Statistics in Psychology (PSY516)

Table of Content

Lesson 1

INTRODUCTION TO STATISTICS IN PSYCHOLOGY-I

Introduction to Statistics

The study of statistics is gaining recognition in a great many fields. In particular, researchers in the social, behavioral and health sciences note its importance for problem solving and its practical importance in their areas. Statistics is a science that enables us to draw conclusions about various phenomena on the basis of real data collected on sample-basis. It is a tool for data-based research and also known as Quantitative Analysis. It is the science of collecting and learning from data. It is a branch of mathematics concerned with the collection, classification, analysis, and interpretation of numerical facts, for drawing inferences on the basis of their quantifiable likelihood (probability). Statistics also allows the researcher to interpret grouped data, too large to be intelligible by ordinary observation.

Statistics has a lot of application in a wide variety of disciplines i.e. Agriculture, Anthropology, Astronomy, Biology, Economic, Engineering, Environment, Geology, Genetics, Medicine, Physics, Psychology, Sociology, Zoology…. Virtually every single subject from Anthropology to Zoology …. A to Z!

Any scientific enquiry in which you would like to base your conclusions and decisions on real-life data, you need to employ statistical techniques. Now a days, in the developed countries of the world, there is an active movement for Statistical Literacy.

Defining Statistics

The term **statistics** refers to a set of mathematical procedures for organizing, summarizing, and interpreting information. One goes through the four stages of statistics:

1. Collection of data,
2. Organizing and summarizing the data,
3. Analysis of data, and
4. Making inferences, or decisions and predictions.

Stage four is the core objective of statistics in psychology, i.e., making inferences about a population based on information contained in a representative sample taken from that population. Statistics consist of facts and figures such as average income, crime rate, birth rate, baseball batting averages, and so on. These statistics are usually informative and time saving because they condense large quantities of information into a few simple figures.

Statistics is the science of learning from data. Statistical procedures help to ensure that the information or observations are presented and interpreted in an accurate and informative way. In somewhat grandiose terms, statistics help researchers bring order out of chaos. Specifically, statistics serve two general purposes:

1. It is used to organize and summarize the information so that the researcher can see what happened in the research study and can communicate the results to others.
2. It helps the researcher to answer the questions that initiated the research by determining exactly what general conclusions are justified based on the specific results that were obtained.

Research in psychology (and other fields) involves gathering information. To determine, for example, whether violence on TV has any effect on children's behavior, you would need to gather information about children's behaviors and the TV programs they watch. When researchers finish the task of gathering information, they typically find themselves with pages and pages of measurements such as IQ scores, personality scores, reaction time scores, and so on.

Importance Of Statistics

As it is such an important area of knowledge, it is definitely useful to have a fairly good idea about the way in which it works, and this is exactly the purpose of this introductory course.

The following points indicate some of the main functions of this science:

- Statistics assists in summarizing the larger set of data in a form that is easily understandable.
- Statistics assists in the efficient design of laboratory and field experiments as well as surveys.
- Statistics assists in a sound and effective planning in any field of inquiry.
- Statistics assists in drawing general conclusions and in making predictions of how much of a thing will happen under given conditions.

Importance Of Statistics In various fields

As stated earlier, Statistics is a discipline that has finds application in the most diverse fields of activity. It is perhaps a subject that should be used by everybody. Statistical techniques being powerful tools for analyzing numerical data are used in almost every branch of learning. In all areas, statistical techniques are being increasingly used, and are developing very rapidly.

- A modern administrator whether in public or private sector leans on statistical data to provide a factual basis for decision.
- A politician uses statistics advantageously to lend support and credence to his arguments while elucidating the problems he handles.
- A businessman, an industrial and a research worker all employ statistical methods in their work. Banks, Insurance companies and Government all have their statistics departments.
- A social scientist uses statistical methods in various areas of socioeconomic life of a nation. It is sometimes said that "*a social scientist without an adequate understanding of statistics, is often like the blind man groping in a dark room for a black cat that is not there".*

Myths About Statistics

When we are studying statistics, several questions and myths arises in our mind. Following are some common myths about statistics:

- Statistics is very Hard.
- Statistics is Math.
- Why it is necessary to study Statistics in Psychology?
- You Lie with Statistics.
- If something is not statistically significant, it is not important.

Lesson 2

INTRODUCTION TO STATISTICS IN PSYCHOLOGY-II

Terms and Definitions

The field of statistics is subdivided into *descriptive* statistics and *inferential* statistics. Following are some statistical jargon used in psychology:

Descriptive statistics

Descriptive statistics are statistical procedures used to summarize, organize, and simplify data. Descriptive statistics are techniques that take raw scores and organize or summarize them in a form that is more manageable. Often the scores are organized in a table or a graph so that it is possible to see the entire set of scores. Another common technique is to summarize a set of scores by computing an average. Typically, there are two general types of statistics that are used to describe data"

- Measures of central tendency
- Measure of spread/dispersion

Inferential statistics

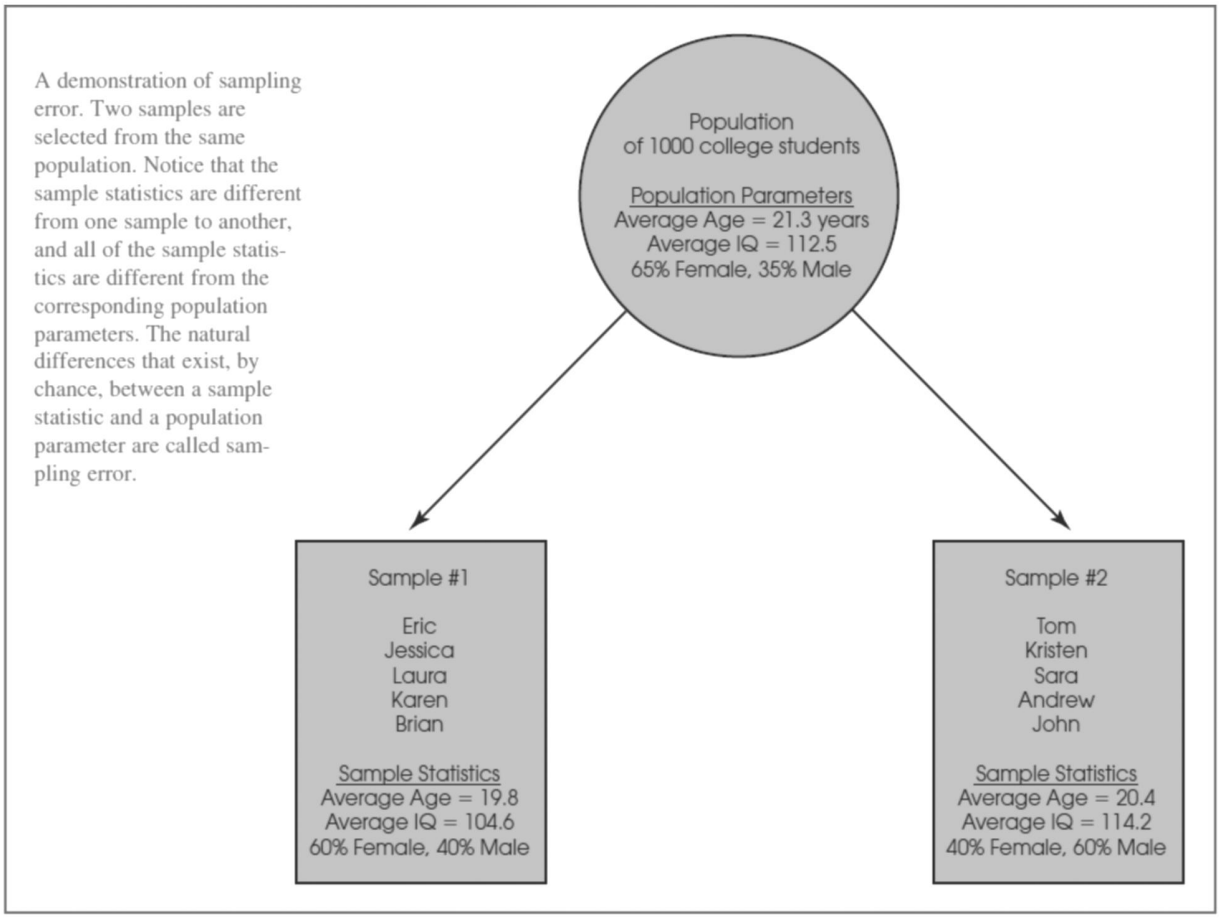
Whereas descriptive statistics is the branch of statistics that involves organizing, displaying, and describing data, inferential statistics is the branch of statistics that involves drawing conclusions about a population based on information contained in a sample taken from that population. Inferential statistics consist of techniques that allow us to study samples and then make generalizations about the populations from which they were selected. Most research uses statistical models called the Generalized Linear model and include Student's t-tests, ANOVA (Analysis of Variance), regression analysis.

Sampling error

Sampling error is the naturally occurring discrepancy, or error, that exists between a sample statistic and the corresponding population parameter.

Because populations are typically very large, it usually is not possible to measure everyone in the population. Therefore, a sample is selected to represent the population. By analyzing the results from the sample, we hope to make general statements about the population. One problem with using samples, however, is that a sample provides only limited information about the population. Although samples are generally representative of their populations, a sample is not expected to give a perfectly accurate picture of the whole population. There usually is some discrepancy

between a sample statistic and the corresponding population parameter. This discrepancy is called sampling error.



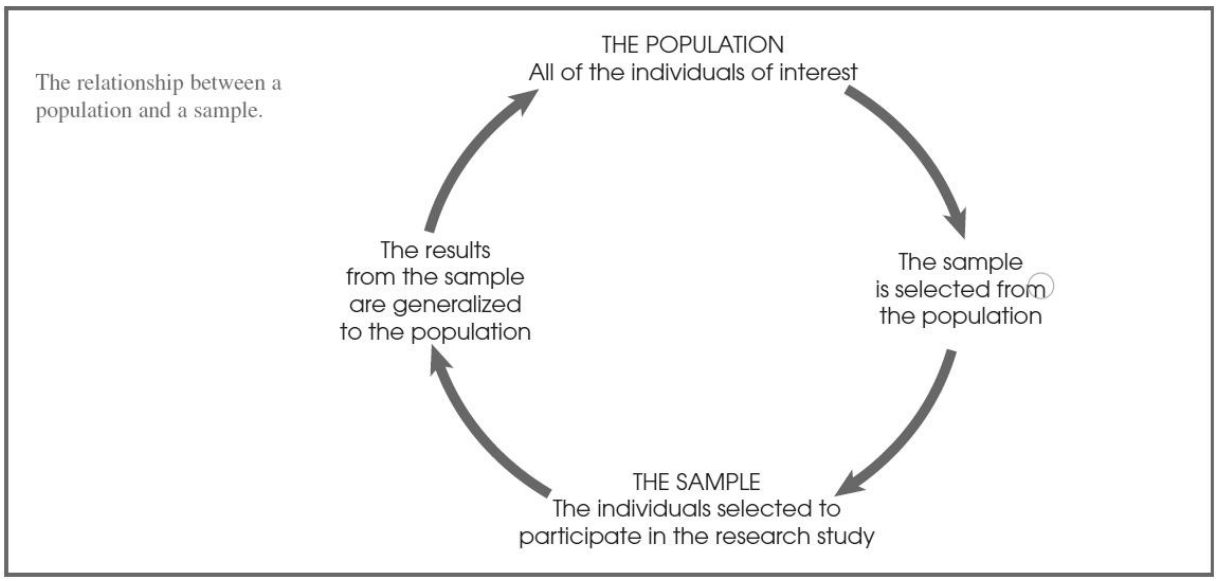
Population

A population is the set of all the individuals of interest in a particular study. For example, a researcher may be interested in the effect of divorce on the self-esteem of preteen children. Or a researcher may want to examine the amount of time spent in the bathroom for men compared to women. In the first example, the researcher is interested in the group of preteen children. In the second example, the researcher wants to compare the group of men with the group of women. In statistical terminology, the entire group that a researcher wishes to study is called a population.

Sample

A sample is a set of individuals selected from a population, usually intended to represent the population in a research study. Because populations tend to be very large, it usually is impossible for a researcher to examine every individual in the population of interest. Therefore, researchers

typically select a smaller, more manageable group from the population and limit their studies to the individuals in the selected group. In statistical terms, a set of individuals selected from a population is called a sample. A sample is intended to be representative of its population, and a sample should always be identified in terms of the population from which it was selected.



The relationship between a population and a sample.

Population Parameter and Sample Statistics

When describing data, it is necessary to distinguish whether the data come from a population or a sample. A characteristic that describes a population—for example, the average score for the population—is called a parameter. A characteristic that describes a sample is called a statistic. A **parameter** is a value, usually a numerical value, that describes a population. A parameter is usually derived from measurements of the individuals in the population. A **statistic** is a value, usually a numerical value, that describes a sample. A statistic is usually derived from measurements of the individuals in the sample.

Every population parameter has a corresponding sample statistic, and most research studies involve using statistics from samples as the basis for answering questions about population parameters.

Variables and their Data

Typically, researchers are interested in specific characteristics of the individuals in the population (or in the sample), or they are interested in outside factors that may influence the individuals. For example, a researcher may be interested in the influence of the weather on

people's moods. As the weather changes, do people's moods also change? Something that can change or have different values is called a variable. A **variable** is a characteristic or condition that changes or has different values for different individuals.

To demonstrate changes in variables, it is necessary to make measurements of the variables being examined. The measurement obtained for each individual is called a datum or, more commonly, a score or raw score. The complete set of scores is called the data set or simply the data. **Data** (plural) are measurements or observations. A data set is a collection of measurements or observations. A datum (singular) is a single measurement or observation and is commonly called a score or raw score.

Before we move on, we should make one more point about samples, populations, and data. Earlier, we defined populations and samples in terms of individuals. For example, we discussed a population of college sophomores and a sample of preschool children. Be forewarned, however, that we will also refer to populations or samples of scores. Because research typically involves measuring each individual to obtain a score, every sample (or population) of individuals produces a corresponding sample (or population) of scores.

What to Measure?

Some variables, such as height, weight, and eye color are well-defined, concrete entities that can be observed and measured directly. On the other hand, many variables studied by behavioral scientists are internal characteristics that people use to help describe and explain behavior. Variables like intelligence, anxiety, and hunger are called constructs. The external behaviors can then be used to create an operational definition for the construct.

Constructs are internal attributes or characteristics that cannot be directly observed but are useful for describing and explaining behavior.

An **operational definition** identifies a measurement procedure (a set of operations) for measuring an external behavior and uses the resulting measurements as a definition and a measurement of a hypothetical construct. Note that an operational definition has two components: First, it describes a set of operations for measuring a construct. Second, it defines the construct in terms of the resulting measurements.

Discrete Variable

A discrete variable consists of separate, indivisible categories. No values can exist between two neighboring categories. Also known as a categorical variable, because it has separate, invisible categories.

