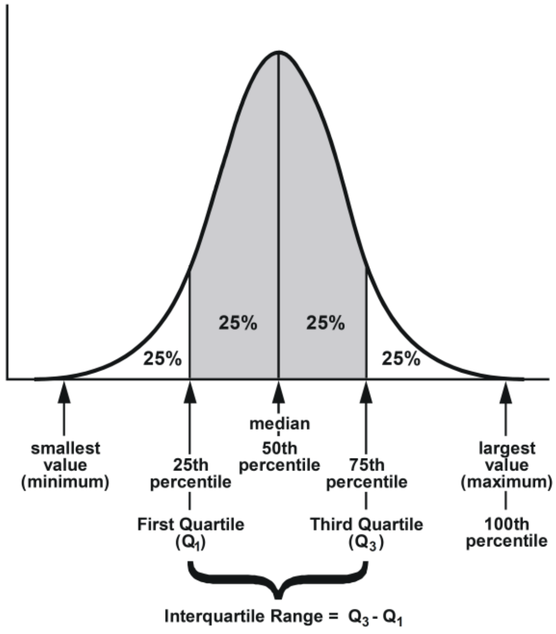


Figure 4-4. The middle half of the observations in a frequency distribution lie within the interquartile range

Calculating the interquartile range from individual data

1. Arrange the observations in increasing order.
2. Find the position of the 1st and 3rd quartiles.
3. Identify the value of the 1st and 3rd quartiles.
4. Calculate the interquartile range as Q_3 minus Q_1.

Example:

Still using the data on the incubation period of the 10 patients affected by the dengue fever outbreak, find the interquartile range (x_i : 3, 4, 5, 5, 6, 7, 8, 9, 12, 14).

The position of Q_1, which is equivalent to P_{25} is given by:

$$\frac{nk}{100} = \frac{10(25)}{100} = \frac{250}{100} = 2.5$$

Since 2.5 is not an integer, the position of Q_1 is the 3rd observation and is equal to 5 days.

The position of Q_3, which is equivalent to P_{75} is given by:

$$\frac{nk}{100} = \frac{10(75)}{100} = \frac{750}{100} = 7.5$$

Since 7.5 is not an integer, the position of Q_3 is the 8th observation and is equal to 9 days.

Therefore, the interquartile range is $Q_3 - \; Q_1 = 9 - 5 = \mathbf{4}$

Now that we have learned how to calculate the measures of central tendency, dispersion, and location for ungrouped data, let us now discuss how to calculate these measures when the data are arranged in a frequency distribution.

MEASURES OF CENTRAL TENDENCY (Grouped Data)

Mean

- In calculating the mean from grouped data, we assume that all values falling into a particular class interval are located at the *midpoint* of the interval.

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

where:

- x_i is the midpoint of the i^{th} interval
- f_i is the frequency of observations in the i^{th} interval
- *k* is the number of categories
- *n* is the total number of observations

- **The midpoint of a class interval is also called the class mark. It may be calculated by dividing by taking the sum of either the stated or true lower and upper limits of the same interval and dividing it by 2.**

$$class\ mark = \frac{stated\ lower\ class\ limit + stated\ upper\ class\ limit}{2}$$

$$class\ mark = \frac{true\ lower\ class\ limit + true\ upper\ class\ limit}{2}$$

We will use the following data in all examples for grouped data calculations:

Clasen et al. studied sparteine and mephenytoin oxidation in a group of participants who inhabit Greenland. Two populations were represented in their study: East Greenlanders and West Greenlanders. The investigators were interested in genetic polymorphisms between the two groups. The ages of the participants in this study are summarized in the following frequency distribution table:

Table 4-5. Frequency distribution of ages of the 169 participants in a study of sparteine and mephenytoin oxidation

Class Interval	Frequency
10 – 19	4
20 – 29	66
30 – 39	47
40 – 49	36
50 – 59	12
60 – 69	4
Total	169

SOURCE: Knud Clasen, Laila Madsen, Kim Brøsen, Kurt Albøge, Susan Misfeldt, and Lars F. Gram, "Sparteine and Mephenytoin Oxidation: Genetic Polymorphisms in East and West Greenland," *Clinical Pharmacology & Therapeutics*, *49*, (1991), 624-631

***Example*:**

Class Interval	Frequency (f_i)	Midpoint (x_i)	f_ix_i
10 – 19	4	14.5	58.0
20 – 29	66	24.5	1617.0
30 – 39	47	34.5	1621.5
40 – 49	36	44.5	1602.0
50 – 59	12	54.5	654.0
60 – 69	4	64.5	258.0
Total	169		5810.5

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

$$\bar{x} = \frac{5810.5}{169} = 34.4\ years$$

Interpretation: The average age of the participants in the sparteine and mephenytoin oxidation study is 34.4 years.

Median

The steps for calculating the median from grouped data is as follows:

1. Construct the cumulative frequency distribution (CFD)
2. Calculate n/2, where n is the number of observations
3. Starting from the top, locate the value in the CFD column that is ≥ n/2 for the first time. The class interval corresponding to that value is the median class.
4. Approximate the median using the formula:

$$Median = l_m + \frac{w_m(n/2 - cf_{m-1})}{f_m}$$

where:

- l_m is the true lower limit of the median class
- w_m is the class width of the median class
- n is the total no. of observations
- cf_{m-1} is the cumulative frequency of the class before the median class
- f_m is the frequency of the median class

- Unlike computing the mean from grouped data where it is assumed that the values within a class interval are located at the midpoint, in computing the median, we assume that they are evenly distributed through the interval.

Example:

Class Interval	Class Boundaries	Frequency (f_i)	Cumulative Frequency (cf)	$cf \geq n/2 = 84.5$?
10 – 19	9.5 – 19.5	4	4	No
20 – 29	19.5 – 29.5	66	70	No
30 – 39	29.5 – 39.5	47	117	Yes
40 – 49	39.5 – 49.5	36	153	
50 – 59	49.5 – 59.5	12	165	
60 – 69	59.5 – 69.5	4	169	
Total		169		

*The class shaded pink is the median class

$$Median = \ l_m + \frac{w_m(n/2 - cf_{m-1})}{f_m}$$

$$Median = 29.5 + \frac{10({}^{169}/_2 - 70)}{47} = \boldsymbol{32.6\ years}$$

Interpretation: Half of the participants in the sparteine and mephenytoin oxidation study have an age less than 32.6 years while the other half have an age greater than 32.6 years.*

*Note that the interpretation uses the less than $k\%$ statement rather than the less than or equal to $k\%$ statement. This is the appropriate interpretation for the median of grouped data and is true for the percentiles of grouped data as well.

Mode

The steps for calculating the mode from grouped data is as follows:

1. Locate the modal class. For a frequency distribution with equal class sizes, the modal class is the class with the highest frequency.
2. Approximate for the mode using the formula:

$$Mode = \ l_{mo} + w_{mo}\left(\frac{f_{mo} - f_1}{2f_{mo} - f_1 - f_2}\right)$$

where:

- l_{mo} is the true lower limit of the modal class
- w_{mo} is the class width
- f_{mo} is the frequency of the modal class
- f_1 is the frequency of the class preceding the modal class
- f_2 is the frequency of the class following the modal class

Example:

Class Interval	Class Boundaries	Frequency (f_i)
10 – 19	9.5 – 19.5	4
20 – 29	19.5 – 29.5	66
30 – 39	29.5 – 39.5	47
40 – 49	39.5 – 49.5	36
50 – 59	49.5 – 59.5	12
60 – 69	59.5 – 69.5	4
Total		169

*The class shaded pink is the modal class

$$Mode = \ l_{mo} + w_{mo}\left(\frac{f_{mo} - f_1}{2f_{mo} - f_1 - f_2}\right)$$

$$Mode = 19.5 + 10\left(\frac{66 - 4}{2(66) - 4 - 47}\right) = \boldsymbol{27.15\ years}$$

Interpretation: The usual age of the participants involved in the study of sparteine and mephenytoin oxidation is 27.15 years.

MEASURES OF DISPERSION (Grouped Data)

Range

The range is simply the lower class limit of the first class subtracted from the upper class limit of the last class.

$$Range = UCL_{HCl} - LCL_{LCl}$$

where:

- UCL_{HCl} is the upper class limit of the last class
- LCL_{LCl} is the lower class limit of the first class

Example:

Class Interval	Frequency
10 – 19	4
20 – 29	66
30 – 39	47
40 – 49	36
50 – 59	12
60 – 69	4
Total	169

$$Range = 69 - 10 = 59\ years$$

Interpretation: The age difference between the oldest and the youngest sparteine and mephenytoin oxidation study participant is 59 years.

Properties of the Range

- It is a simple measure (easy to compute and understand)

Weaknesses:

- It fails to communicate any information about the clustering or the lack of clustering of values in the middle of the distribution since it uses only the extreme values (minimum and maximum)
- An outlier can greatly affect its value
- It tends to be smaller for smaller collections than for larger collections
- It cannot be approximated from frequency distributions with an open-ended class
- It is not algebraically tractable

Variance

- In calculating the variance (and standard deviation) from grouped data, we assume that all the values falling into a particular class interval are located in the midpoint of the interval.

$$s^2 = \frac{\sum_{i=1}^{k} f_i x_i^2 - \frac{\left(\sum_{i=1}^{k} f_i x_i\right)^2}{n}}{n-1}$$

where:

- s^2 = sample variance
- k = no. of categories
- f_i = frequency of the ith category
- x_i = midpoint of the ith category
- n = total no. of observations

Example:

Class Interval	Frequency (f_i)	Midpoint (x_i)	f_ix_i	$f_ix_i^2$
10 – 19	4	14.5	58.0	841
20 – 29	66	24.5	1617.0	39616.5
30 – 39	47	34.5	1621.5	55941.75
40 – 49	36	44.5	1602.0	71289
50 – 59	12	54.5	654.0	35643
60 – 69	4	64.5	258.0	16641
Total	169		5810.5	219972.3

$$s^2 = \frac{\sum_{i=1}^{k} f_i x_i^2 - \frac{\left(\sum_{i=1}^{k} f_i x_i\right)^2}{n}}{n-1}$$

$$s = \frac{219972.3 - \frac{5810.5^2}{169}}{168} = \mathbf{120.22}\ \boldsymbol{years^2}$$

As mentioned earlier, the variance is difficult to interpret due to the units of measurement being squared. Thus, let us compute for the standard deviation which is simply the square root of the variance.

$$s = \sqrt{s^2} = \sqrt{120.22\ years^2} = \mathbf{10.96}\ \boldsymbol{years}$$

Interpretation: The ages of the participants in the sparteine and mephenytoin oxidation study vary by, on average, ±10.96 years from the mean of 34.4 years (34.4±10.96 years).

MEASURES OF LOCATION (Grouped Data)

Percentile

The steps for calculating the percentile from grouped data is as follows:

1. **Construct the CFD**
2. **Compute for** $nk/100$
3. **Locate the** k^{th} **percentile class (**P_k^{th} **class). The** P_k^{th} **class is the class interval where the less than cumulative frequency is** $\geq nk/100$**for the first time, starting from the top**
4. **Use the following formula to approximate** P_k**:**

$$P_k = l_k + \frac{w_k\left[\frac{nk}{100} - cf_{k-1}\right]}{f_k}$$

where:

- ***P_k*** **=** ***k^{th}*** **percentile to be computed**
- ***l_k*** **= true lower limit of the percentile class**
- ***w_k*** **= width of the percentile class**
- ***n*** **= total no. of observations**
- ***cf_{k-1}*** **= cumulative frequency of the class before the percentile class**
- ***f_k*** **= frequency of the percentile class**

***Example*:**

Using the previous data, compute for Q_3

Class Interval	Class Boundaries	Frequency (f_i)	Cumulative Frequency (cf)	$cf \geq nk/100 = 126.75$?
10 – 19	9.5 – 19.5	4	4	No
20 – 29	19.5 – 29.5	66	70	No
30 – 39	29.5 – 39.5	47	117	No
40 – 49	39.5 – 49.5	36	153	Yes
50 – 59	49.5 – 59.5	12	165	
60 – 69	59.5 – 69.5	4	169	
Total		169		

***The class shaded pink is the percentile class**

$$Q_3 = P_{75}$$

$$P_k = l_k + \frac{w_k\left[\frac{nk}{100} - cf_{k-1}\right]}{f_k}$$

$$P_{75} = 39.5 + \frac{10\left[\frac{169(75)}{100}\right] - 117}{36} = \mathbf{42.21}\ \boldsymbol{years}$$

Interpretation: 75% of the participants in the sparteine and mephenytoin oxidation study had an age less than 42.21 years while 25% had an age greater than 42.21 years.

LABORATORY EXERCISE 5
Descriptive Measures (50 points)

NAME:____ DATE:______________

SECTION:______________________________ INSTRUCTOR:__________________________

1. When computing for the sample variance or standard deviation, what is the rationale behind dividing by $n - 1$ instead of n? (5 points)

____

____

____

____

____

2. The following table shows the number of hours 45 hospital patients slept following the administration of a certain anesthetic.

7	10	12	4	8	7	3	8	5
12	11	3	8	1	1	13	10	4
4	5	5	8	7	7	3	2	3
8	13	1	7	17	3	4	5	5
3	1	17	10	4	7	7	11	8

a. Determine the mean, median, and mode. Interpret the results (6 points)

Mean (2 pts)	Interpretation (1 pt)
Median (2 pts)	Interpretation (1 pt)

Mode (2 pts)	**Interpretation (1 pt)**

b. **Determine the variance and standard deviation. Interpret the standard deviation (5 points)**

Variance (3 pts)	
Standard Deviation (1 pts)	**Interpretation (1 pt)**

c. **Determine P_{60}, D_1, and the interquartile range. Interpret the results (11 points)**

$\boldsymbol{P_{60}}$ **(2 pts)**	**Interpretation (1 pt)**

D_1 (2 pts)	**Interpretation (1 pt)**
Interquartile Range (4 points)	**Interpretation (1 pt)**

3. **The following table shows the frequency distribution of serum cholesterol levels (mg/dl) in a 4,462 individuals who reported for hypercholesterolemia. Use the supplied columns to write down the necessary information that you need to solve the following problems.**

Class Interval	Class Boundaries	Midpoint (x_i)	Freq. (f_i)	$f_i x_i$	$f_i x_i^2$	Cum. freq. (*cf*)	*cf* ≥ *n*/2?	*cf* ≥ *nk/100*?		
								P_{88}	D_3	Q_1
60-99			9							
100-139			111							
140-179			811							
180-219			1677							
220-259			1285							
260-299			464							
300-339			88							
340-379			11							
380-419			6							
Total			4462							

a. **Determine the mean, median, and mode. Interpret the results (9 points)**

Mean (2 pts)	**Interpretation (1 pt)**
Median (2 pts)	**Interpretation (1 pt)**
Mode (2 pts)	**Interpretation (1 pt)**

b. Determine the variance and standard deviation. Interpret the results (5 points)

Variance (3 pts)	
Standard Deviation (1 pts)	**Interpretation (1 pt)**

c. Determine the 88th percentile, 3rd decile, and 1st quartile. Interpret the results (9 points)

P_{88} **(2 pts)**	**Interpretation (1 pt)**

D_3 **(2 pts)**	**Interpretation (1 pt)**
Q_1 **(2 pts)**	**Interpretation (1 pt)**

5 Basic Concepts in Probability

Probability *is a mathematical concept that determines the likelihood of occurrence of events that are subject to chance*. When we say an event is subject to chance, we mean the outcome is in doubt and there are at least two possible outcomes.

The origins of probability can be traced to gambling. Games of chance provide good demonstrations of what the possible events are. Typical statements that you might see concerning probability include: the chance of getting a flush in a 5-card poker hand is about 2 in 1000 or 0.2%; the chance of throwing a sum of 10 with two dice is 1 in 12 or 8.33%, or the probability that a ball will land on red in a roulette wheel is 1 in 2 or 50%. Meanwhile, in the health sciences setting, one might be interested in the probability that a patient who receives a novel medical treatment will live for two or more years. We may hear a physician say that a patient has a 50-50 chance of surviving a particular operation. Knowledge of the probability of these outcomes can help in making better informed decisions, for example, whether or not the patient should undergo the operation if the chance of surviving is small. Another demonstration of probability lies on the fact that many events in life are uncertain. We do not know whether it will rain tomorrow or when the next disaster would strike. Probability is a formal way to measure the chance of these uncertain events.

As these examples suggest, probabilities are usually expressed in terms of percentages or fractions (percentages are the result of fractions multiplied by 100). Therefore, the probability of occurrence of some event is measured by a number between zero and one. The more likely the event, the closer it is to one and the more unlikely the event, the closer it is to zero. An event that cannot occur has a probability of zero, and an event that is certain to occur has a probability of one.

The Two Views of Probability

1. Objective probability – the view that the likelihood of the outcome of any event is an *objective* phenomenon derived from *objective* processes. This view of probability can be further classified into classical (or *a priori*) probability and the relative frequency (or a *posteriori*) probability.

Classical probability

This treatment of probability originated from the work of two mathematicians, Pascal and Fermat, and dates back to the 17th century. Much of this theory culminated out of attempts to solve problems regarding games of chance, such as those involving cards or the rolling of dice. The principles involved in classical probability are very well illustrated by examples from games of chance. For example, if a fair six-sided die is rolled, the probability that a 1 will be observed is equal to 1/6 and is the same for the other five faces. If a card is drawn at random from a well-shuffled deck of standard playing cards, the probability of drawing a spade is 13/52 or ¼. In this view of probability, probabilities are calculated by the process of abstract reasoning. Rolling a die or drawing cards from a deck is not required to compute for these probabilities. In the rolling of the die, there is an *equal likelihood* of observing any of the six sides if there is no reason to favor any one side. Likewise, if there is no reason to favor the drawing of a specific card from a deck of cards, it can be said that each of the 52 cards has an *equal likelihood* of being drawn. Probability in the classical sense can be defined as follows:

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

If *P*(*E*) is read as "the probability of *E*," this definition can be expressed as

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

Relative frequency probability

Also known as the frequentist approach to probability, this relies on the repeatability of some process and the ability to count the number of repetitions, as well as the number of instances that some event of interest takes place. In this context, the probability of observing some characteristic, *E*, of an event can be defined as follows:

> If an event can occur in *N* mutually exclusive and equally likely ways, and if *m* of these possess a characteristic, *E*, the probability of occurrence of *E* is approximately *m*/*N*.

In compact form, this definition can be expressed as

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

For example, we are interested in knowing the probability of some event in some process. This event could be the number of times a head is obtained in the process of tossing a coin. Suppose that we toss the coin many times, where we are careful to toss the coin in the same manner each time the process is repeated. Suppose this experiment was repeated 50 times and the following results were recorded:

Table 5-1. The result of 50 coin tosses

Toss No.	Result	Toss No.	Result	Toss No.	Result	Toss No.	Result	Toss No.	Result
1	H	11	T	21	T	31	T	41	T
2	T	12	H	22	H	32	T	42	H
3	H	13	T	23	H	33	H	43	T
4	T	14	H	24	T	34	H	44	H
5	H	15	T	25	H	35	T	45	T
6	H	16	H	26	H	36	H	46	H
7	T	17	H	27	T	37	H	47	T
8	H	18	T	28	T	38	T	48	H
9	H	19	T	29	H	39	T	49	H
10	T	20	T	30	H	40	H	50	H

To approximate the probability of obtaining a head or *P*(*H*) during a coin toss, we can count the number of heads (H) obtained in the trials (27) and divide by the total number of trials (50). That is, the probability of observing a head is roughly the relative frequency of heads observed in the experiment.

$$P(H) \approx \frac{m}{n} \approx \frac{\#\, of\, heads}{\#\, of\, trials} \approx \frac{27}{50} \approx 0.54$$

Comments about the frequentist definition of probability:

a. **The observed relative frequency is only an *approximation* to the true probability of an event. However, as the number of trials is increased, one might expect the relative frequency to become a better approximation of the true probability. If the coin was tossed 100 times, 200 times, 300 times, and so on, we would observe that the proportion of heads observed would become closer and closer to the true probability of 0.50. A controversial claim of this approach is that in the long run, as the number of trials approaches infinity, the relative frequency will converge *exactly* to the true probability:**

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

b. **This view of probability depends on the important assumption that the process or experiment can be repeated many times under similar circumstances. In the case where this assumption does not hold true, the subjective interpretation of probability is useful.**

2. **Subjective probability – also called the *personalistic* concept of probability, this view holds that probability measures the confidence that a particular individual has in the truth of a particular proposition or occurrence of event. Unlike objective probability, this view does not rely on the repeatability of any process. In fact, one may evaluate the probability of an event that can only happen once by applying this concept of probability. For example, you are likely interested that you will get a grade of 1.00 in this course. Most likely, you will take this course only one time; even if you retake this course next semester, you would not be taking it in the same conditions as this semester. You will have a different instructor, a different set of quizzes and exams, and possibly different work conditions. In the case where the process occurs only once, how do we view probabilities? In the subjective view of probability, a person assigns a number to this event which reflects his or her personal belief in the likelihood of this event occurring. If you are doing well in this course and you think that a grade of 1.00 is a certainty, then you would assign a probability of 1 to this event. Meanwhile, if you are experiencing difficulties in this course, you might think that getting a grade of 1.00 is close to impossible and so you would assign a probability close to 0. What if you are not sure about the grade that you will get? In this case, you would assign a number to this event between 0 and 1.**

Comments about this view of probability:

a. Subjective probability reflects a person's opinion about the likelihood of an event. If the event of interest is "you will get a 1.00 in this class", then your opinion about the likelihood of this event is probably different from your instructor's or your classmate's view about this event. Subjective probabilities are *personal* and they will differ between people.

b. Can I assign any number to events? The numbers you assign must be proper probabilities. That is, they must satisfy some basic rules that all probabilities obey. In addition, they should reflect your opinion about the likelihood of the events.

c. Assigning probabilities to events is not an easy task, especially when you are uncertain whether the event will occur or not. However, comparing the likelihoods of different events can be used as a guide in assigning subjective probabilities in a process called a *calibration experiment*.

Events and Elementary Events

- An event is the basic element to which probability can be applied. It is the result of an observation or experiment, or the description of some potential outcome.
- Elementary events are the building blocks of a probability model. They are events that cannot be broken down or decomposed into smaller sets of events.

For example, we roll two dice at the same time. Assume that the two dice are fair (they do not favor any face or number) and are independent of one another (that is, the outcome in one die would not affect the outcome in the other). We sum the two faces and are interested in the event that the faces add up to 7. For each die there are 6 faces numbered 1 to 6 with dots. Each face is assumed to have an equal 1/6 chance of landing up. In this case, there are 36 equally likely elementary events or outcomes for a pair of dice. These elementary events are denoted by pairs, such as {2, 3}, which denotes a roll of 2 on one die and 3 on the other. The 36 elementary events are: {1, 1}, {1, 2}, {1, 3}, {1, 4}, {1, 5}, {1, 6}, {2, 1}, {2, 2}, {2, 3}, {2, 4}, {2, 5}, {2, 6}, {3, 1}, {3, 2}, {3, 3}, {3, 4}, {3, 5}, {3, 6}, { 4, 1}, {4, 2}, {4, 3}, {4, 4}, {4, 5}, {4, 6}, {5,1}, {5, 2}, {5, 3},{5, 4},{5, 5}, {5, 6}, {6, 1}, {6, 2}, {6, 3}, {6, 4}, {6, 5}, and {6, 6}

These 36 elementary events constitute the sample space (commonly denoted as S, Ω, or U) which is the set of all possible outcomes for the experiment. Meanwhile, the sample space for a coin flip is $S = \{H, T\}$. The probability of an event E is determined by first defining the set of all possible elementary events, associating a probability with each elementary event, and then summing the probabilities of all elementary events that imply the occurrence of E. The elementary events are *distinct* and *mutually exclusive*.

The term mutually exclusive means that for elementary events E_1 and E_2, if E_1 happens then E_2 cannot happen and vice versa. For example, you can pass this course but not fail at the same time and vice versa. This property is necessary to sum probabilities, as we will discuss later.

In this example, seven occurs if we have {1, 6}, {2, 5}, {3, 4}, {4, 3}, {5, 2}, or {6, 1}. That is, the probability of observing a sum of seven is $6/36 = 1/6 \approx 0.167$.

Elementary Properties of Probability

1. Given some process (or experiment) with *n* mutually exclusive outcomes (called events), $E_1, E_2, \ldots, E_n$, the probability of any event E_i is assigned a nonnegative number. That is,

$$P(E_i) \geq 0$$

In other words, all events must have a probability greater than or equal to zero, a reasonable condition in view of the difficulty of thinking of negative probability. A key concept in the statement of this property is the concept of *mutually exclusive* outcomes. Two events are said to be mutually exclusive if they cannot occur simultaneously.

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

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

This is the property of *exhaustiveness* and denotes the fact that the observer of a probabilistic process must consider all possible events, and when all are taken together, their total probability is 1. The requirement that the events be mutually exclusive is specifying that the events $E_1, E_2, \ldots, E_3$ do not overlap.

3. Consider any two mutually exclusive events, E_i and E_j. The probability of the occurrence of either E_i or E_j is equal to the sum of their individual probabilities.

$$P(E_i \text{ or } E_j) = P(E_i) + P(E_j)$$

Suppose the two events were not mutually exclusive; that is, suppose they could occur at the same time. In attempting to compute for the probability of the occurrence of either E_i or E_j the problem of overlapping would be discovered, and the process could become rather complicated.

Operations with Events

- Intersection – the intersection between two events A and B, denoted as A ∩ B, is defined as the event "both A and B".

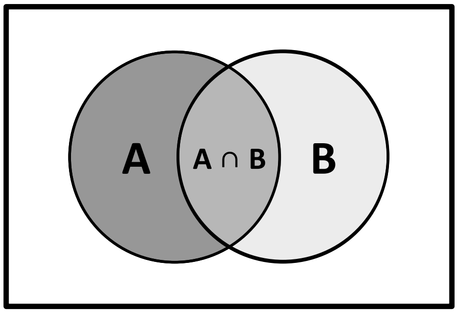


Figure 5-1. The intersection of A and B is represented only by the overlap in both sets, A ∩ B

- **Union – The union of two events A and B, denoted as A ∪ B, is defined as the event "either A or B".**

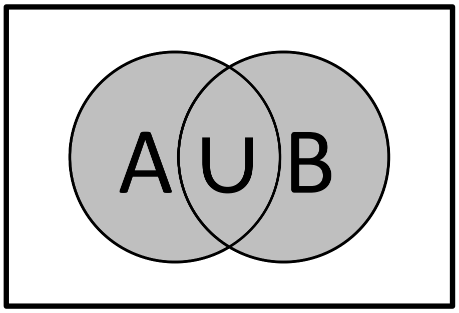


Figure 5-2. The union of A and B is represented by the combined areas of both sets, A ∩ B

- **Complement – The complement of an event A, denoted as A^c or $\overline{A}$, is defined as the event "not A"**

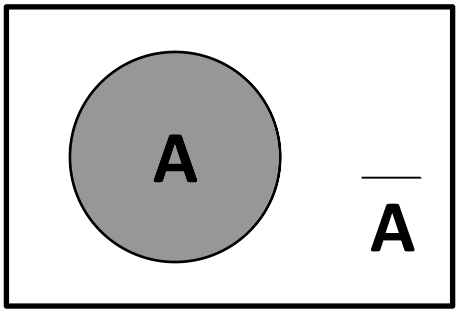


Figure 5-3. The complement of A is represented by the area outside the shaded circle.

***Sample Problem*: Calculating the probability of an event**

In an article in *The American Journal of Drug and Alcohol Abuse*, Erickson and Murray state that women have been identified as a group at particular risk for cocaine addiction and that it has been suggested that their problems with cocaine are greater than those of men. Based on their review of scientific literature and their analysis of the results of an original research study, the authors argue that there is no evidence that women's cocaine use exceeds that of men, that women's rate of cocaine use are growing faster than men's, or that female cocaine users experience more problems than male cocaine users.

The subjects in the study by Erickson and Murray consisted of a sample of 75 men and 36 women. The authors state that the subjects are a fairly representative sample of 'typical' adult users who were neither in treatment nor in jail. Table 4.1 shows the lifetime frequency of cocaine use and the gender of these subjects. Suppose we pick a person at random from this sample. What is the probability that this person will be a male?

SOURCE: Patricia G. Erickson and Glenn F. Murray, "Sex Differences in Cocaine Use and Experiences: A Double Standard?" *American Journal of Drug and Alcohol Abuse*, *15* (1989), 135-132 as printed in *Biostatistics*: *A Foundation for Analysis in The Health Sciences* 6e by Wayne W. Daniel (1995)

Table 5-2. Frequency of Cocaine Users by Gender Among Adult Cocaine Users

Lifetime Frequency of Cocaine Use	Male (*M*)	Female (*F*)	Total
1-19 times (*A*)	32	7	39
20-99 times (*B*)	18	20	38
100 + times (*C*)	25	9	34
Total	75	36	111

Solution:

We assume that male and female are mutually exclusive categories and the likelihood of selecting any one person is equal to the likelihood of selecting any other person. We define the desired probability as the number of subjects with the characteristic of interest (male) divided by the total number of subjects. We may write the result in probability notation as follows:

$P(M)$ = Number of males/Total number of subjects
$= 75/111 = 0.6757$

Conditional Probability

In some instances, the set of "all possible outcomes" may comprise a subset of the total group. In other words, the size of the group of interest may be diminished by conditions not applicable to the total group. When probabilities are computed with a *subset of the total group as the denominator*, the result is a conditional probability.

We may think of the probability calculated in the previous example as an unconditional probability since the size of the total group served as the denominator. No conditions were enforced to restrict the size of the denominator. This probability can also be thought of as a *marginal probability* since one of the marginal totals was used as the numerator. We may illustrate the concept of conditional probability by referring again to Table 6-2.

Sample Problem:

Suppose we pick a subject randomly from the 111 subjects and find that he is a male (*M*). What is the probability that he will be one who has used cocaine 100 times or more during his lifetime (*C*)?

Solution:

In this particular problem, the total number of subjects is not of concern anymore, since, with the selection of a male, the females are eliminated. The desired probability can then be defined as follows: Given that the selected subject is a (*M*), what is the probability that the subject used cocaine 100 times or more (*C*) during his lifetime? This is a conditional probability written as $P(C|M)$ in which the vertical line is read "given". In this conditional probability, the 75 males become the denominator, and 25, the number of males who have used cocaine 100 times or more during their lifetime, becomes the numerator. Therefore, our desired probability is

$$P(C|M) = 25/75 = 0.3333$$

Joint Probability

- The probability of two events in conjunction
- The probability that a subject picked at random from a group of subjects possesses two characteristics at the same time

Sample Problem:

What is the probability that a person picked at random from the 111 subjects will be a male (*M*) and a person who has used cocaine 100 times or more during his lifetime (*C*)? [Refer to Table 6-2]

Solution:

The probability we are looking for may be expressed in symbolic notation as $P(M \cap C)$ in which the symbol ∩ is read either as "intersection" or "and". The expression $M \cap C$ denotes the joint occurrence of conditions *M* and *C*. The number of subjects who satisfy both of the desired conditions is found in Table 6-2 at the intersection of the column labeled *M* and the row labeled *C* and is equivalent to 25. Since the selection will be made from the total set of subjects, the denominator is 111. Therefore, this joint probability may be written as

$$P(M \cap C) = 25/111 = 0.2252$$

Multiplication Rule of Probability

- When two independent events occur simultaneously, the combined probability of the two outcomes is equal to the product of their individual probabilities of occurrence.
- The probability that two events A and B will both occur is equal to the probability of A multiplied by the probability of B given that A has already occurred.

In addition, the multiplication rule of probability is a relationship where a joint probability may be computed as the product of an appropriate marginal probability and an appropriate conditional probability. This relationship can be illustrated by the following example:

Sample Problem:

We wish to compute the joint probability of male (*M*) and a lifetime frequency of cocaine use of 100 times or more (*C*). [Refer once again to Table 6-2]

Solution:

The probability we are looking for is $P(M \cap C)$. We have already calculated a marginal probability, $P(M) = 75/111 = 0.6757$, and a conditional probability, $P(C|M) = 25/75 = 0.3333$. Coincidentally, these are the appropriate marginal and conditional probabilities for calculating the desired joint probability. We can now calculate $P(M \cap C) = P(M)\ P(C|M) = (0.6757)(0.3333) = 0.2252$. Note that as expected, this is the same result we calculated earlier for $P(M \cap C)$.

The multiplication rule can be stated in general terms as follows:

- **For any two events *A* and *B*,**

$$P(A \cap B) = P(B)P(A|B), \text{ if } P(B) \neq 0$$

- **For the same two events A and B, the multiplication rule may also be written as**

$$P(A \cap B) = P(\mathrm{A})P(B|A), \text{ if } P(A) \neq 0$$

Through algebraic manipulation, these equations can be used to find any one of the three probabilities if the other two are known. For example, we may find the conditional probability *P*(*A*|*B*) by dividing *P*(*A*∩*B*) by *P*(*B*). This relationship allows us to formally define conditional probability as follows:

> **The *conditional probability* of *A* given *B* is equal to the probability of *A*∩*B* divided by the probability of *B*, provided the probability of *B* is not zero.**

That is,

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, P(B) \neq 0$$

Addition Rule of Probability

The third property of probability previously stated that the probability of the occurrence of either one or the other of two mutually exclusive events is equal to the sum of their individual probabilities. For example, suppose that we pick a person at random from the 111 subjects represented in Table 6-2. What is the probability that this person will be a male (*M*) or a female (*F*)? This probability can be stated as *P*(*M* ∪ *F*) where the symbol ∪ is read either as "union" or "or". Since the two genders are mutually exclusive, *P*(*M* ∪ *F*) = *P*(*M*) + *P*(*F*) = (75/111) + (36/111) = 0.6757 + 0.3243 = 1.