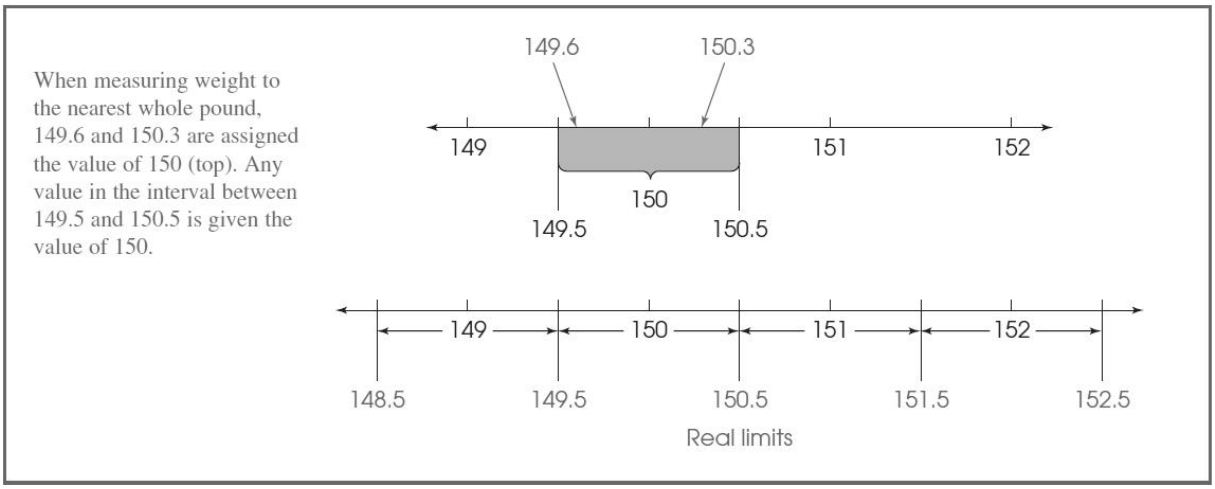
Discrete variables are commonly restricted to whole, countable numbers—for example, the number of children in a family or the number of students attending class. If you observe class attendance from day to day, you may count 18 students one day and 19 students the next day. However, it is impossible ever to observe a value between 18 and 19. A discrete variable may also consist of observations that differ qualitatively.

Continuous Variable

For a continuous variable, there are an infinite number of possible values that fall between any two observed values. A continuous variable is divisible into an infinite number of fractional parts. Also known as a ratio/interval variable, consists of ordered categories that are all intervals of exactly the same size with or without absolute zero. Variables such as time, height, and weight are not limited to a fixed set of separate, indivisible categories. You can measure time, for example, in hours, minutes, seconds, or fractions of seconds. Two other factors apply to continuous variables:

1. When measuring a continuous variable, it should be very rare to obtain identical measurements for two different individuals. Because a continuous variable has an infinite number of possible values, it should be almost impossible for two people to have exactly the same score.
2. When measuring a continuous variable, each measurement category is actually an interval that must be defined by boundaries. These boundaries are called real limits and are positioned exactly halfway between adjacent scores. Thus, a score of X 150 pounds is actually an interval bounded by a lower real limit of 149.5 at the bottom and an upper real limit of 150.5 at the top. Any individual whose weight falls between these real limits will be assigned a score of X 150.

Real Limits are the boundaries of intervals for scores that are represented on a continuous number line. The real limit separating two adjacent scores is located exactly halfway between the scores. Each score has two real limits. The upper real limit is at the top of the interval, and the lower real limit is at the bottom.



Independent Variable

A variable thought to be the cause of some effect. The independent variable is the variable that is manipulated by the researcher. In behavioral research, the independent variable usually consists of the two (or more) treatment conditions to which subjects are exposed. The independent variable consists of the antecedent conditions that were manipulated prior to observing the dependent variable. E.g. two conditions, or control and experimental group or pre/post test. Note that the independent variable always consists of at least two values. (Something must have at least two different values before you can say that it is "variable."). Predictor variable: A variable thought to predict an outcome variable. This is basically another term for independent variable.

Dependent Variable

The dependent variable is the variable that is observed to assess the effect of the treatment. A variable thought to be affected by changes in an independent variable. You can think of this variable as an outcome. The variable that is observed and measured to obtain scores is the dependent variable. Outcome variable: A variable thought to change as a function of changes in a predictor variable. This term could be synonymous with 'dependent variable'.

Control Variables

Variables that are held constant throughout the experiment. The temperature and light in the room the plants are kept in, and the volume of water given to each plant.

Extraneous Variables

Extraneous variables are any variables that you are not intentionally studying in your experiment or test but they effect the results of the study.

Extraneous Variables can be: Demand characteristics: environmental clues which tell the participant how to behave, like features in the surrounding or researcher's non-verbal behavior. Experimenter / Investigator Effects: where the researcher unintentionally affects the outcome by giving clues to the participants about how they should behave. Participant variables, like prior knowledge, health status or any other individual characteristic that could affect the outcome. Situational variables, like noise, lighting or temperature in the environment.

Confounding Variables

A variable that hides the true effect of another variable in your experiment. This can happen when another variable is closely related to a variable you are interested in, but you haven't controlled it in your experiment.

Scales of Measurement

It should be obvious by now that data collection requires that we make measurements of our observations. Measurement involves assigning individuals or events to categories. The categories can simply be names such as male/female or employed/unemployed, or they can be numerical values such as 68 inches or 175 pounds. The categories used to measure a variable make up a scale of measurement, and the relationships between the categories determine different types of scales.

The Nominal Scale

The word nominal means "having to do with names." A nominal scale consists of a set of categories that have different names. Measurements on a nominal scale label and categorize observations, but do not make any quantitative distinctions between observations.

The measurements from a nominal scale allow us to determine whether two individuals are different, but they do not identify either the direction or the size of the difference. A sub-type of nominal scale with only two categories (e.g. male/female) is called "dichotomous." If you are a student, you can use that to impress your teacher. Other sub-types of nominal data are "nominal with order" (like cold, warm, hot, very hot) and nominal without order (like male/female).

Although the categories on a nominal scale are not quantitative values, they are occasionally represented by numbers. For example, the rooms or offices in a building may be identified by numbers. You should realize that the room numbers are simply names and do not reflect any quantitative information. Room 109 is not necessarily bigger than Room 100 and certainly not 9 points bigger.

The Ordinal Scale

An ordinal scale consists of a set of categories that are organized in an ordered sequence. Measurements on an ordinal scale rank observation in terms of size or magnitude.

With measurements from an ordinal scale, you can determine whether two individuals are different and you can determine the direction of difference. However, ordinal measurements do not allow you to determine the size of the difference between two individuals. For example, if Billy is placed in the low-reading group and Tim is placed in the high-reading group, you know that Tim is a better reader, but you do not know how much better.

Ordinal scale consists of a series of ranks (first, second, third, and so on) like the order of finish in a horse race. Occasionally, the categories are identified by verbal labels like small, medium, and large drink sizes at a fast-food restaurant. In either case, the fact that the categories form an ordered sequence means that there is a directional relationship between categories.

Interval Scale

An interval scale consists of ordered categories that are all intervals of exactly the same size. Equal differences between numbers on a scale reflect equal differences in magnitude. However, the zero point on an interval scale is arbitrary and does not indicate a zero amount of the variable being measured.

For example, a temperature of 0° Fahrenheit does not mean that there is no temperature, and it does not prohibit the temperature from going even lower. Furthermore, you know that a measurement of 80° Fahrenheit is higher than a measure of 60°, and you know that it is exactly 20° higher.

Interval scales are numeric scales. Interval scales don't have a "true zero." Hence, zero and negative numbers also have meaning. Without a true zero, it is impossible to compute ratios.

The Ratio Scale

A ratio scale is an interval scale with the additional feature of an absolute zero point. With a ratio scale, ratios of numbers do reflect ratios of magnitude.

For example, a score of 16 on an anxiety scale means that the person is, in reality, twice as anxious as someone scoring 8.

A ratio scale is anchored by a zero point that is not arbitrary but rather is a meaningful value representing none (a complete absence) of the variable being measured. The existence of an

absolute, non-arbitrary zero point means that we can measure the absolute amount of the variable; that is, we can measure the distance from 0. This makes it possible to compare measurements in terms of ratios.

Due to absolute zero–which allows for a wide range of both descriptive and inferential statistics to be applied. These variables can be meaningfully added, subtracted, multiplied, divided (ratios). Central tendency can be measured by mode, median, or mean; measures of dispersion, such as standard deviation and coefficient of variation can also be calculated from ratio scales.

Lesson 3

CORRELATION METHODS

Most research, are intended to examine relationships between two or more variables. For example, is there a relationship between the amount of violence that children see on television and the amount of aggressive behavior they display? Is there a relationship between the quality of breakfast and level of academic performance for elementary school children? To establish the existence of a relationship, researchers must make observations—that is, measurements of the two variables. The resulting measurements can be classified into two distinct data structures that also help to classify different research methods and different statistical techniques. In the following section we identify and discuss these two data structures.

The Correlational Method

One method for examining the relationship between variables is to observe the two variables as they exist naturally for a set of individuals. That is, simply measure the two variables for each individual. Most common analysis used in correlational studies is Pearson Product Moment Correlation (r).

In the **correlational method**, two different variables are observed to determine whether there is a relationship between them.

Correlational research allows researchers to:

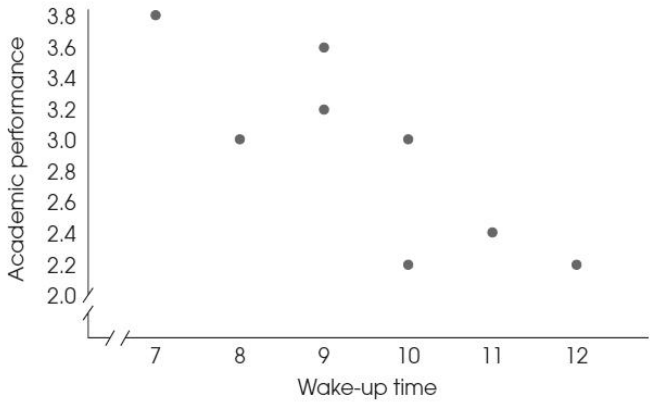
- establish reliability and validity
- provide converging evidence
- describe relationships
- make predictions

For example, research has demonstrated a relationship between sleep habits, especially wake-up time, and academic performance for college students (Trockel, Barnes, and Egget, 2000). The researchers used a survey to measure wake-up time and school records to measure academic performance for each student. The researchers then look for consistent patterns in the data to provide evidence for a relationship between variables. For example, as wake-up time changes from one student to another, is there also a tendency for academic performance to change?

(a)

Student	Wake-up Time	Academic Performance
A	11	2.4
B	9	3.6
C	9	3.2
D	12	2.2
E	7	3.8
F	10	2.2
G	10	3.0
H	8	3.0

(b)



One of two data structures for studies evaluating the relationship between variables. Note that there are two separate measurements for each individual (wake-up time and academic performance). The same scores are shown in a table (a) and in a graph (b).

Characteristics of the Relationship of Correlation:

- A ***positive relationship*** depicts, where low (or high) scores on one variable relate to low (or high) scores on a second variable.
- A ***negative relationship*** results, where low scores on one variable relate to high scores on the other variable.
- A ***zero relationship*** *of scores:* In this distribution, the variables are independent of each other. A particular score on one variable does not predict or tell us any information about the possible score on the other variable.

Degree Of Association means that the association between two variables or sets of scores is a correlation coefficient of –1.00 to +1.00, with 0.00 indicating no linear association at all.

Types of Correlational Designs:

Cross-Sectional Designs - One on more samples are drawn from a population, which are studied at one time on the same variables. E.g. effect of gender on depression of people born in 60s and 90s. (year makes two different samples).

Longitudinal Design - Same respondents or samples are surveyed over a period of time and is useful for assessing changes in behavior seen in individuals over time. The design is also the best form of survey for studying the effect of a naturally occurring event or phenomena.

Limitations of the Correlational Method:

The results from a correlational study can demonstrate the existence of a relationship between two variables, but they do not provide an explanation for the relationship. In particular, a

correlational study cannot demonstrate a cause-and-effect relationship. In the above example a systematic relationship between wake-up time and academic performance for a group of college students; those who sleep late tend to have lower performance scores than those who wake early. However, there are many possible explanations for the relationship and we do not know exactly what factor (or factors) is responsible for late sleepers having lower grades.

Lesson 4

EXPERIMENTAL AND NON-EXPERIMENTAL METHOD

The Experimental Method

One specific research method that involves comparing groups of scores is known as the experimental method or the experimental research strategy. The goal of an experimental study is to demonstrate a cause-and-effect relationship between two variables. Specifically, an experiment attempts to show that changing the value of one variable causes change to occur in the second variable.

In the **experimental method**, one variable is manipulated while another variable is observed and measured. To establish a cause-and-effect relationship between the two variables, an experiment attempts to control all other variables to prevent them from influencing the results.

Three conditions needed to make causal inferences

1. ***Covariation:*** Covariation should exist i.e. when a relationship is seen between the IV and DV.
2. ***Time order relationship:*** Time order relationship: when researchers manipulate an independent variable and then see a difference in the subsequent behavior or DV. The cause should always come before the effect.
3. ***Elimination of confounding variables*** or alternative causes that can affect the outcome of experiment of DV.

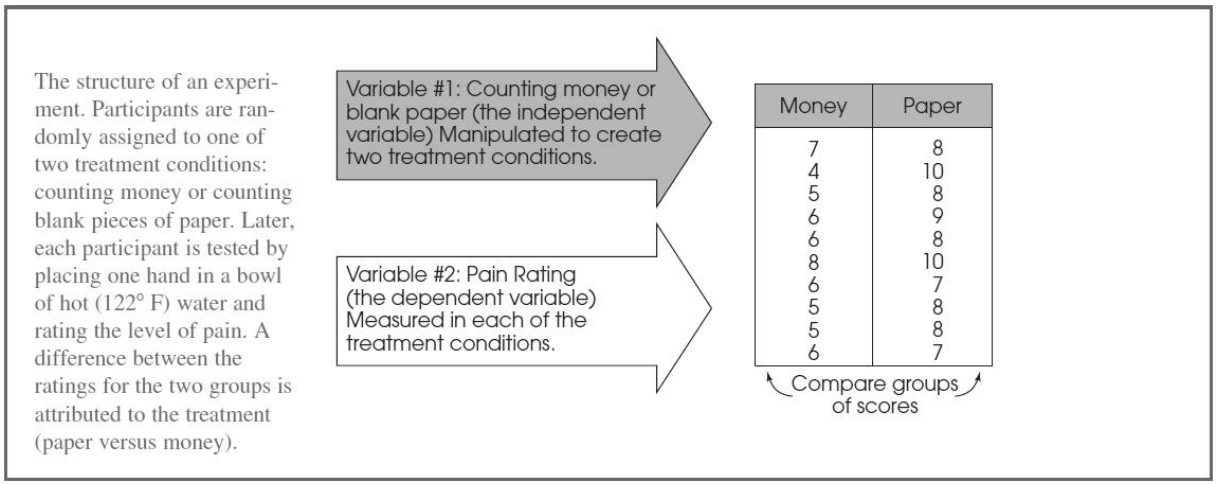
Two methods can be used to accomplish the goal of establishing cause-and-effect relationship in experimental method:

1. ***Manipulation***: The researcher manipulates one variable by changing its value from one level to another. A second variable is observed (measured) to determine whether the manipulation causes changes to occur.
2. ***Control***: The researcher must exercise control over the research situation to ensure that other, extraneous variables do not influence the relationship being examined.

To demonstrate these two characteristics, consider an experiment in which researchers demonstrate the pain-killing effects of handling money (Zhou & Vohs, 2009). In the experiment, a group of college students was told that they were participating in a manual dexterity study. The researcher then manipulated the treatment conditions by giving half of the students a stack of money to count and the other half a stack of blank pieces of paper. After the counting task, the

participants were asked to dip their hands into bowls of painfully hot water (122° F) and rate how uncomfortable it was. Participants who had counted money rated the pain significantly lower than those who had counted

paper.



There are two general categories of variables that researchers must consider:

1. ***Participant Variables***: These are characteristics such as age, gender, and intelligence that vary from one individual to another. Whenever an experiment compares different groups of participants (one group in treatment A and a different group in treatment B), researchers must ensure that participant variables do not differ from one group to another.
2. ***Environmental Variables***: These are characteristics of the environment such as lighting, time of day, and weather conditions. A researcher must ensure that the individuals in treatment A are tested in the same environment as the individuals in treatment B.

Researchers typically use three basic techniques to control other variables:

1. Random assignment, which means that each participant has an equal chance of being assigned to each of the treatment conditions. The goal of random assignment is to distribute the participant characteristics evenly between the two groups so that neither group is noticeably smarter (or older, or faster) than the other. Random assignment can also be used to control environmental variables.

2. Second, the researcher can use matching to ensure equivalent groups or equivalent environments. For example, the researcher could match groups by ensuring that every group has exactly 60% females and 40% males.
3. Finally, the researcher can control variables by holding them constant. For example, if an experiment uses only 10-yearold children as participants (holding age constant), then the researcher can be certain that one group is not noticeably older than another.

Types of Experimental Research Design:

- ***Between Subjects or Independent Measures Design*** - The two or more groups of participants take part in different treatment conditions, a between-subjects design allows only one score per participant (every score represents a separate, unique participant).

  Main Components: Manipulation: It is used to check whether the change in the DV is because of the change or manipulation of the IV and not due to any other factor. *Holding* ***Conditions Constant:*** is a control techniques experimenter use to eliminate the effects of confounding factors. ***Balancing:*** The main assumption of the experiment method is to form comparable or similar groups.

  Block randomization: works by randomizing participants within blocks such that an equal number are assigned to each treatment to avoid selection bias.

- ***Within Subjects or Repeated Measures Design*** - Repeated measures design is the one in which each participant is exposed to all treatment conditions unlike independent groups design in which different participants take part in different treatment conditions.

Within subjects or repeated measures design is used:

- When there are small number of participants
- For controlling confounding
- For increasing the sensitivity of an experiment
- Less time consuming and more convenient to arrange
- No need for a separate control group

However, there are certain ***threats to internal validity*** of such experiments e.g. practice effect. These effects can be reduced through counterbalancing or giving participants the treatment conditions by changing their order like ABBA, ABCCAB, ABCCBA, BACABC etc. Practice effects can also be balanced using block randomization.

Non-Experimental Methods

Non-experimental designs encompass all the designs that does not come in true experimental designs. From such, Correlational Designs has already been discussed. Quasi-Experiments, Case Study, Observational Designs.

Quasi-Experiments

Such experiments involve a manipulation of an independent variable or variables but there is no random assignment of participants to different treatment conditions. It is used in contexts when randomization is not possible. In a non-experimental study, the "independent variable" that is used to create the different groups of scores is often called the quasi-independent variable. Following are the two sub types of quasi experiments:

Nonequivalent Groups: study comparing boys and girls. Notice that this study involves comparing two groups of scores (like an experiment). This type of research compares preexisting groups, the researcher cannot control the assignment of participants to groups and cannot ensure equivalent groups.

Nonequivalent group studies include comparing 8-year-old children and 10-year-old children, people with an eating disorder and those with no disorder, and comparing children from a single-parent home and those from a two-parent home. Because it is impossible to use techniques like random assignment to control participant variables and ensure equivalent groups, this type of research is not a true experiment.

Pre–Post Study: The two groups of scores are obtained by measuring the same variable twice for each participant; once before and again after applying treatment. In a pre–post study, however, the researcher has no control over the passage of time. The "before" scores are always measured earlier than the "after" scores. Although a difference between the two groups of scores may be caused by the treatment, it is always possible that the scores simply change as time goes by.

Single Case Study

These involve intensive and detailed descriptions and analysis of a single case. Multiple methods for data collection including: interviews, psychological tests, observations etc. So, the data obtained can be both quantitative and qualitative in nature and it differs from experiments due to lack of control. ***Advantages:*** Case studies can provide new ideas and hypothesis. It is best method for studying rare and individual phenomena or personalities. Opportunity to try out new therapeutic interventions on patients. For challenging theories and formulating new theories.

Disadvantages: Difficult to make cause and effect assumptions. Findings are often subjective because of the subjective bias of the experimenter and hence cannot be generalized.

Advantages of Non-Experimental Methods

- Can provide new ideas and hypothesis
- For studying rare phenomena
- Opportunity to try out new therapeutic interventions on patients
- Can be used for studying rare phenomena
- For challenging theories
- Formulating new theories

Disadvantages of Non-Experimental Methods

- Difficult to make cause and effect assumptions
- Findings are often subjective
- Problems in generalizing the results
- Selection biases

Observational Designs

They involve data collection using direct or indirect observations. *Direct observations* involve looking at a behavior or phenomena directly. These include: observations without intervention and observations with intervention. *Indirect observations* looking at evidence of past behavior using archival records like birth certificates, Facebook entries, marriage licenses, college degrees etc. and physical traces like drawings, textbooks, products used by an individual etc.

Direct observations have 3 variations:

1. ***Participant observation*** *(without Intervention):* Researcher is present physically at the scene and observing from far. ***Merits:*** It allows researchers to gain a closer look at a behavior or phenomena and record information. ***De-merits:*** But it might negatively effect the behavior of participants because when people are aware that they are being watched, their behaviors change
2. ***Structured Observations or controlled observations*** *(without Intervention):* The researchers do not get involved with the participants. The behaviors to be recorded and analyzed are ore-determined. Data collection and analysis techniques are structured. The researcher decides where, when and how the observation will occur.

Advantages: Are reliable, Easy to replicate, less time consuming and can explore multiple behaviors. ***Disadvantages:*** Can lack validity if participants become aware that they are being observed. The researcher's biases can influence the results

3. ***Field Experiment or social experiments*** *(direct observation with intervention):* Researcher mix among the participants and try to manipulate circumstance to produce some effect in a natural setting. ***Merits:*** It gives opportunity to document real life behaviors, the results are generalizable and the findings have mundane realism (having real life application). ***De-merits:*** low internal validity, no control on confounding factors, results can be biased etc.

Lesson 5

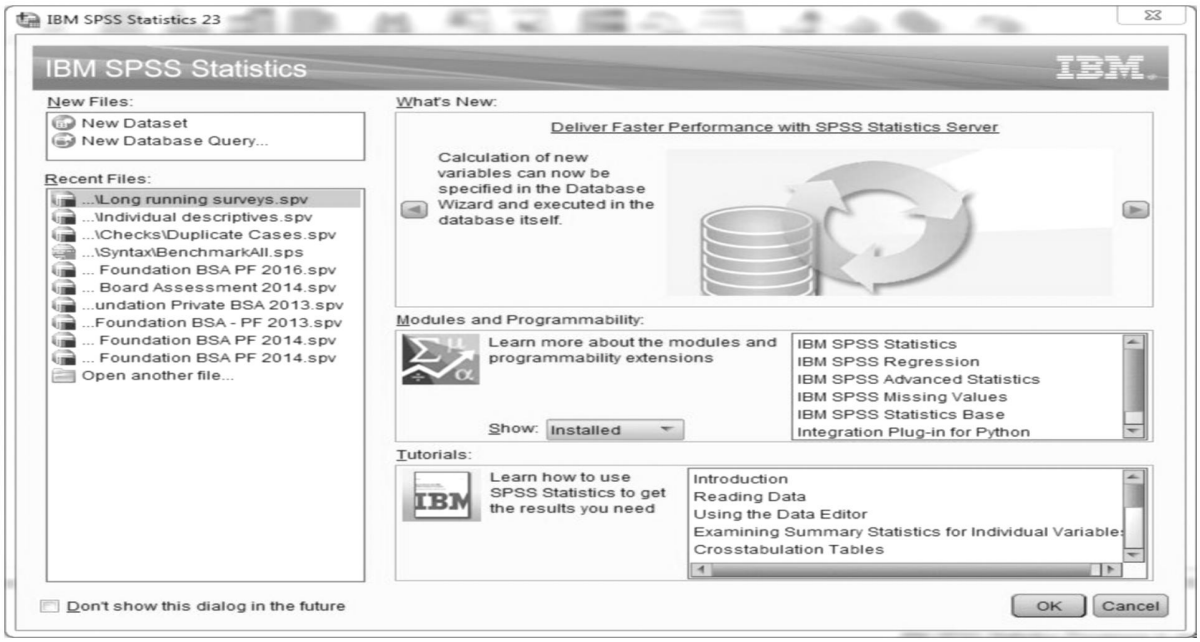
INTRODUCTION TO SPSS

Introduction to SPSS - Statistical Package for Social Science

SPSS for Windows is a popular and comprehensive data analysis package containing a multitude of features designed to facilitate the completion of a wide range of statistical analyses. It was developed for the analysis of data in the social sciences.

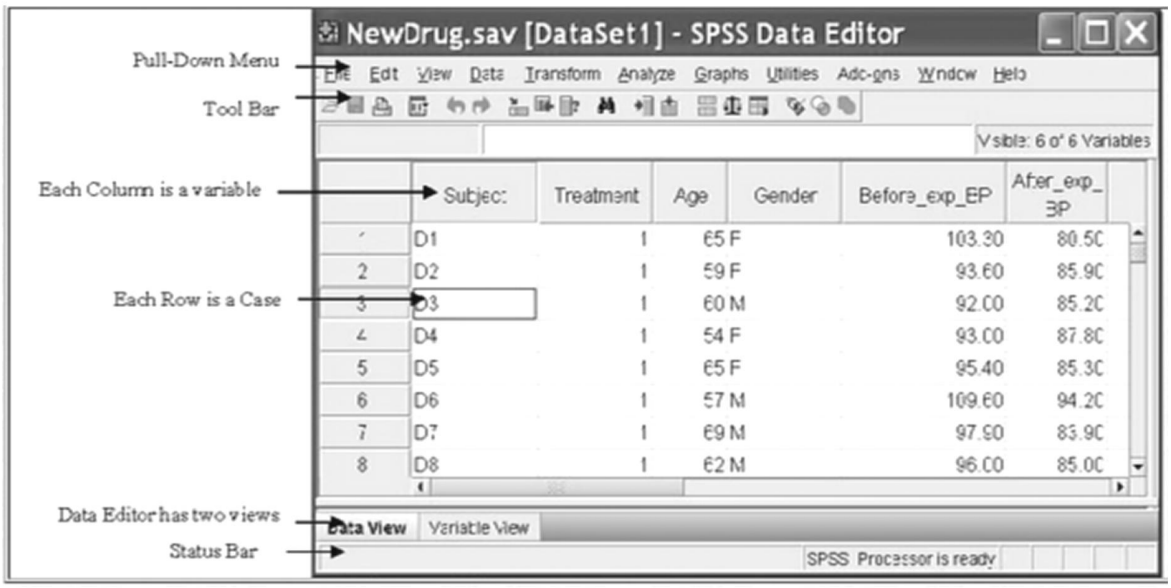
SPSS datasets always have 2-dimensional table structure where the rows typically represent cases (such as individuals or households) and the columns or variables represent measurements (such as age, sex or household income). SPSS can read and write data from ASCII text files, other statistics packages, spreadsheets and databases.

Once SPSS has been activated, a start-up window will appear, which allows you to select various options. You can open and create new file or open an existing file. It also shows the previously opened files.



After opening the file main screen shows spread sheet with two bars i.e. Menu Bar and Tool icon bar. Also, at the right bottom of the sheet show two views for data entry i.e. Data view and Variable view.

SPSS has three windows for working with data. 1. The Data Editor Window (.sav) i.e. Data view and Variable view, 2. The Output Viewer Window

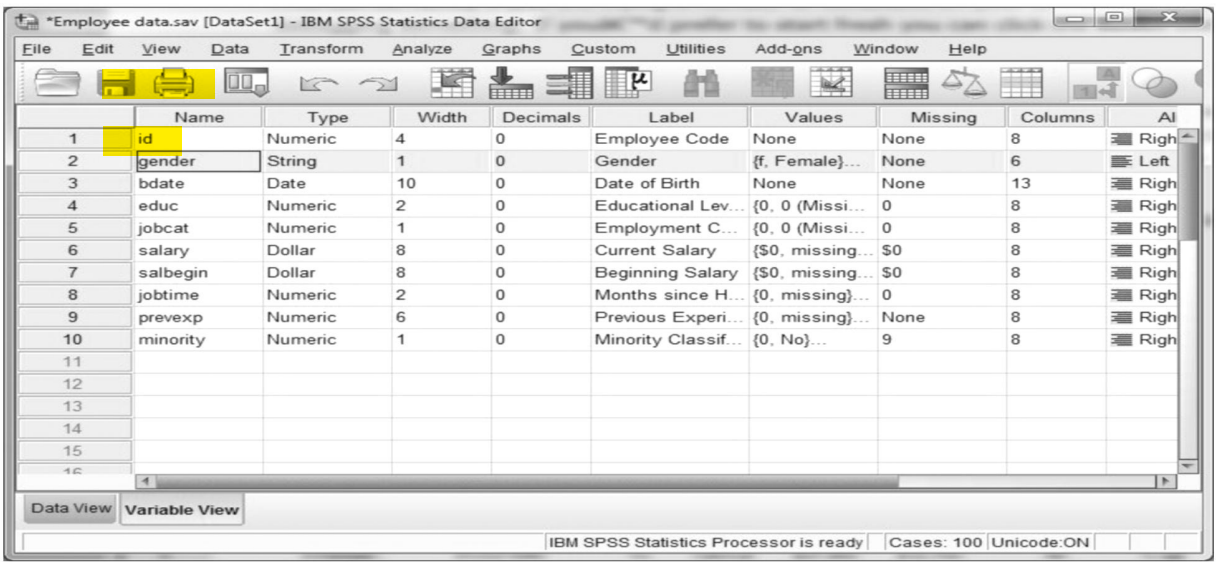


Variable View

Rows define the variable characteristics: Name, Type, Width, Decimals, Label, Values, Missing, Columns, Align, Measure.

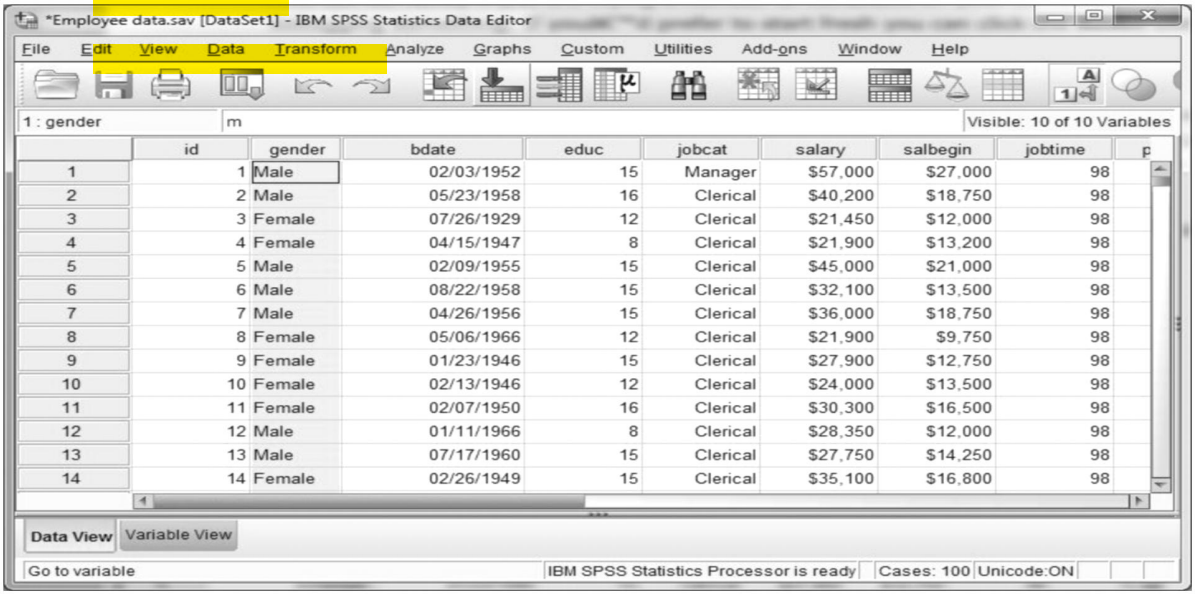
The Measure of variables in the dataset is important:

- Scale = "continuous" – e.g., age, weight, income
- Ordinal = "ordered" – categories that can be ranked (e.g., level of satisfaction or agreement, Likert-type scales)
- Nominal = "names" – categories that cannot be ranked (e.g., ID number)
- String = Letter plus numbers like Type 2.