What if two events are not mutually exclusive? This case is covered by what is known as the addition rule, which may be stated as follows:

> Given two events *A* and *B*, the probability that event *A*, or event *B*, or both occur is equal to the probability that event *A* occurs, plus the probability that event *B* occurs, minus the probability that the events occur simultaneously.

The addition rule may be written as

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

***Sample Problem*:**

If we select a person at random from the 111 subjects represented in Table 6-2, what is the probability that this person will be a male (*M*) or will have used cocaine 100 times or more during his lifetime (*C*) or both? [Refer to Table 6-2]

***Solution*:**

The probability we are looking for is $P(M \cup C)$. Based on the addition rule, this probability may be expressed as $P(M \cup C) = P(M) + P(C) - P(M \cap C)$. We previously calculated that $P(M) = 75/111 = 0.6757$ and $P(M \cap C) = 25/111 = 0.2252$. From the information in Table 6-2, we can calculate $P(C) = 34/111 = 0.3063$. Substituting these results into the equation for $P(M \cup C)$, we have $P(M \cup C) = 0.6757 + 0.3063 - 0.2252 = 0.7568$.

Note that the 25 subjects who are *both* male *and* have used cocaine 100 times or more are included in the 75 who are male as well as the 34 who have used cocaine 100 times or more. Since, in calculating the probability, these 25 have been added in the numerator twice, they have to be subtracted out once to cancel the effect of duplication or overlapping.

Independent Events

Suppose that we are told that event *B* has occurred, but this fact has no effect on the probability of *A*. That is, the probability of event *A* is the same regardless of whether or not *B* occurs. In this situation, $P(A|B) = P(A)$. In cases such as these, we say that *A* and *B* are independent events.

- **Two events are said to be independent, if the outcome of one event has no effect on the occurrence of the other. If A and B are independent events,**
$$P(A|B) = P(A) \text{ and } P(B|A) = P(B)$$
- **In this special case of independence, the multiplicative rule of probability may be written as**
$$P(A \cap B) = P(A)P(B)$$

Therefore, we can observe that if two events are independent, the probability of their joint occurrence is equal to the probabilities of their individual occurrences.

Note that when two events with nonzero probabilities are independent, each of the following statements is true:

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

Two events are not independent unless all these statements are true.

***Sample Problem*:**

In a certain high school class, consisting of 60 girls and 40 boys, it is observed that 24 girls and 16 boys wear eyeglasses. If a student is picked at random from this class, the probability that the student wears eyeglasses $P(E)$ is 40/100 or 0.4.

a. **What is the probability that a student picked at random wears eyeglasses, given that the student is a boy?**
b. **What is the probability of the joint occurrence of the events of wearing eyeglasses and being a boy?**

Solution:

a.

$$P(E|B) = \frac{P(E \cap B)}{P(B)} = \frac{16/100}{40/100} = \mathbf{0.4}$$

We may also demonstrate that the event of wearing eyeglasses, *E*, and *not* being a boy, $\overline{\mathrm{B}}$, are also independent as follows:

$$P(E|\bar{B}) = \frac{P(E \cap \bar{B})}{P(\bar{B})} = \frac{24/100}{60/100} = \mathbf{0.4}$$

b. **$P(E \cap B) = P(B)P(E|B) = P(B)P(E)$**
= (40/100)(40/100)
= 0.16

Permutations and Combinations

In the previous sections, we discussed a common method for finding probabilities: we calculated the probability of an event by counting the number of possible ways an event can occur and dividing the resulting number by the total number of equally likely elementary outcomes [*P*(*E*) = *m*/*N*]. Because we used simple examples, such as the rolling of two dice which has 36 different possibilities at most, we did not encounter any problem applying this formula.

However, there are situations when the number of ways that an event can occur is so large that complete and exhaustive enumeration becomes tedious and impractical. The combinatorial methods discussed in this section will facilitate the calculation of the numerator and denominator for the probability of interest. *Combinatorics*, also called *combinatorial mathematics*, is the field of mathematics involved with the problems of selection, arrangement, and operation within a finite or discrete system. Its objective is to apply methods that allow you to count without counting. Thus, one of the fundamental problems of combinatorics is to determine the number of possible configurations of objects or events of a given type.

Again, let us consider the experiment where we roll dice. On any roll of a die, there are six elementary outcomes. Suppose we roll the die three times so that each roll is independent of the other rolls. We want to know how many ways we can roll a 4 or less on all three rolls of the die without repeating a number.

Direct enumeration is difficult and tedious since there are a total of 6 x 6 x 6 = 216 possible outcomes. Moreover, the number of successful outcomes may not be obvious. There is a shortcut solution that becomes even more vital as the space of possible outcomes, and possible successful outcomes, become even larger than in this example.

Thus far, our problem is not well defined. We must also specify whether or not the order of distinct numbers maters. When *order matters* we are dealing with permutations. When order does not matter we are dealing with combinations.

First, let us consider the case in which order is important; therefore, we will be determining the number of permutations. If order matters, then the triple {4, 3, 2} is a successful outcome but differs from the triple {4, 2, 3} because order matters. In fact, the triples {4, 3, 2}, {4, 2, 3}, {3, 4, 2}, {3, 2, 4}, {2, 4, 3}, and {2, 3, 4} are six distinct outcomes when order matters but count only as one outcome when order does not matter since they all correspond to an outcome in which the three numbers 2, 3, and 4 each occur once.

A successful outcome is when a 4 or lower number is rolled. In this case, there are four objects – the numbers 1, 2, 3, and 4 – to choose from because a choice of 5 or 6 on any trial leads to a failed outcome. Because there are only three rolls of the die, and a successful roll requires a different number on each trial, we want to determine the number of ways of selecting three objects out of four when order matters. This type of selection is called the number of possible permutations for selecting three objects out of four.

In general, let *r* be the number of objects to choose from and *n* the number of objects available. The number of permutations of *r* objects chosen out of *n* can be determined using the following formula:

$$P(n,r) = \frac{n!}{(n-r)!}$$

The symbol "!" represents the function called the factorial. The notation *n*! (read as, *n* factorial) means by definition the product:

$$n! = (n)(n-1)(n-2)(n-3)\ldots(3)(2)(1)$$

For example, 3! = (3)(2)(1) = 6; while 4! = (4)(3)(2)(1) = 24; and 5! (5)(4)(3)(2)(1) = 120. Note that 0! exists and is equal to 1.

The problem of determining the number of ways we can roll a 4 or less on three rolls of a die without repeating any number is solved as:

$$P(4{,}3) = \frac{4!}{(4-3)!} = \frac{24}{1} = 24$$

Now let us consider combinations. In combinations, only distinct subsets but not their order, are considered. In the example of distinct outcomes of three rolls of the die where success means three distinct numbers less than 5 without regard to order, the triplets {2, 3, 4}, {2, 4, 3}, {3, 2, 4}, {3, 4, 2}, {4, 3, 2}, and {4, 2, 3} differ only in order and not in the objects included.

Observe that for each different set of three distinct numbers, the common number of permutations is always 6. For example, the 1, 2, and 3 contains the six triplets {1, 2, 3}, {1, 3, 2}, {2, 1, 3}, {2, 3, 1}, {3, 1, 2}, and {3, 2, 1}. Notice that the number six occurs because it is equal to *P*(3,3) = 3!/0! = 6.

The formula for the number of combination of *r* objects taken out of *n* is:

$$C(n,r) = \frac{n!}{(n-r)!\,r!}$$

In our example of three rolls of the die resulting to three distinct numbers less than 5, the number of combinations for choosing 3 objects out of 4 is:

$$C(4,3) = \frac{4!}{(4-3)!\,3!} = \frac{24}{1(6)} = 4$$

These four distinct combinations are the sets consisting of (1) 1, 2, and 3; (2) 1, 2, and 4; (3) 1, 3, and 4; and (4) 2, 3, and 4.

LABORATORY EXERCISE 6
Basic Concepts in Probability (38 points)

NAME:____ DATE:______________

SECTION:____________________________ INSTRUCTOR:________________________

1. **Why is it that science is not always certain? How do scientists cope with uncertainty? (3 points)**

____

____

____

____

2. **One hundred married women were asked to specify which type of birth control method they preferred. The following table shows the 100 responses cross-classified by educational level of the respondent. (7 points)**

Table 5-3. Birth control method preferred by women according to educational level

Birth Control Method	Educational Level			
	High School (*A*)	College (*B*)	Graduate School (*C*)	Total
S	15	8	7	30
T	3	7	20	30
V	5	5	15	25
W	10	3	2	15
Total	33	23	44	100

Specify the number of members of each of the following sets:

a. S ________

b. $V \cup C$ ________

c. A ________

d. $\overline{W}$ ________

e. $\overline{B}$ ________

f. $T \cap B$ ________

g. $\overline{T \cap C}$ ________

Solve the following problems on the space provided. Write the pertinent probability expression(s) for each one.

3. **Laveist and Nuru-Jeter conducted a study to determine if doctor-patient race concordance was associated with greater satisfaction with care. Toward the end, they collected a national sample of African-American, Caucasian, Hispanic, and Asian-American respondents. The following table classifies the race of the subjects as well as the race of their physician:**

Table 5-4. Doctor-patient race concordance of a national sample of 2720 respondents

	Patient's Race				
Physician's Race	**Caucasian**	**African-American**	**Hispanic**	**Asian-American**	**Total**
White	779	436	406	175	1796
African-American	14	162	15	5	196
Hispanic	19	17	128	2	166
Asian/Pacific-Islander	68	75	71	203	417
Other	30	55	56	4	145
Total	910	745	676	389	2720

Source: Thomas A. Laveist and Amani Nuru-Jeter, "Is Doctor-patient Race Corcondance Associated with Greater Satisfaction with Care?" *Journal of Health and Science Behavior*, 43 (2002), 296-306 as printed in *Biostatistics: A Foundation for Analysis in the Health Sciences* by Wayne W. Daniel (1995)

a. **What is the probability that a randomly selected subject will have an Asian/Pacific-Islander physician? (2 points)**

b. **What is the probability that an African-American subject will have an African-American physician? (3 points)**

c. **What is the probability that a randomly selected subject in the study will be Asian-American and have an Asian/Pacific-Islander physician? (2 points)**

d. **What is the probability that a subject chosen at random will be Hispanic or have a Hispanic physician? (3 points)**

e. **Use the concept of complementary events to find the probability that a subject chosen at random in the study does not have a white physician. (2 points)**

4. **The probability is 0.6 that a patient selected at random from the current residents of a certain hospital will be a male. The probability that the patient will be a male who is in for surgery is 0.2. A patient randomly selected from current residents is found to be a male; what is the probability that the patient is in the hospital for surgery? (3 points)**

5. **The probability that a person selected at random from a population will exhibit the classic symptom of a certain disease is 0.2, and the probability that a person selected at random has the disease is 0.23. The probability that a person who has the symptom also has the disease is 0.18. A person selected at random from the population does not have the symptom; what is the probability that the person has the disease? (5 points)**

6. Systematically enumerate all 24 permutations of rolling a die three times and observing a number less than 5 on each separate occasion without any number repeating. Explain the method that you used to systematically enumerate the permutations. (5 points)

(1) ________

(2) ________

(3) ________

(4) ________

(5) ________

(6) ________

(7) ________

(8) ________

(9) ________

(10) ________

(11) ________

(12) ________

(13) ________

(14) ________

(15) ________

(16) ________

(17) ________

(18) ________

(19) ________

(20) ________

(21) ________

(22) ________

(23) ________

(24) ________

Explain the method you used:

____

____

____

____

7. How many ways can a researcher select a distinct set of 10 participants from a pool of 50 volunteers? Show your solution. (3 points)

6 Sampling Methods

The term population refers to a collection of people or objects that share common observable characteristics. For example, a population could be all the inhabitants of your city, all of the students enrolled in a certain university, or all of the people who are afflicted by a particular disease (e.g., all men diagnosed with prostate cancer during the last five years). Meanwhile, a sample is a subset of the population. One approach of statistics, called *inferential statistics* relies upon the use of samples in making inferences about populations. Sampling is the act of studying a subset of the population to make inferences on the whole.

The Rationale

If a researcher wishes to gather information regarding a population through questioning or testing, he/she has two basic options:

1. Every member of the population can be questioned or tested (i.e., a census); or
2. A sample can be drawn from the population of interest where only selected members of the population are questioned or tested.

Understandably, choosing the first option means that the processes of contacting, questioning, and gathering information from a large population, such as all of the households within Metro Manila, is extremely expensive, challenging, and time-consuming. However, a properly designed sampling method provides a reliable way of inferring information about a population without examining each of its members or elements.

Another advantage of sampling is that it is more accurate than a census of the entire population. The smaller sampling operation allows the application of more rigorous controls which guarantees better accuracy. These rigorous controls give researchers the ability to reduce nonsampling errors like interviewer bias and mistakes, nonresponse problems and attrition, questionnaire design flaws, as well as data processing and analysis errors.

These nonsampling errors are partly reduced through pretesting which gives the researcher information on whether or not his/her tools of data gathering are accurately and appropriately designed by administering the test to a small subset of respondents. When conducting a census, pretesting cannot be done without risking possible contamination of some of the respondents. In addition, the scope and detail of information that can be asked in a sample is greater that in a census due to cost and time limitations under which most researchers are operating. Administering a relatively long and difficult survey to a sample is easier than administering a concise questionnaire to the entire population. However, be mindful that not all samples are accurate or the appropriate vehicle for obtaining information or testing a hypothesis about a population. In addition, there are cases where sampling is the only possible method for destructive procedures. The following sections in this chapter will discuss the advantages and disadvantages of various sampling procedures.

Sampling Terminologies

- Population – the totality of individuals or objects of interest → *parameter* is measured
- Target population – a subset of the population where representative information is desired and to which inferences will be made
- Sampling population – a subset of the population where the *sample* is actually taken → statistic is measured
    - Sampling unit – the units which are chosen in selecting the sample, and may be comprised of a non-overlapping collection of elements.
    - Elementary unit/element – the sample where observations or data will be acquired
- Sampling frame – listing of all sampling units from which a sample will be drawn or the collection of all the sampling units
- Sampling error – the difference between the population value (parameter) and the estimate of this value based on the different samples (statistics)

Example:

- Population: Filipino college students
- Target population: state university students
- Sampling population: UP Manila students
- Elementary unit: BS Biology majors
- Sampling frame: OCS list of enrolled students

Criteria of a good sampling design

- Representative of the population
- Adequate sample size
- Practical and feasible
- Economical and efficient

Types of Sampling Designs

- Non-Probability Sampling
    - Any sampling scheme in which the probability of a population element being chosen is unknown
    - Data normally analyzed using non-parametric statistical tests (normal distribution <u>not</u> assumed)
    - Appropriate when the researcher has no intention of generalizing beyond the sample
    - Advantages:
        - More easily administered
        - Samples tend to be less complicated and less time consuming
        - May occasionally serve as the only possible means of getting a sample (especially for "hidden populations", e.g. MSM, drug users, etc.)

- **Disadvantages:**
    - **Does not allow the study's findings to be generalized from the sample to the population**
    - **When discussing the results of a nonprobability sample, the researcher must limit his/her findings to the persons or elements sampled**
    - **More likely to produce biased results**
    - **No defined rules to compute for estimates**
    - **Cannot compute the reliability of estimates**

- **Probability Sampling**
    - **Any sampling scheme wherein each population element has a known non-zero chance of being included in the sample**
    - **Analyzed using parametric statistical tests (normal distribution assumed)**
    - **Uses random selection procedures to ensure that each unit of the sample is chosen on the basis of chance where all units of the study population should have an equal or at least a known chance of being included in the sample.**
    - **Requires that a listing of all study units (sampling frame) exists or can be compiled.**
    - **Advantages:**
        - **Probability samples are the only type of samples where the results can be generalized from the sample to the population**
        - **Allows the researcher to calculate the precision of the estimate as well as the sampling error**
    - **Disadvantages**
        - **More difficult and costly to conduct**

Non-Probability Sampling Methods

- **Judgmental/Purposive – a representative sample of the population is selected based on a researcher's "expert" judgment. Prior knowledge and research skill are used in choosing the respondents or elements to be sampled.**
    - **selection of patients in a clinical trial by a medical specialist**
    - **choice of participants based on a pre-test questionnaire or focus-group discussion (FGD)**

- **Haphazard/Accidental – also known as convenience sampling, is a sampling design where the samples are selected by an arbitrary method that is easy to carry out. Convenience samples have been used when it is very difficult or impossible to draw a random sample. Studies based on convenience samples have results that are descriptive and may be used to recommend future research but should not be used to draw inferences about the population under study.**
    - **friends as a sample of college students**
    - **ambush interviews of random people in the area**
    - **households of relatives or friends as sampling sites**

- **Quota – dividing the population into predetermined classes then obtaining *haphazard* samples of a fixed size (quota) within each class. As each class fills or reaches its quota, additional respondents that would have fallen into these classes are rejected or excluded from the results.**

- Interviewing the first 10 males and females in a public restroom regarding shampoo preference
- Obtaining the RH bill opinion of 20 people per region in a municipality
- A researcher desires to obtain a certain number or respondents from different income categories. Generally, researchers have no knowledge of the incomes of the people they are sampling until they ask about income. Thus, the researcher is unable to subdivide the population from which the sample is drawn into mutually exclusive income categories prior to drawing the sample. In this type of sample, bias can be introduced when the respondents who are rejected (because the class to which they belong has already reached its quota) differ from those who are used.

Probability Sampling Methods

- Simple Random Sampling – the most basic type of sampling design wherein every element in the population has an equal chance of being included in the sample.

  Simple random sampling is carried out by following these steps:

  - Prepare an exhaustive list (sampling frame) of all members of the population of interest.
  - Decide on the size of the sample
  - Select the necessary number of sampling units, using a "lottery" method, a table of random numbers, or the RAN function of a calculator.

  - Advantages:
    - Simple design, easy to analyze
    - Random sampling guarantees unbiased estimates of population parameters. Unbiased means that the average of the sample estimates over all possible samples is equal to the population parameter.
  - Disadvantages:
    - Not cost efficient because elementary units may be too widespread
    - Requires a sampling frame or listing of all elementary units of the population which might be costly and tedious to prepare
    - Does not guarantee balance in any particular sample drawn at random. Even though the probability is very small, it is possible that nonrepresentative samples can be drawn from the population. For example, suppose a catheter ablation treatment is known to have a 95% chance of success. Thus, we can expect only about one failure in a sample size of 20. However, even though it is highly unlikely, it is possible that we could select a random sample of 20 individuals with the result that all 20 individuals have failed ablation procedures.

- **Systematic Sampling – in this method, the researcher selects samples at regular intervals (every 1st, 2nd, 3rd,… or kth). The sampling interval is computed as $k = N/n$ where k is the sampling interval, N is the population size, and n is the desired sample size.**
    - **Example:** $N = 20; n = 10$
        - **Sampling interval = 20/10 = 2 or every 2nd unit**
        - **Randomly draw a number from 1 to 20 = 3 for instance**
        - **Samples are selected every 2nd unit (example: households 3, 5, 7… and so on) until $n = 10$ is reached.**
    - **Advantages:**
        - **Tend to be easier to draw and execute compared to simple random sampling since the researcher does not have to jump forward and backward through the sampling frame to draw the members to be sampled.**
        - **Compared to simple random sampling, a systematic sample may spread the members selected for measurement more evenly across the entire population. Thus, in some cases, systematic sampling may have better representativeness and be more precise.**
        - **It can allow the researcher to draw a probability sample without complete prior knowledge of the sampling frame.**
    - **Disadvantages:**
        - **Can lead to difficulties when the variable of interest is periodic (with period *n*). For example, when conducting a sample of financial records, or other objects that follow a calendar schedule, setting the sampling interval to 7 would mean that all observations would fall on the same day of the week. This introduces bias into the sample since an inappropriate interval was chosen.**
        - ***Periodic* or *list effect* – a foregoing source of sampling error. For example, if we used a very long list such as a telephone directory for our sampling frame and needed to sample only a few names using a short sampling interval, there is a possibility that by accident, we select a sample from a portion of the list wherein a certain ethnic group is concentrated. Thus, the representativeness of the sample would not be very good. If the characteristics we are interested in varied considerably by ethnic group (such as average lifespan, allelic frequencies, genetic polymorphisms, etc.), our estimate of the population parameter could be very biased.**

- **Stratified Random Sampling – a modification of simple random sampling that is used when we want to guarantee that each *stratum* (subgroup) constitutes an appropriate portion or representation in the sample. It involves categorizing the members of the population into mutually exclusive and collectively exhaustive groups. Simple random sampling is then independently carried out for each group.**

  Stratified random sampling is carried out by following these steps:
    - **Define m subgroups of strata**
    - **For the i^{th} subgroup, we select a simple random sample of size n_i**
    - **Follow this procedure for each subgroup.**

  - The total sample size n is then $\sum_{i=1}^{n} n_i$. The notation Σ stands for the summation of the individual n_i's. For example, if three groups, then $\sum_{i=1}^{3} n_i = n_1 + n_2 + n_3$. In general, we have a total sample size "n" in mind.
- **Advantages:**
    - Can be used to improve the accuracy of the sample estimates when there is prior knowledge that the variability in the data is not constant across the subgroups.
    - It can enable the researcher to determine the desired level of sampling precision for each stratum.
    - It can be demonstrated by statistical theory that in many situations, stratified random sampling produces an unbiased estimate of the population mean with better precision compared to simple random sampling with the same total sample size *n*. Choosing large values of n_i for the subgroups with the largest variability and small values for the subgroups with the least variability improves the precision of the estimate.
- **Disadvantage:**
    - May require a very large n if reliable estimates for each stratum are desired

- Cluster Sampling – a method of sampling in which the element selected is a group (rather than an individual), called a *cluster*. For instance, the clusters could be city blocks. Cluster sampling is similar to stratified sampling because the population to be sampled is subdivided into mutually exclusive groups. However, in cluster sampling the groups are defined so as to maintain the heterogeneity of the population. The goal of the researcher is to establish clusters that are representative of the population as a whole, although in practice this may be difficult to achieve. Once the clusters are established, a simple random sample of the clusters is drawn and the members of the chosen clusters are sampled.
    - One-stage cluster sampling – all of the elements (members) of the clusters selected are sampled
    - Two-stage cluster sampling – a random sample of the elements of each selected cluster is drawn.

    - Advantages:
        - Can be employed in the absence of a sampling frame. For example, a researcher is interested to measure the age distribution of people residing in Metro Manila. It is easier to compile a list of residential addresses in Metro Manila than to compile a list of every person residing in Metro Manila. In this case, each address would represent a cluster of elements (people) to be sampled.
        - Since the clusters are randomly selected, the samples can be representative of the population and unbiased estimates of the population total or mean value for a particular parameter can be obtained.
        - Reduced cost of data collection
    - Disadvantage:
        - Sometimes, there is loss of precision for the estimate relative to the simple random sampling if the heterogeneity within the clusters is not similar to the heterogeneity within the population.

- **Multistage Sampling – a procedure carried out in phases involving a combination of probability sampling designs. The population is divided into sets of primary or first stage sampling units and then a random sample of secondary stage units is obtained from each of the selected units in the first stage. This type of sampling method is appropriate for very large and diverse populations where the selection of elementary units may be done in two or more stages.**

  Example: Nationwide survey of all 15 regions (stratified)
  - **1 province/region – primary sampling unit (simple random sampling)**
  - **1 urban & 1 rural barangay – secondary sampling unit (stratified random sampling)**
  - **1 cluster of 35 households – tertiary sampling unit (cluster sampling)**
  - **Choose the households per cluster – elementary unit (systematic sampling)**

  Sampling design: 4-stage, stratified, systematic, cluster, simple random sampling.
  - **Advantages:**
    - **Cost-efficient**
    - **Samples are easier to select if a sampling frame is available**
  - **Disadvantages:**
    - **May require complicated analyses**
    - **Sample size must be large enough to obtain representative estimates of the parameters**

LABORATORY EXERCISE 7
Sampling Methods (35 points)

NAME:____ **DATE:**____________

SECTION:__________________________ **INSTRUCTOR:**______________________

1. **How can bias affect a sample design? Explain by using the terms selection bias, response bias, and periodic effects (5 points)**

____

____

____

____

____

2. **How is sampling with replacement different from sampling without replacement? (2 points)**

____

____

____

____

____

3. **Why would a convenience sample of college students on vacation at Boracay not be representative of students at a particular college or university? (3 points)**

____

____

____

____

____

4. **Identify the possible sampling frame that could be used for each of the following surveys (2 points each)**

   a. **A survey among the alumni of the BS Biology program of UP Manila on the relevance of the BS Biology curriculum in their present work**

   ____

   b. **A survey to determine the percentage of CAS students who intend to go to medical school.**

   ____

c. **A survey to find out the distribution of mammals at the Manila Zoo.**

____

5. **On the next page is a map of a certain community with 20 houses. Using the following random numbers,**

1402 1185 0920 3610 0879 1304

Draw or obtain:

a. **A simple random sample of 6 households (write the household numbers that you were able to draw) (5 points).**

____

b. **A systematic sample with $k = 5$ (write the household numbers that you were able to draw) (5 points).**

____

c. **Based on the simple random sample you acquired in (a), compute for the mean age of the household heads (2 points).**

Place your computations here:

1. **How does your computed mean age compare with that of your classmate's result? (Indicate your classmate's name and her/his computed mean) Is it larger, lower or of the same value? (2 points)**

____

____

2. **How does your computed mean age compare with the mean age of the household heads for the entire population? Is it larger, lower or of the same value? (2 points)**

____

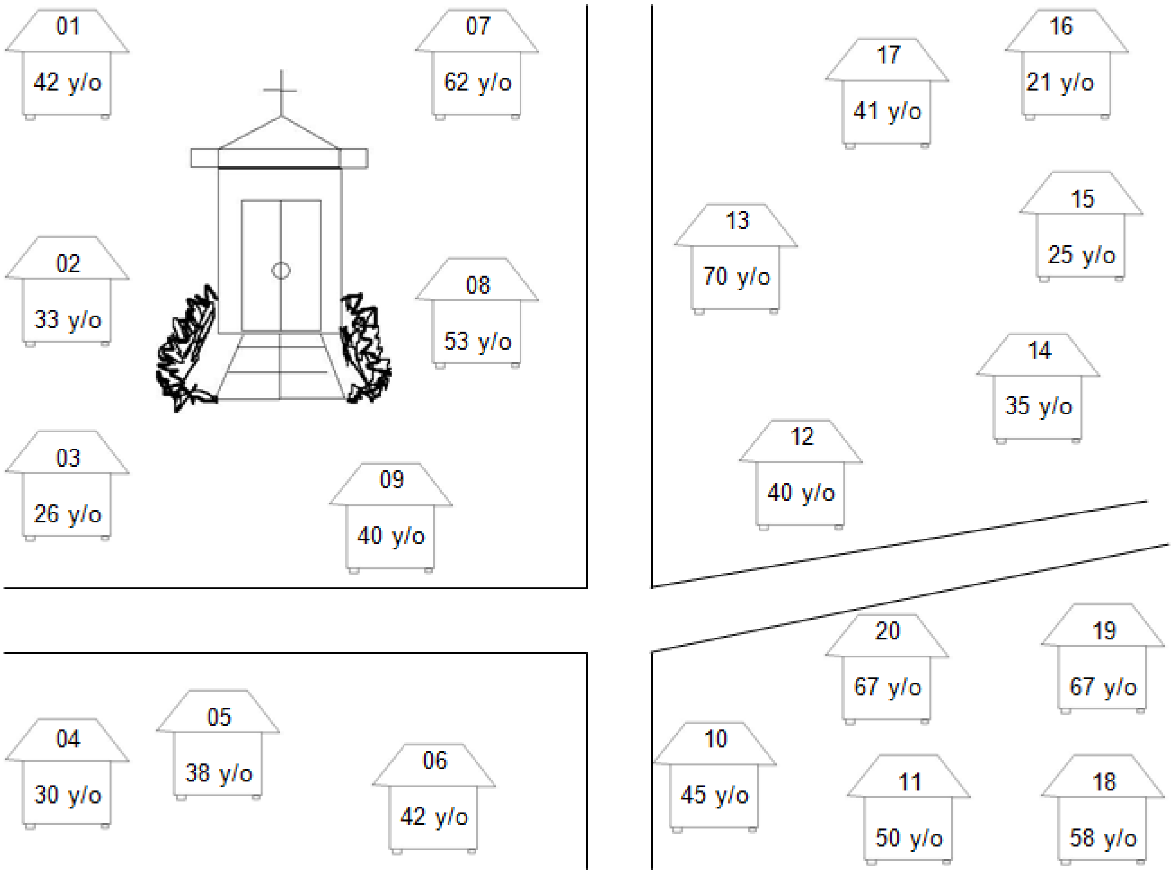
____

3. **To what do you attribute the difference between your computed sample mean, your classmate' result and that of the total population? (3 points)**

____

____

____



7 Normal Distribution

Probability Theory is an essential mathematical concept that serves as a foundation of statistics. Statistical tests which are used to make inferences related to a set of hypotheses are based on probabilistic models or distributions. A probability distribution is a function, that is, a graphical relationship between all possible outcomes of an event (*x*, domain) and the probabilities of occurrence of each outcome (*y*, codomain).

- The "event" can be considered in the light of a quantitative variable (e.g. count, weight)
- The "possible outcomes of an event" as any possible value of the quantitative variable; also called a "random variable"
- The "probabilities of occurrence" corresponds to the frequency of occurrence of the value
- If all "possible outcomes of an event" are tabulated in a frequency distribution table (where frequency/ relative frequency is the probability of a specific outcome), a corresponding histogram or frequency polygon serves as the graphical representation of the probability distribution

Introduction to Probability Distributions

There are two general kinds of probability distributions: probability mass functions and probability density functions. Without going into mathematical rules and set notations, for the purpose of this general biostatistics course:

1. Probability mass function (PMF)
   - probability distribution of values of quantitative <u>discrete</u> random variables; a histogram
   - especially useful in genetics, e.g. determining the probability of a combination of a certain number of boys and girls in a 3-child family
     - this specific PMF is a binomial distribution (i.e. only two possible outcomes: girl and boy)
     - given that the probability of having a boy is 50% and a girl is also 50%, then to calculate the different combinations for a 3-child family, binomial expansion of the equation $(boys + girls)^3 or\ (x + y)^3$ can provide a guide in calculating the probabilities of the combinations: $(x + y)^3 = x^3 + 3x^2y + 3xy^2 + y^3$
       - probability of 3 boys = $x^3 = (0.5)^3$=**0.125**
       - probability of 2 boys = $3x^2y = 3(0.5)^2(0.5) = 0.375$
       - probability of 1 boy = $3xy^2 = 0.375$
       - probability of 0 boys = $y^3 = 0.125$

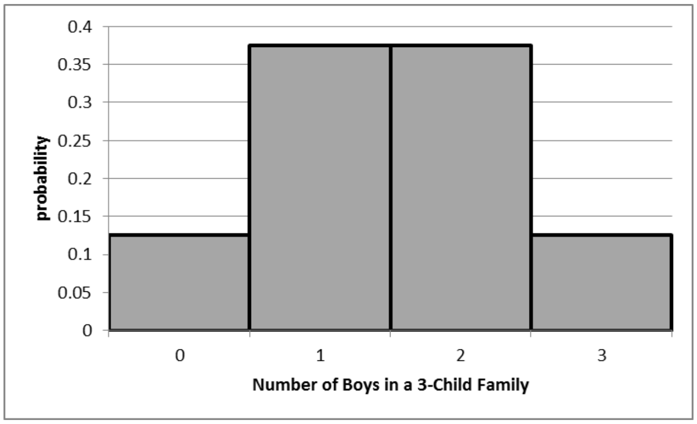


Figure 7-1. Probability Mass Function for Number of Boys in a 3-Child Family

2. **Probability density function (PDF)**
    - **probability distribution of values of quantitative continuous random variables; a curve or frequency polygon with small class widths so as to appear smooth**
    - **the probability of values of a quantitative continuous random variable correspond to the areas under a curve and where the total area under the curve is equal to 1 or the integral:** $\int_{-\infty}^{\infty} f(x)dx = 1$
    - **A specific type of PDF or curve is the normal curve or the normal distribution**

The Normal Distribution

The normal distribution is the most important probability distribution in statistics. The idea of the normal curve was first introduced by the French mathematician Abraham De Moivre in the early 1700s whose probability theory focused on the binomial distribution and its applications on his side job as a gambling consultant (i.e. coin flips in most probability theory problems).

Pierre-Simon de Laplace rediscovered the normal curve in the 1780s through his central limit theorem. An important concept of Laplace's central limit theorem states that for a random variable of interest, when a large enough sample size (n) is acquired from a population (N), the mean of all samples $(\bar{x})$ **from the population approximates the population mean** $(\mu)\ or\ (\bar{x} \approx \mu)$**. The population mean,** μ**, is at the center of the symmetrical distribution.**

In the 1830s a German mathematics and astronomy professor, Carl Friedrich Gauss, determined that the distribution of accidental errors in astronomical measurements also followed a normal curve. His Gaussian Law of Errors in the Theory of Additive Functions was studied and applied to other situations that the normal distribution was occasionally called the Gaussian distribution. However, in France, the normal distribution is mostly called the Laplacian distribution in reference to de Laplace.