Data View

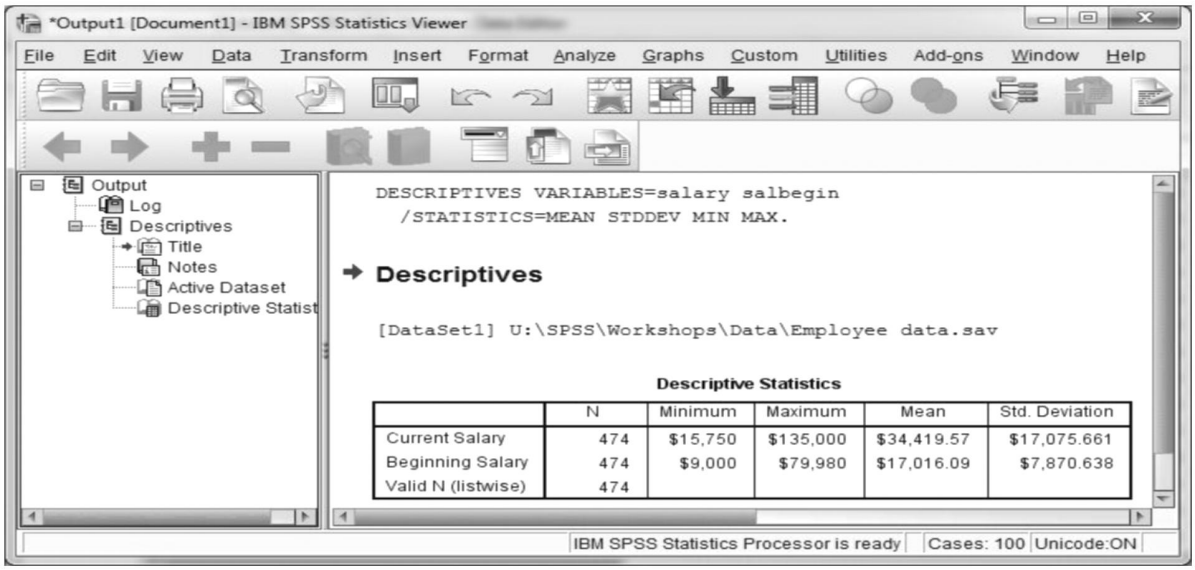
- Rows are cases (participants in the study).
- Columns are variables. A variable could be the answer to a question or any other piece of information recorded on each case.



	id	gender	bdate	educ	jobcat	salary	salbegin	jobtime
1	1	Male	02/03/1952	15	Manager	$57,000	$27,000	98
2	2	Male	05/23/1958	16	Clerical	$40,200	$18,750	98
3	3	Female	07/26/1929	12	Clerical	$21,450	$12,000	98
4	4	Female	04/15/1947	8	Clerical	$21,900	$13,200	98
5	5	Male	02/09/1955	15	Clerical	$45,000	$21,000	98
6	6	Male	08/22/1958	15	Clerical	$32,100	$13,500	98
7	7	Male	04/26/1956	15	Clerical	$36,000	$18,750	98
8	8	Female	05/06/1966	12	Clerical	$21,900	$9,750	98
9	9	Female	01/23/1946	15	Clerical	$27,900	$12,750	98
10	10	Female	02/13/1946	12	Clerical	$24,000	$13,500	98
11	11	Female	02/07/1950	16	Clerical	$30,300	$16,500	98
12	12	Male	01/11/1966	8	Clerical	$28,350	$12,000	98
13	13	Male	07/17/1960	15	Clerical	$27,750	$14,250	98
14	14	Female	02/26/1949	15	Clerical	$35,100	$16,800	98

The Output Viewer Window (.spv)

shows results of data analysis. The left-hand side is an outline of all of the output in the file. The right side is the actual output. To shrink or enlarge either side put your cursor on the line that divides them. When the double headed arrow appears, hold the left mouse button and move the line in either direction. Release the button and the size will be adjusted.

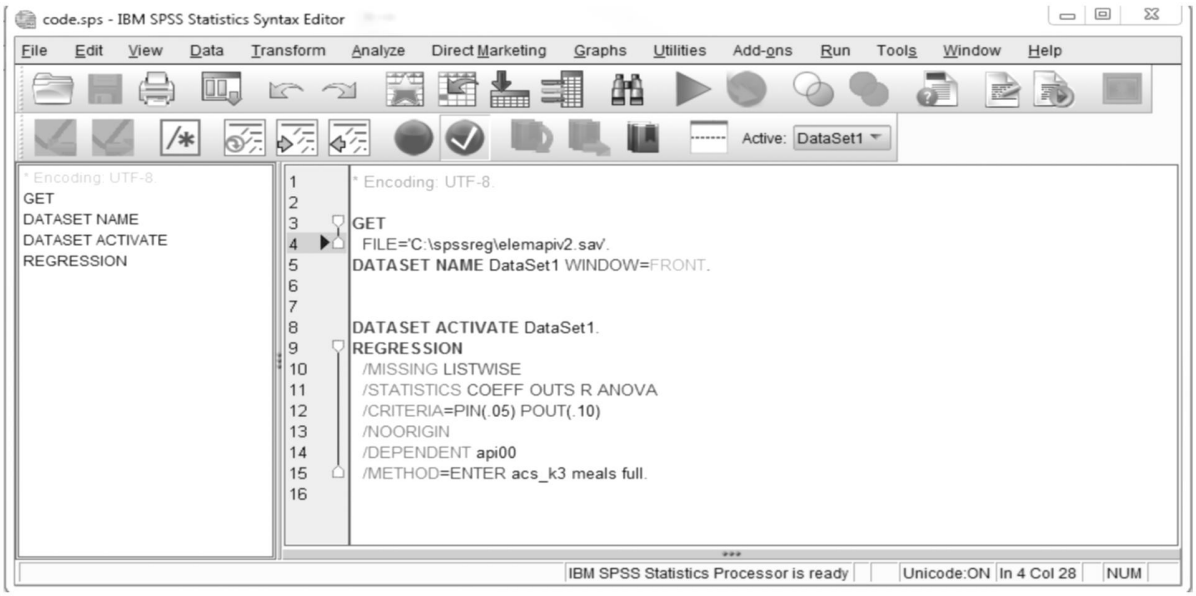


	N	Minimum	Maximum	Mean	Std. Deviation
Current Salary	474	$15,750	$135,000	$34,419.57	$17,075.661
Beginning Salary	474	$9,000	$79,980	$17,016.09	$7,870.638
Valid N (listwise)	474				

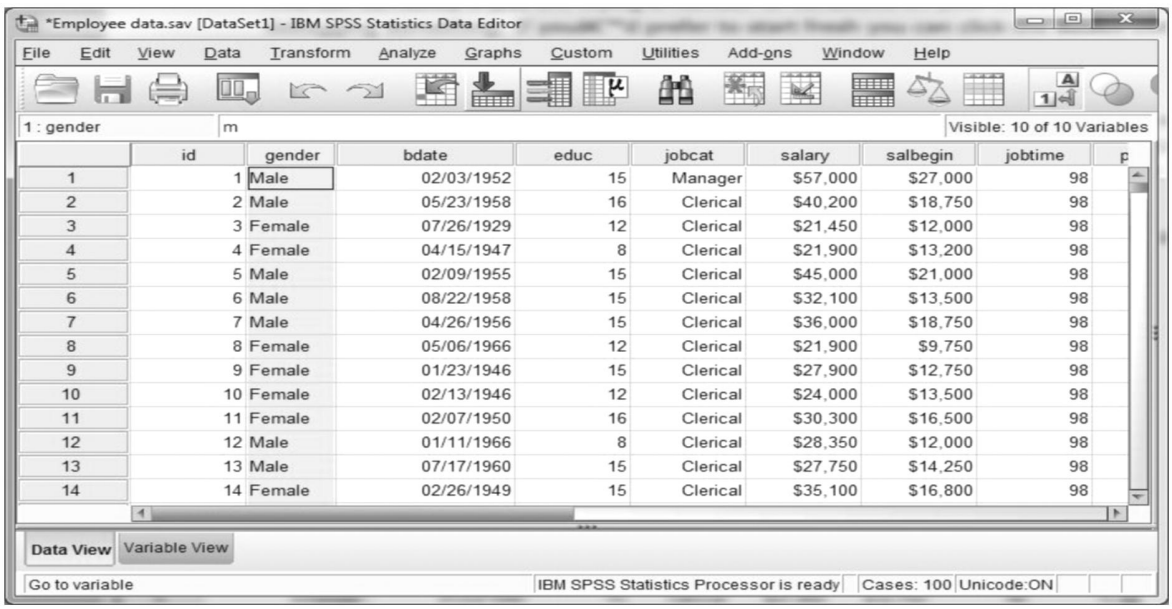
The Syntax Editor Window (.sps)

shows the syntax command script. This is also where you can type and run your own syntax commands. The Syntax window would be activated if you pasted the commands from the dialog box to it, or if you wrote you own syntax--something we will not focus on here. Syntax files end in the

extension .sps.



Data view in SPSS



	id	gender	bdate	educ	jobcat	salary	salbegin	jobtime
1	1	Male	02/03/1952	15	Manager	$57,000	$27,000	98
2	2	Male	05/23/1958	16	Clerical	$40,200	$18,750	98
3	3	Female	07/26/1929	12	Clerical	$21,450	$12,000	98
4	4	Female	04/15/1947	8	Clerical	$21,900	$13,200	98
5	5	Male	02/09/1955	15	Clerical	$45,000	$21,000	98
6	6	Male	08/22/1958	15	Clerical	$32,100	$13,500	98
7	7	Male	04/26/1956	15	Clerical	$36,000	$18,750	98
8	8	Female	05/06/1966	12	Clerical	$21,900	$9,750	98
9	9	Female	01/23/1946	15	Clerical	$27,900	$12,750	98
10	10	Female	02/13/1946	12	Clerical	$24,000	$13,500	98
11	11	Female	02/07/1950	16	Clerical	$30,300	$16,500	98
12	12	Male	01/11/1966	8	Clerical	$28,350	$12,000	98
13	13	Male	07/17/1960	15	Clerical	$27,750	$14,250	98
14	14	Female	02/26/1949	15	Clerical	$35,100	$16,800	98

Menu Bar Commands

Below is a brief reference guide to each of the menus and some of the options that they contain

- *File* includes all of the options you typically use in other programs, such as open, save, exit.
- *Edit* includes the typical cut, copy, and paste commands, and allows you to specify various options for displaying data and output.
- *View:* The option most frequently used is value labels.
- *Data* allows you to select several options ranging from displaying data that is sorted by a specific variable to selecting certain cases for subsequent analyses. For example, sorting cases, merging or aggregating files, and selecting or weighting cases.
- *Transform* includes several options to change current variables. It is used for recoding, computing new variables and dealing with missing values. example, you can change continuous variables to categorical variables, change scores into rank scores, add a constant to variables, etc.
- *Analyze* includes all of the commands to carry out statistical analyses and to calculate descriptive statistics.
- *Graphs* includes the commands to create various types of graphs including box plots, histograms, line graphs, and bar charts.
- *Utilities* allows you to list file information which is a list of all variables, there labels, values, locations in the data file, and type.
- *Add-ons* are programs that can be added to the base SPSS package. You probably do not have access to any of those.
- *Window* can be used to select which window you want to view (i.e., Data Editor, Output Viewer, or Syntax). Since we have a data file and an output file open.
- *Help* has many useful options including a link to the SPSS homepage, a statistics coach, and a syntax guide. Using topics, you can use the index option to type in any key word and get a list of options, or you can view the categories and subcategories available under contents. This is an excellent tool and can be used to troubleshoot most problems.

Toolbar Icons

The 19 Icons directly under the Menu bar provide shortcuts to many common commands that are available in specific menus.

 This icon gives you the option to open a previously saved file (if you are in the data editor, SPSS assumes you want to open a data file; if you are in the output viewer, it will offer to open a viewer file).

 This icon allows you to save files. It will save the file you are currently working on (be it data or output). If the file hasn't already been saved it will produce the Save Data As dialog box.

 This icon activates a dialog box for printing whatever you are currently working on (either the data editor or the output). By default, SPSS will print everything in the output window, so a useful way to save trees is to print only a selection of the output.

 Clicking on this icon will activate a list of the last 12 dialog boxes that you used. You can select any box from the list and it will appear on the screen. This icon makes it easy for you to repeat parts of an analysis.

 This icon enables you to go directly to a case (a row in the data editor). This button is useful if you are working on large data files: if you were analyzing a survey with 3000 respondents it would get pretty tedious scrolling down the data sheet to find the responses of participant 2407. By clicking on this icon, you can skip straight to the case by typing the case number required.

 Similar to the previous icon, clicking this button activates a function that enables you to go directly to a variable (i.e., a column in the data editor).

 Clicking on this icon opens a dialog box that shows you the variables in the data editor and summary information about each one.

 Click this button to search for words or numbers in your data file and output window. In the data editor it will search within the variable (column) that is currently active.

 Clicking on this icon inserts a new case in the data editor (so it creates a blank row at the point that is currently highlighted in the data editor). This function is very useful if you need to add new data at a particular point in the data editor.

 Clicking on this icon creates a new variable to the left of the variable that is currently active (to activate a variable simply click once on the name at the top of the column).

 Clicking on this icon is a shortcut to the function Split-1file. There are often situations in which you might want to analyze groups of cases separately. In SPSS we differentiate groups of cases by using a coding variable, and this function lets us divide our output by such a variable.

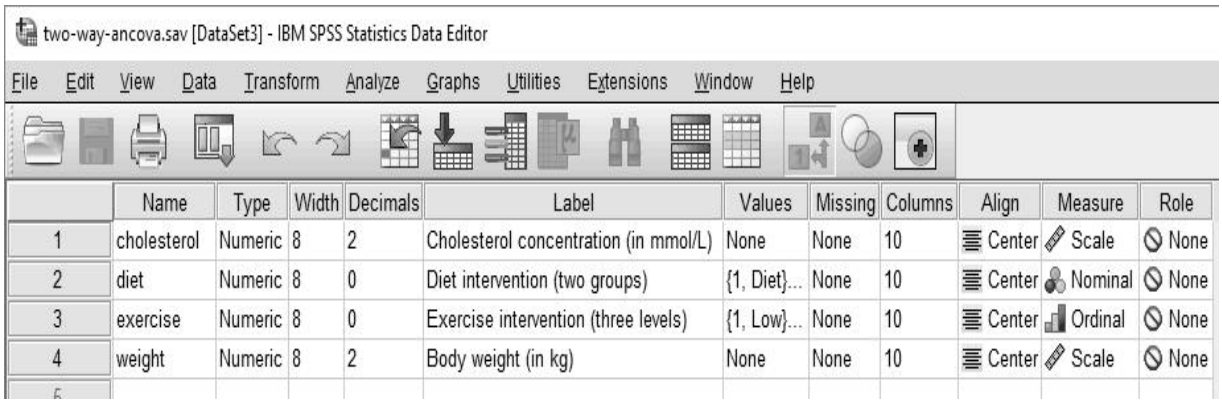
 This icon shortcuts to the function. This function is necessary when we come to input frequency data (see Section 18.5.2.2) and is useful for some advanced issues in survey sampling.

 This icon is a shortcut to the function. If you want to analyze only a portion of your data, this is the option for you. This function allows you to specify what cases you want to include in the analysis.

 Clicking on this icon will either display or hide the value labels of any coding variables. For example, if we coded gender as 1 = female, 0 = male then the computer knows that every time it comes across the value 1 in the Gender column, that person is a female. If you press this icon, the coding will appear on the data editor rather than the numerical values; so, you will see the words male and female in the Gender column rather than a series of numbers.

Variable View in SPSS

Every row of the variable view represents a variable, and you set characteristics of a particular variable by entering information into the following labelled columns (play around and you'll get the hang of it):



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	cholesterol	Numeric	8	2	Cholesterol concentration (in mmol/L)	None	None	10	Center	Scale	None
2	diet	Numeric	8	0	Diet intervention (two groups)	{1, Diet}...	None	10	Center	Nominal	None
3	exercise	Numeric	8	0	Exercise intervention (three levels)	{1, Low}...	None	10	Center	Ordinal	None
4	weight	Numeric	8	2	Body weight (in kg)	None	None	10	Center	Scale	None

Name You can enter a name in this column for each variable. This name will appear at the top of the corresponding column in the data view, and helps you to identify variables in the data view.

Type You can have different types of data. Mostly you will use numeric variables (which means that the variable contains numbers and is the default). You will come across string variables, which consist of strings of letters.

Width By default, when a new variable is created, SPSS sets it up to be numeric and to store 8 digits/characters, but you can change this value by typing a new number in this column in the dialog box. For numeric variables 8 digits is fine (unless you have very large numbers), but for string variables you will often make this value bigger (you can't write a lot in only 8 characters).

Decimals Another default setting is to have 2 decimal places displayed. (You'll notice that if you don't change this option then when you type in whole numbers to the data editor SPSS adds a decimal place with two zeros after it, which can be disconcerting.) If you want to change the number of decimal places for a given variable then replace the 2 with a new value or increase or decrease the value using.

Label The name of the variable (see above) has some restrictions on characters, and you also wouldn't want to use huge long names at the top of your columns (they become hard to read). Therefore, you can write a longer variable description in this column.

Values This column is for assigning numbers to represent groups of people.

Missing This column is for assigning numbers to missing data.

Columns Enter a number into this column to determine the width of the column, that is, how many characters are displayed in the column. (This characteristic differs from, which determines the width of the variable itself – you could have a variable of 10 characters but by setting the column width to 8 you would see only 8 of the 10 characters of the variable in the data editor.) It can be useful to increase the column width if you have a string variable that exceeds 8 characters, or a coding variable with value labels that exceed 8 characters.

Align You can use this column to select the alignment of the data in the corresponding column of the data editor. You can choose to align the data to the right, left or center.

Measure This is where you define the level at which a variable was measured (Nominal, Ordinal or Scale).

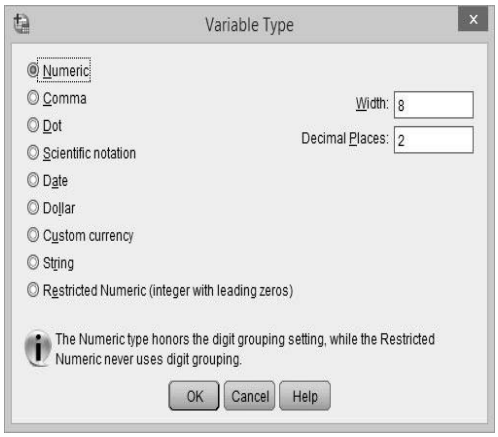
Role

There are some procedures in SPSS that attempt to run analyses automatically without you needing to think about what you're doing (one example is the Automatic Linear Modeling option in the Regression part of the Analyze menu). To think on your behalf, SPSS needs to know whether a variable is a predictor an outcome both, a variable that splits the analysis by different groups a variable that selects out part of the data or a variable that has no predefined role. These roles can be useful if you're chugging out huge numbers of analyses and want to automate them.

Computing And Recording Techniques

Creating a String Variable

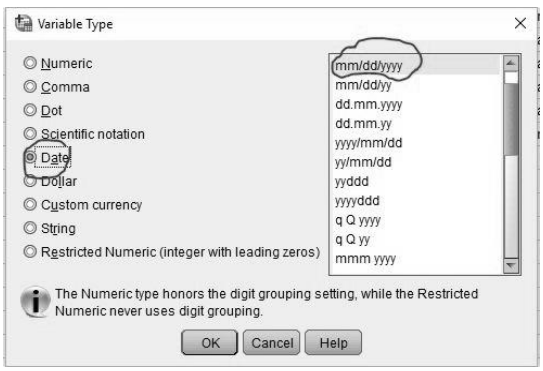
1. Click in the first white cell in the column labeled Name.
2. Type the word 'Name'.
3. Move off this cell using the arrow keys on the keyboard (you can also just click on a different cell, but this is a very slow way of doing it).
4. Move into the column labeled TYPE. Click on it and a dialogue box will open.
5. Choose the STRING option and click OK to make the variable string. SPSS assume that we want a numeric variable (i.e., numbers), therefore to create string variable we have to change the variable type.



6. Finally, to specify Measure and selecting either Nominal, Ordinal or Scale from the drop-down list.

Creating a Date Variable

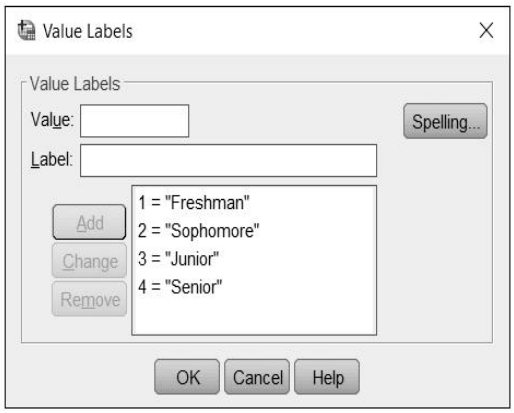
1. To enter date variables into SPSS we use the same procedure as with the string variable, except that we need to change the variable type.
2. Move into the column labelled TYPE. Click on it and a dialogue box will open.



3. Choose the DATE option and click OK to make the variable string.

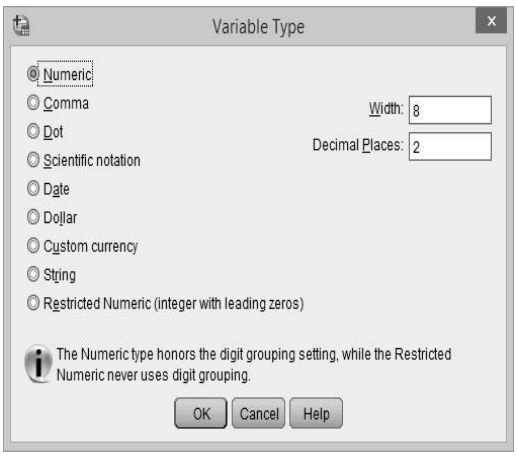
Creating Coding Variables

1. A coding variable (also known as a grouping variable) uses numbers to represent different groups of data. As such, it is a numeric variable, but these numbers represent names (i.e., it is a nominal variable).
2. For coded categorical variables, the value label(s) that should be associated with each category abbreviation. Value labels are useful primarily for categorical (i.e., nominal or ordinal) variables, especially if they have been recorded as codes (e.g., 1, 2, 3).
3. Example: In the sample dataset, the variable Rank represents the student's class rank. The values 1, 2, 3, 4 represent the categories Freshman, Sophomore, Junior, and Senior, respectively.



Creating a Numeric Variable

1. Numeric variables are the easiest ones to create because SPSS assumes this format for data. e.g. Age, No. of Friends, total income etc.
2. Click in the first white cell in the column labelled Name and type Age.
3. Move into the column labelled TYPE. Click on it and a dialogue box will open. Choose Numeric and click ok.
4. Specify the level at which a variable was measured by going to the column labelled Measure and selecting SCALE.



Lesson 6

DESCRIPTIVE STATISTICS

Although researchers have developed a variety of different statistical procedures to organize and interpret data, these different procedures can be classified into two general categories: descriptive and inferential. Each of these segments is important, offering different techniques that accomplish different objectives. The first category, descriptive statistics, consists of statistical procedures that are used to simplify and summarize data.

Descriptive statistics

Descriptive statistics are statistical procedures used to summarize, organize, and simplify data. Descriptive statistics are techniques that take raw scores and organize or summarize them in a form (graphically or numerically) that is more manageable. Often the scores are organized in a table or a graph so that it is possible to see the entire set of scores. Another common technique is to summarize a set of scores by computing an average. Note that even if the data set has hundreds of scores, the average provides a single descriptive value for the entire set.

Two objectives for descriptive statistics are:

We want to choose a statistic that shows how different units seems similar i.e. Central tendency.

We want to choose another statistic that shows how they differ i.e. Variability.

For example, you want to study the status of different leisure activities by gender. You distribute a survey and ask participants how many times they did each of the following in the past year:

- Go to a library
- Watch a movie at a theater
- Visit a national park

Your data set is the collection of responses to the survey. Now you can use descriptive statistics to find out the overall frequency of each activity (distribution), the averages for each activity (central tendency), and the spread of responses for each activity (variability).

Descriptive statistics usually include:

- Tables/Distributions
- Central Tendency
- Measures of Dispersion
- Graphs/Charts

Presenting the data: Graphs

Descriptive statistics are very important because if we simply presented our raw data it would be hard to visualize what the data was showing, especially if there was a lot of it. Descriptive statistics therefore enables us to present the data in a more meaningful way, which allows simpler interpretation of the data. For example, if we had the results of 100 pieces of students' coursework, we may be interested in the overall performance of those students. We would also be interested in the distribution or spread of the marks.

When we use descriptive statistics, it is useful to summarize the group of data using a combination of tabulated description (i.e., tables), graphical description (i.e., graphs and charts) and statistical commentary (i.e., a discussion of the results).

Histograms

A histogram is a graphical display of a distribution. It presents the same information as a frequency table but in a way that is even quicker and easier to grasp.

Lesson 7

FREQUENCY DISTRIBUTION-I

The results from a research study usually consist of pages of numbers corresponding to the measurements, or scores, collected during the study. The immediate problem for the researcher is to organize the scores into some comprehensible form so that any patterns in the data can be seen easily and communicated to others. This is the job of descriptive statistics: to simplify the organization and presentation of data. One of the most common procedures for organizing a set of data is to place the scores in a frequency distribution.

Frequency Distribution

A frequency distribution is an organized tabulation of the number of individuals located in each category on the scale of measurement.

A frequency distribution takes a disorganized set of scores and places them in order from highest to lowest, grouping together individuals who all have the same score. If the highest score is *X* _ 10, for example, the frequency distribution groups together all the 10s, then all the 9s, then the 8s, and so on. Thus, a frequency distribution allows the researcher to see "at a glance" the entire set of scores. It shows whether the scores are generally high or low, whether they are concentrated in one area or spread out across the entire scale, and generally provides an organized picture of the data.

A frequency distribution can be structured either as a table or as a graph, but in either case, the distribution presents the same two elements:

1. The set of categories that make up the original measurement scale.
2. A record of the frequency, or number of individuals in each category.

Thus, a frequency distribution presents a picture of how the individual scores are distributed on the measurement scale—hence the name *frequency distribution*.

Frequency Distribution Tables

The simplest frequency distribution table presents the measurement scale by listing the different measurement categories (*X* values) in a column from highest to lowest. Beside each *X* value, we indicate the frequency, or the number of times that particular measurement occurred in the data.

It is customary to use an *X* as the column heading for the scores and an *f* as the column heading for the frequencies.

It is customary to list categories from highest to lowest, but this is an arbitrary arrangement. Many computer programs list categories from lowest to highest.

Proportions And Percentages

In addition to the two basic columns of a frequency distribution, there are other measures that describe the distribution of scores and can be incorporated into the table. The two most common are proportion and percentage. Proportion measures the fraction of the total group that is associated with each score. In Example 2.2, there were two individuals with $X = 4$. Thus, 2 out of 10 people had $X = 4$, so the proportion would be 2/10= 0.20. In general, the proportion associated with each score is

$$Proportion = P = f/N$$

Because proportions describe the frequency (*f*) in relation to the total number (*N*), they often are called *relative frequencies*. Although proportions can be expressed as fractions (for example, 2/10), they more commonly appear as decimals. A column of proportions, headed with a *p,* can be added to the basic frequency distribution table.

In addition to using frequencies (*f*) and proportions (*p*), researchers often describe a distribution of scores with percentages. For example, an instructor might describe the results of an exam by saying that 15% of the class earned *A*s, 23% earned *B*s, and so on. To compute the percentage associated with each score, you first find the proportion (*p*) and then multiply by 100:

Percentage = p(100) = f/N (100)

Percentages can be included in a frequency distribution table by adding a column headed with %.

X	f	p = f/N	% = p(100)
5	1	1/10 = 0.10	10%
4	2	2/10 = 0.20	20%
3	3	3/10 = 0.30	30%
2	3	3/10 = 0.30	30%
1	1	1/10 = 0.10	10%

Grouped Frequency Distribution

If we were to list all of the individual scores from $X = 96$ down to $X = 41$, it would take 56 rows to complete the frequency distribution table. Although this would organize the data, the table would be long and cumbersome. Remember: The purpose for constructing a table is to obtain a relatively simple, organized picture of the data. This can be accomplished by grouping the scores into intervals and then listing the intervals in the table instead of listing each individual score. For example, we could construct a table showing the number of students who had scores in the 90s, the number with scores in the 80s, and so on. The result is called a *grouped frequency distribution table* because we are presenting groups of scores rather than individual values. The groups, or intervals, are called *class intervals*.

Grouped frequency distribution tables—group the scores into intervals and list these intervals in the frequency distribution table. Remember, when the scores are whole numbers, the number of rows is determined by highest scores– lowest scores+ 1

Rule of thumb is to have 5 to 10 intervals or less (of equal width)

Example

An instructor has obtained the set of N=25 exam scores shown here. To help organize these scores, we will place them in a frequency distribution table. The scores are:

82, 75, 88, 93, 53, 84, 87, 58, 72, 94, 69, 84, 61, 91, 64, 87, 84, 70, 76, 89, 75, 80, 73, 78, 60

The first step is to determine the range of scores. For these data, the smallest score is X=53 and the largest score is X=94, so a total of 42 rows would be needed for a table that lists each individual score. Because 42 rows would not provide a simple table, we have to group the scores into class intervals.