Properties of the Normal Distribution

1. **Bell-shaped curve that is symmetrical at the mean and extends from** $-\infty\ to + \infty$ **(asymptotic horizontally)**
2. **The measures of central tendency (mean, median and mode) are equal**

3. $\int_{-\infty}^{\infty} f(x)dx = 1$, **that is, the total area under the curve is equal to 1**
4. **Determined by the mean, μ, and the standard deviation, σ, so that every specific pair of mean and standard deviation corresponds to a specific normal curve**

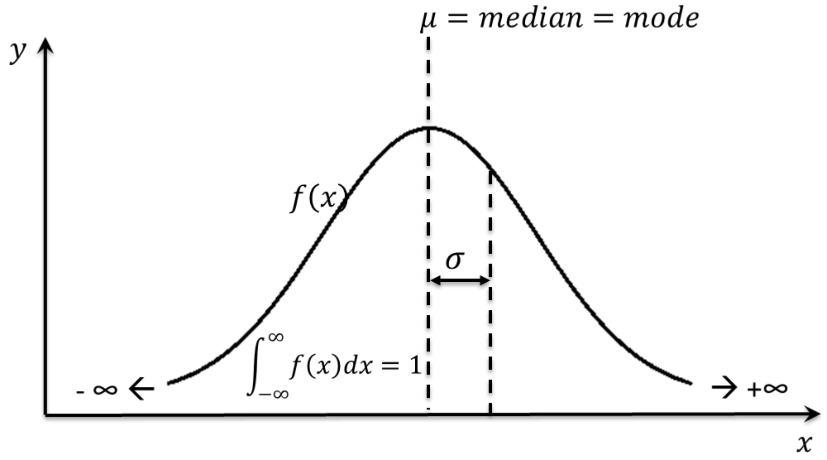


Figure 7-2. The Normal Distribution

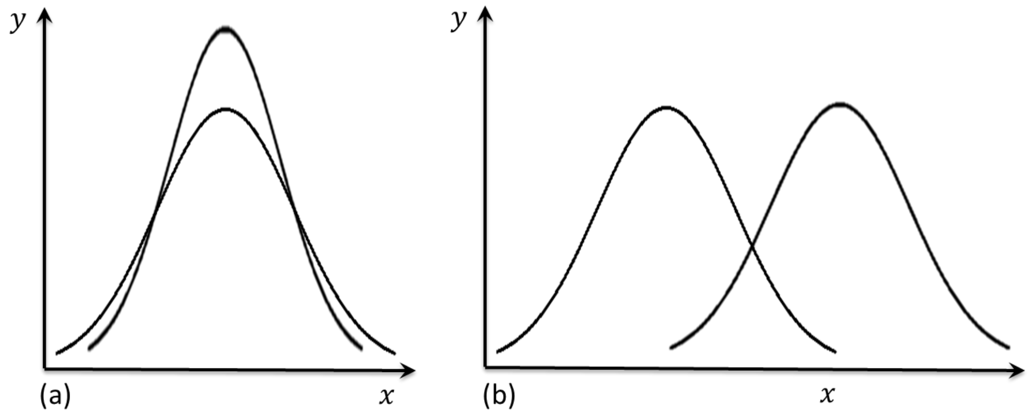


Figure 7-3. Normal Distributions: (a) Same μ but different σ, (b) Same σ but different μ,

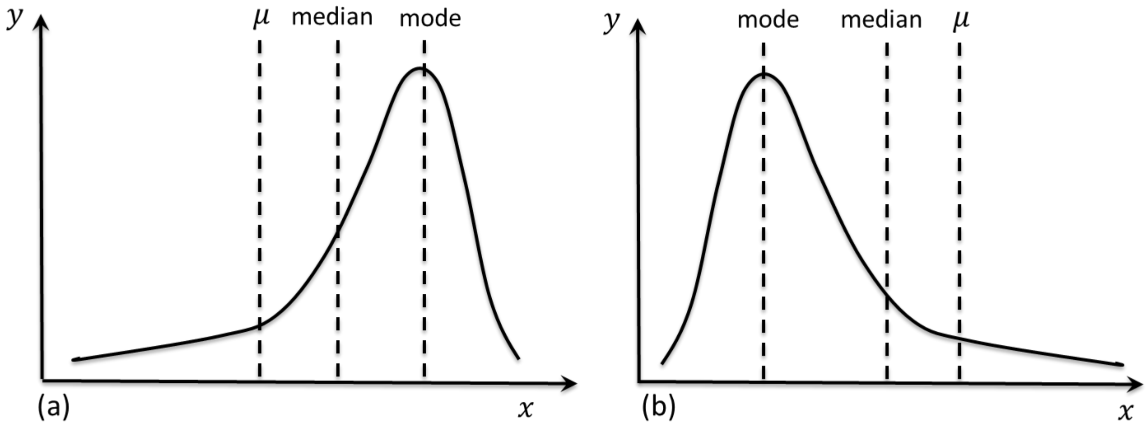


Figure 7-4. Skewed (Not Normal) Distributions: (a) Skewed to the Left/ Negatively Skewed and (b) Skewed to the Right/ Positively Skewed.

A more specific kind of normal distribution is the standard normal distribution, wherein the μ=0 and σ=1. A standard normal curve is obtained by transforming any value of the random variable (x) into a z-score and is calculated as:

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

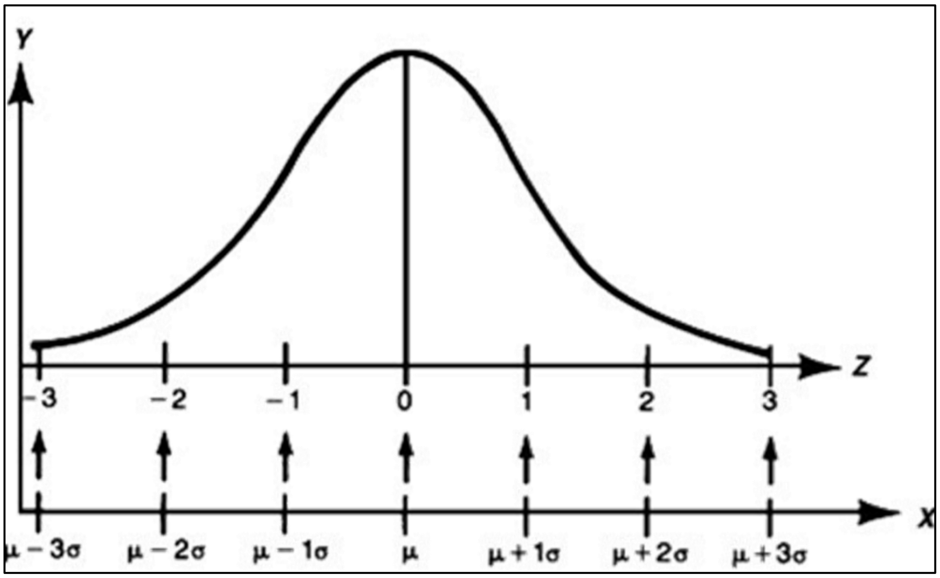


Figure 7-5. Transformation of the distribution of a random variable (X) into the standard normal distribution (z)

Finding the area under the curve

A table that provides the results of all the integrations of $\int_{-\infty}^{\infty} f(x)dx = 1$ that we might be interested in is provided at the end of this module. In the body of the table are found the areas under the curve between $-\infty$ and the values of *z* shown in the left-most column of the table. The shaded area of Figure 7-6 represents the area listed in the table being between $-\infty$ and z_0 where z_0 is the specified value of *z*.

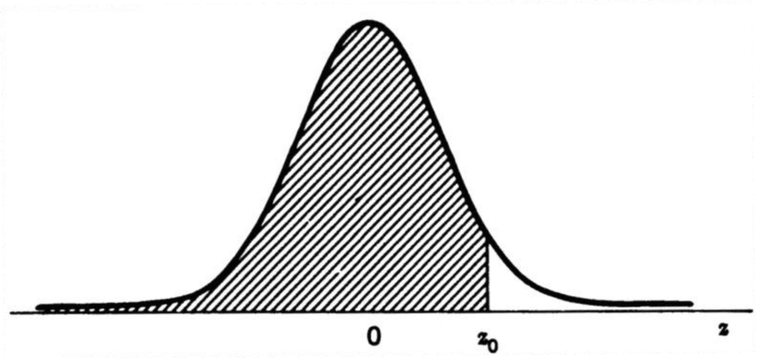


Figure 7-6. Area given by Table C

Example:

Given the standard normal distribution, find the area under the curve, above the *z*-axis between $z = -\infty$ and $z = 2$

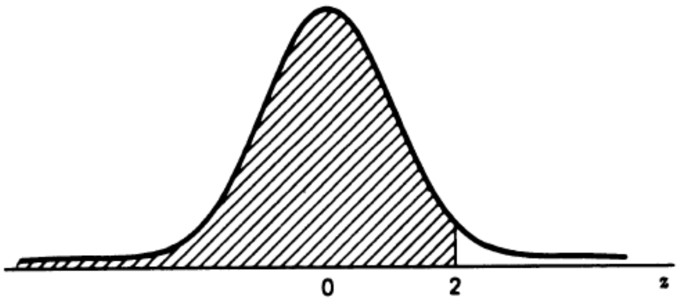


Figure 7-7. Graph of the standard normal distribution showing area between $z = -\infty$ and $z = 2$

Solution:

It will be helpful to draw a graph of the standard normal distribution and shade the desired area as in Figure 7-7. If we locate $z = 2$ in Table C and read the corresponding entry in the body of the table, we find the desired area to be .9772. We can interpret this value in the following ways:

- 97.72% is the probability that a *z* picked at random from a population of *z*'s will have a value between $-\infty$ and 2.
- 97.72% is the relative frequency of occurrence (or proportion) of values of *z* between -$-\infty$ and 2, or we may say that 97.72% of the *z*'s have a value between $-\infty$ and 2.

Alternatively, instead of looking up the desired areas on the table, you may use Microsoft Excel's *NORM.S.DIST* function. The syntax for the function is =NORM.S.DIST(z,cumulative) where z is the value for which you want the distribution and cumulative is a logical value that determines the form of the function. If cumulative is TRUE, NORM.S.DIST returns the cumulative distribution function or the probability that the value will be less than or equal to *z*. If cumulative is FALSE, NORM.S.DIST returns the probability mass function or the probability that the value *z* will occur. We will mostly use TRUE as the argument for our purposes since we desire to find the area to the left of *z*.

Therefore, from the example above, the correct way to calculate the area between $z = -\infty$ and $z = 2$ is to input =NORM.S.DIST(2,TRUE) into a cell. The output 0.97725 will be generated.

Note that for versions of Excel earlier than 2010, the analogous function is *NORMSDIST* whose syntax is =NORMSDIST(z). Notice that the argument for *cumulative* is not required since this function always returns the cumulative distribution function. When using Excel 2007 or earlier, the correct way to input the syntax into a cell is =NORMSDIST(2). This will give the same output of 0.97725

Thought to ponder

Recall that the NORM.S.DIST function returns the area from $-\infty$ to the specified value of *z*, similar to the tabulated values. How should you write your syntax if you want to find areas from *z* to $+\infty$ instead? How about when you want to find the probability that a z picked at random from the population of z's will have a value between two z's such as $P(-2.55 < z < 2.55)$?

Another useful function is NORM.S.INV. The NORM.S.INV function works in reverse as the NORM.S.DIST function (i.e., it returns the inverse of the standard normal cumulative distribution). Its syntax is =NORM.S.INV(probability) where the argument probability refers to a probability corresponding to the normal distribution (i.e., an area under the curve). NORM.S.INV considers the area from $-\infty$ to the specified probability to calculate the *z* value. The analogous function for Excel versions 2007 and earlier is NORMSINV with syntax =NORMSINV(probability).

Normal Distribution Applications

Example:

As part of a study of Alzheimer's disease, Dusheiko reported data that are compatible with the hypothesis that brain weights of victims of the disease are normally distributed. From the reported data, we may compute a mean of 1076.80 grams and an SD of 105.76 grams. If we assume that these results are applicable to all victims of Alzheimer's disease, find the probability that a randomly selected victim of the disease will have a brain that weighs less than 800 grams.

SOURCE: S.D. Dusheiko, "Some Questions Concerning the Pathological Anatomy of Alzheimer's Disease," *Soviet Neurological Psychiatry*, *7* (1974), 56-64 as printed in Wayne Daniel, *Biostatistics: A Foundation for Analysis in the Health Sciences"* 6e (1995).

Solution:

1. Draw a graph of the distribution and shade the area corresponding to the probability of interest.

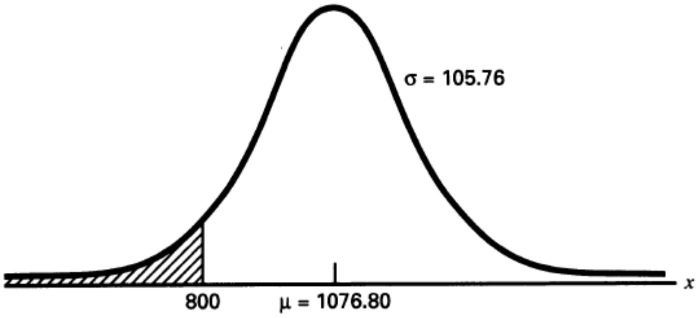


Figure 7-8. Normal distribution to approximate distribution of brain weights of patients with Alzheimer's disease (mean and standard deviation estimated)

2. If this were a standard normal distribution, with a mean of 0 and a standard deviation of 1, we could look up the probability in the tabulated values of *z* and find it with ease. Fortunately, it is possible for any normal distribution to be transformed to the standard normal with little effort. What we do is transform all values of *X* to corresponding values of *z*. This means that the mean of *X* must become 0, the mean of *z*. In this particular problem, we must determine what value of *z*

corresponds to an *x* of 800. This can be accomplished by using this formula which was presented earlier:

$$z = \frac{x - \mu}{\sigma} = \frac{800 - 1076.80}{105.76} = -2.62$$

Therefore, *P*(*x* < 800) = *P*(*z* < -2.62)

Alternatively, you can use Excel's *STANDARDIZE* function to transform values of *x* into *z*. The syntax for this function is =STANDARDIZE(x,mean,standard_dev) where x is the value you want to normalize, mean is the arithmetic mean of the distribution, and standard_dev is the standard deviation of the distribution. In this example, the proper way to input the syntax into a cell is =STANDARDIZE(800,1076.8,105.76). This will return a result of -2.61725.

3. Find the area to the left of *z* = -2.62 by looking at the tabulated values of the standard normal distribution of by using the NORM.S.DIST function.

$$P(z < -2.62) = \mathbf{0.0044}$$

However, there exists a simpler, one-step solution to solve for the desired probability when the given values are not normalized such as the case in this problem. This is through the use of Excel 2010's *NORM.DIST* function which returns the normal distribution for the specified mean and standard deviation. Its syntax is is =NORM.DIST(x,mean,standard_dev,cumulative) where all four arguments are as defined earlier. For this particular problem, the function should be used as =NORM.DIST(800,1076.8,105.76,TRUE). The result of using this function is 0.004432. The analogous function to NORM.DIST for Excel 2007 and earlier is NORMDIST which uses the same combination and sequence of arguments.

Simply put, using =NORM.DIST(x,mean,standard_dev,cumulative) is similar to using =NORM.S.DIST(STANDARDIZE(800,1076.8,105.76),TRUE).

Interpretation: The probability that a randomly selected victim of Alzheimer's disease will have a brain that weighs less than 800 grams is 0.44%

Importance of the Normal Distribution

- In inferential statistics, statistical tests used (e.g. t-test, analysis of variance) assume a normal distribution. Statistical tests that assume a normal distribution are also termed as *parametric* tests.
- Basis for estimation of parameters and calculation of the probability of occurrence of random variables based on areas under the normal curve in a *sampling distribution*
    1. To determine the proportion (%) of values or proportion of the population that belong to certain categories (given the μ & σ)
    2. To estimate the probability that a member of the population will belong to a category (given the μ & σ)
    3. To determine the bounding variables/ values (i.e. x1 & x2) given the proportion or probability

The Sampling Distribution

An important goal of data analysis is to distinguish between features of the data that reflect real facts and features that may reflect only chance effects. Because it is impossible to actually acquire values of a random variable from all members of a population so as to measure the population mean, μ, and standard deviation, σ. It is essential to acquire values of a random variable from a random sample of a population in order to get the sample mean, $\bar{x}$, and sample standard deviation, *s*.

- **Parameter versus Statistic**
  - **Parameter = numerical constant acquired by observing the population (e.g. μ, σ)**
  - **Statistic = numerical variable acquired by observing a sample from the population (e.g. $\bar{x}, s$)**
- **Sampling Distribution: probability distribution of a statistic that would be obtained under repeated sampling**
  - **Sampling Distribution of the Mean: probability distribution of sample means ($\bar{x}$) obtained from all possible samples of size n.**
  - **Sampling Distribution of Proportions: probability distribution of sample proportions (*p*) obtained from all possible samples of size n. (note: Population Proportion = *P* and Sample Proportion = *p*)**

1. **Sampling Distribution of the Mean**
   - **Properties:**
     - **It is approximately normally distributed**
     - **$\mu_x = \mu$, mean of all sample means is equal to the population mean**
     - ***s* (a.k.a. standard error of the mean) $\approx \frac{\sigma}{\sqrt{n}}$**
   - **Applications:**
     - **Determining the probability of occurrence of *x* given a pre-specified magnitude (or bounding values under the normal curve) from the population**
     - **Estimation of the μ**
     - **Hypothesis testing about μ**
   - **Sample Problem: If the distribution of systolic blood pressure (SBP) of non-hypertensive chimpanzees (*Pan* sp.) has a mean of 110 mmHg & standard deviation of 12 mmHg:**

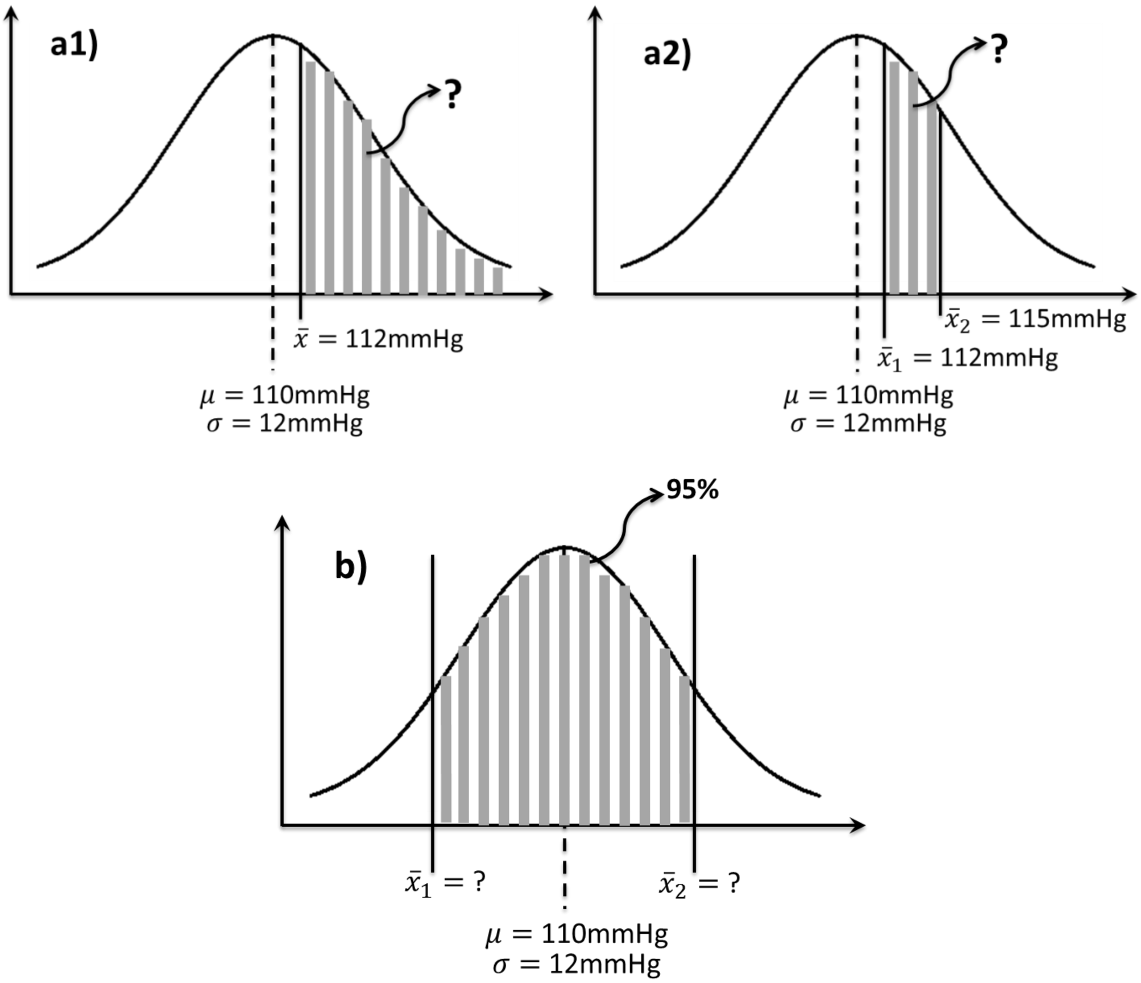
a) What is the probability that a sample of 49 chimpanzees will yield a mean that is:
(1) Greater than or equal to 112 mmHg?
(2) Between 112 and 115 mmHg?
b) Within what range will the middle 95% of the sample means fall?

Given: $\mu = 110$**mmHg** **n = 49 chimpanzees**
σ **= 12 mmHg** $\bar{x}$ **= 112 mmHg**

Required: **a1) Probability of a mean that is ≥ 112mmHg**
a2) Probability of a mean that is between 112mmHg and 115mmHg
b) The bounding values ($\bar{x}_1$ *and* $\bar{x}_2$) for the middle 95% of the sample means

Solution:

- **Step 1 – sketch the distribution as a guide; shade areas under the curve that are required or given**



- **Step 2 (for a1 and a2) – compute for the z-deviate**

a1) $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} = \frac{112-110}{12/\sqrt{49}} = 1.17$

a2) $z_1 = \frac{\bar{x}_1-\mu}{\sigma/\sqrt{n}} = \frac{112-110}{12/\sqrt{49}} = 1.17$

$z_2 = \frac{\bar{x}_2-\mu}{\sigma/\sqrt{n}} = \frac{115-110}{12/\sqrt{49}} = 2.92$

Alternatively, you may take advantage of Excel's STANDARDIZE function and substitute the value of σ/n (the standard error) for the standard_dev argument. Take note however that standard error is not the same as standard deviation. We will discuss this in a further lesson. We are only exploiting the algorithm employed by Excel to easily compute for the *z*-deviate of the sampling distribution.

Example for a1: =STANDARDIZE(112,110,12/SQRT(49))
The result is 1.166667

- **Step 2 (for b) – Since the area under the curve is = 1 or 100%, then the total remaining area under the tail regions is 0.05 or 5% where each tail has an area of 0.025 or 2.5%**

- **Step 3 (for a1 and a2) – refer to a z-table (or table for the areas under the standard normal curve) that may either provide the probabilities for the tail region or middle region of the curve. Assuming that the z-table used for this problem provides areas under the tail based on the z deviate calculated:**

  a1) 1.17 → 0.1210 or 12.10%

Alternatively, you may also exploit the NORM.DIST function for a one-step solution. The appropriate syntax would be =1−NORM.DIST(112,110,12/SQRT(49),TRUE). Once again, we have substituted the value of the standard error σ/n into the standard_dev argument. We began the expression with "1−" since we want to get the area to the right of the specified value of *x*.

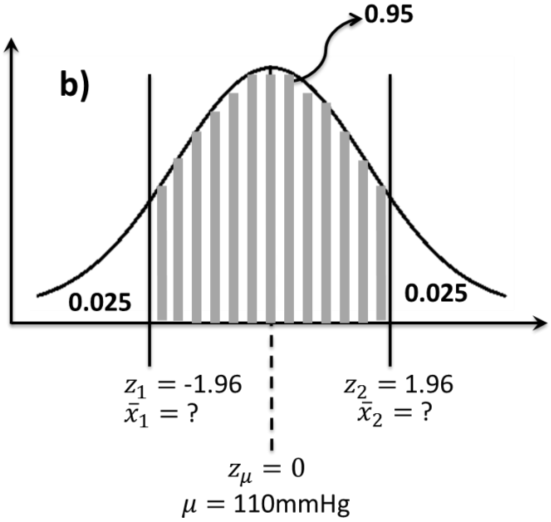
Interpretation: ***The probability of a mean that is ≥ 112 mmHg is 12.10%***

a2) z1 = 1.17 → 0.1210 or 12.10% and z2 = 2.92 → 0.0018 or 0.18% (since these probabilities are tail end probabilities and the middle is need then subtract these from 100%)

$$\therefore \text{Probability}(112\text{mmHg} \geq \bar{x} \leq 115\text{mmHg}) = 100\% - (12.10\% + 0.18\%) = \mathbf{11.92}\%$$

Interpretation: ***The probability of a mean that is from 112 mmHg to 115 mmHg is 11.92%***

- **Step 3 (for b) – Also based on a z-table that provides areas under the tail region of a standard normal distribution; look for the area that is 0.025. This time, the corresponding z-deviate for the area 0.025 must be acquired, which in this case is 1.96. Based on the symmetrical nature of the standard normal distribution:**



In order to get the appropriate $\bar{x}_1$ and $\bar{x}_2$, simply use the given z-deviates:

$$z_1 = \frac{\bar{x}_1-\mu}{\sigma/\sqrt{n}} \rightarrow -1.96 = \frac{\bar{x}_1-110}{12/\sqrt{49}} \rightarrow \bar{x}_1 = (-1.96)\left(12/\sqrt{49}\right) + 110$$

= -3.36 + 110
= 106.64 mmHg

$$z_2 = \frac{\bar{x}_2-\mu}{\sigma/\sqrt{n}} \rightarrow 1.96 = \frac{\bar{x}_2-110}{12/\sqrt{49}} \rightarrow \bar{x}_1 = (1.96)\left(12/\sqrt{49}\right) + 110$$

= 3.36 + 110
= 113.36 mmHg

Interpretation: ***The middle 95% of the sample means fall between 106.64 mmHg to 113.36 mmHg***

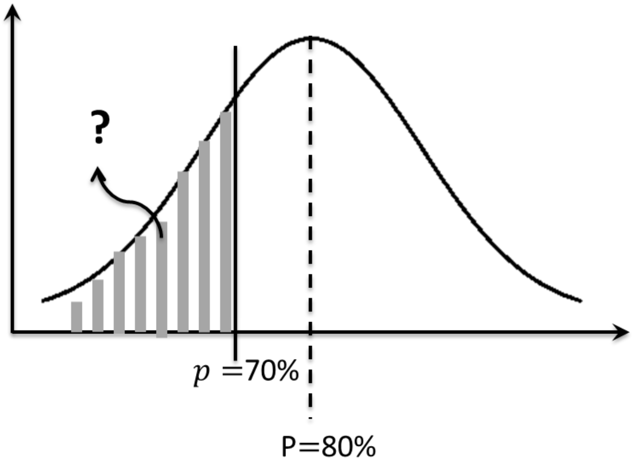
2. Sampling Distribution of Proportions

- **Properties (in a large sample size, n, both n(P) and n(1-P) must be ≥ 5):**
  - **It is approximately normally distributed**
  - **μ_p = P, sample mean proportion is equal to the population proportion**
  - $\sigma_p = \sqrt{\frac{P(1-P)}{n}}$
- **Applications:**
  - **Determining the probability of occurrence of *p* given a pre-specified magnitude from the population**
  - **Estimation of the P**
  - **Hypothesis testing about P**
- **In calculating the z-deviate, the following formula is used:** $z = \frac{p-P}{\sqrt{\frac{P(1-P)}{n}}}$
- **Sample Problem: If the cure rate for a new intestinal parasite drug for dogs is 80%, what is the probability that up to 70% of 50 dogs in an animal shelter given the drug will be cured?**
  - **Note: nP and n(1-P) must both be greater than or equal to 5**
  - **nP = (0.80)(50) = 40 ☑ n(1-P) = (0.20)(50) = 10 ☑**

Given: P = 80% *p* = 70% n = 50 dogs

Required: Probability that up to 70% of 50 dogs will be cured by the parasite drug

Solution: Sketch the graph, transform to the standard normal distribution (z deviate) and get the probability of the computed z from the z table.



$$z = \frac{p-P}{\sqrt{\frac{P(1-P)}{n}}} = \frac{0.70-0.80}{\sqrt{\frac{(0.80)(0.20)}{50}}} = -1.77$$

For a z table with tail areas, z= -1.77 → 0.0384 → 3.84%

Interpretation: ***The probability that up to 70% of 50 dogs in a shelter will be cured by the drug.***

LABORATORY EXERCISE 8
Normal Distribution (50 Points)

NAME:____ **DATE:**_____________

SECTION:________________________________ **INSTRUCTOR:**__________________________

1. Given the values of z, find the area under the normal curve that lies between. (2.5 points each)

a. z = -1.85 and z = 1.85 **b. z = -0.76 and z = 1.13**

______________________________ ______________________________

______________________________ ______________________________

______________________________ ______________________________

a. Answer: ____________________ **b. Answer:** ______________________

2. Determine the area under the normal curve towards the tail end from the z deviates below. (2.5 points each)

a. z = -2.41 **b. z = 1.73**

______________________________ ______________________________

______________________________ ______________________________

______________________________ ______________________________

Answer: ____________________ **b. Answer:** ______________________

3. **If the age of at onset of a hypothetical disease Y has a normal distribution with a mean of 55 years old and a standard deviation of 10 years. What is the probability that that onset of disease Y for a person is before 40 years old? (5 points)**

4. **If the nursing board exam passing rate of a new nursing school is 75%, what is the likelihood that half of 350 nursing students who will take the boards this year will pass? (5 points)**

5. **If the mean serum cholesterol level of Manila residents is 217mg/dL and the variance is 750, calculate the probability that a random person will have a cholesterol value that is < 150 mg/dL. (5 points)**

____

____

____

____

____

____

____

____

____

____

____

6. **Use the NORM.S.INV function to solve for z_1 given the following probabilities (write all pertinent syntax as well as probability notations):**

***Example*:**

$P(z \leq z_1) = 0.1234$

Syntax: =NORM.S.INV(0.1234)

$z_1 = -1.16$

Probability Notation(s): $P(-\infty \leq z \leq -1.16) = 0.1234$

a. **$P(z \leq z_1) = 0.0077$ (2 points)**

Syntax: ____

z_1: ______________

Probability Notation(s):

____

b. **$P(z > z_1) = 0.5250$ (2 points)**

Syntax: ____

z_1: ______________

Probability Notation(s):

____

c. **$P(-2.75 \le z \le z_1) = 0.9885$ (4 points)**

Syntax: ____

z_1: ____________

Probability Notation(s):

____

d. **$P(-z_1 \le z \le z_1) = 0.6263$ (5 points)**

Syntax: ____

z_1: ____________

Probability Notation(s):

____

e. **$P(z_1 \le z \le 3.60) = 0.1071$ (4 points)**

Syntax: ____

z_1: ____________

Probability Notation(s):

____

7. **Use all Excel functions at your disposal to solve the following problem:**

Suppose the average length of stay in a chronic disease hospital of a certain type of patient is 45 days with a standard deviation of 10. If it is reasonable to assume an approximately normal distribution of lengths of stay, find the probability that a randomly selected patient from this group will have a length of stay:

a. **Greater than 30 days (2 points) Answer:** __________

b. **Between 20 and 45 days (3 points) Answer:** __________

c. **Less than 10 days (1 points) Answer:** __________

d. **Greater than 70 days (2 points) Answer:** __________

8 Estimation

On occasion, we calculate descriptive measures to describe a particular set of data. At other times, when the data represent a sample from a larger population, we might be interested in drawing inferences from the data. A large collection of statistical techniques is available to allow us to perform this and constitutes the branch of statistics called *inferential statistics*. Statistical inference is the process by which we reach a conclusion about a population based on the information present in a sample drawn from that population.

When inferences are drawn from normally distributed data, conclusions are based on the relationships of the standard deviation and the mean to the normal curve. When the graph of a frequency distribution seems normal, we can assume that the population of data where our sample originated is normally distributed as well given that the sample size is sufficiently large. In most practical situations, a sample size of 30 is satisfactory. We then assume that if we possessed all possible observations from that population of data, we would discover that 68.3%, 95.5%, and 99.7% of the population would lie between the mean and ±1, 2, and 3 standard deviations. In addition, we assume that 95% of the population would like between the mean and ±1.96 standard deviations.

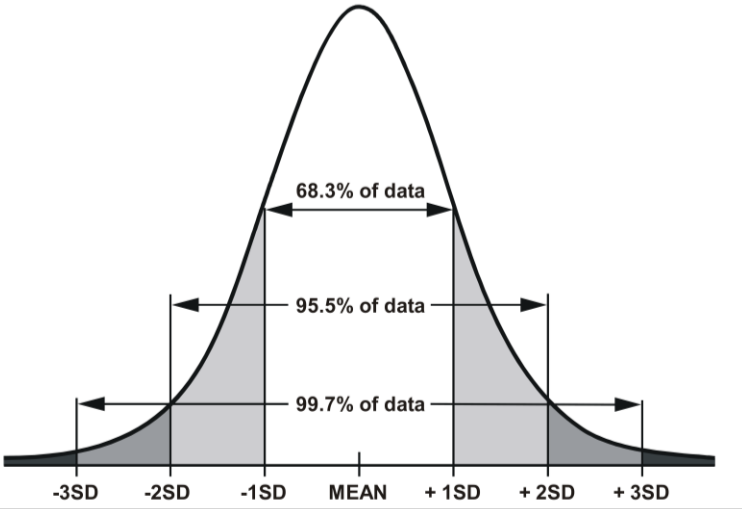


Figure 8-1. Areas under the normal curve that lie between 1, 2, and 3 standard deviations on each side of the mean

SOURCE: *Principles of Epidemiology*, Centers for Disease Control and Prevention (1992)

In this chapter, we will discuss *estimation*, the first of the two general areas of statistical inference, the other being *hypothesis testing*. Estimation is the process by which a statistic computed for a random sample is used to approximate the corresponding parameter.

Here are some examples of situations when estimation is useful:

- A biochemist is interested in estimating the average concentration of a certain protein in a patient population.
- A geneticist is interested in estimating the allele frequencies of certain genes in a target population.

The interests above are to estimate a certain numerical quantity associated with a particular population.

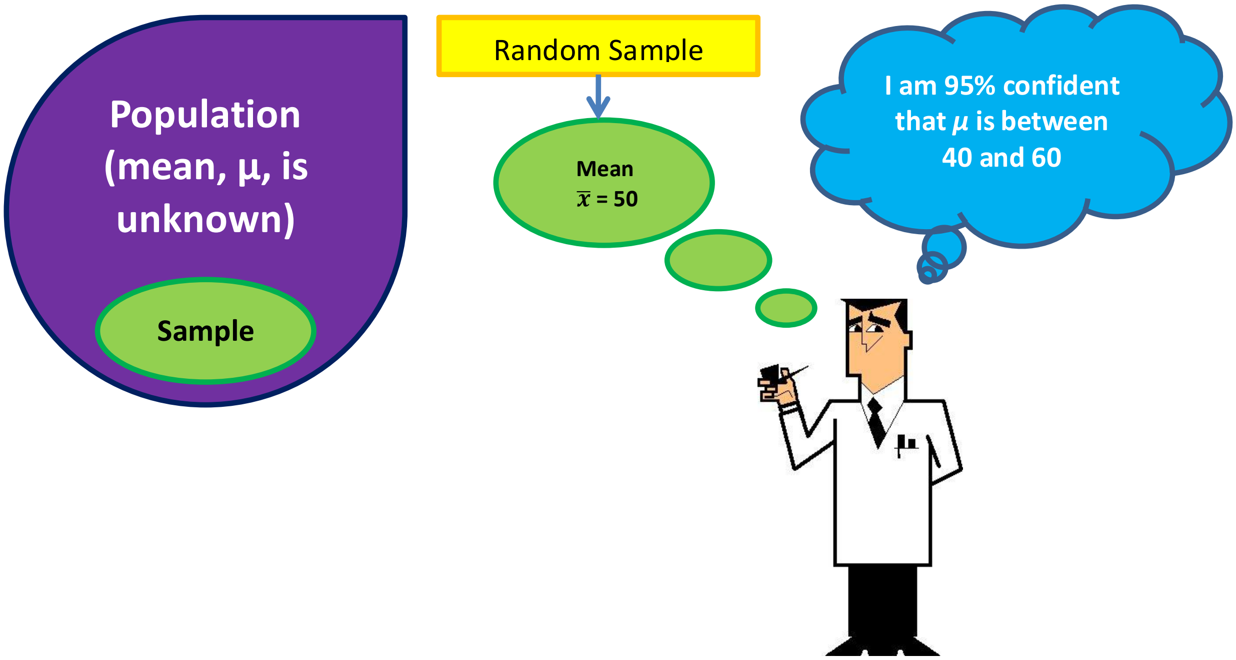

Population (mean, μ, is unknown)
Sample
Random Sample
Mean
$\bar{x}$ = 50
I am 95% confident that μ is between 40 and 60

Figure 8-2. The estimation process relies on a sample of subjects drawn from the intended population. Then, the sample data is used to estimate the unknown population parameter.

Rationale for Estimation in the Health Sciences

The rationale behind estimation in the health sciences field rests on the assumption that professionals in this field are interested in parameters such as means and populations of various populations. If this is the case, there are two good reasons why one must rely on estimation to gather information about these parameters.

1. Many populations of interest, although finite, are so large that a 100% examination would be prohibitive from the standpoint of cost.
2. Populations that are infinite are incapable of complete examination

The Two Types of Estimates

- Point estimate – a *single* numerical value used to estimate the corresponding population parameter.
- Interval estimate – consists of *two* numerical values defining a range of values that, with a specified degree of confidence, includes the parameter being estimated.

Estimator

The estimator is the sample statistic used to make inferences about an unknown parameter.

Example:

$$\bar{x} = \frac{\sum x_i}{n}$$

where the sample mean $\bar{x}$ is an estimator of the population mean μ.

Table 8-1. Examples of Unbiased Estimators

Summarizing Figure	Population Parameter	Sample Statistic (Estimator)
Mean	μ	$\bar{x}$
Variance	σ^2	s^2
Proportion	P	p
Difference between two means	$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$
Difference between to proportions	$P_1 - P_2$	$p_1 - p_2$

Characteristics of a Good Estimator

1. Unbiased – its expected value is equal to the parameter being estimated (regardless of sample size *n*). It should neither neither consistently overestimate nor underestimate the parameter. The sample mean and sample variance have this property and are therefore unbiased. However, the sample standard deviation is not.
2. Precise – it is repeatable because its standard error is small; should not vary too much from sample to sample
3. Consistent – its deviation from the parameter being estimated decreases as sample size increases.

Point Estimate of the Population Mean μ

- Simply the mean computed from a sample
- $\mu = x$

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

Example:

Prostate specific antigen (PSA) is a protein produced by the cells from the prostate. The blood concentration of PSA is often used as a biomarker of prostate cancer. Results under 4 ng/mL are usually considered normal. The higher the PSA level, the more likely a patient has prostate cancer. Because of this relationship, postproctectomy PSA has also been used to measure the success of the operation.

Table 8-2. Prostate specific antigen levels (ng/mL) in 30 patients measured 6 months after proctectomy

0.2	0.1	0.0	0.0	0.1	0.1	0.1	0.1	0.1	0.0
0.4	0.0	0.0	0.2	0.2	0.1	2.7	0.1	0.0	0.2
1.3	0.0	0.2	0.0	0.1	0.3	0.1	0.0	0.0	0.1

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{0.2 + 0.1 + 0.0 + \cdots 0.1}{30} = \mathbf{0.2267\ ng/mL}$$

The point estimate for PSA levels in patients measured 6 months after proctectomy is 0.2267 ng/mL.

Point Estimate of the Difference Between Two Population Means

- Simply the difference between two sample means, $\bar{x}_1 - \bar{x}_2$
- $\mu_1 - \mu_2 = \bar{x}_1 - \bar{x}_2$

Example:

A study aims to determine the relationship between salt intake and blood pressure of persons aged 15 years and older. The mean SBP of 20 subjects with low salt diet was 120 mmHg while the mean SBP of those with high salt diet was 138 mmHg.

$$\bar{x}_1 - \bar{x}_2 = 138 - 120 = \mathbf{18\ mmHg}$$

The point estimate for the difference in mean systolic blood pressure between patients with high and low salt diet is 18 mmHg

Point Estimate of the Population Proportion, *P*

- simply its corresponding statistic, p
- the proportion of the sample possessing the characteristic of interest

$$P = p$$

$$p = \frac{number\ who\ possess\ the\ characteristic}{total\ number\ examined}$$

Example:

In a prospective cohort study, troponin T levels were obtained for a sample of 801 patients who had been hospitalized with acute myocardial ischemia. Whether or not the patients died within 30 days was then obtained and is summarized in the following table. What is an overall estimate of dying within 30 days for the patients with high troponin T levels?

Table 8-3. Survival versus Troponin T Levels

	Troponin T Level		
Status	**>0.1 ng/mL**	**≤0.1 ng/mL**	**Total**
Alive	255	492	747
Dead	34	20	54
Total	289	512	801

$$\hat{p} = \frac{34}{289} = 0.12$$

The proportion of dying within 30 days for patients with high troponin T levels is 0.12.

Point Estimate of the Difference Between Two Population Proportions ($P_1 - P_2$)

- Simply the difference between two sample proportions, $p_1 - p_2$
- $(P_1 - P_2) = (p_1 - p_2)$

Example:

From the previous table, what is an estimate of the difference in proportion of deaths between the patients with high versus low troponin T levels?

$$\hat{p}_1 - \hat{p}_2 = {}^{34}/_{289} - {}^{20}/_{512} = 0.12 - 0.04 = \mathbf{0.08}$$

The point estimate for the difference in proportion of deaths between patients with high versus low troponin T levels is 0.08.

Interval Estimation

An interval estimate gives a range of values, taking into consideration the variation in sample statistics form sample to sample.

- It conveys information about the probable magnitude of the population parameter.
- It is stated in levels of confidence. It can never be 100% confident.
- Express the probability that a prescribed interval will contain the true parameter.

Components of the interval estimate

- Estimator – point estimate of the population parameter in the center
- Reliability or confidence coefficient – the degree of certainty that the population parameter being estimated is within the computed confidence interval
- Standard error of the mean – or simply standard error, it is the standard deviation of the sampling distribution of the statistic

The standard deviation and standard error of the mean should not be confused. The standard deviation is a measure of the variability or dispersion of a set of observations about the mean. Meanwhile, the standard error of the mean is a measure of the variability or dispersion of sample means about the true population mean.

The general form of an interval estimate is expressed as:

Estimator ± (Reliability Factor X Standard Error)

Notice that the center of the interval estimate is the point estimate of parameter of interest. The quantity obtained by multiplying the reliability factor by the standard error of the sample statistic is called the precision of the estimate or the margin of error.

Table 8-4. Commonly used confidence levels and their corresponding two-tailed reliability coefficients

Reliability or Confidence Coefficient	Reliability Factor or *z*-deviate
90%	1.645
95%	1.960
99%	2.576

Take note that researchers may use any confidence coefficient they wish but the most frequently used values are presented above.

Formula for estimating the standard error of the mean:

$$Standard\ error\ of\ the\ mean = SE = \frac{s}{\sqrt{n}}$$

Note that the standard error of the mean is affected by two components, the standard deviation and the sample size. The greater the observations vary about the mean, the greater the uncertainty of the mean, and the greater the standard error of the mean. The larger the sample size, the more confidence we have in the mean, and the standard error of the mean becomes less.

Example:

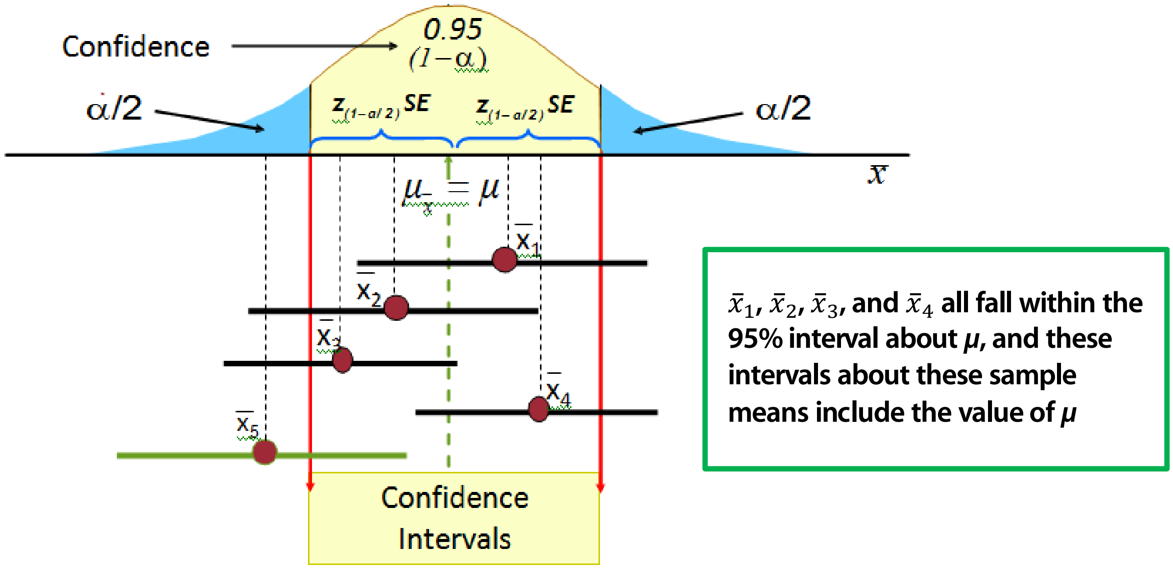
Occupational health researchers measured the weights of a random sample of 120 male workers at an industrial factory, factory F. The mean weight was 92.546 kg, with a standard deviation of 5.265 kg. Calculate the standard error of the mean for the height of workers at factory F.

$$SE = \frac{s}{\sqrt{n}} = \frac{5.265}{\sqrt{120}} = 0.481$$

Interpreting Confidence Intervals

Suppose the reliability coefficient is 95% with a corresponding reliability factor of 1.96. We can say that in repeated sampling, approximately 95% of the intervals constructed will include the population parameter being estimated. This interpretation is based on the probability of occurrence of different values of the sampling statistic. This interpretation can be generalized if we designate the total area under the curve of the point estimate that is outside the interval [estimator ± (reliability factor X standard error)] as α and the area within the interval as 1 – α and give the following *probabilistic interpretation*:

Probabilistic interpretation: *In repeated sampling, from a normally distributed population with a known standard deviation,* 100(*1* – α) *of all intervals constructed will in the long run include the population parameter of interest*.



$\bar{x}_1$, $\bar{x}_2$, $\bar{x}_3$, and $\bar{x}_4$ all fall within the 95% interval about μ, and these intervals about these sample means include the value of μ

Figure 8-3. Normal distribution of a sample mean showing the region of confidence when α = 0.05

Suppose that in the figure above, 15 additional confidence intervals are constructed so that there are a total of 20 confidence intervals for the population mean. When the confidence level is 95%, we can expect that only one out of the 20 confidence intervals that we constructed would not include the population parameter being estimated. In the case above, $\bar{x}_5$ did not include the true value of the population mean, μ.

The quantity 1 – α, in this case is called the *confidence coefficient* (or confidence level), and the interval [estimator ± (reliability factor X standard error)] is called a *confidence interval* for the population parameter of interest. When (1 – α) = 0.95, the interval is called the 95% confidence interval for the population parameter. Another way of interpreting confidence intervals is through the *practical interpretation* which may be expressed as follows:

Practical interpretation: ***When sampling is from a normally distributed population with known standard deviation, we are*** **100(*1 –* α)** ***percent confident that the single computed interval,*** **[estimator ± (reliability factor X standard error)]*, contains the population parameter of interest*.**

Confidence Interval for the Population Mean

Known population variance

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

where $\frac{\sigma}{\sqrt{n}}$ = standard error (SE)

***Sample Problem*:**

A physical therapist wished to estimate, with 99% confidence, the mean maximal strength of a particular muscle in a certain group of individuals. He is willing to assume that strength scores are normally distributed with a variance of 144. A sample of 15 subjects who participated in the study yielded a mean of 84.3

***Solution*:**

The *z* value corresponding to a confidence (or reliability) coefficient of 0.99 is 2.576.

The 99% confidence interval for *μ* is

$$84.3 \pm 2.576\left({}^{12}/_{\sqrt{15}}\right) = 84.3 \pm 2.576(3.0984) = 84.3 \pm 7.98$$

$Lower\ limit: 76.3$

$Upper\ limit: 92.3$

Interpretation (Practical): We are 99% confident that the mean maximal strength of the particular muscle in the population is between 76.3 and 92.3.

Sampling from nonnormal populations

It will not always be possible or prudent to assume that the population of interest is normally distributed. However, based on the *central limit theorem*, this will not deter us if the sample size is sufficiently large. For sufficiently large samples $(n \geq 30)$, the sampling distribution of $\bar{x}$ is approximately normally distributed regardless of how the parent population is distributed.

***Sample Problem*:**

Punctuality of patients in keeping appointments is of interest to a research team. In a study of patient flow through the offices of general practitioners, it was found that a sample of 35 patients were 17.2 minutes late for appointments, on the average. Previous research had shown the standard deviation to be about 8 minutes. The population distribution was felt to be nonnormal. What is the 90% confidence interval for *μ*, the true mean amount of time late for appointments?

Solution:

Since the sample size is satisfactorily large (greater than 30), and since the population standard deviation, σ, is known, we draw on the central limit theorem and assume the sampling distribution of $\bar{x}$ to be approximately normally distributed.

$17.2 \pm 1.645\left({}^{8}\!/_{\sqrt{35}}\right) = 17.2 \pm 1.645(1.3522) = 17.2 \pm 2.2$

$Lower\ limit: 15.0$

$Upper\ limit: 19.4$

Interpretation (Practical): We are 90% confident that within the population, the average amount of time patients are late for appointments with their general practitioners is between 15.0 to 19.4 minutes.

Unknown population variance

In the previous section, we outlined the procedure for constructing a confidence interval for the population mean which requires the knowledge of the variance of the population from which the sample is drawn. It may seem somewhat strange that one can have knowledge of the population variance but not know the value of the population mean. Indeed, the usual case is that in most situations, both the population variance as well as the population mean are unknown. For example, although the statistic

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

is normally distributed when the population is normally distributed, and is at least approximately normally distributed when *n* is large, regardless of whether or not the population is normally or nonnormally distributed, we cannot use the *z* statistic because σ is unknown. However, all is not lost, and the most logical solution to the problem is to use the sample standard deviation, *s*, in place of σ. When the sample size is sufficiently large $(n \geq 30)$, our faith in *s* as an approximation of σ is usually substantial, and we may feel justified in using normal distribution theory to construct a confidence interval for the population mean just like we have done previously.

$$\bar{x} \ \pm z_{(1-\alpha/2)}\frac{s}{\sqrt{n}}, if\ n \geq 30$$

It is when we do not have a sufficiently large sample size that it becomes necessary for us to find an alternative procedure for constructing confidence intervals.

This alternative is the *Student's t distribution*, usually shortened to *t distribution* which is the result of the work of Gosset writing under the pseudonym of "Student". The following quantity follows the *t* distribution:

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

Properties of the *t* Distribution

1. It has a mean of 0.
2. It is symmetrical about the mean.
3. Generally, it has a variance greater than 1, but the variance approaches 1 as the sample size becomes large. For *df* > 2, the variance of the *t* distribution is *df*/(*df* – 2), where *df* is the degrees of freedom. Alternatively, since here *df* = *n* – 1 for *n* > 3, the variance of the *t* distribution can be written as (*n* – 1)/(*n* – 3).
4. The variable ranges from $-\infty$ to $+\infty$.
5. The *t* distribution is actually a family of distributions, since there is a different distribution for each sample value of *n* – 1, the divisor used in computing s^2. Recall that *n* – 1 is referred to as degrees of freedom.

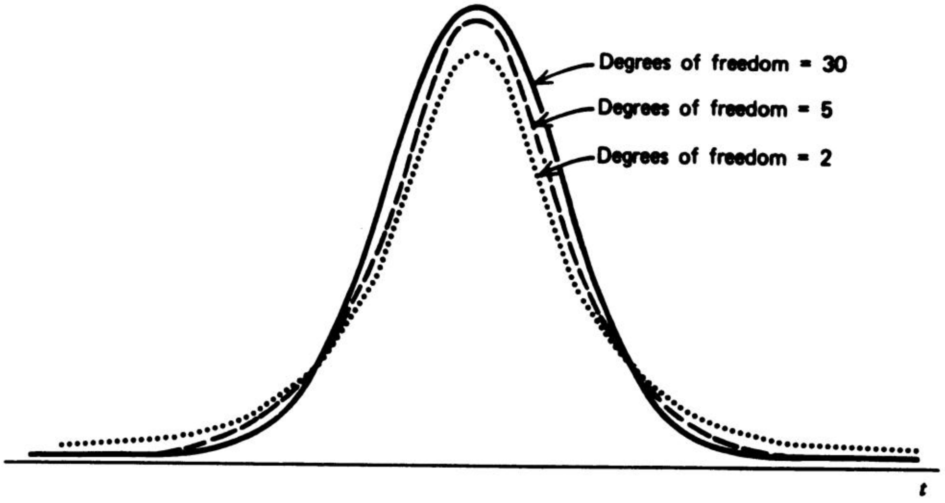


Figure 8-4. The *t* distribution for different degrees-of-freedom values
SOURCE: Daniel, Wayne. *Biostatistics: A Foundation For Analysis in the Health Sciences* 6e (1995).

6. Compared to the normal distribution, the *t* distribution is less peaked in the center and has higher tails.

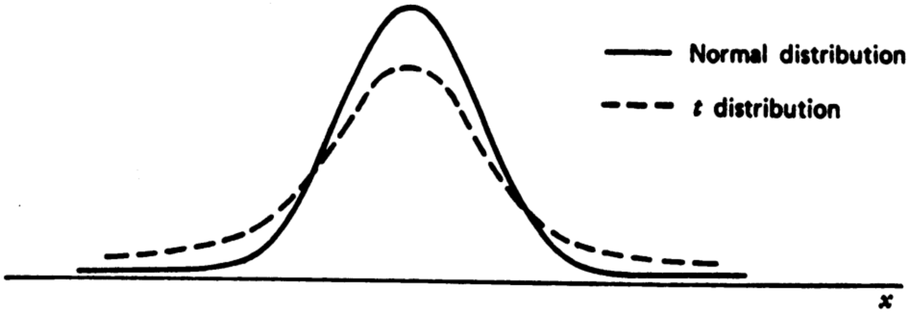


Figure 8-5. Comparison of normal distribution and *t* distribution
SOURCE: Daniel, Wayne. *Biostatistics: A Foundation For Analysis in the Health Sciences* 6e (1995).

Confidence Intervals Using *t*

Despite the need to use the *t* distribution rather than the standard normal distribution, the general procedure for constructing confidence intervals is still the same. The expression

Estimator ± (Reliability Factor X Standard Error)

still applies. The only difference is the source of the reliability coefficient. It is now obtained from the table of the *t* distribution instead of the table of the standard normal distribution. To be more specific, *when sampling is from a normal distribution whose standard deviation*, σ, *is unknown*, *the* 100(1 – α) *percent confidence interval for the population mean*, *μ*, is given by

$$\bar{x} \pm t_{(\alpha/2,df)}\frac{s}{\sqrt{n}}, if\ n < 30, df = n - 1$$

Note that the sample must still be drawn from a normal distribution to justify the valid use of the *t* distribution. However, it has been empirically demonstrated that moderate departures from this requirement can be tolerated. Therefore, the *t* distribution is used even when there is knowledge that the parent population somewhat deviates from normality. Most researchers consider the assumption of at least a mound-shaped population to be acceptable.

Sample Problem:

Maureen McCauley conducted a study to evaluate the effect of on-the-job body mechanics instruction on the work performance of newly employed young workers. She used two randomly selected groups of subjects, an experimental group and a control group. The experimental group received one hour of back school training provided by an occupational therapist. The control group did not receive this training. A criterion-referenced Body Mechanics Evaluation Checklist was used to evaluate each worker's lifting, lowering, pulling, and transferring of objects in the work environment. A correctly performed task received a score of 1. The 15 control subjects made a mean score of 11.53 on the evaluation with a standard deviation of 3.681. We assume that these 15 controls behave as a random sample from a population of similar subjects. We wish to use these sample data to estimate the mean score for the population.

SOURCE: Maureen McCauley, "The Effect of Body Mechanics Instruction on Work Performance Among Young Workers," *The American Journal of Occupational Therapy*, *44* (1990), 402-407 as printed in Wayne Daniel, *Biostatistics: A Foundation for Analysis in the Health Sciences"* 6e (1995).

Solution:

1. Since the level of confidence was not stated, assume that a 95% confidence interval is desired.

2. Find the reliability coefficient, the value of *t* associated with a confidence coefficient of .95 and *n* – 1 = 14 degrees of freedom. Since a 95% confidence interval leaves .05 of the area under the curve of *t* to be equally divided between the two tails, we need the value of *t* to the right of which lies .025 of the area. Alternatively, Microsoft Excel's *T.INV.2T* function can be used. The syntax for this function is as follows: =T.INV.2T(probability, deg_freedom) where probability

stands for α and deg_freedom stands for degrees of freedom or *n* – 1. The T.INV.2T function returns the two-tailed inverse of the *t* distribution. Note that a similar function, *T.INV* exists but returns the left-tailed inverse of the *t* distribution. For Excel versions earlier than 2010, the analogous function to T.INV.2T is the *TINV* function with the syntax =TINV(probability, deg_freedom) where probability and deg_freedom are as previously defined. The value of *t*, which is our reliability coefficient, is found to be 2.1448.

3. **Construct the 95% confidence interval as follows:**

$$\bar{x} \pm t_{(\alpha/2,df)} \frac{s}{\sqrt{n}}$$

$$11.53 \pm 2.1448\left({}^{3.681}\big/_{\sqrt{15}}\right) = 11.53 \pm 2.1448(.9504) = 11.53 \pm 2.04$$

$Lower\ limit$: 9.49

$Upper\ limit$: 13.57

Interpretation (Practical): We are 95% confident that the mean score for the population lies between 9.49 and 13.57

Factors to consider when deciding between *z* and *t*

- **sample size**
- **functional form of the sampled population (whether it is normally or nonnormally distributed)**
- **knowledge of the population variance**

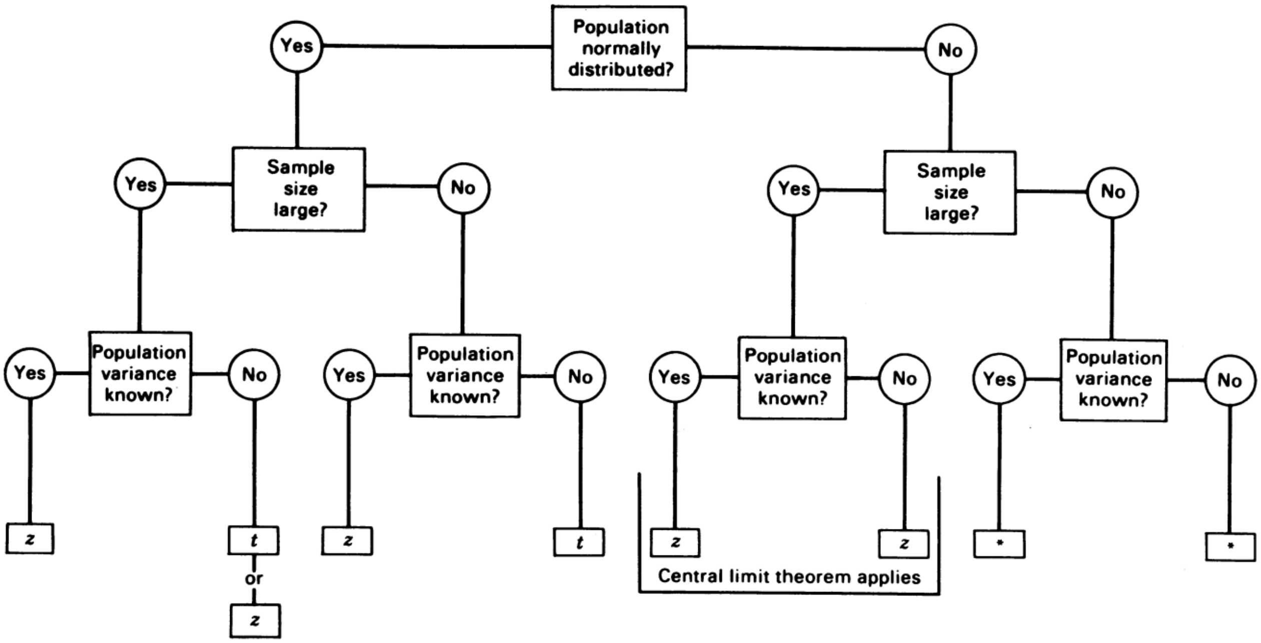

Population normally distributed?
Yes
No
Sample size large?
Yes
No
Sample size large?
Yes
No
Population variance known?
Yes
No
Population variance known?
Yes
No
Population variance known?
Yes
No
Population variance known?
Yes
No
z
t
or
z
z
t
z
z
Central limit theorem applies
*
*

Figure 8-6. Flowchart for use in deciding between *z* and *t* when making inferences about population means. (*Use a nonparametric procedure)

SOURCE: Daniel, Wayne. *Biostatistics*: *A Foundation for Analysis in the Health Sciences* 6e (1995).

Confidence Interval for the Difference Between Two Population Means

Known population variances

$$(\bar{x}_1 - \bar{x}_2) \pm z_{(1-{}^{\alpha}/_{2})}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

***Sample Problem*:**

A research team is interested in the difference between serum uric acid levels in patients with and without Down's syndrome. In a large hospital for the treatment of the mentally retarded, a sample of 12 individuals with Down's syndrome yielded a mean of $\bar{x}_1$ = 4.5 mg/100 ml. In a general hospital a sample of 15 normal individuals of the same age and sex were found to have a mean value of $\bar{x}_2$ = 3.4 mg/ 100 ml. If it is reasonable to assume that the two populations of values are normally distributed with variances equal to 1 and 1.5, find the 95% confidence interval for $\mu_1 - \mu_2$.

***Solution*:**

1. **Since the level of confidence was not stated, assume that a 95% confidence interval is desired.**

2. **Find the reliability coefficient, the value of *z* associated with a confidence coefficient of .95 from the table of the standard normal distribution or using Microsoft Excel's *NORM.S.INV* function.**

3. **Construct the 95% confidence interval as follows:**

$$(\bar{x}_1 - \bar{x}_2) \pm z_{(1-{}^{\alpha}/_{2})}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = (4.5 - 3.4) \pm 1.96\left(\sqrt{\frac{1}{12} + \frac{1.5}{15}}\right)$$

$1.1 \pm 1.96(0.4282) = 1.1 \pm 0.84$

$Lower\ limit\!: 0.26$

$Upper\ limit\!: 1.94$

Interpretation (practical and probabilistic): We are 95% confident that the true difference, $\mu_1 - \mu_2$, is somewhere between 0.26 and 1.94, because, in repeated sampling 95% of the intervals constructed in this manner would include the difference between the true means.

Sampling from nonnormal populations

Constructing confidence intervals for the difference between two population means when sampling is from nonnormal populations is similarly to what we have previously done in constructing confidence intervals for a single population mean when the samples were taken from a nonnormal population. Remember that both *n_1* and *n_2* must be sufficiently large for the central limit theorem to be valid.

Unknown population variances

In cases when the population variances are unknown, and the confidence interval for the difference between two population means is desired, the *t* distribution can be used as a source of the reliability factor if certain assumptions are met. We must know, or be willing to assume, that both sampled populations have a normal distribution.

A. *Equal variances assumed*

If the assumption of equal population variances is justified, the two sample variances that we compute from our two independent samples may be considered as estimates of the same quantity, the common variance. To capitalize on this, a *pooled estimate* of the common variance is calculated. The pooled estimate is obtained by calculating the weighted average of the two sample variances where each sample variance is weighted by its degrees of freedom. If *n* is equal for both populations, the pooled estimate is the arithmetic mean of the two sample variances. However, if *n* is unequal for both populations, the weighted average takes advantage of the additional information provided by the larger sample. The pooled estimate is given by the formula:

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

Thus, the 100(1 – α) percent confidence interval for $\mu_1 - \mu_2$ is given by

$$(\bar{x}_1 - \bar{x}_2)\ \pm t_{(1-\alpha/2)}\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}, \text{where } df = n_1 + n_2 - 2$$

Sample Problem:

The purpose of a study by Stone et al. was to determine the effects of long-term exercise intervention on corporate executives enrolled in a supervised fitness program. Data were collected on 13 subjects (the exercise group) who voluntarily entered a supervised exercise program and remained active for an average of 13 years and 17 subjects (the sedentary group) who elected not to join the fitness program. Among the data collected on the subjects was maximum number of sit-ups completed in 30 seconds. The exercise group had a mean and standard deviation for this variable of 21.0 and 4.9, respectively. The mean and standard deviation for the sedentary group were 12.1 and 5.6, respectively. We assume that the two populations of overall muscle condition measures are approximately normally distributed and that the two population variances are equal. We wish to construct a 95% confidence interval for the difference between the means of the populations represented by these two samples.

SOURCE: William J. Stone, Debra E. Rothstein, and Cynthia L. Shoenhair, "Coronary Health Disease Risk Factors and Health Related Fitness in Long-Term Exercising Versus Sedentary Corporate Executives," *American Journal of Health Promotion*, 5, (1991), 169-173 as printed in Wayne Daniel, *Biostatistics: A Foundation for Analysis in the Health Sciences"* 6e (1995).

Solution:

1. **Calculate the pooled estimate of the common population variance.**

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(13-1)(4.9^2) + (17-1)(5.6^2)}{13 + 17 - 2} = 28.21$$

2. **Substitute the given values into the formula for constructing the confidence interval for the difference between two means when the population variances are unknown and equal population variances are assumed.**

$$(21.0 - 12.1) \ \pm 2.0484\sqrt{\frac{28.21}{13} + \frac{28.21}{17}}$$

$$8.9 \ \pm 4.0085$$

$Lower\ limit$: 4.9

$Upper\ limit$: 12.9

Interpretation (practical and probabilistic): We are 95% confident that the difference between population means is somewhere between 4.9 and 12.9. We can say this because we know that if we were to repeat the study many, many times, and construct confidence intervals in the same manner, around 95% of the intervals would include the difference between the population means.