The best method for finding a good interval width is a systematic trial-and-error approach that uses guidelines 1 and 2 simultaneously. Specifically, we want about 10 intervals and we want the interval width to be a simple number. For this example, the scores cover a range of 42 points, so we will try several different interval widths to see how many intervals are needed to cover this range. For example, if each interval is 2 points wide; it would take 21 intervals to cover a range of 42 points. This is too many, so we move on to an interval width of 5 or 10 points. The following table shows how many intervals would be needed for these possible widths:

Width	Number of Intervals Needed to Cover a Range of 42 Points	
2	21	(too many)
5	9	(OK)
10	5	(too few)

Notice that an interval width of 5 will result in about 10 intervals, which is exactly what we want. The next step is to actually identify the intervals. The lowest score for these data is X=53, so the lowest interval should contain this value. Because the interval should have a multiple of 5 as its bottom score, the interval should begin at 50. The interval has a width of 5, so it should contain 5 values: 50, 51, 52, 53, and 54. Thus, the bottom interval is 50–54. The next interval would start at 55 and go to 59. Note that this interval also has a bottom score that is a multiple of 5, and contains exactly 5 scores (55, 56, 57, 58, and 59). The complete frequency distribution table showing all of the class intervals is presented in Table below:

X	*f*
90–94	3
85–89	4
80–84	5
75–79	4
70–74	3
65–69	1
60–64	3
55–59	1
50–54	1

This grouped frequency distribution table shows the data from Example 2.4. The original scores range from a high of X=94 to a low of X=53. This range has been divided into 9 intervals with each interval exactly 5 points wide. The frequency column (f) lists the number of individuals with scores in each of the class intervals.

Frequency Distribution Graphs/Charts

When the data consist of numerical scores that have been measured on an interval or ratio scale, there are two options for constructing a frequency distribution graph. The two types of graphs are called *histograms* and *polygons.*

Histograms To construct a histogram, you first list the numerical scores (the categories of measurement) along the *X*-axis. Then you draw a bar above each *X* value so that

a. The height of the bar corresponds to the frequency for that category.
b. For continuous variables, the width of the bar extends to the real limits of the category. For discrete variables, each bar extends exactly half the distance to the adjacent category on each side.

Polygons The second option for graphing a distribution of numerical scores from an interval or ratio scale of measurement is called a polygon. To construct a polygon, you begin by listing the numerical scores (the categories of measurement) along the *X*-axis. Then,

a) A dot is centered above each score so that the vertical position of the dot corresponds to the frequency for the category.
b) A continuous line is drawn from dot to dot to connect the series of dots.
c) The graph is completed by drawing a line down to the *X*-axis (zero frequency) at each end of the range of scores. The final lines are usually drawn so that they reach the *X*-axis at a point that is one category below the lowest score on the left side and one category above the highest score on the right side.

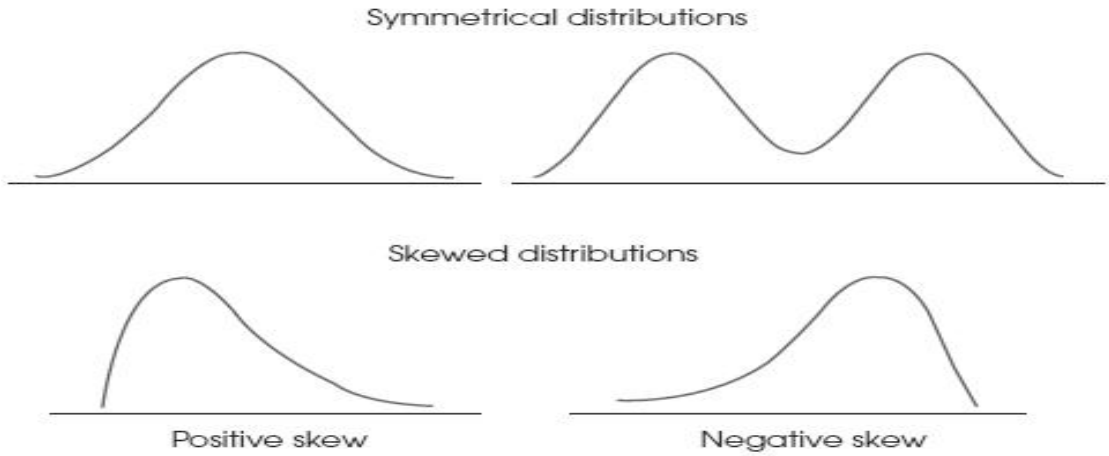
When the scores are measured on a nominal or ordinal scale (usually non-numerical values), the frequency distribution can be displayed in a *bar graph.*

Bar Graphs A bar graph is essentially the same as a histogram, except that spaces are left between adjacent bars. For a nominal scale, the space between bars emphasizes that the scale consists of separate, distinct categories. For ordinal scales, separate bars are used because you cannot assume that the categories are all the same size.

To construct a bar graph, list the categories of measurement along the *X*-axis and then draw a bar above each category so that the height of the bar equals the frequency for the category.

Nearly all distributions can be classified as being either ***symmetrical* or *skewed*.** In a symmetrical distribution, it is possible to draw a vertical line through the middle so that one side of the distribution is a mirror image of the other.

In a skewed distribution, the scores tend to pile up toward one end of the scale and taper off gradually at the other end.



Lesson 8

FREQUENCY DISTRIBUTION-II

Percentiles and Percentile Ranks

Although the primary purpose of a frequency distribution is to provide a description of an entire set of scores, it also can be used to describe the position of an individual within the set. Individual scores, or *X* values, are called *raw scores*. By themselves, raw scores do not provide much information. For example, if you are told that your score on an exam is *X* _ 43, you cannot tell how well you did relative to other students in the class. To evaluate your score, you need more information, such as the average score or the number of people who had scores above and below you. With this additional information, you would be able to determine your relative position in the class. Because raw scores do not provide much information, it is desirable to transform them into a more meaningful form. One transformation that we consider changes raw scores into *percentiles*.

The rank or percentile rank of a particular score is defined as the percentage of individuals in the distribution with scores equal to or less than the particular value.

When a score is identified by its percentile rank, the score is called a percentile.

Suppose, for example, that you have a score of $X = 43$ on an exam and that you know that exactly 60% of the class had scores of 43 or lower. Then your score $X = 43$ has a percentile rank of 60%, and your score would be called the 60^{th} percentile. Notice that *percentile rank* refers to a percentage and that *percentile* refers to a score. Also notice that your rank or percentile describes your exact position within the distribution.

To determine percentiles or percentile ranks, the first step is to find the number of individuals who are located at or below each point in the distribution. This can be done most easily with a frequency distribution table by simply counting the number who are in or below each category on the scale. The resulting values are called ***cumulative frequencies*** because they represent the accumulation of individuals as you move up the scale.

The cumulative frequencies show the number of individuals located at or below each score. To find percentiles, we must convert these frequencies into percentages. The resulting values are called ***cumulative percentages*** because they show the percentage of individuals who are accumulated as you move up the scale.

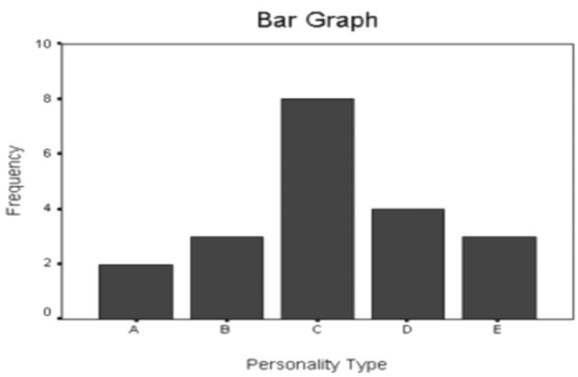
Stem and Leaf Displays

In 1977, J.W. Tukey presented a technique for organizing data that provides a simple alternative to a grouped frequency distribution table or graph (Tukey, 1977). This technique, called a *stem and leaf display*, requires that each score be separated into two parts: The first digit (or digits) is called the *stem*, and the last digit is called the *leaf*. For example, $X = 85$ would be separated into a stem of 8 and a leaf of 5. Similarly, $X = 42$ would have a stem of 4 and a leaf of 2.

Data			Stem and Leaf Display	
83	82	63	3	23
62	93	78	4	26
71	68	33	5	6279
76	52	97	6	283
85	42	46	7	1643846
32	57	59	8	3521
56	73	74	9	37
74	81	76		

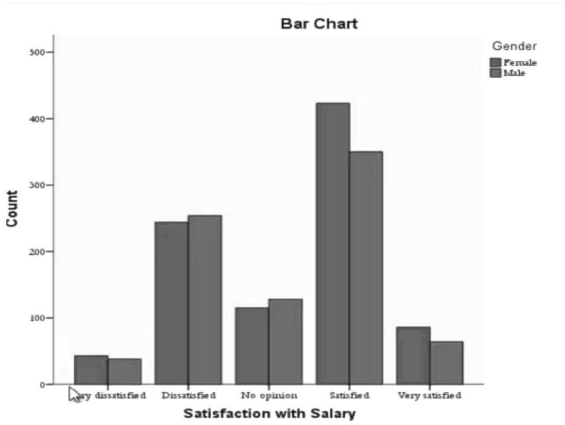
Art of Presenting Data

Bar Graph— like a histogram, a bar is drawn above each X value, so that the height of the bar corresponds to the frequency of the score. Usually is from discrete (nominal or ordinal level data).



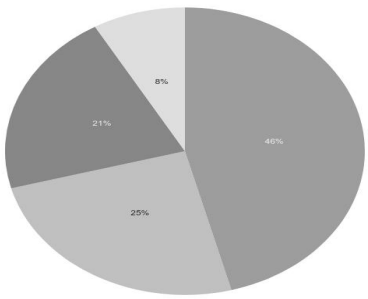
Clustered Bar Graph— can be used when you have two or more nominal or ordinal variables and want to illustrate the differences in the categories of these two variables based on some statistic.

e.g. comparing satisfaction with salary in males in females.



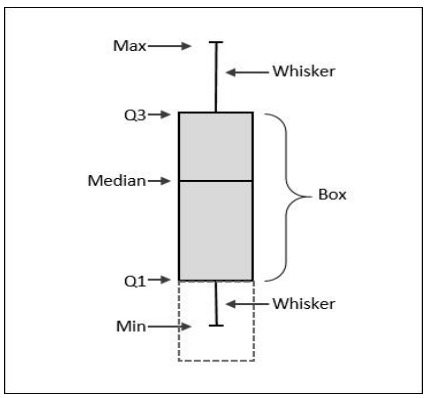
Pie Chart— is a special chart that uses Pie slices to show the relative sizes of the data.

Uses Nominal data; e.g. which movie types are most liked, (you can see in a glance that Romantic movies (blue slice) are liked more.



Romantic	Action	Horror	Drama
46%	25%	21%	8%

Box Plot— A boxplot is a graph that gives you a good indication of how the values in the data are spread out. They useful when comparing distributions between many groups or datasets.



Exercise: Constructing Frequency Distribution

For Frequency Distribution:

Enter data in data view→analyze→descriptive statistics→frequencies→drag scores in variable(s) box→ok. So here you can find frequency distribution table

For the grouped distribution got to transform→ visual binning→ send scores in variables to bin box→continue→name variable in binned variable box which will add a variable in your data set (I.e. groups), to make cut points click→ make cut points→ define value for first cut point (I.e. 5 for data starting from 2), define width (it will automatically define number of cut points)→apply→make labels→ok→ok

If you want exact frequency distribution that we calculated manually, click analyze→descriptive statistics→frequencies→drag binned variable(s) box→ok. So here you can find frequency distribution table of grouped data

For Graphs:

analyze→descriptive statistics→frequencies→ select variable (scores)→statistics tab→define percentile, minimum, maximum→continue→ click on charts→histograms→continue→ok

For Cumulative Percentage Curve/Ogive Percentile Graph

Graphs→legacy dialogue→histogram→drop variable in variable box→ok

You can also make it through legacy chart, for this click on Graphs→chart builder→ok→select histogram→select, hold and drag to the chart preview box→define variable on X-Axis→ok

For Polygon

Graphs→legacy dialogues→line chart→simple→define→drag score to category Axis→ok

For Ogive

Graphs→legacy dialogues→line chart→simple→define→drag score to category Axis→select cum%→ok

For Bar Diagram

Graphs→legacy dialogues→bar→simple→define→drag score to category Axis→ok

Similar procedure can be used for nominal or ordinal data

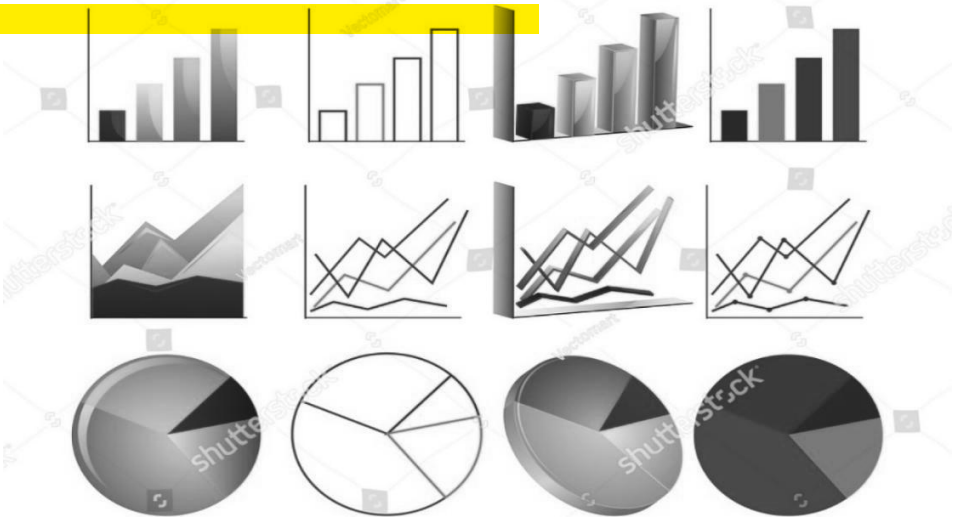
For Pie Chart

Graphs→legacy dialogues→pie→define→drag binned data to "define slices by"box→% of cases

Chart can also be edit by double clicking on chart, values and titles can also be placed.

For Cluster Bar Chart

First add categorical data, Graphs→legacy dialogues→bar→clustered→define→define clusters by gender→drag binned score in category axis→ok



Lesson 9

MEASURE OF CENTRAL TENDENCY-I

Central tendency is a statistical measure to determine a single score that defines the center of a distribution. The goal of central tendency is to find the single score that is most typical or most representative of the entire group.

In everyday language, central tendency attempts to identify the "average" or "typical" individual. This average value can then be used to provide a simple description of an entire population or a sample. In addition to describing an entire distribution, measures of central tendency are also useful for making comparisons between groups of individuals or between sets of figures. For example, weather data indicate that for Seattle, Washington, the average yearly temperature is 53° and the average annual precipitation is 34 inches. By comparison, the average temperature in Phoenix, Arizona, is 71° and the average precipitation is 7.4 inches. The point of these examples is to demonstrate the great advantage of being able to describe a large set of data with a single, representative number.

Unfortunately, there is no single, standard procedure for determining central tendency. To deal with these problems, statisticians have developed three different methods for measuring central tendency: the mean, the median, and the mode.

MEAN

The mean for a distribution is the sum of the scores divided by the number of scores.

The *mean*, also known as the arithmetic average, is computed by adding all the scores in the distribution and dividing by the number of scores. The mean for a population is identified by the Greek letter mu, μ (pronounced "mew"), and the mean for a sample is identified by M or $\bar{x}$, (read "x-bar").