B. *Equal variances not assumed*

When it cannot be determined whether the variances of two populations of interest are equal, even though both populations may be assumed to have normal distributions, it is not correct to use the *t* distribution as we have discussed for constructing some of the confidence intervals. The problem lies in the fact that the quantity

$$\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

does not follow a *t* distribution with $n_1 + n_2 - 2$ degrees of freedom when the population variances are not equal. Cochran proposed a solution which consists of computing the reliability factor, $t'_{(1-\alpha/_2)}$ by the following formula:

$$t'_{(1-\alpha/_2)} = \frac{w_1 t_1 + w_2 t_2}{w_1 + w_2}$$

where

- $w_1 = s_1^2/n_1$
- $w_2 = s_2^2/n_2$
- $t_1 = t_{(1-\alpha/_2)}$ **for $n_1 - 1$ degrees of freedom**
- $t_2 = t_{(1-\alpha/_2)}$ **for $n_2 - 1$ degrees of freedom**

An approximate 100(1 – α) percent confidence interval for $\mu_1 - \mu_2$ is given by

$$(\bar{x}_1 - \bar{x}_2)\ \pm t'_{(1-{}^{\alpha}/_{2})}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$$

***Sample Problem*:**

In the study by Stone et al. described in the previous example, the investigators also reported the following information on a measure of overall muscle condition scores by the subjects:

Sample	*n*	Mean	Standard Deviation
Exercise group	13	4.5	.3
Sedentary group	17	3.7	1.0

We assume that the two populations of overall muscle condition scores are approximately normally distributed. We are unwilling to assume, however, that the two population variances are equal. We wish to construct a 95% confidence interval for the difference between the mean overall muscle condition scores of the two populations represented by the samples.

***Solution*:**

1. **Find the values of *t*$_1$ and *t*$_2$. Using the tabulated values of the *t* distribution or Microsoft Excel's T.INV.2T function, it can be seen that with 12 degrees of freedom and α = 0.05, *t*$_1$ = 2.1788. Similarly, with 16 degrees of freedom and α = 0.05, *t*$_2$ = 2.1199.**
2. **Compute for the value of *t'***

$$t' = \frac{({.}^{3}/_{13})(2.1788) + \left(\frac{1.0^2}{17}\right)(2.1199)}{\left(\frac{.3^2}{13}\right) + \left(\frac{1.0^2}{17}\right)} = \frac{.139784}{.065747} = 2.1261$$

3. **Substitute the values into the formula for constructing the confidence interval for the difference between two means when the population variances are unknown and equal population variances are not assumed.**

When constructing a confidence interval for the difference between two population means, use the following figure do decide quickly whether the reliability factor should be *z*, *t*, or *t'*.

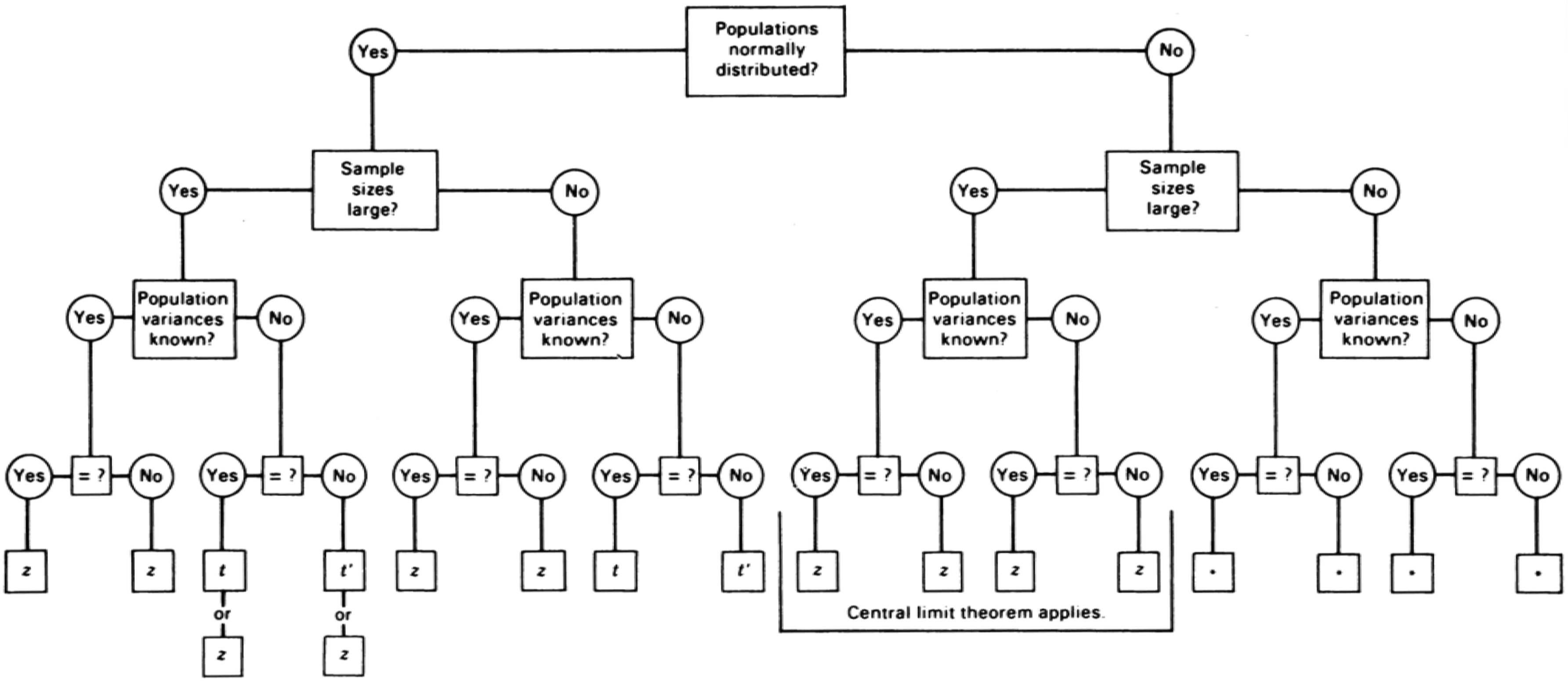

Populations normally distributed?
Yes
No
Sample sizes large?
Yes
No
Sample sizes large?
Yes
No
Population variances known?
Yes
No
Population variances known?
Yes
No
Population variances known?
Yes
No
Population variances known?
Yes
No
= ?
Yes
No
z
z
= ?
Yes
No
t
or
z
t'
or
z
= ?
Yes
No
z
z
= ?
Yes
No
t
t'
= ?
Yes
No
z
z
= ?
Yes
No
z
z
Central limit theorem applies.
= ?
Yes
No
•
•
= ?
Yes
No
•
•

Figure 8-7. Flowchart for use in deciding whether the reliability factor should be *z*, *t*, or *t′* when making inferences about the difference between two population means

SOURCE: Daniel, Wayne. *Biostatistics*: *A Foundation For Analysis in the Health Sciences* 6e (1995).

Confidence Interval for a Population Proportion

Health workers often ask questions related to population proportions. What proportion of patients who receive a particular type of treatment recover? What proportion of some population if affected by a certain disease? What proportion of a population is immune to a particular disease?

The estimation of a population proportion is similar to what we have done with estimating population means or the difference between two population means. The expression

Estimator ± (Reliability Factor X Standard Error)

still applies. A sample is taken from the population of interest, and the sample proportion, p, is calculated. The sample proportion serves as the point estimate of the population proportion.

When both np and $n(1-p)$ are greater than 5, the sampling distribution of p may be considered to be quite close to the normal distribution. When this condition is satisfied, a *z*-value from the standard normal distribution is used as the reliability factor. The confidence interval for a population can be constructed by using the following formula:

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

Sample Problem:

Mathers et al. found that in a sample of 591 admitted to a psychiatric hospital, 204 admitted to using cannabis at least once in their lifetime. We wish to construct a 95% confidence interval for the proportion of lifetime cannabis users in the sampled population of psychiatric hospital admissions.

SOURCE: D.C. Mathers, A.H. Ghodse, A.W. Caan, and S.A. Scott, "Cannabis Use in a Large Sample of Acute Psychiatric Admissions," *British Journal of Addiction*, *86* (1991), 779-784 as printed in Wayne Daniel, *Biostatistics: A Foundation for Analysis in the Health Sciences"* 6e (1995).

Solution:

1. Calculate the point estimate of the population proportion, p
   $p = 204/591 = .3452$
2. Determine the corresponding reliability factor for the desired confidence interval. In this case, the reliability factor that should be used is 1.96.
3. Construct the confidence interval for *p* using the formula
   $.3452 \pm 1.96\sqrt{(.3452)(.6548)/591}$
   $.3452 \pm 1.96(.01956)$
   $.3452 \pm .0383$
   $Lower\ limit: 4.9$
   $Upper\ limit: 12.9$
   Interpretation (practical and probabilistic): We are 95% confident that the population proportion *p* is between .3069 and .3835, since, in repeated sampling, about 95% of the intervals constructed in the manner of the current single interval would include the true *p*. On the basis of these results we would expect, with 95% confidence, to find somewhere between 30.69% and 38.35% of psychiatric hospital admissions to have a history of cannabis use.

Confidence Interval for the Difference Between Two Population Proportions

There are occasions when we might be interested in the magnitude of the difference between two proportions. For example, we may want to compare men and women, two age groups, two socioeconomic groups, or two diagnostic groups with respect to the proportion possessing some characteristic of interest. A 100(1 – α) percent confidence interval for $P_1 - P_2$ is given by

$$(p_1 - p_2) \pm z_{(1-\alpha/2)}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

***Sample problem*:**

Borst et al. investigated the relation of ego development, age, gender, and diagnosis to suicidality among adolescent psychiatric inpatients. Their sample consisted of 96 boys and 123 girls between the ages of 12 and 16 years selected from admissions to a child and adolescent unit of a private psychiatric hospital. Suicide attempts were reported by 18 of the boys and 60 of the girls. Let us assume that the girls behave like a simple random sample from a population of similar girls and that the boys likewise may be considered a simple random sample from a population of similar boys. For these two populations, we wish to construct a 99% confidence interval for the difference between the proportions of suicide attempters.

SOURCE: Sophie R. Borst, Gil G. Noam, and John A. Bartok, "Adolescent Suicidality: A Clinical-Development Approach," *Journal of the American Academy of Child and Adolescent Psychiatry*, *30* (1991), 796-803 as printed in Wayne Daniel, *Biostatistics: A Foundation for Analysis in the Health Sciences"* 6e (1995).

Solution:

1. **Calculate the sample proportions for the girls and boys**
   $p_G = 60/123 = .4878$
   $p_B = 18/96 = .1875$
2. **Determine the corresponding reliability factor for the desired confidence interval. In this case, the reliability factor that should be used is 2.576.**
3. **Construct the confidence interval for $p_1 - p_2$ using the formula**
   $$(p_G - p_B) \pm z_{(1-\alpha/2)}\sqrt{\frac{p_G(1-p_G)}{n_G} + \frac{p_B(1-p_B)}{n_B}}$$
   $$(.4878 - .1875) \pm 2.576\sqrt{\frac{(.4878)(.5122)}{123} + \frac{(.1875)(.8125)}{96}}$$
   $.3003 \pm 2.576(.0602)$
   $Lower\ limit: .1450$
   $Upper\ limit: .4556$
   Interpretation (practical): We are 99% confident that for the sampled proportions, the proportion of suicide attempts among girls exceeds the proportion of suicide attempts among boys by somewhere between .1450 and .4556.

LABORATORY EXERCISE 9
Estimation (90 points)

NAME:____ DATE:_______________

SECTION:_________________________________ INSTRUCTOR:____________________________

A. Software Application (35 points)

Obtain a template of the spreadsheet that you will be using for this exercise from your instructor. Using the NORM.S.INV and T.INV.2T functions of Microsoft Excel 2010 (NORMSINV and TINV for earlier versions), design an algorithm that will return the lower and upper limits of the confidence interval when the necessary data are inputted into the cells of the spreadsheet.

B. Problem Solving (55 points)

Use the spreadsheet that you have just designed to help you answer the following questions:

1. We wish to estimate the mean serum indirect bilirubin level of 4-day-old infants. The mean for a sample of 16 infants was found to be 5.98 mg/100 cc. Assuming bilirubin levels in 4-day-old infants are approximately normally distributed with a standard deviation of 3.5 mg/100 cc find:

   a. The 90% confidence interval for μ (2 points)

      Lower limit: _______________

      Upper limit: _______________

   b. The 95% confidence interval for μ (2 points)

      Lower limit: _______________

      Upper limit: _______________

   c. The 99% confidence interval for μ (2 points)

      Lower limit: _______________

      Upper limit: _______________

2. In an investigation of the flow and volume dependence of the total respiratory system in a group of mechanically ventilated patients with chronic obstructive pulmonary disease (COPD), Tantucci et al. collected the following baseline values on constant inspiratory flow (L/s): .90, .97, 1.03, 1.10, 1.04, 1.00. Assume that the six subjects constitute a simple random sample from a normally distributed population of similar subjects.

   SOURCE: C. Tantucci, C. Corbeil, M. Chassé, J. Braidy, N. Matar, and J. Milic-Emili, "Flow Resistance in Patients with Chronic Obstructive Pulmonary Disease in Acute Respiratory Failure," *American Review of Respiratory Disease*, *144* (1991), 384-389 as printed in Wayne Daniel, *Biostatistics: A Foundation for Analysis in the Health Sciences"* 6e (1995).

   a. What is the point estimate of the population mean? (1 point)

      ____

   b. What is the standard deviation of the sample? (2 points)

      ____

c. **What is the estimated standard error of the sample mean? (2 points)**

____

d. **Construct a 95% confidence interval for the population mean constant inspiratory flow (2 points)**
Lower limit: ________________

Upper limit: ________________

e. **What is the precision of the estimate? (2 points)**

____

f. **State the probabilistic interpretation of the confidence interval you constructed. (2 points)**

____

____

____

g. **State the practical interpretation of the confidence interval you constructed. (2 points)**

____

____

____

3. **Zucker and Archer state that N-NITROSOBIS (2-oxopropyl)amine (BOP) and related β-oxidized nitrosamines produce a high incidence of pancreatic ductular tumors in the Syrian golden hamster. They studied the effect on body weight, plasma glucose, insulin, and plasma glutamate-oxaloacetate transaminase (GOT) levels of exposure of hamsters *in vivo* to BOP. The investigators reported the following plasma glucose levels for 8 treated and 12 untreated animals:**

Subject Group	Sample Mean	Sample Standard Deviation
Untreated	101 mg/dl	5 mg/dl
Treated	74 g/dl	6 mg/dl

SOURCE: Peter F. Zucker and Michael C. Archer, "Alterations in Pancreatic Islet Function Produced by Carcinogenic Nitrosamines in the Syrian Hamster," *American Journal of Pathology*, *133* (1998), 573-577 as printed in Wayne Daniel, *Biostatistics: A Foundation for Analysis in the Health Sciences"* 6e (1995).

a. **State the necessary assumptions for this problem (2 points)**

____

____

____

Construct the following confidence intervals:

b. **The 90% confidence interval for $\mu_1 - \mu_2$ (2 points)**

Lower limit: ________________

Upper limit: ________________

c. The 95% confidence interval for $\mu_1 - \mu_2$ (2 points)

Lower limit: _______________

Upper limit: _______________

d. The 99% confidence interval for $\mu_1 - \mu_2$ (2 points)

Lower limit: _______________

Upper limit: _______________

4. The average length of stay of a sample of 20 patients discharged from a general hospital was 7 days with a standard deviation of 2 days. A sample of 24 patients discharged from a chronic disease hospital has an average length of stay of 36 days with a standard deviation of 10 days. Assuming normally distributed populations with unequal variances, construct the following confidence intervals:

a. The 90% confidence interval for $\mu_1 - \mu_2$ (2 points)

Lower limit: _______________

Upper limit: _______________

b. The 95% confidence interval for $\mu_1 - \mu_2$ (2 points)

Lower limit: _______________

Upper limit: _______________

c. The 99% confidence interval for $\mu_1 - \mu_2$ (2 points)

Lower limit: _______________

Upper limit: _______________

5. Rothberg and Lits studied the effect on birth weight of maternal stress during pregnancy. Participants were 86 Caucasian mothers with a history of stress who had no known medical or obstetric risk factors for reduced birth weight. The investigators found that 11 of the mothers in the study gave birth to babies satisfying the criterion for low birth weight.

SOURCE: Alan D. Rothberg and Bernice Lits, "Psychosocial Support for Maternal Stress During Pregnancy: Effect on Birth Weight," *American Journal of Obstetrics and Gynecology, 165* (1991), 403-407 as printed in Wayne Daniel, *Biostatistics: A Foundation for Analysis in the Health Sciences"* 6e (1995).

a. What is the point estimate of the population proportion? (1 point)

____

b. What is the standard deviation of the sample? (2 points)

____

c. What is the estimated standard error of the sample proportion? (3 points)

____

d. Construct a 99% confidence interval for the population proportion of mothers with a history of stress that gives birth to underweight babies (2 points).

Lower limit: _______________

Upper limit: _______________

e. **What is the precision of the estimate? (2 points)**

____

f. **State the probabilistic interpretation of the confidence interval you constructed. (2 points)**

____

____

____

g. **State the practical interpretation of the confidence interval you constructed. (2 points)**

____

____

____

6. **Research by Lane et al. assessed differences in breast cancer screening practices between samples of predominantly low-income women aged 50 to 75 using county-funded health centers and women in the same age group residing in the towns where the health centers are located. Of the 404 respondents selected from the community at large, 59.2% agreed with the following statement about breast cancer: "Women live longer if the cancer is found early." Among the 795 in the sample of health center users, 44.9% agreed with the statement.**

SOURCE: Etta Williams, Leclair Bissell, and Eleanor Sullivan, "The Effects of Co-dependence on Physicians and Nurses," *British Journal of Addiction, 86* (1991), 37-42 as printed in Wayne Daniel, *Biostatistics: A Foundation for Analysis in the Health Sciences"* 6e (1995).

a. **State the assumptions that you think are appropriate for calculating the interval estimate: (2 points)**

____

____

____

Construct the following confidence intervals:

b. **The 90% confidence interval for $P_1 - P_2$ (2 points)**

Lower limit: _______________

Upper limit: _______________

c. **The 95% confidence interval for $P_1 - P_2$ (2 points)**

Lower limit: _______________

Upper limit: _______________

d. **The 99% confidence interval for $P_1 - P_2$ (2 points)**

Lower limit: _______________

Upper limit: _______________

9 Hypothesis Testing

In the previous lesson, the term hypothesis has already been introduced. To review, a hypothesis is a belief concerning a parameter. As we have indicated previously, a parameter may be a population mean, proportion, variance, standard deviation, etc. We also mentioned that there are two forms of hypothesis: null hypothesis and alternative hypothesis. The null hypothesis is the prevalent opinion, previous knowledge, basic assumption, or prevailing theory while the alternative hypothesis is the rival opinion. The null hypothesis is assumed to be true as long as we find evidence against it. If a sample gives strong enough evidence against the null hypothesis, then the alternative hypothesis comes into force.

Below are different examples of null and alternative hypotheses. Study how these hypotheses were formulated and try to identify the differences between the two in the manner of how these two hypotheses were formulated.

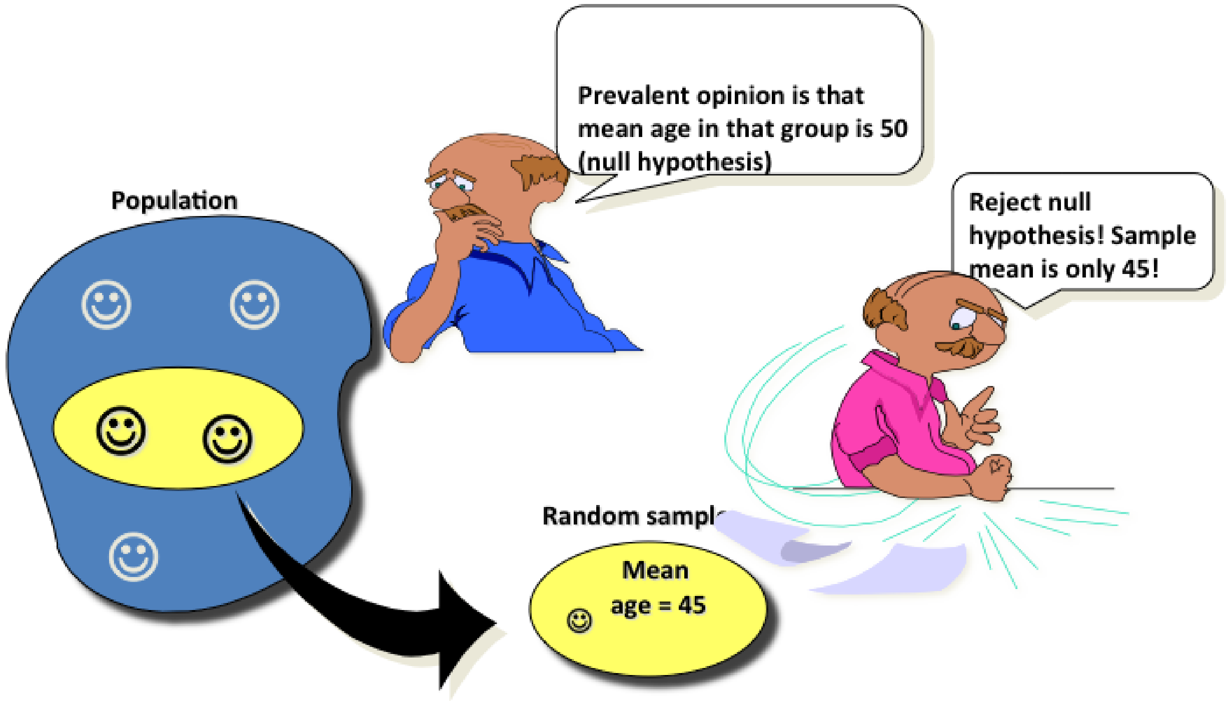
H_0: The mean height of males equals 174 cm.
H_A: The mean height of males is taller than 174 cm.

H_0: Half of the population is in favor of the nuclear power plant operation.
H_A: More than half of the population is in favor of the nuclear power plant operation.

H_0: The amount of overtime work is equal for males and females in Community X.
H_A: The amount of overtime work is not equal for males and females in Community X.

H_0: There is no correlation between the interest rate and gold price in the market.
H_A: There is a correlation between the interest rate and gold price in the market.

What is the logic behind hypothesis testing?


Prevalent opinion is that mean age in that group is 50 (null hypothesis)
Population
Reject null hypothesis! Sample mean is only 45!
Random sampl
Mean age = 45

What is the risk that our hypothesis is wrong?

The risk that our hypothesis is wrong is like asking the risk of a judge giving a wrong judgment to a case. There may be a 50% probability that the hypothesis we give is wrong, just like a judge giving a trialled person a not guilty verdict when the defendant is in fact guilty. What we need to remember is that the null hypothesis remains valid until it is proven otherwise. Sometimes it happens that an innocent person is proven guilty, so this is the same case that may happen to our hypothesis testing. We may reject a null hypothesis although it is true because there is always a risk of being wrong when we reject a null hypothesis. The occurrence of this risk is due to what we call sampling error.

What would be the basis of our decision?

The basis of our decision on whether we should reject or accept our hypothesis is through a *p* value (probability value). Let us start with the basic assumption that our null hypothesis is true. The *p*-value is the probability of getting a value equal to or more extreme than the sample result, given that the null hypothesis is true. Our decision rule enables us to see whether we should reject or not reject the null hypothesis and whether we should accept or reject our alternative hypothesis. If the *p*-value is less than 5%, then we reject the null hypothesis and if the p-value is 5% or more, then the null hypothesis remains valid. In any case, one must give the *p*-value as a justification for our decision in hypothesis testing.

What are the Steps in Hypothesis Testing?

1. State the null hypothesis, H_0; and the alternative hypothesis, H_A or H_1;
2. State the level of significance, α;
3. Choose the test statistic and determine its sampling distribution;
4. Determine the critical region;
5. Calculate the test statistic;
6. Make a statistical decision. (Note: This is where you make the decision of whether or not to reject the null hypothesis);
7. Draw the conclusion about the population.

Notes on the Decision Rule:

Take note, if the *p*-value is less than your α which may be 0.01 with a 99% level of significance; 0.05 with a 95% level of significance; or 0.1 with a 90% level of significance) then reject the null hypothesis. Otherwise, the null hypothesis remains valid. In any case, you must give the *p*-values as a basis for your decision.

Let us look at the steps of hypothesis testing in detail.

Statement of the Null and Alternative Hypotheses

The hypotheses that we must formulate must be something that we can "test" so that it opens the way for a statistical assessment. There are three general statements to keep in mind when framing the null and alternative hypotheses. These three general statements are as follows:

- The null hypothesis is the hypothesis of "no difference." This means that the null hypothesis is a statement of equality.
- H_1 is usually the research hypothesis, meaning it is the hypothesis the investigator believes in.
- H_0 is always framed in hopes of being able to reject it so that H_1 could be accepted.

The formulation of our hypothesis is dependent on whether we want to perform a one-tailed hypothesis test or a two-tailed hypothesis test. Obviously, when we indicate that we are performing a one-tailed hypothesis test, we must use either the left tail or the right tail of the normal distribution curve. Meanwhile, when we want to perform a two-tailed hypothesis test, we are interested in using both the left and right tails of the normal distribution curve.

In a one-tailed test, we must know beforehand that only deviations to one direction are possible. An example would be to say that our H_o : Pa = Pb, while the alternative hypothesis takes the form of being either "less than" or "greater than" the other parameter or statistic. This is either expressed as H_A: Pa > Pb, or Pa < Pb.

In a two-tailed test, the assumption is to use both the left and right tails of the normally distributed curve. We typically use the two-tailed test if we do not have knowledge regarding the direction of the data set. In the two-tailed test, the deviations from the null hypothesis are in both directions of the normally distributed curve. In formulating the alternative hypothesis in a two-tailed test, it takes the form "different than." An example of our null hypothesis is H_0: Pa is equal to Pb while an example of the alternative hypothesis is H_A: Pa is not equal to Pb.

Statement of Significance Level

In this step, we determine our probability level. It is denoted by the Greek letter α (alpha) and is conventionally taken as 5%, 1%, or 10%. The significance level that the researcher chooses is the risk that he/she is willing to take of making the wrong decision of dismissing the null hypothesis as being very unlikely and to favor the null hypothesis instead.

There are two types of errors: Type 1 error or α error and Type II error or β error. We commit a Type 1 error if the null hypothesis is in fact true and it is just unfortunate that our sample yields an unlikely result that we reject the null hypothesis. This occurs when there is really no difference between the population parameters being tested, but the investigator is misled by chance differences in the sample data. Therefore, the type I error is the error of rejecting a true hypothesis. On the other hand, the Type II error is the error that we commit when we do not reject a false null hypothesis. It occurs when there really is a difference between the population parameters being tested, but the investigator misses the difference. It can result from either too much sampling variability or the insensitivity of the test employed, or both, and depends on the number of observations (sample size) included in the study.

Table 9-1. Conditions under which type I and type II errors may be committed.

Possible Action	Condition of Null Hypothesis	
	True	False
Fail to reject H_0	Correct action Probability = $1 - \alpha$	Type II error Probability = β
Reject H_0	Type I error Probability = α	Correct action Probability = $1 - \beta$

Choosing the test statistic

In this step, we select the appropriate tool to test our particular hypothesis. When we choose the test statistic, it has its own sampling distribution that we use to assess the probability of occurrence of sample results under our null hypothesis. There is a wide array of statistical tests that we can use to test the hypothesis before we can make any decision of rejecting the null hypothesis or not rejecting it.

What is the criterion for test selection?

The choice of a particular test statistic depends on several criteria including the types of variables (qualitative or quantitative), level of measurement (nominal, ordinal, interval, or ratio), whether samples are dependent (related or before-after measurements) or independent (different groups).

There is a wide array of reasons for doing the tests. We may want to determine whether the sample could have come from a population with a stipulated mean or proportion, or from a population of some pre-specified distribution (one sample case). It could be that we want to do the test because we want to compare two means or two proportions (two sample cases). We may also want to use the test because we are interested in comparing more than two means or proportions (*k* sample cases). Better yet, we may want to determine whether a relationship exists between the variables that we are studying. The assumption of selecting the test to use also depends on whether we want to do a parametric test or a non-parametric test. A parametric test is appropriate when the data we have is obtained through the random selection of the sample. The normal distribution of the population to which the samples were drawn exists, and when more than one population is sampled, there is equality of variances (homoscedasticity). Also, the numerical data measured must either be on an interval or a ratio scale. The non-parametric test is a type of a distribution-free test. This is a test in which no hypothesis is made about the specific values of population parameters. If the researcher doubts the validity of whether the study satisfies the parametric assumptions, then the non-parametric test should be employed. All sets of data, which are not truly numerical, are tested through non-parametric tests.

Determining the critical region

In this step, the critical region is our region of rejection. The critical region is the set of values of the test statistic which leads to the rejection of the null hypothesis. The critical region indicates the values whose probability of occurrence is less than or equal to the level of significance. The critical region is usually found at the tail end of the distribution. It is also similar to what comprises the region of acceptance. The size of the critical region is determined by the researcher's chosen level of significance. The location of the critical region is also determined by the nature of the alternative hypothesis and whether the researcher opted to do a one-tailed or a two-tailed test of hypothesis.

Calculation of the Test Statistic

In this step, the test statistic chosen is calculated to help us decide on whether or not the null hypothesis should be rejected. It is presumed that the level of significance, test statistic, and the critical region have been determined prior to doing this step. The formula to be used by the researcher varies depending on what test statistic is chosen.

Making a statistical decision

In this step, we either decide whether to reject or not reject the null hypothesis. We reject the null hypothesis if the computed value of our test statistic falls within the critical region. Otherwise, it is not rejected. When the null hypothesis is rejected, the results are statistically significant and the observed difference may not be attributed to sampling variation. If the hypothesis is not rejected, the results are not statistically significant and the sampling variation is the likely explanation of the observed differences.

Drawing Conclusions

The rejection of the null hypothesis leads to a conclusion stated in the form of the alternative hypothesis. If a statistical decision is to not reject the null hypothesis, we do not necessarily accept the null hypothesis. Instead, we say that there is no sufficient evidence to conclude whatever is stated in the alternative hypothesis. The table below shows an example of how we draw conclusions based on our statistical decision.

Statistical Decision	Conclusion
H_0 rejected	(We state the alternative hypothesis) The proportion of students who obtain a grade of 2.0 or better among those using modules for demonstration is greater than the proportion among those using cadavers.
H_0 not rejected	("We do not have sufficient evidence to say that (state the alternative hypothesis)" NOTE: We do not "accept" the null hypothesis, we can only "reject" it.) There is no sufficient evidence to say that the proportion of students who obtain a grade of 2.0 or better among those using modules for demonstration is greater than the proportion among those using cadavers.

General Comments on Hypothesis Testing

- The null hypothesis (H_0) which cannot be rejected is not necessarily "accepted" especially if the sample size is small. We simply do not have enough evidence to reject it.
- Statistical significance is not the same as practical or clinical significance.
- Statistical inference is not valid for badly designed studies. Therefore, when we do a study, before we collect the data, we have to make sure that our methods are accurate and reliable. The sources of biases should be removed or minimized.
- Statistical analysis is not applicable for studies involving the total population.
- One of the purposes of hypothesis testing is to assist administrators and clinicians in making decisions. However, the outcome of statistical tests is only one piece of evidence that should influence the administrative or clinical decision. The statistical decisions must be interpreted along with all the other relevant information to the decision maker.

10 Analysis of Qualitative Data

What is qualitative data?

Normally, when we talk about qualitative data, they are not numerical data. The type of data may not be possible to be applied in basic mathematical operations. There is a need to categorize this kind of data and determine the frequency per category. Working with the frequencies of qualitative data in different categories is the most common way of presenting the data or the results of a study. When working with the frequencies, the usual process is to compare the observed and the expected frequencies or relate the observed frequency to the total number of observations to get a sample proportion to which is now comparable to a pre-specified population proportion. A common example of qualitative data is gender, where we have males and females. Another example could be blood type where it is categorized into different categories like blood type A, B, AB, and O.

What qualitative tests can we use?

Due to the nature of the data that we have oftentimes we are limited in terms of methods or tests that we can use to deal with this kind of data. The common statistical tests that we will be taking up and are indicated in this module are: z test for 1 proportion; z test for 2 proportions; and the chi square test of homogeneity.

In most qualitative tests that we employ, the data that we have is usually categorized into what we call as a binomial population. A binomial population is one in which the elements of our data set belong to either one of the two mutually exclusive and collectively exhaustive categories. There are 2 categories that are mutually exclusive because the elements can only belong to one and only one of the categories. Just like when we talk of gender, it is either a subject is categorized as a male or a female. We cannot have one belonging in both categorizations. Collectively exhaustive as what we indicated earlier means that we literally exhaust the population after distributing the elements to the categories to which they all belong. Looking again at our parameter and statistic, let us subdivide the population into those who possess the characteristic and those who do not. This is shown on the table below.

Table 10-1. Population proportion parameters and their corresponding estimators

	Parameter	Statistic
Proportion who possess the characteristic	P	p
Proportion who do not possess the characteristic	$Q = 1 - p$	$q = 1 - p$
Standard error of proportions	$\sigma = \sqrt{PQ/N}$	$s = \sqrt{pq/n}$

What is the concept of estimation of population parameters?

The concept of estimation of the population parameters focuses on either the point estimation or the interval estimation. Point estimate for a population parameter is the corresponding statistic. The point estimate of a population proportion, possessing a characteristic of interest as denoted by *P*, is the sample proportion, *p*. This proportion is expressed below:

$$proportion\ who\ possess\ the\ characteristic = \frac{number\ who\ possess\ the\ characteristic}{total\ number\ examned}$$

The point estimate is used if we are definitely sure that it will hit the parameter but if not, we can use the interval estimate. Through the interval estimate, we can set bounding values of the statistic, which will include the parameter with a specified degree of confidence. This is just like showing the interval or the possible values where the statistic will lie depending on the degree of confidence. For example, if we know that approximately 95% of the possible values of a statistic, let's say proportion lie within the ± 1.96 standard deviations from the parameter proportion. The 1.96 here is the z-deviate score of our 95% degree of confidence.