The formula for the *population mean* is

$$\mu = \Sigma X/N$$

First, add all the scores in the population, and then divide by N. For a sample, the computation is exactly the same, but the formula for the *sample mean* uses symbols that signify sample values:

$$\text{Sample mean} = M = \Sigma x/n$$

Example:

For a population of $N = 4$ scores,

3, 7, 4, 6

the mean is

$$\mu = \frac{\Sigma X}{N} = \frac{20}{4} = 5$$

Weighted Mean

When we have two or more sets of data and want to find overall mean for the combined group.

Sample 1= 4, 5, 6, 7, 8

Sample 2= 4, 6, 8, 10, 12

$$\text{overall mean} = M = \frac{\Sigma X \text{ (overall sum for the combined group)}}{n \text{ (total number in the combined group)}}$$

$$= \frac{\Sigma X_1 + \Sigma X_2}{n_1 + n_2}$$

Mean from Grouped Data

The first step is to determine the midpoint of each interval, or class. These midpoints must then be multiplied by the frequencies of the corresponding classes. The sum of the products divided by the total number of values will be the value of the **mean.**

$$\overline{x} = \frac{\sum fx}{n}$$

MEDIAN

The second measure of central tendency we consider is called the *median*. The goal of the median is to locate the midpoint of the distribution. Unlike the mean, there are no specific symbols or notation to identify the median. Instead, the median is simply identified by the word *median*. In addition, the definition and the computations for the median are identical for a sample and for a population.

If the scores in a distribution are listed in order from smallest to largest, the **median** is the midpoint of the list. More specifically, the median is the point on the measurement scale below which 50% of the scores in the distribution are located. Defining the median as the *midpoint* of a distribution means that the scores are divided into two equal-sized groups.

This example demonstrates the calculation of the median when *n* is an odd number. With an odd number of scores, you list the scores in order (lowest to highest), and the median is the middle score in the list. Consider the following set of N =5 scores, which have been listed in order:

3, 5, 8, 10, 11

The middle score is X =8, so the median is equal to 8. Using the counting method, with N =5 scores, the 50% point would be 2 ½ scores. Starting with the smallest scores, we must count the 3, the 5, and the 8 before we reach the target of at least 50%. Again, for this distribution, the median is the middle score, X=8.

Finding Median with Grouped Data

$$M_m = l + \left(\frac{\frac{n}{2} - cf}{f}\right)h$$

Where

l= Lower limit of median class,

n= number of observations,

cf= cumulative frequency of class preceding the median class,

f= frequency of median class,

h= class size (assuming class size to be equal)

MODE

The final measure of central tendency that we consider is called the *mode*. In its common usage, the word *mode* means "the customary fashion" or "a popular style." The statistical definition is similar in that the mode is the most common observation among a group of scores.

In a frequency distribution, the mode is the score or category that has the greatest frequency.

As with the median, there are no symbols or special notation used to identify the mode or to differentiate between a sample mode and a population mode. In addition, the definition of the mode is the same for a population and for a sample distribution. The mode is a useful measure of central tendency because it can be used to determine the typical or average value for any scale of measurement, including a nominal scale.

Extreme Scores or Skewed Distributions

When a distribution has a few extreme scores, scores that are very different in value from most of the others, then the mean may not be a good representative of the majority of the distribution.

The problem comes from the fact that one or two extreme values can have a large influence and cause the mean to be displaced. In this situation, the fact that the mean uses all of the scores equally can be a disadvantage. Consider, for example, the distribution of n =10 scores. For this sample, the mean is

$$M = \Sigma x/n = 203/10 = 20.3$$

Notice that the mean is not very representative of any score in this distribution. Although most of the scores are clustered between 10 and 13, the extreme score of X =100 inflates the value of ΣX and distorts the mean.

The median, on the other hand, is not easily affected by extreme scores. For this sample, n _ 10, so there should be five scores on either side of the median. The median is 11.50. Notice that this is a very representative value. Also note that the median would be unchanged even if the extreme score were 1000 instead of only 100.

Because it is relatively unaffected by extreme scores, the median commonly is used when reporting the average value for a skewed distribution. For example, the distribution of personal incomes is very skewed, with a small segment of the population earning incomes that are astronomical. These extreme values distort the mean, so that it is not very representative of the salaries that most of us earn.

The median is the preferred measure of central tendency when extreme scores exist.

Lesson 10

MEASURE OF CENTRAL TENDENCY-II

When To Use Which Measure Of Central Tendency?

The goal of central tendency is to find out one single value that best represents the entire distribution. Mean is preferred measure as it uses every score in the distribution and is related to standard deviation so is valuable measure in inferential statistics. But then when and why to use median and mode?

When To Use The Median?

- When there are extreme scores in data/ skewed distribution
- For undetermined or incomplete data points
- Open ended distribution (e.g. 5 or less, 20 or more)
- When data is ordinal scale

When To Use The Mode?

We consider three situations in which the mode is commonly used as an alternative to the mean, or is used in conjunction with the mean to describe central tendency.

1. **Nominal Scales** The primary advantage of the mode is that it can be used to measure and describe central tendency for data that are measured on a nominal scale. Recall that the categories that make up a nominal scale are differentiated only by name. Because nominal scales do not measure quantity (distance or direction), it is impossible to compute a mean or a median for data from a nominal scale. Therefore, the mode is the only option for describing central tendency for nominal data.
2. **Discrete Variables** Recall that discrete variables are those that exist only in whole, indivisible categories. Often, discrete variables are numerical values, such as the number of children in a family or the number of rooms in a house. When these variables produce numerical scores, it is possible to calculate means. In this situation, the calculated means are usually fractional values that cannot actually exist. For example, computing means generates results such as "the average family has 2.4 children and a house with 5.33 rooms." On the other hand, the mode always identifies the most typical case and, therefore, it produces more sensible measures of central tendency. Using the mode, our conclusion would be "the typical, or modal, family has 2 children and a house with 5

rooms." In many situations, especially with discrete variables, people are more comfortable using the realistic, whole-number values produced by the mode.

3. **Describing Shape** Because the mode requires little or no calculation, it is often included as a supplementary measure along with the mean or median as a no-cost extra. The value of the mode (or modes) in this situation is that it gives an indication of the shape of the distribution as well as a measure of central tendency. Remember that the mode identifies the location of the peak (or peaks) in the frequency distribution graph. For example, if you are told that a set of exam scores has a mean of 72 and a mode of 80, you should have a better picture of the distribution than would be available from the mean alone.

Central Tendency And The Shape Of The Distribution

For a ***symmetrical distribution,*** the right-hand side of the graph is a mirror image of the left-hand side. If a distribution is perfectly symmetrical, the median is exactly at the center because exactly half of the area in the graph is on either side of the center. The mean also is exactly at the center of a perfectly symmetrical distribution because each score on the left side of the distribution is balanced by a corresponding score (the mirror image) on the right side. As a result, the mean (the balance point) is located at the center of the distribution. Thus, for a perfectly symmetrical distribution, the mean and the median are the same (Figure 10.1). If a distribution is roughly symmetrical, but not perfect, the mean and median are close together in the center of the distribution.

If a symmetrical distribution has only one mode, it is also in the center of the distribution. Thus, for a perfectly symmetrical distribution with one mode, all three measures of central tendency, the mean, the median, and the mode, have the same value. For a roughly symmetrical distribution, the three measures are clustered together in the center of the distribution. On the other hand, a bimodal distribution that is symmetrical [see Figure 10.1(b)] has the mean and median together in the center with the modes on each side. A rectangular distribution [see Figure 10.1(c)] has no mode because all X values occur with the same frequency. Still, the mean and the median are in the center of the distribution.

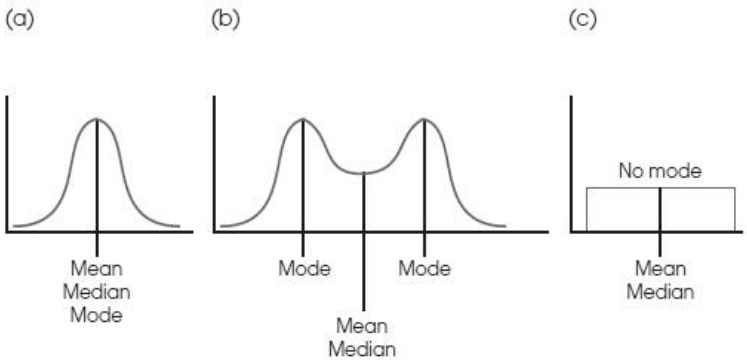


Figure 10.1 Measures of central tendency for three symmetrical distributions: normal, bimodal, and recntangular

In ***skewed distributions,*** especially distributions for continuous variables, there is a strong tendency for the mean, median, and mode to be located in predictably different positions. Figure 10.2(a), for example, shows a positively skewed distribution with the peak (highest frequency) on the left-hand side. This is the position of the mode. However, it should be clear that the vertical line drawn at the mode does not divide the distribution into two equal parts. To have exactly 50% of the distribution on each side, the median must be located to the right of the mode. Finally, the mean is located to the right of the median because it is the measure influenced most by the extreme scores in the tail and is displaced farthest to the right toward the tail of the distribution. Therefore, in a positively skewed distribution, the order of the three measures of central tendency from smallest to largest (left to right) is the mode, the median, and the mean.

Negatively skewed distributions are lopsided in the opposite direction, with the scores piling up on the right-hand side and the tail tapering off to the left. The grades on an easy exam, for example, tend to form a negatively skewed distribution [see Figure 10.2(b)]. For distribution with negative skew, the mode is on the right-hand side (with the peak), whereas the mean is displaced toward the left by the extreme scores in the tail. As before, the median is located between the mean and the mode. In order from the smallest value to the largest value (left to right), the three measures of central tendency for a negatively skewed distribution are the mean, the median, and the mode.

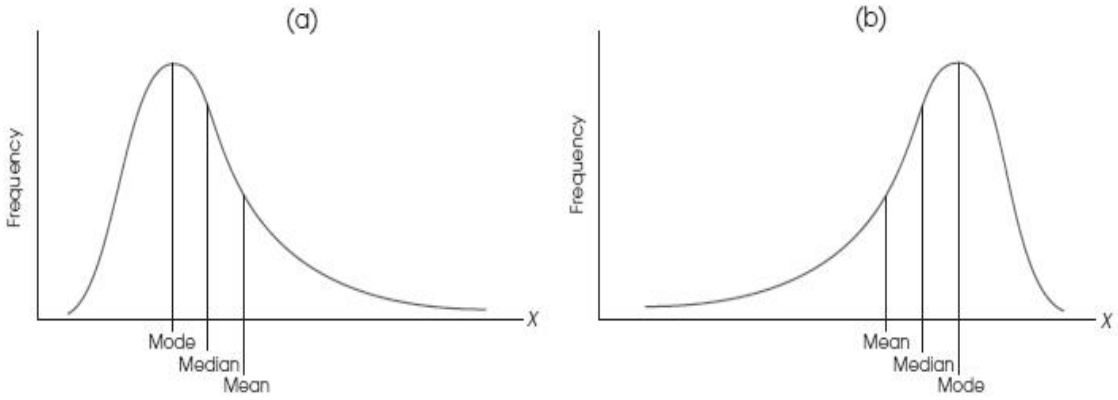


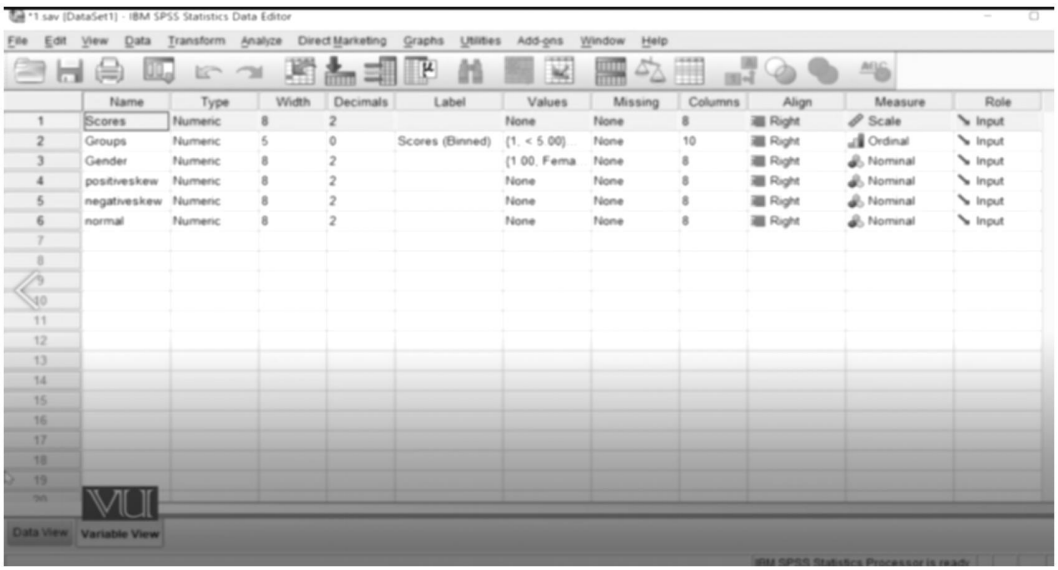
Figure 10.2 Measures of central tendency for skewed distribution

Exercise: Measure Of Central Tendency in SPSS

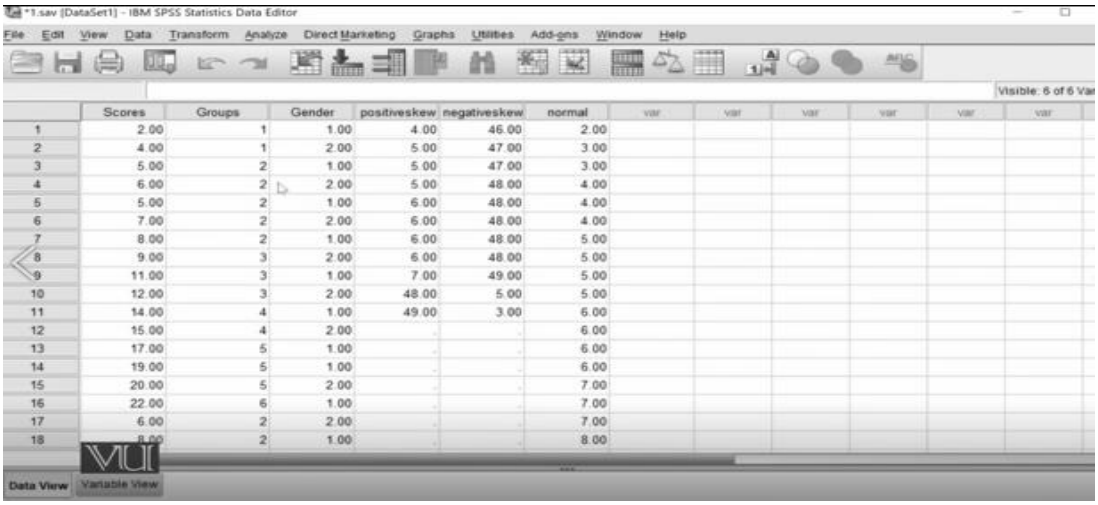
Analyze → Descriptive statistics→ Descriptive→ calculate mean, Standard Deviation, Minimum, Maximum

To calculate all together, go to Analyze → Frequencies→ Drag score in variables(s) → click statistics tab→ tick mean, median & mode→continue

Then select chart→ histogram→ tick show normal curve in histogram→continue→ok



Lets go to Graphs→legacy dialogue→line chart→simple→Define→send variable to category axis→ok



Lesson 11

MEASURE OF VARIABILITY-I

The term variability has much the same meaning in statistics as it has in everyday language; to say that things are variable means that they are not all the same. In statistics, our goal is to measure the amount of variability for a particular set of scores, a distribution. In simple terms, if the scores in a distribution are all the same, then there is no variability. If there are small differences between scores, then the variability is small, and if there are large differences between scores, then the variability is large.