Take note that the common z-deviate scores for the following commonly used degree of confidence assuming a 2-tailed test is performed are shown on the table below:

Table 10-2. List of commonly used 2-tailed *z*-values

Degree of Confidence	Z-deviate score
90%	1.64
95%	1.96
99%	2.58

If we want to do a 1-tailed test, the following are the degree of confidence and the z-deviate scores.

Table 10-3. List of commonly used1-tailed *z*-values

Degree of Confidence	Z-deviate score
90%	1.28
95%	1.64
99%	2.33

Meanwhile, the interval estimate is expressed on the illustration below:

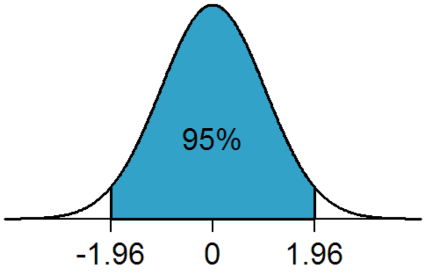


Figure 10-1. The shaded portion represents 95% of the total area under the curve and is bounded by -1.96σ and $+1.96\sigma$ to the left and right respectively.

The 2 points that are ± 1.96 standard deviations from *P* are:

$$P1 = P - 1.96\sigma \, , \sigma = \sqrt{PQ/N}$$
$$P2 = P + 1.96\sigma$$

The equation is expressed as:

$$P \pm 1.96\sqrt{PQ/N}$$

Since the P is unknown, we substitute its point estimator, p in equation as:

$$p \pm 1.96\sqrt{pQ/N}$$

Let us look at an example that uses the concept of estimation. Let's say that a survey was conducted and that study dealt with the dental health practices of children in a certain school. Of the 300 school children interviewed, 123 school children indicated that they had a regular dental check-up twice a year. Now, the problem is this: (1) What percent of the school children in the sample had regular dental check-ups? (2) Give an estimate of the school children population who had regular dental check-ups? (3) Compute and interpret the 95% confidence interval estimate of the population proportion.

First, let us answer the first problem. In this problem, we are asked what percent of the school children in the sample had regular check-ups. Thus, we are interested in getting the proportion of school children with regular check-up over the total number of school children included in the study. We can express this using the formula below:

$$p = \frac{\text{Number of school children with regular check-up}}{\text{Total number of school children examined}} \times 100$$

$$p = 123/300 \times 100$$

$$p = 41\%$$

The proportion of school children with regular check-up over those school children examined is 41%. Let's proceed with the next problem. We are now asked to give an estimate on the school children population who had regular dental check-ups. The population proportion P is unknown, so we use the point estimator, p to estimate for the school children population who had regular dental check-ups. In this case, the point estimate of P is 41%. For the third problem, we are asked to compute and interpret the 95% confidence interval estimate of the population proportion of school children with regular dental check-up. The 95% confidence interval estimate is calculated using the formula below.

$$p \pm 1.96\sqrt{pq/n}$$

$$\begin{aligned} p1 &= 0.41 - 1.96\sqrt{(0.41)\,(0.59)\,/\,300} \\ &= 0.354 \; or \; 35.4\% \\ p2 &= 0.41 + 1.96\sqrt{(0.41)\,(0.59)\,/\,300} \\ &= 0.466 \; or \; 46.6\% \end{aligned}$$

In this problem, we can interpret the values we've obtained by saying that we are 95% confident that the proportion of school children in the population who submit themselves to regular dental check-ups is anywhere in between 35.4% and 46.6%.

In the determination for interval estimate, 90%, 95% and 99% are the commonly used confidence coefficients. The confidence coefficient is a degree of certainty that the parameter being estimated is within the computed confidence interval.

Using the same problem indicated above, compute and interpret the 90% and 99% confidence interval estimates. How would these confidence interval estimates compare with the 95% confidence interval estimate? Using the same formula but changing the confidence coefficients, we can calculate for the p1 and the p2 for both the confidence coefficients. Try to see how the succeeding values were derived. For the 90% confidence interval estimate, the values are in between 36.3% and 45.6% whereas in the 99% confidence interval estimate, the values are in between 33.7% and 48.3%. Take note that as we have a higher confidence coefficient, the wider our estimates are. Hence, the less precise, is our confidence interval.

The aim of performing statistical tests for 1 sample is to determine whether a sample comes from a standard population or norm. The norm or standard population is either known based on the past experiences, or it may be specified as the way one wants or claims the population to be, or it may obey a basic law. Let us look at some examples of what we call as a standard population or norm.

- How the standard population gets known through past experience:
  - Example1: It is known through previous Operation Timbang Projects that 30% of the children less than 6 years of age are malnourished.
  - Example2: It is known that through previous study, 90% of the individuals in the community suffered from dengue.
  - Example 3: It is known through previous surveys that common intestinal worms affect 90% of preschoolers.

- How the standard populations are specified or claimed by interested parties:
  - Example 1: The aim of the Local Govenment is to cover 90% of the target population for vaccination of the ailment X.
  - Example2: A newly developed drug for a particular disease is claimed by the manufacturer to be 95% effective.

- Obeys a basic law:
  - Example1: The basic law states that the probability of a male birth is equal to the probability of a female birth is true, then we should expect 50% of births to be males and 50% females.

In our previous lesson on hypothesis testing, we had discussed the z test for single proportions. Recall the z- test formula for estimating population proportions. The formula is shown below:

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

What if we have a problem and our problem is about the Operation Timbang conducted in a community. The Operation Timbang was conducted to monitor and help out the children from malnutrition. The survey was conducted among 200 randomly selected children to determine if the aim of the Health Office to cover 80% of the target population was attained. It was found that 176 out of the 200 children surveyed were provided with feeding. Did the Health office meet their objective?

To answer the problem, let us first analyze what type of test we can use to answer the problem. Looking at it, we can say that it requires a test of hypothesis for a single proportion. We want to find out if the data collected from a sample of 200 children support the hypothesis that 80% of the population is addressed. The sample proportion *p* and the sample size requirement that both *np* and *nq* must be greater than or equal to 5 must be initially satisfied if we want to employ the *z* statistic. If it satisfies the condition, then we can proceed with the statistic. Let us now employ the steps of hypothesis testing that we have learned from the previous lesson.

Steps:

1. **Stating the hypothesis:**
   Null hypothesis: P is equal to 0.8
   Alternative hypothesis: P is not equal to 0.8

2. **Identify level of significance:**
   Level of significance = 0.05

3. **Choose Test Statistic:**
   The test statistic is the z test for estimating population proportions.

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

4. **Determine the critical region:**
   Since the level of significance is 0.05 and we want to perform a 2-tailed test, our critical region will be:

   $z \geq 1.96$ and $z \leq -1.96$

5. **Do the computations:**
   p = 176/200 = 0.88

   Substitute the computed p and the given *P* and compute for the *z* value using the test statistic indicated above. The *z* value will be equal to 28.28

6. **Make a Decision:**
   The calculated *z* value is 28.28 and this value is greater than the 95% confidence *z* value, which is 1.96.

$$z = 28.28 > z_{0.05} = 1.96$$

From this, we can reject our null hypothesis and accept our alternative hypothesis.

What if we are interested to make inferences for two proportions?

In making inferences for two proportions, we basically consider that there are two independent samples to make inferences for two proportions. There is a need to randomly select the subjects from each of the two populations or we randomly allocate the volunteers to two groups to come up with two independent samples. From the two independent samples, we can estimate the two proportions and estimate the differences between the two proportions by the use of point estimate and interval estimate.

The point estimate of the difference between two population proportions, $(P_1 - P_2)$, is the difference between the sample proportions, $(p_1 - p_2)$. If the sample is sufficiently large, the distribution is approximately normal with the mean equal to $(P_1 - P_2)$and the standard error equal to

$$\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

Since P_1 and P_2 are unknown, we use the estimates p_1 and p_2 instead. From the general formula for computing confidence interval estimates, we derive the formula for the confidence interval estimate of the difference between two proportions as shown below:

$$(p_1 - p_2)\ \pm z_{(1-\alpha/2)}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Let us look at an example where we can apply the given formula above. Suppose that a study was conducted and it determined the prevalence of cholera in the 6 barangays of Manila City. Only the results for the Barangays 1 and 2 are given. Of the 414 respondents in Barangay 2, 46 (11.1%) were positive for cholera as compared to 62 (15.1%) of the 410 respondents in Barangay 1. The following are the problems: (1) What is the point estimate for the difference in the proportion of positive for cholera in Barangays 1 and 2? (2) Construct the 90%, 95% and 99% confidence interval estimates of the difference in the proportions of positive in cholera in Barangays 1 and 2. (3) How does the confidence coefficient affect the interval estimate calculated?

Let us begin by solving the first problem. This problem is asking us to determine the point estimate for the difference between those who are positive for cholera in Barangays 1 and 2. The point estimate is determined by getting the difference between the two sample proportions, p_1 and p_2. The

point estimate is expressed as $(p_1 - p_2)$. We can assume that p_1 is the proportion of respondents who are positive for cholera in Barangay 1 and p_2 as the proportion of respondents who are positive for cholera in Barangay 2. Getting the difference will give us a value of 4%. We will obtain a different point estimate if we assign p_1 as the proportion of respondents who are positive for cholera in Barangay 2 and p_2 as the proportion of respondents who are positive for cholera in Barangay 1. In this case, the point estimate is -4%. Note that the negative sign just indicates that the proportion of respondents who are positive for cholera is higher in Barangay 1.

To solve the second problem, we are asked to construct the 90%, 95% and 99% confidence interval estimates of the difference in the proportions of the respondents who are positive for cholera in Barangays 1 and 2. Remember the z-deviate scores for the 90%, 95% and 99% confidence levels because these values will be substituted into the formula. The 90% confidence interval can be calculated using the formula below:

$$(p_1 - p_2) \pm 1.64 \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

The z-deviate score for a 90% two-tailed hypothesis is 1.64. This value has to be changed if the confidence level is 95% or 99%. Using the formula above, the confidence estimate will be -7.8% and -0.2%. The solution is shown below:

$$(11.1 - 15.1) \pm 1.64 \sqrt{\frac{11.1(88.9)}{414} + \frac{15.1(84.9)}{410}}$$

Try to solve for the 95% and the 99% confidence interval estimates. It is expected that the 95% confidence interval estimate will be -8.6% and 0.6% while the 99% confidence interval estimate will be -10.0% and 2.0%. For the third problem, you are asked how the confidence coefficient correlates with the interval estimate calculated. So, what do you think? If your answer is that the higher the confidence coefficient, the wider the confidence interval estimate becomes, then you got it right!

Hypothesis testing is not limited to testing a single proportion but can also be applied for two proportions similar to testing single means and the difference between two means. In the test of hypothesis for two proportions, the most commonly tested hypothesis for population proportions is:

$$H_0 : P1 = P2 \, ; \, (P1 - P2) = 0$$

Under the null hypothesis, $P_1 = P_2$, two possible conditions may arise:

a) $H_0 : P_1 = P_2 \; or \; P_1 - P_2 = 0$;

b) $H_0: P_1 = P_2 = P$, where *P* is the common value to which both P_1 and P_2 are equal.

The pooled estimate P indicated above is more precise than either one of the 2 distinct estimates. Because of the condition set in (b), the p_1 and p_2 are values that are both estimating *P*. We can combine the samples to derive a pooled estimate of *P* (*p*), which is more precise than either one of the two distinct estimates. The formulas are shown on the next page.

$$\bar{p} = \frac{p_1 + p_2}{n_1 + n_2}$$

The *z*-statistic is therefore calculated as:

$$z = \frac{p_1 - p_2}{\sqrt{\frac{\bar{p}\bar{q}}{n_1} + \frac{\bar{p}\bar{q}}{n_2}}}$$

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

The sample size requirement for the application of the previous equation is, that both n_1pq and n_2pq should be greater than or equal to 5.0. Let us look at an example. A nutritionist screened 465 males and 656 females and classified them as to whether they had normal or elevated blood uric acid (BUA) levels. Results showed that 143 males and 133 females had elevated BUA. Do these findings suggest that there is a higher proportion of males with elevated BUA? Set alpha to 0.10.

Given:

- n_1 (♂) = 465; p_1 (♂ w/ hyperuricemia) = 143/465 or 30.75%
- n_2 (♀) = 656; p_2 (♀ w/ hyperuricemia) = 133/656 or 20.27%

H_0: $p_1 = p_2$

H_A: $p_1 > p_2$

Level of Significance (α) = 0.10

Test statistic: *z*-test for 2 proportions

$$z = \frac{p_1 - p_2}{\sqrt{\frac{\bar{p}\bar{q}}{n_1} + \frac{\bar{p}\bar{q}}{n_2}}} \; where\ \bar{p} = \frac{p_1 + p_2}{n_1 + n_2}$$

Critical Region (α = 0.1): $\geq$ 1.28 (at the 'tail')

Calculations: p = (143 + 133)/(465 + 656) = 276/1121 = 0.2462 or (24.62%)

q = 100 – 24.62 = 75.38%

$$z = \frac{30.75 - 20.27}{\sqrt{\frac{(24.62)(75.38)}{465} + \frac{(24.62)(75.38)}{656}}} = \frac{10.48}{\sqrt{6.8201}} = 4.0129$$

Decision: Since 4.0129 is within the critical region, reject Ho.

"The proportion of males with hyperuricemia is greater than the proportion of females with hyperuricemia."

Take note, that the alternative hypothesis shows that proportion 1 is greater than proportion 2. The calculations and the hypothesis of going about it are shown above. Now, let us look at how the hypothesis testing is to be undertaken if we say that P1 is not equal to P2. The calculations and the steps are shown on the next page:

Given:

- **n_1 (♂) = 465; p_1 (♂ w/ hyperuricemia) = 143/465 or 30.75%**
- **n_2 (♀) = 656; p_2 (♀ w/ hyperuricemia) = 133/656 or 20.27%**

H_0: $p_1 = p_2$
H_A: $p_1 = p_2$
Level of Significance (α) = 0.10
Test statistic: *z*-test for 2 proportions

$$z = \frac{p_1 - p_2}{\sqrt{\frac{\bar{p}\bar{q}}{n_1} + \frac{\bar{p}\bar{q}}{n_2}}} \; where \; \bar{p} = \frac{p_1 + p_2}{n_1 + n_2}$$

Critical Region ($\alpha/2 = 0.05$): $z \geq 1.64$ and $z \leq -1.64$ (at the 'tails')
Calculations: p = (143 + 133)/(465 + 656) = 276/1121 = 0.2462 or (24.62%)
q = 100 – 24.62 = 75.38%

$$z = \frac{30.75 - 20.27}{\sqrt{\frac{(24.62)(75.38)}{465} + \frac{(24.62)(75.38)}{656}}} = \frac{10.48}{\sqrt{6.8201}} = 4.0129$$

Decision: Since 4.0129 is within the critical region, reject Ho.

"The proportion of males with hyperuricemia is not equal to the proportion of females with hyperuricemia. It is significantly higher."

At the same level of significance, a one-tailed hypothesis test has a higher chance of rejecting the null hypothesis compared to a two-tailed hypothesis test. In one-tailed tests of hypotheses, the possibility of rejecting the null hypothesis increases as you increase the value of α.

The Chi-Square Test

The chi-square test is another useful tool for the analysis of qualitative data. The chi-square distribution is the most frequently employed statistical technique for the analysis of count or frequency data. The test compares whether the observed and expected frequencies fall under the same category if the null hypothesis were true. The chi-square test has three types: the goodness of fit test, the test of independence, and the test of homogeneity. We will discuss the test of independence and test of homogeneity in more detail on Chapter 13: Investigating Relationships Between Variables.

LABORATORY EXERCISE 10

Analysis of Qualitative Data (30 points)

NAME:____ **DATE:**__________

SECTION:________________________ **INSTRUCTOR:**____________________

For each of the following problems, carry out the steps in hypothesis testing at the designated level of significance.

1. **A researcher conducted a study to examine the reasons why occupational therapists have left the field of occupational therapy. Her sample consisted of female certified occupational therapists who had left the profession either permanently or temporarily. Out of 696 participants who responded to the data-gathering survey, 63% had planned to take time off from their jobs to have and raise children. On the basis of these data, can we conclude that, in general, more than 60% of the subjects in the sampled population had planned to take time off to have and raise children? Let** $\alpha = 0.05$.

   a. **H_0 (1 point):**____

   b. **H_A (1 points):**____

   c. **Level of Significance:** $\alpha = 0.05$

   d. **Test statistic (2 points):**

   e. **Critical region (2 points):**

   ____

   f. **Calculation of the test statistic (5 points):**

   g. **Statistical Decision (2 points):**___________________________________

   h. **Conclusion (2 points):**____

2. **Research has suggested a high rate of alcoholism among patients with primary unipolar depression. A study further explored this possible relationship. In 210 families of females with primary unipolar major depression, they found that alcoholism was present in 89. Of 299 control families, alcoholism was present in 94. Do these data provide sufficient evidence for us to conclude that alcoholism is more likely to be present in families of subjects with unipolar depression? Let** $\alpha = 0.05$

   a. **H_0 (1 point):** ____

   b. **H_A (1 points):** ____

   c. **Level of Significance:** $\alpha = 0.05$

   d. **Test statistic (2 points):**

   e. **Critical region (2 points):**

   ____

   f. **Calculation of the test statistic (5 points):**

   g. **Statistical Decision (2 points):** ____

   h. **Conclusion (2 points):** ____

11 Analysis of Quantitative Data

In this section, emphasis is applied on the hypothesis testing of means. Recall that before we performed hypothesis testing, we initially estimated for population means by either calculating the point estimate or constructing the interval estimate. Remember that in point estimation, a single numerical value is used to estimate for the corresponding population parameter whereas in interval estimation, two numerical values define an interval ranging the degrees of confidence of the corresponding population parameter.

Hypothesis Testing: A Single Population Mean

In a test of hypothesis of the population mean, our objective is to determine whether the results obtained from the sample supports long established norms or whether it is consistent with what is claimed to be the existing population value. Just like in hypothesis testing, we start by defining the null and alternate hypotheses. The test statistic that we use depends on whether or not the population variance is known.

Known variance:

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

Unknown variance:

$$z = \frac{x - \mu}{s/\sqrt{n}} \; if \; n \geq 30$$

$$t = \frac{x - \mu}{s/\sqrt{n}} \; if \; n < 30$$

The critical region of the computed *z* value depends on the level of significance that we establish in the beginning. Meanwhile, the critical region of the *t* test depends on both the level of significance and the degrees of freedom. The computed *t* values, which fall within the critical region gives us reason to reject our null hypothesis. Let us look at an example of testing for the hypothesis of a single mean. The average number of persons per household for the whole country is based on a census made and was found out to be 5.6. If a random sample of 25 households in a survey showed a mean household size of 5.2 persons with a standard deviation of 1.56, does this result indicate that there has been a change in the mean household size in the Philippines since the last census? In this example, let us assume that we want to use a 90% level of significance.

H$_0$: $\mu = 5.6$
H$_A$: $\mu \neq 5.6$
Level of Significance (α) = 0.10

Test statistic: *t*-test

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

Critical Region: $t_{(\alpha/2,df)}$ with $df = n - 1 = 24$; $t_{(0.05,\ 24)} = 1.711$.
Therefore, the critical region is $t \geq 1.711$ and $t \leq -1.711$

Calculations:

$$t = \frac{5.2 - 5.6}{1.56/\sqrt{25}} = -1.28$$

Decision: Since the computed value does not fall within the critical region, do not reject H$_0$.
Conclusion: There is no sufficient evidence to conclude that there has been a change in the mean household size in the Philippines since the last census.

Hypothesis Testing: The Difference Between Two Population Means

Remember that two samples may either be independent or related. If we want to do a test of hypothesis for two means, these means must be obtained from two independent samples. To perform the test of hypothesis for two means, there is a need to know the population variances. If we come across a situation where the population variances for both groups are unknown, we can assume them to be equal before we can perform the test. The selection of our statistical test here depends on whether or not the population variances or standard deviation is known. If it is known, then we perform the *z* test and if it is unknown then we can assume that they are equal and perform the *t* test. Let us show an example. In this case, assume that the null and alternate hypothesis for this problem is:

Hypothesis:

H$_0$: $\mu_1 = \mu_2$ or $\mu_1 - \mu_2 = 0$
H$_A$: if two-tailed: simply a statement of inequality between the two groups
$\mu_1 \neq \mu_2$
if one-tailed: specify direction of relationship
$\mu_1 > \mu_2$ or $\mu_1 < \mu_2$

Known population variances for both groups

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

Unkown population variances for both groups but assumed to be equal

- in practice, population variances are rarely known
- when the population variance is unknown but the sample size is satisfactorily large for both groups (*n* ≥ 30), the sample variance (s) may be substituted for the population variance (σ) in the *z*-test.

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- when the sample size is small for both groups (*n* < 30), the independent *t*-test is the appropriate test statistic
  - population variances can be estimated by the sample variances s_1^2 and s_2^2
  - when variances are assumed to be equal, a pooled variances, s_p^2 can be computed

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

  - the test statistic follows the *t*-distribution with $df = n_1 + n_2 - 2$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Sample Problem:

A study aims to determine the association of salt intake to the blood pressure of persons aged 15 years and over. The mean systolic blood pressure (SBP) of 20 subjects with low salt diet was compared to that of an equal number of subjects with a high salt diet. The following data were obtained:

Statistics	High salt diet	Low salt diet
Mean SBP	138 mmHg	120 mmHg
SD of SBP	11.9 mmHg	12.2mmHg

Is there a difference between the mean SBP of subjects with high and low salt diets? Let $\alpha = 0.05$

Solution:

Hypotheses:
H_0: $\mu_H = \mu_L$ (There is no difference between the mean SBP of subjects with high and low salt diets.)
H_A: $\mu_H \neq \mu_L$ (There is a difference between the mean SBP of subjects with high and low salt diets.)

Level of significance: $\alpha = 0.05$
Test statistic:

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

We chose this test statistic because both population variances are unknown and both groups have a small sample size.

Critical region: $t_{\alpha/2,(20+20-2)} = t_{\alpha/2,38}$
$t \leq -2.021$ **or** $t \geq 2.021$

Calculation of the test statistic:

$$s_p^2 = \frac{s_1^2(n_1-1) + s_2^2(n_2-1)}{n_1+n_2-2} = \frac{(11.9)^2(20-1) + (12.2)^2(20-1)}{20+20-2} = 145.225$$

$$t = \frac{138 - 120}{\sqrt{145.225}\left(\sqrt{\frac{1}{20} + \frac{1}{20}}\right)} = \frac{18}{3.811} = 4.723$$

Statistical Decision: Since 4.72 > 2.021, reject H_0
Conclusion: There is a significant difference between the mean SBP of subjects with a high salt diet and those with a low salt diet.

Paired Comparisons

If we are performing a test of hypothesis for two means that are paired or matched, we must keep in mind that matching may only be achieved if we use the same subjects in the both samples such as in "before and after studies" or "brand X and brand Y" comparisons. The pairing of the subjects must be with respect to any extraneous variables in order to minimize the unwanted effects of these variables (an example of which is using twins in a study). Paired or matched sampling is used to overcome the difficulty imposed by extraneous differences between two groups when we are testing the differences between two means.

Hypotheses:

H_0: $\mu_d = d_0$
where *d* is the "difference"

Example:

	Sample				
Before	**X1**	**X2**	**X3**	**…**	**Xn**
After	**Y1**	**Y2**	**Y3**	**…**	**Yn**
Difference	**d1**	**d2**	**d3**	**…**	**Dn**

H_A: **if two tailed:** $\mu_d \neq d_0$
if one-tailed: $\mu_d > d_0$ **or** $\mu_d < d_0$

Test Statistic:

$$t = \frac{\bar{d} - \mu_d}{s_d / \sqrt{n}}$$

with $df = n - 1$

where:

- $\bar{d}$ = mean difference
- s_d = standard deviation of difference

$$= \sqrt{\frac{\sum d_i^2 - (\sum d_i)^2 / n}{n-1}}$$

- n = number of pairs

Critical region:

If two-tailed: $|t| \geq t_{(\alpha/2,\ df)}$

If one-tailed: $t \geq t_{(\alpha,\ df)}$ or $t \leq -t_{(\alpha,\ df)}$

Decision rule:

If two-tailed: reject H_0 if $|t| \geq t_{(\alpha/2,\ df)}$

If one-tailed: reject H_0 if $t \geq t_{(\alpha,\ df)}$ or if $t \leq -t_{(\alpha,\ df)}$

***Sample Problem*:**

The women's weight before and after the 12 weeks of treatment with a very low calorie diet (VLCD) are shown below. We wish to know if these data provide sufficient evidence to conclude that the treatment is effective in causing weight reduction in obese women.

Before	117.3	111.4	98.6	104.3	105.4	100.4	81.7	89.5	78.2
After	83.3	85.9	75.8	82.9	82.3	77.7	62.7	69.0	63.0

Hypotheses:

H_0: $\mu_B = \mu_A$ (There is no significant difference in the mean weights of women before and after the diet.)

H_A: $\mu_B > \mu_A$ (The mean weights of the women before the diet is greater than their mean weights after the diet.)

Level of significance: $\alpha = 0.05$ (if level of significance is not mentioned, assume at 5%)

Critical region: $t_{(\alpha,df)} = t_{(0.05,8)} = 1.860$; Reject H_0 if $t > 1.860$

Test statistic:

$$t = \frac{\bar{d} - \mu_d}{s_d / \sqrt{n}}$$

with $df = 9 - 1 = 8$

Calculation of the test statistic:

Before	117.3	111.4	98.6	104.3	105.4	100.4	81.7	89.5	78.2
After	83.3	85.9	75.8	82.9	82.3	77.7	62.7	69.0	63.0
Difference	34.0	25.5	22.8	21.4	23.1	22.7	19.0	20.5	15.5

Sum of differences: 22.70

Standard deviation of differences: 5.10

$$t = \frac{22.70}{5.10/\sqrt{9}} = 13.35$$

Statistical Decision: Since 13.35 > 1.86, reject H_0

Conclusion: The mean weight of the women before the diet is greater than their mean weights after the diet. We therefore conclude that the treatment is effective in causing weight reduction in obese women.

Hypothesis Testing with Microsoft Excel 2010

Here are some relevant functions to help you calculate test statistics with ease:

- Z.TEST – returns the one-tailed *p*-value of a *z*-test (ZTEST for Excel 2007 and earlier)
  - Syntax: =Z.TEST(array,X,sigma)
    - array – the array or range of data to which to test x
    - x – the value to test
    - Sigma (optional) – the population (known) standard deviation. If omitted, the sample standard deviation is used
    - **To do a 2-tailed *z*-test, use the following syntax: =2*MIN(Z.TEST(array,x,sigma),1-Z.TEST(array,x,sigma))**
  - **Use NORM.S.INV (NORMSINV for Excel 2007 and earlier) on the Z.TEST result to obtain the value of the test statistic.**

- **T.TEST – returns the probability associated with a Student's *t*-test (TTEST for Excel 2007 and earlier)**
  - **Syntax: =T.TEST(array1,array2,tails,type)**
    - **array1 – the first data set**
    - **array 2 – the second data set**
    - **tails – If tails = 1, uses the one-tailed *t*** distribution. If tails = 2, uses the two-tailed *t* distribution
    - type – the kind of *t*-test to perform.
      - If type = 1, performs a paired *t*-test
      - If type = 2, performs an independent *t*-test (variances assumed equal or homoscedastic)
      - If type = 3, performs an independent *t*-test (unequal variances or heteroscedastic)
  - **Use T.INV.2T (TINV for Excel 2007 and earlier) on the T.TEST result to obtain the value of the test statistic.**

LABORATORY EXERCISE 11

Analysis of Quantitative Data (60 points)

NAME:____ DATE:________________

SECTION:________________________________ INSTRUCTOR:__________________________

For each of the following problems, carry out the steps in hypothesis testing at the designated level of significance. When possible, use Microsoft Excel to calculate the value of the test statistic. Write the syntax to substitute for calculations by hand. Submit a copy of your spreadsheet to your instructor.

1. **Suppose it is known that the IQ scores of a certain population of adults are approximately normally distributed with a standard deviation of 15. A simple random sample of 25 adults drawn from this population had a mean IQ score of 105. On the basis of these data can we conclude that the mean IQ score for the population is not 100? Let the probability of committing a type I error be 0.05.**

   a. **H_0 (1 point):** ____

   b. **H_A (1 points):** ____

   c. **Level of Significance:** $\alpha = 0.05$

   d. **Test statistic (2 points):**

   e. **Critical region (2 points):**

   ____

   f. **Calculation of the test statistic (5 points):**

   g. **Statistical Decision (2 points):** ____

   h. **Conclusion (2 points):** ____

2. **A researcher conducted a study to examine prospectively collected data on gentamicin pharmacokinetics in three populations over 18 years of age: patients with acute leukemia, patients with other nonleukemic malignancies, and patients with no underlying malignancy or pathophysiology other than renal impairment known to alter gentamicin pharmacokinetics. Among other statistics reported by the investigators were a mean initial calculated creatinine clearance value of 59.1 with a standard deviation of 25.6 in a sample of 211 patients with malignancies other than leukemia. We wish to know if we may conclude that the mean for a population of similar subjects is less than 60. Let $\alpha = 0.10$.**

   a. **H_0 (1 point):** ____

   b. **H_A (1 points):** ____

   c. **Level of Significance:** $\alpha = 0.10$

   d. **Test statistic (2 points):**

   e. **Critical region (2 points):**

      ____

   f. **Calculation of the test statistic (5 points):**

   g. **Statistical Decision (2 points):** ____

   h. **Conclusion (2 points):** ____

3. **Does sensory deprivation have an effect on a person's alpha-wave frequency? Twenty volunteer subjects were randomly divided into two groups. Subjects in group A were subjected to a 10-day period of sensory deprivation, while subjects in group B served as controls. At the end of the experimental period the alpha-wave frequency components of the subjects' electroencephalograms were measured. The results were as follows:**

   Group A: 10.2, 9.5, 10.1, 10.0, 9.8, 10.9, 11.4, 10.8, 9.7, 10.4
   Group B: 11.0, 11.2, 10.1, 11.4, 11.7, 11.2, 10.8, 11.6, 10.9, 10.9

   a. **H_0 (1 point):** ____

   b. **H_A (1 points):** ____

   c. **Level of Significance:** $\alpha = 0.05$

   d. **Test statistic (2 points):**

   e. **Critical region (2 points):**

   ____

   f. **Calculation of the test statistic (5 points):**

   g. **Statistical Decision (2 points):** ____

   h. **Conclusion (2 points):** ____

4. A researcher conducted a study to test the hypotheses that weight loss in apneic patients results in decreases in upper airway critical pressure (*P*crit) and that these decreases are associated with reductions in apnea severity. The study participants were patients referred to the Johns Hopkins Sleep Disorder Center and in whom obstructive sleep apnea was newly diagnosed. Patients were invited to participate in either a weight loss program (experimental group) or a "usual care" program (control group). Among the data collected during the course of the study were the following before and after *P*crit (cm H_2O) scores for the weight-loss participants:

Before	-2.3	5.4	4.1	12.5	0.4	-0.6	2.7	2.7	-0.3	3.1	4.9	8.9	-1.5
After	-6.3	0.2	-5.1	6.6	-6.8	-6.9	-2.0	-6.6	-5.2	3.5	2.2	-1.5	-3.2

May we conclude on the basis of these data that the weight-loss program was effective in decreasing upper airway *P*crit? Let $\alpha = 0.10$

a. H_0 (1 point): ____

b. H_A (1 points): ____

c. Level of Significance: $\alpha = 0.10$

d. Test statistic (2 points):

e. Critical region (2 points):

____

f. Calculation of the test statistic (5 points):

g. Statistical Decision (2 points): ____

h. Conclusion (2 points): ____

12 Sample Size and Power

It is nearly impossible to examine all members of a population in order to come up with population parameters, which is why random sampling and sampling distributions are essential statistical analysis. Whereas the statistical analysis to be performed is highly dependent on the study design, it is also reliant on the sample size. A statistical test can be deemed powerful if the sample size is sufficient.

- Sample Size – the number of elementary units that must be examined in order to arrive at data that may be truly representative of the population. The researcher's discretion must be based on pragmatism, that is, a sample size that is too large may lead to excessive cost, while a sample size that is too small may lead to erroneous results.
- Power – the power of any statistical test is a measure of the sensitivity of the statistical test, whether it can really determine the difference between two means, two proportions, etc. It is calculated as Power = 1 – β, where beta is Type II error (error of not rejecting a null hypothesis that is false). Most often, power is assigned at 80% or 0.80 in the same manner that the significance level (α) is usually 5% or 0.05.

In research proposals, the study design must include the calculated sample size, assigned significance level (α) and occasionally, the power. The method of sample size calculation depends on the type of statistical analysis to be used, the type of variable to be examined, parameters or published data, and even the actual study design. The sample size can be calculated through a (1) formula, (2) published sample size tables or (3) software.

Sample Size Calculation through Formulae

Based on Whitley, E. and Ball, J. (2002) Statistics review 4: Sample size calculations. *Critical Care, Vol.* 6(4): 335-341. Available online: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC137461/

- For Hypothesis Testing for the Difference Between Two Means

$$n = \frac{2}{d^2} \times c_{p,power}$$

Where: n = number of subjects required in each group

d = standardized difference = $\frac{target\ difference}{standard\ deviation}$ (the target difference is assigned by the researcher and the standard deviation is based on previous similar studies)

$c_{p,power}$ = a constant defined by the values chosen for the *P* value and power

Table 12-1. Commonly used values for $c_{p,power}$

P value	Power (%)			
	50	80	90	95
0.05	3.8	7.9	10.5	13.0
0.01	6.6	11.7	14.9	17.8

*Table from Whitley and Ball (2002).

- For Hypothesis Testing for the Difference Between Two Proportions

$$n = \frac{[p_1(1-p_1)+p_2(1-p_2)]}{(p_1-p_2)^2} \times c_{p.power}$$

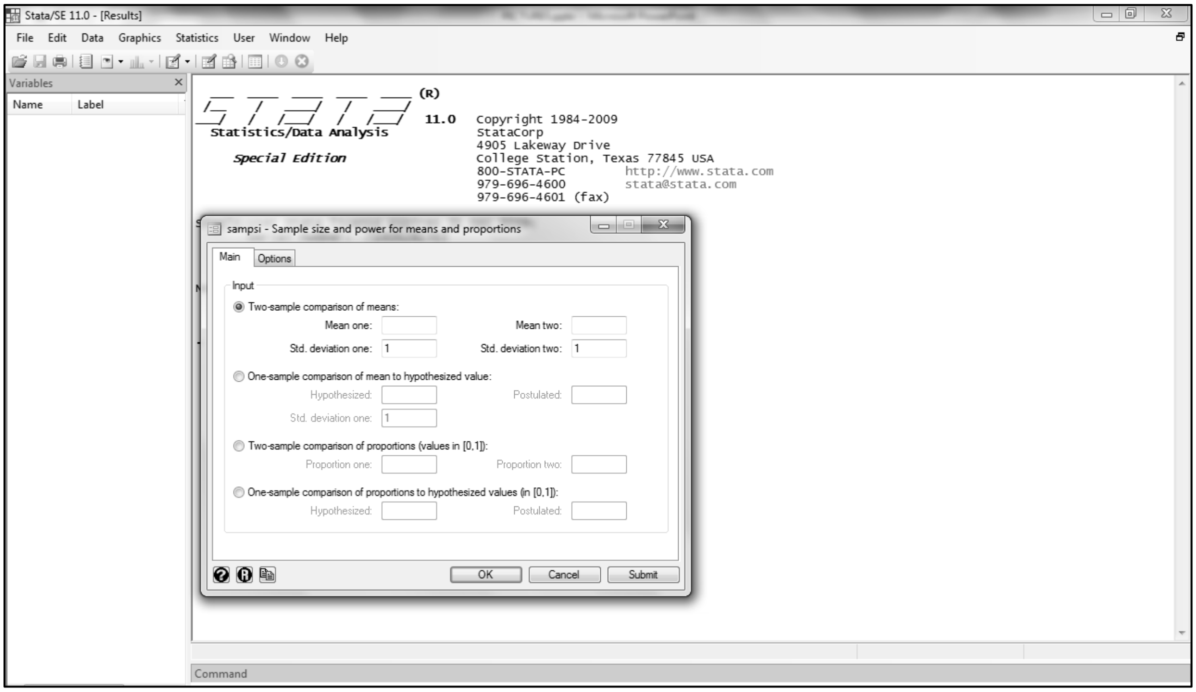
Where: n = number of subjects required in each group
$c_{p,power}$ = a constant defined by the values chosen for the *P* value and power

Use of Sample Size Tables

The use of published sample size tables is especially important in surveys or observational studies. The use of sample size tables may be based on the total population, significance level, p value or standardized difference assigned by the researcher. The World Health Organization has published a public document entitled "Tables of Minumum Sample Size" available for download at http://whqlibdoc.who.int/publications/9241544058_%28p23-p80%29.pdf

Use of Software

Sample size calculation with power is usually performed by researchers using statistical software such as Stata (licensed) and the free, java-run, online and downloadable Piface (http://homepage.cs.uiowa.edu/~rlenth/Power/).


Stata/SE 11.0 - [Results]
File Edit Data Graphics Statistics User Window Help
Variables
Name Label
Command

```
 ___  ____  ____  ____  ____ (R)
 Statistics/Data Analysis      11.0   Copyright 1984-2009
                                      StataCorp
   Special Edition                    4905 Lakeway Drive
                                      College Station, Texas 77845 USA
                                      800-STATA-PC        http://www.stata.com
                                      979-696-4600        stata@stata.com
                                      979-696-4601 (fax)
```

sampsi - Sample size and power for means and proportions

Main | Options

Input

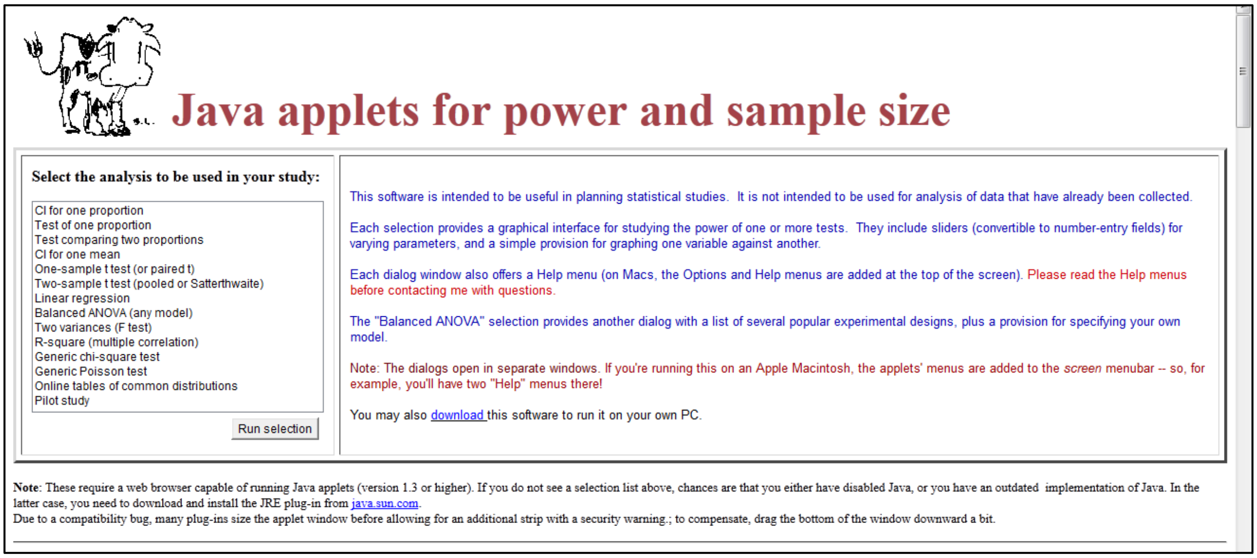
● Two-sample comparison of means:
Mean one: Mean two:
Std. deviation one: 1 Std. deviation two: 1

○ One-sample comparison of mean to hypothesized value:
Hypothesized: Postulated:
Std. deviation one: 1

○ Two-sample comparison of proportions (values in [0,1]):
Proportion one: Proportion two:

○ One-sample comparison of proportions to hypothesized values (in [0,1]):
Hypothesized: Postulated:

Figure 12-1. Screenshot of sample size and power calculator in Stata11.



Java applets for power and sample size

Select the analysis to be used in your study:

CI for one proportion
Test of one proportion
Test comparing two proportions
CI for one mean
One-sample t test (or paired t)
Two-sample t test (pooled or Satterthwaite)
Linear regression
Balanced ANOVA (any model)
Two variances (F test)
R-square (multiple correlation)
Generic chi-square test
Generic Poisson test
Online tables of common distributions
Pilot study

This software is intended to be useful in planning statistical studies. It is not intended to be used for analysis of data that have already been collected.

Each selection provides a graphical interface for studying the power of one or more tests. They include sliders (convertible to number-entry fields) for varying parameters, and a simple provision for graphing one variable against another.

Each dialog window also offers a Help menu (on Macs, the Options and Help menus are added at the top of the screen). Please read the Help menus before contacting me with questions.

The "Balanced ANOVA" selection provides another dialog with a list of several popular experimental designs, plus a provision for specifying your own model.

Note: The dialogs open in separate windows. If you're running this on an Apple Macintosh, the applets' menus are added to the *screen* menubar -- so, for example, you'll have two "Help" menus there!

You may also download this software to run it on your own PC.

Note: These require a web browser capable of running Java applets (version 1.3 or higher). If you do not see a selection list above, chances are that you either have disabled Java, or you have an outdated implementation of Java. In the latter case, you need to download and install the JRE plug-in from java.sun.com.
Due to a compatibility bug, many plug-ins size the applet window before allowing for an additional strip with a security warning.; to compensate, drag the bottom of the window downward a bit.

Figure 12-2. Screenshot of Piface homepage.

Whichever method a student researcher prefers to use for sample size calculation, it is important to remember that the calculated sample size is usually a minimum sample size and that it should not limit the researcher from using more than the calculated sample size. Typically, in the case of surveys on human subjects, an additional number of samples may be studied to account for possible non-response. As long as the sample size is sufficient and is at least equal to the sample size calculated, then the researcher can be confident in the implementation of his/her study design.

13 Investigating Relationships Between Variables

In this section, we are interested in (1) establishing association between two qualitative variables; (2) establishing correlation between two quantitative variables; and (3) predict whether a relationship exists between two or more variables. This section will introduce you to the chi square test of association to answer our first interest; the correlation coefficient for the second; and the regression analysis for the last interest.

Why do we have to investigate for relationships?

We want to investigate the relationships of our variables because we want to know if they are associated or correlated with each other. We want to determine their strength of association and we want to predict a given outcome based on the relationship of the variables.

The choice of our statistical tool depends on the number of variables we are testing, the type of variables we are investigating, and what the intention of the researcher is. There are two approaches that we can use to show relationships between variables: the graphical approach and the statistical test. In the graphical approach, we are interested in demonstrating the nature of the relationship between the variables. Here we can use several graphical representations to show the relationship of the variables. For example, when we want to show the relationship between two variables that are both qualitative, we can construct a comparative bar graph. On the other hand, if we want to show the relationship of two variables that are both quantitative, we can construct a scatter plot diagram. There are endless possibilities of graphical approaches that we can use to depict the relationship of our variables. Usually, the graphical approach is merely a descriptive manner of showing the relationship between variables. It does not put one in position to make generalizations based on the graphs showed. The graphical approach is oftentimes used as the first step or guide in choosing the best statistical tool to analyze the data.

Chi-Square Test

In this section, we will discuss the chi-square test of independence (also called the chi-square test of association) as well as the chi-square test of homogeneity. The test of independence is useful for testing the relationship between two categorical variables while the test of homogeneity allows the investigator to determine whether several groups of samples are homogeneous with respect to a particular classification. Both tests use the same mathematical concepts and are carried out by following the steps in hypothesis testing. However, the first step in performing a chi-square test is to construct the contingency table.

The contingency table

What is a contingency table? A contingency table is a cross-tabulation consisting of *r* number of rows and *c* number of columns and gives frequency of observations that fall under each category. The table below is an example of a contingency table. In this table, note that the O_{11} is the observed frequency in the C_1 category of the variable of interest of the sample stratum 1. The n_1 indicated on the table signifies the row totals while the $n_{\cdot 1}$ is the column total.

Table 10-4. Sample Contingency Table

Samples	Categories			Total
	C_1	C_2	C_3	
S_1	O_{11}	O_{12}	O_{1n}	n_1
S_2	O_{21}	O_{22}	O_{2n}	n_2
S_n	O_{n1}	O_{n2}	O_{nn}	n_n
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot n}$	$n_{\cdot\cdot} = n$

In the contingency table, take note that the categories are assumed to be exhaustive as well as exclusive. The samples of the strata are presumed to be independent of one another. If we are to apply the chi-square test to test a hypothesis, we must follow the steps in hypothesis testing as we have done previously. We still start the hypothesis testing by stating the null and alternative hypotheses. In this case, let us look at an example of how the null and the alternative hypotheses are formulated.

H_0: The proportion of elements falling in each category of the variable of interest is the same for all strata.
H_A: There are differences between strata in the proportion of elements falling in each category of the variable.

After the hypotheses have been stated, the next step is to set the level of significance. Then, identify the appropriate test statistic. In this case, let us assume that we are going to use the chi-square test as our test statistic. Take note that the shape of the chi-square distribution depends on the number of degrees of freedom (*df*) where $df = (r - 1)(c - 1)$. We then proceed to determining the critical region. This is determined by the number of the degrees of freedom and the level of significance. The chi-square values are obtained from the chi-square distribution table listed as Table E at the end of this module.

The test statistic for the chi-square test is:

$$X^2 = \sum \left[\frac{(O_i - E_i)^2}{E_i}\right]$$

with $(r - 1)(c - 1)$ degrees of freedom

where:

- O_i is the observed frequency for the i^{th} category; and
- E_i is the observed frequency given H_0 is true for the i^{th} category

- The quantity X^2 is a measure of the extent to which pairs of observed and expected frequencies agree.
- When there is close agreement between observed and expected frequencies, it is small, and when the agreement is poor, it is large.
- Only a sufficiently large value of X^2 will cause rejection of the H_0.

The Decision Rule

- **The computed value of X^2 is compared with the tabulated value of χ^2 with k – r degrees of freedom.**
- **The decision rule is: Reject H_0 if X^2 is greater than or equal to the tabulated χ^2 for the chosen value of α.**

Calculating the expected frequencies

We have learned in lesson 5 (basic concepts in probability) that if two events are independent, the probability of their joint occurrence is equal to the product of their individual probabilities. For example, under the assumption of independence, we calculate the probability that one of the *o* observations represented in Table 10-4 will be counted in row 1 and column 1 (that is, cell S_1C_1) of the table by multiplying the probability that the subject will be counted in row 1 by the probability that the subject will be counted in column 1. Using the notation of the table, the desired calculation is

$$\left(\frac{n_1}{n}\right)\left(\frac{n_{\cdot 1}}{n}\right)$$

To obtain the expected frequency, we multiply this probability by the total number of subjects, *n*. Therefore, the expected frequency for cell S_1C_1 is given by

$$\left(\frac{n_1}{n}\right)\left(\frac{n_{\cdot 1}}{n}\right)(n)$$

Since the *n* in one of the denominators cancels into numerator *n*, this expression reduces to

$$\frac{(n_1)(n_{\cdot 1})}{n}$$

In general, to obtain the expected frequency for a given cell, multiply the total of the row in which the cell is located by the total of the column in which the cell is located and divide the product by the grand total.

What is the rationale for the computation of expected frequencies?

The rationale on the computation of expected frequencies lies on the assumption that populations are represented in a contingency table that is homogeneous with respect to the variable of interest we are studying. It also depends on the expected frequency in the i^{th} sample that is obtained by multiplying the pooled estimate by the total number of subjects in the sample that we have. Remember that the data in a contingency table is only applicable if the expected frequencies are sufficiently large. In a 2 x 2 table the requirement is for all expected frequencies to be greater than or equal to 5. If it does not satisfy the requirement, then the chi-square test cannot be used. Instead, one can resort to the use of the Fisher's exact probability test. In the case of larger tables, the requirement is for all expected frequencies to be greater than or equal to 1 and not more than 20% of the cells should have expected frequencies less than 5. Again, if the condition is not met the researcher must merge the adjacent categories to meet the condition before the chi-square test can be applied.

Finding the critical value of X^2

The tabulated values of the chi-square distribution are provided in Table E at the end of this module. Note that that in the chi-square test, α is not divided by two. To obtain the desired critical value, simply align the value of α that you have set with its corresponding degrees of freedom. Alternatively, you may use Microsoft Excel 2010's *CHISQ.INV.RT* function which returns the inverse of the right-tailed probability of the chi-square distribution. The appropriate syntax is =CHISQ.INV.RT(probability,deg_freedom) where probability represents α and deg_freedom represents the degrees of freedom given by $(r - 1)(c - 1)$. For those using earlier versions of Excel, the analogous function is CHIINV which uses similar arguments.

The chi-square test of independence

The most frequently used chi-square test, it tests the null hypothesis that two criteria of classification, when applied to the same set of entities, are independent.

- Example: If the socioeconomic status and area of residence of the inhabitants of a certain city are independent, we would expect to find the same proportion of families in the low, medium, and high socioeconomic groups in all areas of the city.
- We say that two criteria of classification are independent if the distribution of one criterion is the same no matter what the distribution of the other criterion.

Sample Problem

The purpose of a study by Vermund et al. (1991) was to investigate the hypothesis that HIV-infected women who are also infected with human papillomavirus (HPV), detected by molecular hybridization, are more likely to have cervical cytologic abnormalities than are women with only one or neither virus. The data shown in Table 9.2 were reported by the investigators. We wish to know if we may conclude that there is a relationship between HPV status and stage of HIV infection.

Table 10-5. HPV status and stage of infection among 96 women

HPV	HIV			Total
	Seropositive, Symptomatic	Seropositive, Asymptomatic	Seronegative	
Positive	23	4	10	37
Negative	10	14	35	59
Total	33	18	45	96

SOURCE: Sten H. Vermund, Karen F. Kelley, Robert S. Klein, Anat R. Feingold, Klaus Schreiber, Gary Munk, and Robert D. Burk, "High Risk of Human Papillomavirus Infection and Cervical Squamous Intraepithelial Lesions Among Women with Symptomatic Human Immunodeficiency Virus Infection," *American Journal of Obstetrics and Gynecology*, 165 (1991), 392 – 400, as printed in *Biostatistics*: *A Foundation for Analysis in The Health Sciences* 6e by Wayne W. Daniel (1995)

Solution:

Hypotheses:

H_0: HPV status and age of HIV infection are independent
H_A: The two variables are not independent

Significance Level: α **= 0.05**

Test Statistic:

$$X^2 = \sum \left[\frac{(O_i - E_i)^2}{E_i} \right]$$

With (r – 1)(c – 1) degrees of freedom = (2 – 1)(3 – 1) = 2

Critical Region:

$X^2 \geq \chi^2_{\alpha,\, df = (r-1)(c-1)}$

Reject H_0 if the computed value of X^2 is greater than or equal to 5.991

Calculation of the test statistic:

$$= \frac{(23 - 12.72)^2}{12.72} + \frac{(4 - 6.94)^2}{6.94} + \cdots \frac{(35 - 27.66)^2}{27.66}$$
$$= 8.30805 + 1.24548 + \cdots 1.94778$$
$$= 20.60081$$

Statistical Decision: Reject H_0 since 20.60081 > 5.991

Conclusion: We conclude that H_0 is false, and that there is a relationship between HPV status and stage of HIV infection.

The chi-square test of homogeneity

The chi-square test of homogeneity finds out whether two or more populations have the same proportions for the different categories of another variable. When only two populations are considered and the variable of interest only has two categories, it can be used interchangeably with the *z* test for two proportions. If this is the case, the data is usually casted in a 2x2 contingency table.

In the chi-square test for independence, the total sample is assumed to have been drawn before the observations were classified according to the two criteria of classification. That is, the number of observations falling into each cell was determined after the sample was drawn. Consequently, the row and column totals are chance quantities not under the control of the researcher. We may think of the sample drawn under these conditions as a single sample drawn from a single population. However, there may be instances when either the row or column totals are under control of the researcher. This means that the researcher may specify that independent samples be drawn from each of several populations. In the case of the test of homogeneity, one set of the marginal totals is said to be *fixed* while the other set, corresponding to the criterion of classification applied to the samples, is said to be *random*. The test of independence and test of homogeneity not only involve different sampling methods but also lead to different questions and null hypotheses. The test of independence is concerned with the question: "*Are two criteria of classification independent?*" while the test of homogeneity is concerned with the question: "*Are the samples drawn from populations that are homogeneous with respect to some criterion of classification?*" In the test of homogeneity, the null hypothesis states that the samples are drawn from the same population. Despite these differences in

concept and sampling method, the two tests are mathematically identical as we will demonstrate in the following example:

***Sample Problem*:**

Kodama et al. studied the relationship between age and several prognostic factors in squamous cell carcinoma of the cervix. Among the data collected were the frequencies of histologic cell types in four age groups. The results are shown in Table 10-6. We want to know if the populations represented by the four age-group samples are not homogeneous with respect to cell type.

Table 10-6. Observed and expected frequencies* of various cell types classified according to age group

*Expected frequencies enclosed in parentheses

		Cell Type		
Age Group (years)	Number of Patients	Large Cell Nonkeratinizing	Keratinizing	Small Cell Nonkeratinizing
30-39	34	18 (19.59)	7 (8.86)	9 (5.55)
40-49	97	56 (55.90)	29 (25.27)	12 (15.83)
50-59	144	83 (82.99)	38 (37.52)	23 (23.49)
60-69	105	62 (60.51)	25 (27.36)	18 (17.13)
Total	380	219	99	62

Solution:

Hypotheses:

H_0: The four populations are homogeneous with respect to cell type.

H_A: The four populations are not homogeneous with respect to cell type.

Significance Level: $\alpha = 0.05$

Test Statistic:

$$X^2 = \sum\left[\frac{(O_i - E_i)^2}{E_i}\right]$$

With $(r-1)(c-1)$ degrees of freedom = $(4 - 1)(3 - 1) = 6$

Critical Region:

$X^2 \geq \chi^2_{\alpha,\, df = (r-1)(c-1)}$

Reject H_0 if the computed value of X^2 is greater than or equal to 12.592

Calculation of the test statistic:

$$= \frac{(18-19.59)^2}{19.59} + \frac{(7-8.86)^2}{8.86} + \cdots \frac{(18-17.13)^2}{17.13}$$

$$= 0.12905 + 0.39047 + \cdots 0.04419$$

$$= 4.444$$

Statistical Decision: We are unable to reject H_0 since 4.444 < 12.592

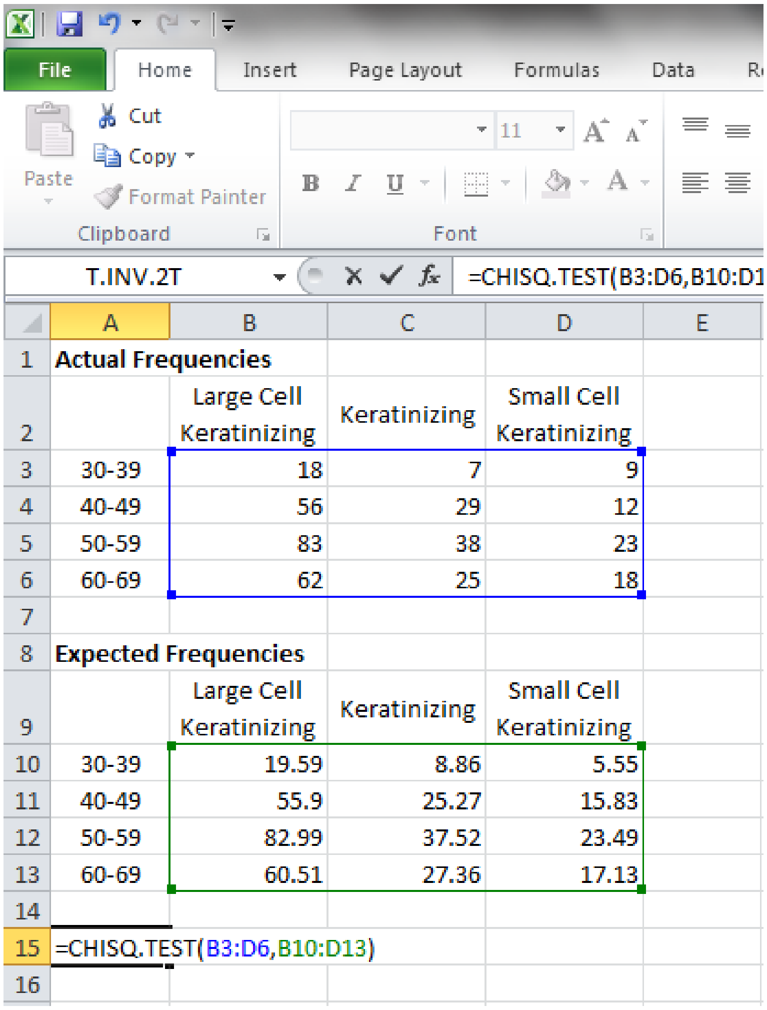
Conclusion: We conclude that the four populations may be homogeneous with respect to cell type.

Calculating the value of X^2 using Microsoft Excel

3. Arrange the data into arrays where the actual or observed frequencies are separated from the expected frequencies.

4. If using Microsoft Excel 2010, use the CHISQ.TEST function with the syntax =CHISQ.TEST(actual_range,expected_range). The analogous function for earlier versions of excel is CHITEST with the same form of arguments.

5. After typing "=CHISQ.TEST(" [without quotation marks] into an empty cell, select the range of actual frequencies, type "," then do the same for the expected frequencies. Finish the syntax by typing ")" and press Enter.



=CHISQ.TEST(B3:D6,B10:D1

	A	B	C	D	E
1	Actual Frequencies				
2		Large Cell Keratinizing	Keratinizing	Small Cell Keratinizing	
3	30-39	18	7	9	
4	40-49	56	29	12	
5	50-59	83	38	23	
6	60-69	62	25	18	
7					
8	Expected Frequencies				
9		Large Cell Keratinizing	Keratinizing	Small Cell Keratinizing	
10	30-39	19.59	8.86	5.55	
11	40-49	55.9	25.27	15.83	
12	50-59	82.99	37.52	23.49	
13	60-69	60.51	27.36	17.13	
14					
15	=CHISQ.TEST(B3:D6,B10:D13)				
16					

6. The CHISQ.TEST function returns the probability value (*p*). You may directly compare this to α to make your statistical decision. To get the value of the X^2 statistic, use the CHISQ.INV.RT function (CHIINV for Excel 2007 and earlier).

Correlation Analysis

This test measures the strength and direction of the relationship between two or more variables. No distinction between the two variables like on how they vary jointly is determined by this test. This test uses a correlation coefficient to show both the strength and the direction of the relationship. If the data is parametric, the Pearson's correlation coefficient is normally used while if the data is non-parametric, the Spearman's rank correlation coefficient should be used.

What are the assumptions of the Pearson's correlation coefficient?

The Pearson's correlation coefficient assumes that for each value of the variable X, the corresponding sub-population of values for the variable Y is normally distributed. It also assumes that for each value of the variable Y, the corresponding sub-population of values for the variable X is normally distributed. The joint distribution of the variables X and Y is also assumed to be normally distributed.

The value of our Pearson's correlation coefficient *r* or rho ranges from -1 to +1. The positive and negative signs show the direction of the relationship. The minimum absolute value is 0 while the maximum absolute value is 1. If *r* or the rho value is 0, it means that no relationship exists between the variables of interest. Meanwhile, if the value is nearer the absolute value of 1, then the stronger the relationship that exists between the variables of interest. A positive sign indicates a direct relationship while a negative sign indicates an inverse relationship. For example, *r*=0.72 and *r*=-0.72 have the same magnitude or strength of relationship. Both values only differ in the direction of the relationship. The Pearson's correlation computational formula is shown below:

$$r = \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{n\sum X^2 - (\sum X)^2 \cdot n\sum Y^2 - (\sum Y)^2}}$$

The strength and direction of the relationship are expressed by means of a correlation coefficient which is mathematically defined as:

$$r = \frac{S_{XY}}{S_X S_Y} = \frac{SCP}{\sqrt{(SSX)(SSY)}}$$

The sum of cross products of deviations

$$SCP = \sum (X_i - \overline{X})(Y_i - \overline{Y}) = \sum X_i Y_i - \frac{(\sum X_i)(\sum Y_i)}{n}$$

The sum of squared deviations for X

$$SSX = \sum (X_i - \overline{X})^2 = \sum X_i^2 - \frac{(\sum X_i)^2}{n}$$

The sum of squared deviations for Y

$$SSY = \sum (Y_i - \overline{Y})^2 = \sum Y_i^2 - \frac{(\sum Y_i)^2}{n}$$

The Pearson's correlation coefficient r

$$r = \frac{SCP}{\sqrt{(SSX)(SSY)}}$$

$$SSX = \sum X_i^2 - \frac{\left(\sum X_i\right)^2}{n}$$

$$SSY = \sum Y_i^2 - \frac{(\sum Y_i)^2}{n}$$

$$SCP = \sum X_i Y_i - \frac{\sum X_i \sum Y_i}{n}$$

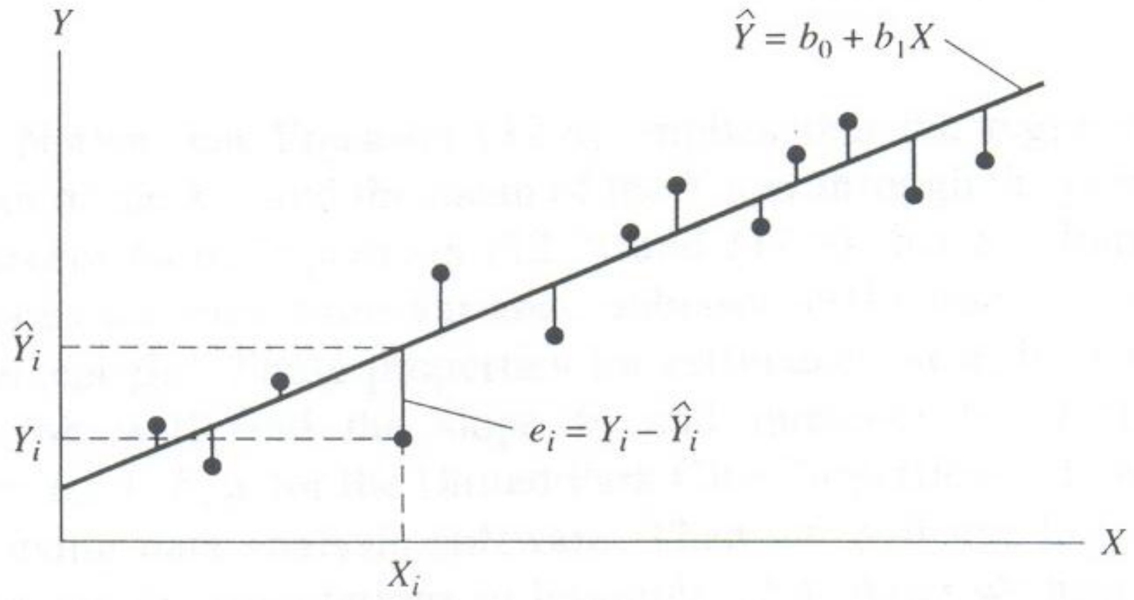
Regression Analysis

This test depends on a dependent variable being affected by an error-free independent variable. This test is primarily concerned with using the relationship between two variables for the purpose of predicting one variable based on the knowledge of the other. Remember that correlation analysis is primarily concerned with discovering whether or not a relationship exists in the first place, and then specifying the strength and direction of the relationship. In regression analysis, we can still determine whether a relationship exists, specify the strength and the direction of the relationship, and in addition, predict the outcome of the variable with the knowledge of the other variable. In other literatures, this test is also referred to as the least squares method. Regression analysis can be performed either as a simple regression analysis or as a multiple regression analysis. In simple regression analysis, we have a single Y and a single X, whereas in multiple regression analysis, we have a single Y and multiple X variables. The simple linear regression equation is given below:

$$\hat{Y} = b_0 + b_1 X$$

Where X = given data
b_0 = intercept of the regression line
b_1 = slope of the regression line

Graphically, this is expressed in the figure below:



In regression analysis, one can also perform a linear regression analysis or a non-linear regression analysis. In linear regression analysis, we assume that for every incremental change in Y, there is a corresponding incremental change in X whereas in non-linear regression analysis, we assume that for every incremental change in Y, there is no corresponding incremental change in X, or that the change in X may not be a linear increment change as expressed by our Y variable.

The coefficient of regression can be calculated using the formulas indicated below:

$$\boxed{b_1 = \frac{SCP}{SSX}} \qquad \boxed{b_0 = \overline{Y} - b_1\overline{X}}$$

$$\boxed{SCP = \sum(X_i - \overline{X})(Y_i - \overline{Y}) = \sum X_iY_i - \frac{(\sum X_i)(\sum Y_i)}{n}}$$

$$\boxed{SSX = \sum(X_i - \overline{X})^2 = \sum X_i^2 - \frac{(\sum X_i)^2}{n}}$$

$$\boxed{b_1 = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sum(X_i - \overline{X})^2}}$$

The coefficient of determination is a measure that is commonly used to describe how well the sample regression line fits the observed data. This is expressed in the formula below:

$$R^2 = \frac{SSR}{SST} = \frac{b_1^2 \sum (X_i - \overline{X})^2}{\sum (Y_i - \overline{Y})^2}$$

The range of the coefficient of determination is: $0 \leq R^2 \leq 1$. A value of 0 indicates the poorest fit while a value closest to 1 indicates the best fit for our regression model.

Analysis of Variance

This test allows us to determine whether significant differences exist between three or more groups. In this section, we will only cover the one-way ANOVA test. If we are to do a one-way ANOVA, there should be more than two levels of our single independent variable. Remember that in performing this test, the variables need to be grouped and the variable must consist of a number of levels. After grouping the independent variables, we can now set the level of significance, sample size, and hypotheses. The null hypothesis here indicates that any observed difference between three or more groups will be attributable to random sampling errors whereas in the alternate hypothesis it indicates that the difference is not attributable to random sampling errors.

The assumptions in performing a one-way ANOVA are the following:

1. Samples should be independent.
2. Each of the *k* populations should be normal
3. The *k* samples should have equal variances

To evaluate the $F_{observed}$ with the F distribution, we need to remember that the distribution of F ratios is not normally distributed as it is positively skewed. The F distribution depends on the number of degrees of freedom in the numerator and the denominator. The $F_{critical}$ is the F value that must be equalled or exceeded to classify a difference among the group means as statistically significant. The $F_{critical}$ depends on the degrees of freedom in the numerator and the denominator and the chosen level of significance, all of which are listed in Table F at the end of this module.

If we want to compute for the ANOVA F statistic manually, an example of performing the test is shown on the next page:

			WITHIN		**BETWEEN**	
			difference:		difference	
		group	data - group mean		group mean - overall mean	
data	group	mean	plain	squared	plain	squared
5.3	1	6.00	-0.70	0.490	-0.4	0.194
6.0	1	6.00	0.00	0.000	-0.4	0.194
6.7	1	6.00	0.70	0.490	-0.4	0.194
5.5	2	5.95	-0.45	0.203	-0.5	0.240
6.2	2	5.95	0.25	0.063	-0.5	0.240
6.4	2	5.95	0.45	0.203	-0.5	0.240
5.7	2	5.95	-0.25	0.063	-0.5	0.240
7.5	3	7.53	-0.03	0.001	1.1	1.188
7.2	3	7.53	-0.33	0.109	1.1	1.188
7.9	3	7.53	0.37	0.137	1.1	1.188
TOTAL				1.757		5.106
TOTAL/df				**0.25095714**		**2.55275**

Overall mean: 6.44 **F = 2.55275/0.25095714 = 10.172**

Once the ANOVA indicates that the groups do not all have the same means, we can compare them two by two by performing a multiple comparisons test or *post hoc* test. A post hoc test determines which specific pairs of means are significantly different. It is a follow up test that is only performed once we have determined a significant ANOVA. The post hoc test is like a collection of little *t*-tests but with control over the overall type I error. There are different types of post hoc tests that one can use. Examples include the Bonferroni procedure, Duncan multiple range test, Dunnett's multiple comparison test, Newman–Keuls test, Scheffe's test, and Tukey's test. No single test is found to be best in all situations, and a major difference between them lies in the manner in which they control the increase in Type I error due to multiple testing. Remember that in hypothesis testing, using the same statistical test repeatedly results to an increase in type I error. The post hoc test may not have as much power as the omnibus test, which is that of the ANOVA. The purpose of this test is just to identify the locus of the effect or which means are significantly different. Among the post hoc tests, the Tukey's HSD (honestly significant difference) test is the commonly used. The Tukey's HSD test uses the formula below:

$$HSD = q\sqrt{\frac{MS_{within}}{n}}$$

The HSD determines the magnitude of the mean difference that must exist to claim that the levels are significantly different. The *q* is the studentized ranged statistic that depends on the number of levels to be compared and the df_{within} and the level of significance. The *q* value is obtained from a table. The Mean Square within is taken from the ANOVA summary table while *n* is the number of subjects in each group.

Table A. Summary of Statistical Tests

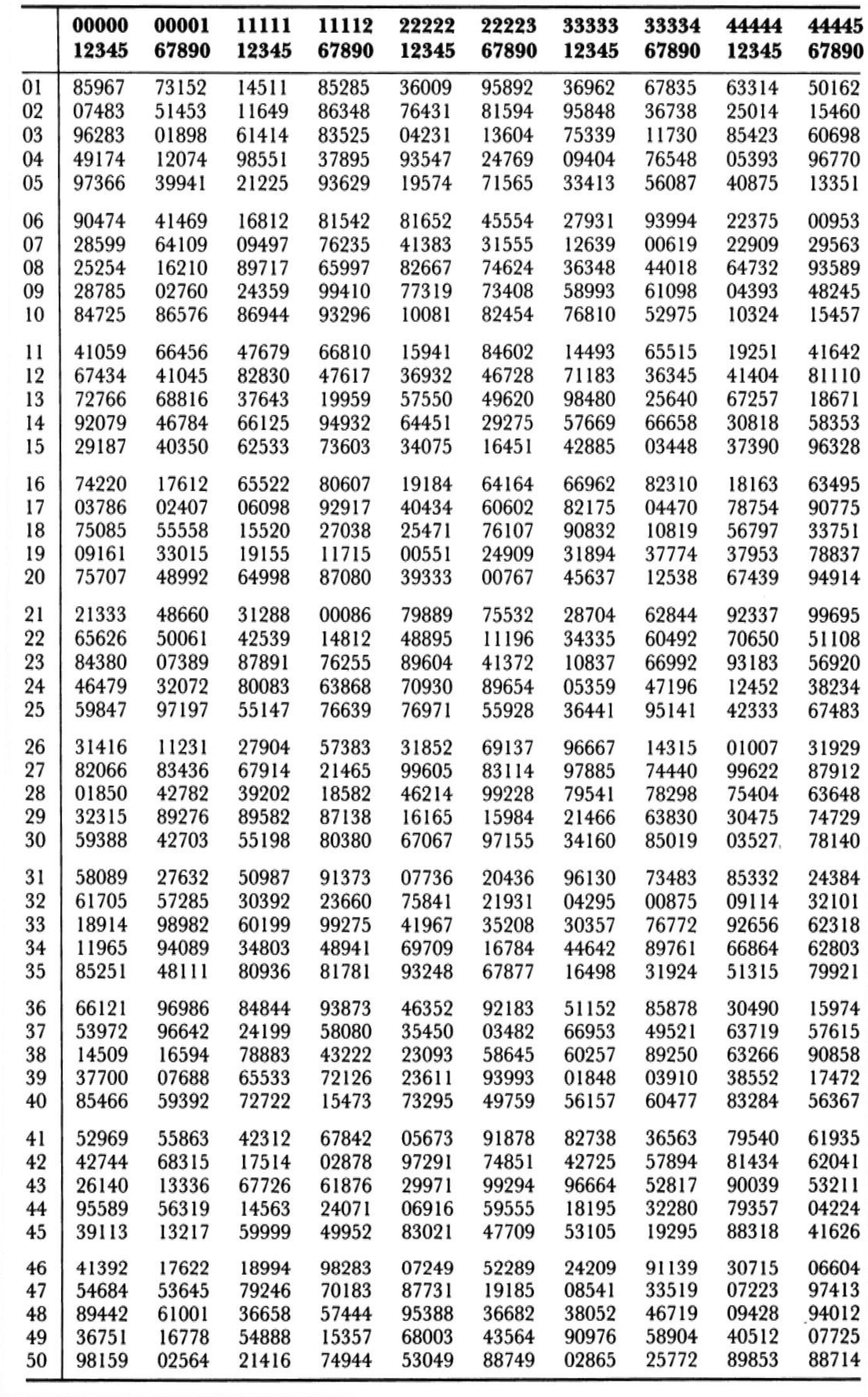
Goal	Quantitative (ratio)	Quantitative (rank, score, numerical)	Qualitative (nominal, binomial)	Survival Time
Describe one group	Mean, Standard deviation	Median, Interquartile range	Proportion	Kaplan-Meier survival curve

Compare one group to a population value	One sample *z*-test (known variance);
	One sample *z*-test (unknown variance & n≥30);
	One-sample t test (unknown variance, n<30)
		Wilcoxon test	Chi-square test or Binomial test	

Compare two independent groups	Two sample *z*-test (known variance);
	Two sample *z*-test (unknown variance, n≥30);
	Independent *t*-test *(*unknown variance, *n<30)*
		Mann-Whitney test	Chi-square test (Fisher's exact test for small samples)	Log-rank test or Mantel-Haenszel

Compare two dependent groups	Paired *t*-test	Wilcoxon test	McNemar's test	Conditional proportional hazards regression
Compare three or more independent groups	One-way ANOVA for single factor; Two-way ANOVA for two factors	Kruskal-Wallis test	Chi-square test	Cox proportional hazards regression
Compare three or more dependent groups	Repeated measures ANOVA	Friedman's test	Cochran's Q	Conditional proportional hazards progression
Quantify association between two variables	Pearson correlation	Spearman correlation	Contingency coefficients	
Predict value from another measured variable	Simple linear regression or nonlinear regression	Non-parametric regression	Simple logistic regression	Cox proportional hazards regression
Predict value from several measured or binomial variables	Multiple linear regression or multiple nonlinear regression	Spearman correlation	Multiple logistic regression	Cox proportional hazards regression

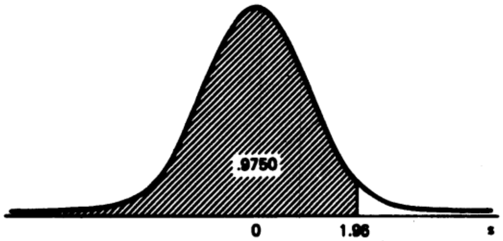
Table B. Random Numbers



	00000	00001	11111	11112	22222	22223	33333	33334	44444	44445
	12345	**67890**	**12345**	**67890**	**12345**	**67890**	**12345**	**67890**	**12345**	**67890**
01	85967	73152	14511	85285	36009	95892	36962	67835	63314	50162
02	07483	51453	11649	86348	76431	81594	95848	36738	25014	15460
03	96283	01898	61414	83525	04231	13604	75339	11730	85423	60698
04	49174	12074	98551	37895	93547	24769	09404	76548	05393	96770
05	97366	39941	21225	93629	19574	71565	33413	56087	40875	13351
06	90474	41469	16812	81542	81652	45554	27931	93994	22375	00953
07	28599	64109	09497	76235	41383	31555	12639	00619	22909	29563
08	25254	16210	89717	65997	82667	74624	36348	44018	64732	93589
09	28785	02760	24359	99410	77319	73408	58993	61098	04393	48245
10	84725	86576	86944	93296	10081	82454	76810	52975	10324	15457
11	41059	66456	47679	66810	15941	84602	14493	65515	19251	41642
12	67434	41045	82830	47617	36932	46728	71183	36345	41404	81110
13	72766	68816	37643	19959	57550	49620	98480	25640	67257	18671
14	92079	46784	66125	94932	64451	29275	57669	66658	30818	58353
15	29187	40350	62533	73603	34075	16451	42885	03448	37390	96328
16	74220	17612	65522	80607	19184	64164	66962	82310	18163	63495
17	03786	02407	06098	92917	40434	60602	82175	04470	78754	90775
18	75085	55558	15520	27038	25471	76107	90832	10819	56797	33751
19	09161	33015	19155	11715	00551	24909	31894	37774	37953	78837
20	75707	48992	64998	87080	39333	00767	45637	12538	67439	94914
21	21333	48660	31288	00086	79889	75532	28704	62844	92337	99695
22	65626	50061	42539	14812	48895	11196	34335	60492	70650	51108
23	84380	07389	87891	76255	89604	41372	10837	66992	93183	56920
24	46479	32072	80083	63868	70930	89654	05359	47196	12452	38234
25	59847	97197	55147	76639	76971	55928	36441	95141	42333	67483
26	31416	11231	27904	57383	31852	69137	96667	14315	01007	31929
27	82066	83436	67914	21465	99605	83114	97885	74440	99622	87912
28	01850	42782	39202	18582	46214	99228	79541	78298	75404	63648
29	32315	89276	89582	87138	16165	15984	21466	63830	30475	74729
30	59388	42703	55198	80380	67067	97155	34160	85019	03527	78140
31	58089	27632	50987	91373	07736	20436	96130	73483	85332	24384
32	61705	57285	30392	23660	75841	21931	04295	00875	09114	32101
33	18914	98982	60199	99275	41967	35208	30357	76772	92656	62318
34	11965	94089	34803	48941	69709	16784	44642	89761	66864	62803
35	85251	48111	80936	81781	93248	67877	16498	31924	51315	79921
36	66121	96986	84844	93873	46352	92183	51152	85878	30490	15974
37	53972	96642	24199	58080	35450	03482	66953	49521	63719	57615
38	14509	16594	78883	43222	23093	58645	60257	89250	63266	90858
39	37700	07688	65533	72126	23611	93993	01848	03910	38552	17472
40	85466	59392	72722	15473	73295	49759	56157	60477	83284	56367
41	52969	55863	42312	67842	05673	91878	82738	36563	79540	61935
42	42744	68315	17514	02878	97291	74851	42725	57894	81434	62041
43	26140	13336	67726	61876	29971	99294	96664	52817	90039	53211
44	95589	56319	14563	24071	06916	59555	18195	32280	79357	04224
45	39113	13217	59999	49952	83021	47709	53105	19295	88318	41626
46	41392	17622	18994	98283	07249	52289	24209	91139	30715	06604
47	54684	53645	79246	70183	87731	19185	08541	33519	07223	97413
48	89442	61001	36658	57444	95388	36682	38052	46719	09428	94012
49	36751	16778	54888	15357	68003	43564	90976	58904	40512	07725
50	98159	02564	21416	74944	53049	88749	02865	25772	89853	88714

SOURCE: *Biostatistics: A Foundation for Analysis in the Health Sciences* 6e by Wayne W. Daniel (1995)

Table C. Normal Curve Areas $P(z \le z_0)$ Entries in the Body of the Table are Areas between $-\infty$ and z



z	−0.09	−0.08	−0.07	−0.06	−0.05	−0.04	−0.03	−0.02	−0.01	0.00	z
−3.80	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	−3.80
−3.70	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	−3.70
−3.60	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0002	.0002	−3.60
−3.50	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	−3.50
−3.40	.0002	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	−3.40
−3.30	.0003	.0004	.0004	.0004	.0004	.0004	.0004	.0005	.0005	.0005	−3.30
−3.20	.0005	.0005	.0005	.0006	.0006	.0006	.0006	.0006	.0007	.0007	−3.20
−3.10	.0007	.0007	.0008	.0008	.0008	.0008	.0009	.0009	.0009	.0010	−3.10
−3.00	.0010	.0010	.0011	.0011	.0011	.0012	.0012	.0013	.0013	.0013	−3.00
−2.90	.0014	.0014	.0015	.0015	.0016	.0016	.0017	.0018	.0018	.0019	−2.90
−2.80	.0019	.0020	.0021	.0021	.0022	.0023	.0023	.0024	.0025	.0026	−2.80
−2.70	.0026	.0027	.0028	.0029	.0030	.0031	.0032	.0033	.0034	.0035	−2.70
−2.60	.0036	.0037	.0038	.0039	.0040	.0041	.0043	.0044	.0045	.0047	−2.60
−2.50	.0048	.0049	.0051	.0052	.0054	.0055	.0057	.0059	.0060	.0062	−2.50
−2.40	.0064	.0066	.0068	.0069	.0071	.0073	.0075	.0078	.0080	.0082	−2.40
−2.30	.0084	.0087	.0089	.0091	.0094	.0096	.0099	.0102	.0104	.0107	−2.30
−2.20	.0110	.0113	.0116	.0119	.0122	.0125	.0129	.0132	.0136	.0139	−2.20
−2.10	.0143	.0146	.0150	.0154	.0158	.0162	.0166	.0170	.0174	.0179	−2.10
−2.00	.0183	.0188	.0192	.0197	.0202	.0207	.0212	.0217	.0222	.0228	−2.00
−1.90	.0233	.0239	.0244	.0250	.0256	.0262	.0268	.0274	.0281	.0287	−1.90
−1.80	.0294	.0301	.0307	.0314	.0322	.0329	.0336	.0344	.0351	.0359	−1.80
−1.70	.0367	.0375	.0384	.0392	.0401	.0409	.0418	.0427	.0436	.0446	−1.70
−1.60	.0455	.0465	.0475	.0485	.0495	.0505	.0516	.0526	.0537	.0548	−1.60
−1.50	.0559	.0571	.0582	.0594	.0606	.0618	.0630	.0643	.0655	.0668	−1.50
−1.40	.0681	.0694	.0708	.0721	.0735	.0749	.0764	.0778	.0793	.0808	−1.40
−1.30	.0823	.0838	.0853	.0869	.0885	.0901	.0918	.0934	.0951	.0968	−1.30
−1.20	.0985	.1003	.1020	.1038	.1056	.1075	.1093	.1112	.1131	.1151	−1.20
−1.10	.1170	.1190	.1210	.1230	.1251	.1271	.1292	.1314	.1335	.1357	−1.10
−1.00	.1379	.1401	.1423	.1446	.1469	.1492	.1515	.1539	.1562	.1587	−1.00
−0.90	.1611	.1635	.1660	.1685	.1711	.1736	.1762	.1788	.1814	.1841	−0.90
−0.80	.1867	.1894	.1922	.1949	.1977	.2005	.2033	.2061	.2090	.2119	−0.80
−0.70	.2148	.2177	.2206	.2236	.2266	.2296	.2327	.2358	.2389	.2420	−0.70
−0.60	.2451	.2483	.2514	.2546	.2578	.2611	.2643	.2676	.2709	.2743	−0.60
−0.50	.2776	.2810	.2843	.2877	.2912	.2946	.2981	.3015	.3050	.3085	−0.50
−0.40	.3121	.3156	.3192	.3228	.3264	.3300	.3336	.3372	.3409	.3446	−0.40
−.030	.3483	.3520	.3557	.3594	.3632	.3669	.3707	.3745	.3783	.3821	−0.30
−0.20	.3859	.3897	.3936	.3974	.4013	.4052	.4090	.4129	.4168	.4207	−0.20
−0.10	.4247	.4286	.4325	.4364	.4404	.4443	.4483	.4522	.4562	.4602	−0.10
0.00	.4641	.4681	.4721	.4761	.4801	.4840	.4880	.4920	.4960	.5000	0.00

Table C (continuation)

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	z
0.00	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359	0.00
0.10	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753	0.10
0.20	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141	0.20
0.30	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517	0.30
0.40	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879	0.40
0.50	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224	0.50
0.60	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549	0.60
0.70	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852	0.70
0.80	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133	0.80
0.90	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389	0.90
1.00	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621	1.00
1.10	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830	1.10
1.20	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015	1.20
1.30	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177	1.30
1.40	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319	1.40
1.50	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441	1.50
1.60	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545	1.60
1.70	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633	1.70
1.80	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706	1.80
1.90	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767	1.90
2.00	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817	2.00
2.10	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857	2.10
2.20	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890	2.20
2.30	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916	2.30
2.40	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936	2.40
2.50	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952	2.50
2.60	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964	2.60
2.70	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974	2.70
2.80	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981	2.80
2.90	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986	2.90
3.00	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990	3.00
3.10	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993	3.10
3.20	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995	3.20
3.30	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997	3.30
3.40	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998	3.40
3.50	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	3.50
3.60	.9998	.9998	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	3.60
3.70	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	3.70
3.80	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	3.80

SOURCE: *Biostatistics: A Foundation for Analysis in the Health Sciences* 6e by Wayne W. Daniel (1995)

Table D. Critical Values (Percentiles) for the *t*-Distribution

The table entries are the critical values (percentiles) for the t-distribution. The column headed d.f. (degrees of freedom) gives the degrees of freedom for the values in that row. The columns are labeled by "percent," "one-sided," and "two-sided." "Percent" is 100 × cumulative distribution function—the table entry is the corresponding percentile. "One-sided" is the significance level for the one-sided upper critical value—the table entry is the critical value. "Two-sided" gives the two-sided significance level—the table entry is the corresponding two-sided critical value.

	Percent											
	75	90	95	97.5	99	99.5	99.75	99.9	99.95	99.975	99.99	99.995
	One-Sided α											
	.25	.10	.05	.025	.01	.005	.0025	.001	.0005	.00025	.0001	.00005
	Two-Sided α											
	.50	.20	.10	.05	.02	.01	.005	.002	.001	.0005	.0002	.0001
d.f.												
1	1.00	3.08	6.31	12.71	31.82	63.66	127.32	318.31	636.62	1273.24	3183.10	6366.20
2	.82	1.89	2.92	4.30	6.96	9.22	14.09	22.33	31.60	44.70	70.70	99.99
3	.76	1.64	2.35	3.18	4.54	5.84	7.45	10.21	12.92	16.33	22.20	28.00
4	.74	1.53	2.13	2.78	3.75	4.60	5.60	7.17	8.61	10.31	13.03	15.54
5	.73	1.48	2.02	2.57	3.37	4.03	4.77	5.89	6.87	7.98	9.68	11.18
6	.72	1.44	1.94	2.45	3.14	3.71	4.32	5.21	5.96	6.79	8.02	9.08
7	.71	1.42	1.90	2.37	3.00	3.50	4.03	4.79	5.41	6.08	7.06	7.88
8	.71	1.40	1.86	2.31	2.90	3.36	3.83	4.50	5.04	5.62	6.44	7.12
9	.70	1.38	1.83	2.26	2.82	3.25	3.69	4.30	4.78	5.29	6.01	6.59
10	.70	1.37	1.81	2.23	2.76	3.17	3.58	4.14	4.59	5.05	5.69	6.21
11	.70	1.36	1.80	2.20	2.72	3.11	3.50	4.03	4.44	4.86	5.45	5.92
12	.70	1.36	1.78	2.18	2.68	3.06	3.43	3.93	4.32	4.72	5.26	5.69
13	.69	1.35	1.77	2.16	2.65	3.01	3.37	3.85	4.22	4.60	5.11	5.51
14	.69	1.35	1.76	2.15	2.63	2.98	3.33	3.79	4.14	4.50	4.99	5.36
15	.69	1.34	1.75	2.13	2.60	2.95	3.29	3.73	4.07	4.42	4.88	5.24
16	.69	1.34	1.75	2.12	2.58	2.92	3.25	3.69	4.02	4.35	4.79	5.13
17	.69	1.33	1.74	2.11	2.57	2.90	3.22	3.65	3.97	4.29	4.71	5.04
18	.69	1.33	1.73	2.10	2.55	2.88	3.20	3.61	3.92	4.23	4.65	4.97
19	.69	1.33	1.73	2.09	2.54	2.86	3.17	3.58	3.88	4.19	4.59	4.90
20	.69	1.33	1.73	2.09	2.53	2.85	3.15	3.55	3.85	4.15	4.54	4.84
21	.69	1.32	1.72	2.08	2.52	2.83	3.14	3.53	3.82	4.11	4.49	4.78
22	.69	1.32	1.72	2.07	2.51	2.82	3.12	3.51	3.79	4.08	4.45	4.74
23	.68	1.32	1.71	2.07	2.50	2.81	3.10	3.49	3.77	4.05	4.42	4.69
24	.68	1.32	1.71	2.06	2.49	2.80	3.09	3.47	3.75	4.02	4.38	4.65
25	.68	1.32	1.71	2.06	2.49	2.79	3.08	3.45	3.73	4.00	4.35	4.62
26	.68	1.32	1.71	2.06	2.48	2.78	3.07	3.44	3.71	3.97	4.32	4.59
27	.68	1.31	1.70	2.05	2.47	2.77	3.06	3.42	3.69	3.95	4.30	4.56
28	.68	1.31	1.70	2.05	2.47	2.76	3.05	3.41	3.67	3.94	4.28	4.53
29	.68	1.31	1.70	2.05	2.46	2.76	3.04	3.40	3.66	3.92	4.25	4.51
30	.68	1.31	1.70	2.04	2.46	2.75	3.03	3.39	3.65	3.90	4.23	4.48
35	.68	1.31	1.69	2.03	2.44	2.72	3.00	3.34	3.59	3.84	4.15	4.39
40	.68	1.30	1.68	2.02	2.42	2.70	2.97	3.31	3.55	3.79	4.09	4.32
45	.68	1.30	1.68	2.01	2.41	2.69	2.95	3.28	3.52	3.75	4.05	4.27
50	.68	1.30	1.68	2.01	2.40	2.68	2.94	3.26	3.50	3.72	4.01	4.23
55	.68	1.30	1.67	2.00	2.40	2.67	2.93	3.25	3.48	3.70	3.99	4.20
60	.68	1.30	1.67	2.00	2.39	2.66	2.91	3.23	3.46	3.68	3.96	4.17
65	.68	1.29	1.67	2.00	2.39	2.65	2.91	3.22	3.45	3.66	3.94	4.15
70	.68	1.29	1.67	1.99	2.38	2.65	2.90	3.21	3.44	3.65	3.93	4.13
75	.68	1.29	1.67	1.99	2.38	2.64	2.89	3.20	3.43	3.64	3.91	4.11
80	.68	1.29	1.66	1.99	2.37	2.64	2.89	3.20	3.42	3.63	3.90	4.10
85	.68	1.29	1.66	1.99	2.37	2.64	2.88	3.19	3.41	3.62	3.89	4.08
90	.68	1.29	1.66	1.99	2.37	2.63	2.88	3.18	3.40	3.61	3.88	4.07
95	.68	1.29	1.66	1.99	2.37	2.63	2.87	3.18	3.40	3.60	3.87	4.06
100	.68	1.29	1.66	1.98	2.36	2.63	2.87	3.17	3.39	3.60	3.86	4.05
200	.68	1.29	1.65	1.97	2.35	2.60	2.84	3.13	3.34	3.54	3.79	3.97
500	.68	1.28	1.65	1.97	2.33	2.59	2.82	3.11	3.31	3.50	3.75	3.92
∞	.67	1.28	1.65	1.96	2.33	2.58	2.81	3.10	3.30	3.49	3.73	3.91

SOURCE: ***Biostatistics*: *A Methodology for the Heatlh Sciences*** **2e by van Belle et al. (2004)**

Table E. Critical Values (Percentiles) for the Chi-Square Distribution

For each degree of freedom (d.f.) in the first column, the table entries are the critical values for the upper one-sided significance levels in the column headings or, equivalently, the percentiles for the corresponding percentages.

	Percentage								
	2.5	5	50	75	90	95	97.5	99	99.9
	Upper One-Sided α								
d.f.	.975	.95	.50	.25	.10	.05	.025	.01	.001
1	.001	.004	.455	1.32	2.71	3.84	5.02	6.63	10.83
2	.051	.103	1.39	2.77	4.61	5.99	7.38	9.21	13.82
3	.216	.352	2.37	4.11	6.25	7.82	9.35	11.34	16.27
4	.484	.711	3.36	5.39	7.78	9.49	11.14	13.28	18.47
5	.831	1.15	4.35	6.63	9.24	11.07	12.83	15.09	20.52
6	1.24	1.64	5.35	7.84	10.64	12.59	14.45	16.81	22.46
7	1.69	2.17	6.35	9.04	12.02	14.07	16.01	18.47	24.32
8	2.18	2.73	7.34	10.22	13.36	15.51	17.53	20.09	26.12
9	2.70	3.33	8.34	11.39	14.68	16.92	19.02	21.67	27.88
10	3.25	3.94	9.34	12.55	15.99	18.31	20.48	23.21	29.59
11	3.82	4.57	10.34	13.70	17.27	19.68	21.92	24.72	31.26
12	4.40	5.23	11.34	14.85	18.55	21.03	23.34	26.22	32.91
13	5.01	5.89	12.34	15.98	19.81	22.36	24.74	27.69	34.53
14	5.63	6.57	13.34	17.12	21.06	23.68	26.12	29.14	36.12
15	6.26	7.26	14.34	18.25	22.31	25.00	27.49	30.58	37.70
16	6.91	7.96	15.34	19.37	23.54	26.30	28.85	32.00	39.25
17	7.56	8.67	16.34	20.49	24.77	27.59	30.19	33.41	40.79
18	8.23	9.39	17.34	21.60	25.99	28.87	31.53	34.81	42.31
19	8.91	10.12	18.34	22.72	27.20	30.14	32.85	36.19	43.82
20	9.59	10.85	19.34	23.83	28.41	31.41	34.17	37.57	45.31
21	10.28	11.59	20.34	24.93	29.62	32.67	35.48	38.93	46.80
22	10.98	12.34	21.34	26.04	30.81	33.92	36.78	40.29	48.27
23	11.69	13.09	22.34	27.14	32.01	35.17	38.08	41.64	49.73
24	12.40	13.85	23.34	28.24	33.20	36.42	39.36	42.98	51.18
25	13.12	14.61	24.34	29.34	34.38	37.65	40.65	44.31	52.62
26	13.84	15.38	25.34	30.43	35.56	38.89	41.92	45.64	54.05
27	14.57	16.15	26.34	31.53	36.74	40.11	43.19	46.96	55.48
28	15.31	16.93	27.34	32.62	37.92	41.34	44.46	48.28	56.89
29	16.05	17.71	28.34	33.71	39.09	42.56	45.72	49.59	58.30
30	16.79	18.49	29.34	34.80	40.26	43.77	46.98	50.89	59.70
35	20.57	22.47	34.34	40.22	46.06	49.80	53.20	57.34	66.62
40	24.43	26.51	39.34	45.62	51.81	55.76	59.34	63.69	73.40
45	28.37	30.61	44.34	50.98	57.51	61.66	65.41	69.96	80.08
50	32.36	34.76	49.33	56.33	63.17	67.50	71.42	76.15	86.66
55	36.40	38.96	54.33	61.66	68.80	73.31	77.38	82.29	93.17
60	40.48	43.19	59.33	66.98	74.40	79.08	83.30	88.38	99.61
65	44.60	47.45	64.33	72.28	79.97	84.82	89.18	94.42	105.99
70	48.76	51.74	69.33	77.58	85.53	90.53	95.02	100.43	112.32
75	52.94	56.05	74.33	82.86	91.06	96.22	100.84	106.39	118.60
80	57.15	60.39	79.33	88.13	96.58	101.88	106.63	112.33	124.84
85	61.39	64.75	84.33	93.39	102.08	107.52	112.39	118.24	131.04
90	65.65	69.13	89.33	98.65	107.57	113.15	118.14	124.12	137.21
95	69.92	73.52	94.33	103.90	113.04	118.75	123.86	129.97	143.34
100	74.22	77.93	99.33	109.14	118.50	124.34	129.56	135.81	149.45

For more than 100 degrees of freedom chi-square critical values may be found in terms of the degrees of freedom and the corresponding two-sided critical value for a standard normal deviate Z by the equation $X^2 = 0.5 \cdot (Z + \sqrt{2 \cdot D - 1})^2$.

SOURCE: ***Biostatistics*: *A Methodology for the Heatlh Sciences* 2e by van Belle et al. (2004)**

Table F. Critical Values (Percentiles) for the *F*-Distribution

Upper one-sided 0.05 significance levels; two-sided 0.10 significance levels; 95% percentiles. Tabulated are critical values for the F-distribution. The column headings give the numerator degrees of freedom and the row headings the denominator degrees of freedom. Lower one-sided critical values may be found from these tables by reversing the degrees of freedom and using the reciprocal of the tabled value at the same significance level (100 minus the percent for the percentile).

	Numerator Degrees of Freedom																		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	6.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96

Table F (Continuation)

	Numerator Degrees of Freedom																		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

Table F (Continuation)

Upper one-sided 0.01 significance levels; two-sided 0.02 significance levels; 99% percentiles.

	Numerator Degrees of Freedom																		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	4052	5000	5403	5625	5764	5859	5928	5982	6022	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.87	26.69	26.60	26.50	26.41	26.32	26.22	26.13
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65

Table F (Continuation)

	Numerator Degrees of Freedom																		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.13
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.03
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

SOURCE: ***Biostatistics*: *A Methodology for the Heatlh Sciences* 2e by van Belle et al. (2004)**

References

Acacio-Claro, PJ (2008). Lectures in Biostatistics. University of the Philippines Manila, Manila.

Albert, J. (1996, November 18). The Subjective Interpretation of Probability. Retrieved October 30, 2012, from http://www-math.bgsu.edu/~albert/m115/probability/subject.html

Albert, J. (1996, November 18). The Relative Frequency Interpretation of Probability. Retrieved October 30, 2012, from http://www-math.bgsu.edu/~albert/m115/probability/relfeq.html

Arsham, H. (n.d.). Combinatorial Mathematics: How to Count Without Counting. Retrieved October 30, 2012, from http://home.ubalt.edu/ntsbarsh/Business-stat/otherapplets/ComCount.htm

Capistrano, TA (n.d.). Course Notes in Stat 114 (Descriptive Statistics). University of the Philippines Diliman, Quezon City.

Centers for Disease Control and Prevention. (1992). *Principles of Epidemiology.*

Chernick, MR & Friis, RH (2003). *Introductory Biostatistics for the Health Sciences.* Hoboken, New Jersey: John Wiley & Sons, Inc.

Daniel, W. (1995). *Biostatistics: A Foundation for Analysis in the Health Sciences* (6th ed.). Hoboken, New Jersey: John Wiley & Sons, Inc.

Doug, J. (2010). Class Notes in Medical Statistics (Fall 2010): Summary Statistics. University of Illinois at Urbana-Champaign.

Fairfax County Department of Neighborhood and Community Services. (2012, April). Overview of Sampling Procedures. Fairfax, Virginia, United States of America. Retrieved November 2, 2012, from http://www.fairfaxcounty.gov/demogrph/pdf/samplingprocedures.pdf

Kuzma, JW & Bohnenblust, SE (2005). Basic Statistics for the Health Sciences International Ed. McGraw-Hill (Asia): Philippines.

Le, CT (2003). *Introductory Biostatistics.* Hoboken, New Jersey: John Wiley & Sons, Inc.

Mapua, CA (2010). Lectures in Molecular Epidemiology and Biostatistics. St. Luke's College of Medicine, Quezon City.

Mendoza, OM; Borja, MP; Sevilla, TL; Ancheta, CA; Saniel, OP; Sarol Jr, JN and Lozano, JP (2009). Foundations of Statistical Analysis for the Health Sciences. Department of Epidemiology and Biostatistics, College of Public Health, University of the Philippines Manila: Manila.

National Institute of Standards and Technology. (n.d.). Percentiles. Retrieved October 30, 2012, from Engineering Statistics Handbook: http://itl.nist.gov/div898/handbook/prc/section2/prc252.htm

Samuels, ML & Witmer, JA (1999). *Statistics for the Life Sciences 2nd Ed.* Prentice-Hall, Inc.: New Jersey.

Schröder, B. (n.d.). Slides in Sample Spaces and Events. Louisiana Tech University, College of Engineering and Science.

van Belle, G; Fisher, LD; Heagerty, PJ; & Lumley, T (2004). *Biostatistics: A Methodology for the Health Sciences* (2nd ed.). Hoboken, New Jersey: John Wiley & Sons, Inc.