Measure of Dispersion/Variability

Variability provides a quantitative measure of the differences between scores in a distribution and describes the degree to which the scores are spread out or clustered together.

1. Variability describes the distribution. Specifically, it tells whether the scores are clustered close together or are spread out over a large distance. Usually, variability is defined in terms of *distance.* It tells how much distance to expect between one score and another, or how much distance to expect between an individual score and the mean. For example, we know that the heights for most adult males are clustered close together, within 5 or 6 inches of the average. Although more extreme heights exist, they are relatively rare.

2. Variability measures how well an individual score (or group of scores) represents the entire distribution. This aspect of variability is very important for inferential statistics, in which relatively small samples are used to answer questions about populations. For example, suppose that you selected a sample of one person to represent the entire population. Because most adult males have heights that are within a few inches of the population average (the distances are small), there is a very good chance that you would select someone whose height is within 6 inches of the population mean. On the other hand, the scores are much more spread out (greater distances) in the distribution of weights. In this case, you probably would *not* obtain someone whose weight was within 6 pounds of the population mean. Thus, variability provides information about how much error to expect if you are using a sample to represent a population.

Types of Measures of Dispersion

- Range
- Inter Quartile Range

- Variance
- Standard Deviation

The Range

The *range* is the distance covered by the scores in a distribution, from the smallest score to the largest score. When the scores are measurements of a continuous variable, the range can be defined as the difference between the upper real limit (URL) for the largest score (*X*max) and the lower real limit (LRL) for the smallest score (*X*min).

Range = URL for *X*max – LRL for *X*min

If the scores have values from 1 to 5, for example, the range is 5.5 – 0.5 = 5 points. When the scores are whole numbers, the range is also a measure of the number of measurement categories. If every individual is classified as either 1, 2, 3, 4, or 5, then there are five measurement categories and the range is 5 points. Defining the range as the number of measurement categories also works for discrete variables that are measured with numerical scores. For example, if you are measuring the number of children in a family and the data produce values from 0 to 4, then there are five measurement categories (0, 1, 2, 3, and 4) and the range is 5 points.

A commonly used alternative definition of the range simply measures the difference between the largest score (*X*max) and the smallest score (*X*min), without any reference to real limits.

range = *X*max – *X*max

Advantages and Disadvantages of Range

Advantages: Quick and give rough estimate of dispersion in the data

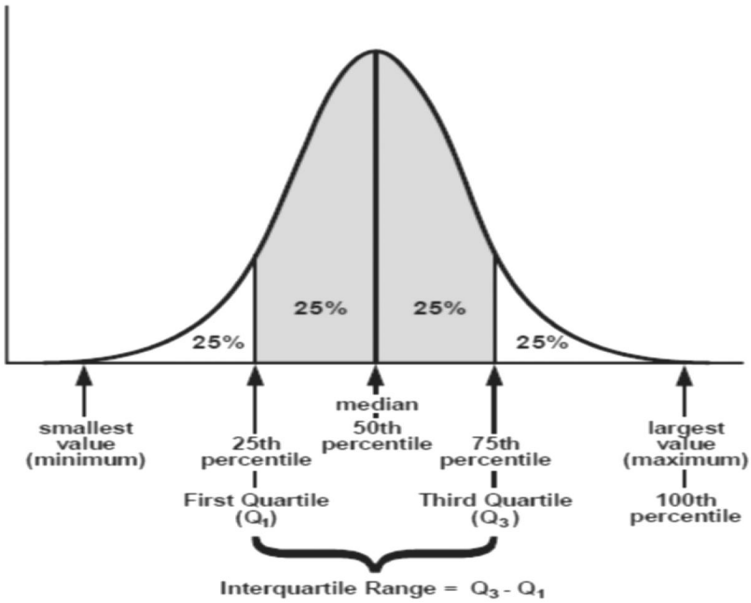
Disadvantages: Considers only extreme values and ignore all the middle values, so it's rough and imprecise and unreliable measure

Inter Quartile Range

The interquartile range (IQR) is a measure of variability, based on dividing a data set into quartiles. Quartiles divide a rank-ordered data set into four equal parts. The values that divide each part are called the first, second, and third quartile; and they are denoted by Q1, Q2, and Q3, respectively.

- Q1 is the "middle" value in the first half of the rank-ordered data set.
- Q2 is the median value in the set.
- Q3 is the "middle" value in the second half of the rank-ordered data set.

IQR = Q3-Q1



Pros of Inter Quartile Range

- Easy to calculate
- Eliminates influence of extreme scores

Cons of Inter Quartile Range

- Discard much of the data

Standard Deviation and Variance For a Population

The standard deviation is the most commonly used and the most important measure of variability. Standard deviation uses the mean of the distribution as a reference point and measures variability by considering the distance between each score and the mean. In simple terms, the standard deviation provides a measure of the standard, or average, distance from the mean, and describes whether the scores are clustered closely around the mean or are widely scattered. The fundamental definition of the standard deviation is the same for both samples and populations, but the calculations differ slightly. Although the concept of standard deviation is straightforward, the actual equations appear complex. Therefore, we begin by looking at the logic that leads to these equations. If you remember that our goal is to measure the standard, or typical, distance from the mean, then this logic and the equations that follow should be easier to remember.

The first step in finding the standard distance from the mean is to determine the *deviation,* or distance from the mean, for each individual score. By definition, the deviation for each score is the difference between the score and the mean.

Deviation is distance from the mean:

deviation score = $X - \mu$

For a distribution of scores with $\mu = 50$, if your score is $X = 53$, then your *deviation score* is

$X - \mu = 53 - 50 = 3$

If your score is $X = 45$, then your deviation score is

$X - \mu = 45 - 50 = -5$

Notice that there are two parts to a deviation score: the sign (_ or –) and the number. The sign tells the direction from the mean—that is, whether the score is located above (+) or below (–) the mean. The number gives the actual distance from the mean. For example, a deviation score of –6 corresponds to a score that is below the mean by a distance of 6 points.

Because our goal is to compute a measure of the standard distance from the mean, the obvious next step is to calculate the mean of the deviation scores. To compute this mean, you first add up the deviation scores and then divide by N. This process is demonstrated in the following example.

We start with the following set of $N = 4$ scores. These scores add up to $\Sigma X = 12$, so the mean is $\mu = \frac{12}{4} = 3$. For each score, we have computed the deviation.

X	$X - \mu$
8	+5
1	–2
3	0
0	–3
	$0 = \Sigma(X - \mu)$

Note that the deviation scores add up to zero. This should not be surprising if you remember that the mean serves as a balance point for the distribution. The total of the distances above the mean is exactly equal to the total of the distances below the mean. Thus, the total for the positive deviations is exactly equal to the total for the negative deviations, and the complete set of deviations always adds up to zero. Because the sum of the deviations is always zero, the mean of the deviations is also zero and is of no value as a measure of variability. The mean of the deviations is zero if the scores are closely clustered and it is zero if the scores are widely scattered. (You should note, however, that the constant value of zero can be useful in other ways. Whenever you are working with deviation scores, you can check your calculations by making sure that the deviation scores add up to zero.)

The average of the deviation scores does not work as a measure of variability because it is always zero. Clearly, this problem results from the positive and negative values canceling each other out. The solution is to get rid of the signs (+ and –). The standard procedure for accomplishing this is to square each deviation score. Using the squared values, you then compute the *mean squared deviation,* which is called *variance.*

Population variance equals the mean squared deviation. Variance is the average squared distance from the mean.

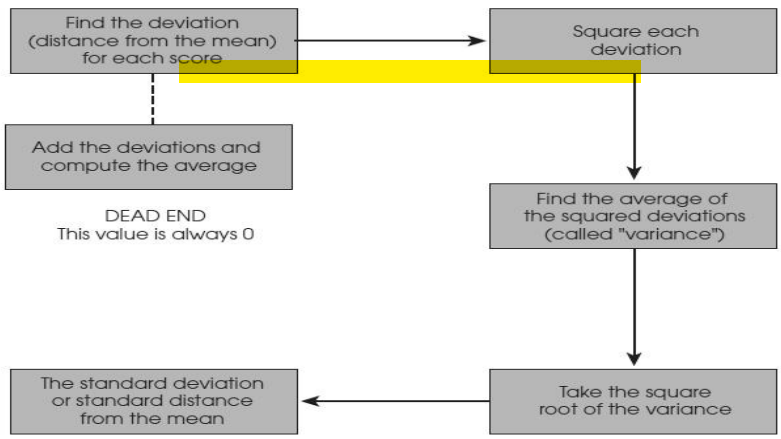
Note that the process of squaring deviation scores does more than simply get rid of plus and minus signs. It results in a measure of variability based on *squared* distances. Although variance is valuable for some of the *inferential* statistical methods covered later, the concept of squared distance is not an intuitive or easy to understand *descriptive* measure. For example, it is not particularly useful to know that the squared distance from New York City to Boston is 26,244 miles squared. The squared value becomes meaningful, however, if you take the square root. Therefore, we continue the process with one more step.

Remember that our goal is to compute a measure of the standard distance from the mean. Variance, which measures the average squared distance from the mean, is not exactly what we want. The final step simply takes the square root of the variance to obtain the *standard deviation,* which measures the standard distance from the mean.

Standard deviation is the square root of the variance and provides a measure of the standard, or average, distance from the mean.

$$\textbf{Standard deviation} = \sqrt{\boldsymbol{Variance}}$$

Figure given below shows the overall process of computing variance and standard deviation.



Formulas For Population Variance And Standard Deviation

The concepts of standard deviation and variance are the same for both samples and populations. However, the details of the calculations differ slightly, depending on whether you have data from a sample or from a complete population.

The sum of squared deviations (*SS*) Recall that variance is defined as the mean of the squared deviations. This mean is computed in exactly the same way you compute any mean: First find the sum, and then divide by the number of scores.

variance = mean squared deviation = sum of squared deviations\number of scores

The value in the numerator of this equation, the sum of the squared deviations, is a basic component of variability, and we focus on it. To simplify things, it is identified by the notation *SS* (for sum of squared deviations), and it generally is referred to as the *sum of squares.*

SS, or sum of squares, is the sum of the squared deviation scores.

You need to know two formulas to compute *SS*. These formulas are algebraically equivalent (they always produce the same answer), but they look different and are used in different situations.

The first of these formulas is called the definitional formula because the symbols in the formula literally define the process of adding up the squared deviations:

Definitional formula: $SS = \Sigma(X - \mu)^2$

Although the definitional formula is the most direct method for computing *SS*, it can be awkward to use. In particular, when the mean is not a whole number, the deviations all contain decimals or fractions, and the calculations become difficult. In addition, calculations with decimal values introduce the opportunity for rounding error, which can make the result less accurate. For these reasons, an alternative formula has been developed for computing *SS*. The alternative, known as the computational formula, performs calculations with the scores (not the deviations) and therefore minimizes the complications of decimals and fractions.

$$\text{computational formula: } SS = \Sigma X^2 - \frac{(\Sigma X)^2}{N}$$

With the definition and calculation of *SS* behind you, the equations for variance and standard deviation become relatively simple. Remember that variance is defined as the mean squared deviation. The mean is the sum of the squared deviations divided by *N*, so the equation for the *population variance* is

variance = SS\N

Standard deviation is the square root of variance, so the equation for the *population standard deviation* is

standard deviation =√SS \ N

To emphasize the relationship between standard deviation and variance, we use σ^2 as the symbol for population variance (standard deviation is the square root of the variance). Thus,

$$\text{population standard deviation} = \sigma = \sqrt{\sigma^2} = \sqrt{\frac{SS}{N}}$$

$$\text{population variance} = \sigma^2 = \frac{SS}{N}$$

Lesson 12

MEASURE OF VARIABILITY-II

Measure of Variability and Inferential Statistics

The goal of inferential statistics is to use the limited information from samples to draw general conclusions about populations. The basic assumption of this process is that samples should be representative of the populations from which they come. This assumption poses a special problem for variability because samples consistently tend to be less variable than their populations. Notice that a few extreme scores in the population tend to make the population variability relatively large. However, these extreme values are unlikely to be obtained when you are selecting a sample, which means that the sample variability is relatively small. The fact that a sample tends to be less variable than its population means that sample variability gives a *biased* estimate of population variability. This bias is in the direction of underestimating the population value rather than being right on the mark. Fortunately, the bias in sample variability is consistent and predictable, which means it can be corrected.

The calculations of variance and standard deviation for a sample follow the same steps that were used to find population variance and standard deviation. Except for minor changes in notation, the first three steps in this process are exactly the same for a sample as they were for a population. That is, calculating the sum of the squared deviations, *SS*, is the same for a sample as it is for a population. The changes in notation involve using *M* for the sample mean instead of μ, and using *n* (instead of *N*) for the number of scores. Thus, to find the *SS* for a sample:

1. Find the deviation from the mean for each score: deviation= $X-M$
2. Square each deviation: squared deviation= $(X-M)^2$
3. Add the squared deviations: $SS= \Sigma (X-M)^2$

These three steps can be summarized in a defintional formula for *SS:*

Definitional formula: $SS= \Sigma (X-M)^2$

The value of *SS* also can be obtained using a computational formula. Except for one minor difference in notation (using *n* in place of *N*), the computational formula for *SS* is the same for a sample as it was for a population (see Equation 4.2). Using sample notation, this formula is:

$$\text{Computational formula: } SS = \Sigma X^2 - \frac{(\Sigma X)^2}{n}$$

Again, calculating *SS* for a sample is exactly the same as for a population, except for minor changes in notation. After you compute *SS*, however, it becomes critical to differentiate between samples and populations. To correct for the bias in sample variability, it is necessary to make an adjustment in the formulas for sample variance and standard deviation. With this in mind, *sample variance* (identified by the symbol *s*2) is defined as

$$\text{sample variance} = s^2 = \frac{SS}{n-1}$$

Sample standard deviation (identified by the symbol *s*) is simply the square root of the variance.

$$\text{sample standard deviation} = s = \sqrt{s^2} = \sqrt{\frac{SS}{n-1}}$$

Notice that the sample formulas divide by $n - 1$, unlike the population formulas, which divide by *N*. This is the adjustment that is necessary to correct for the bias in sample variability. The effect of the adjustment is to increase the value that you obtain. Dividing by a smaller number ($n - 1$ instead of n) produces a larger result and makes sample variance an accurate and unbiased estimator of population variance.

For a sample of n scores, the degrees of freedom, or *df,* for the sample variance are defined as $df = n - 1$. The degrees of freedom determine the number of scores in the sample that are independent and free to vary.

The $n - 1$ degrees of freedom for a sample is the same $n - 1$ that is used in the formulas for sample variance and standard deviation.

Earlier we noted that sample variability tends to underestimate the variability in the corresponding population. To correct for this problem, we adjusted the formula for sample variance by dividing by $n - 1$ instead of dividing by n. The result of the adjustment is that sample variance provides a much more accurate representation of the population variance. Specifically, dividing by $n - 1$ produces a sample variance that provides an *unbiased* estimate of the corresponding population variance. This does not mean that each individual sample variance is exactly equal to its population variance. In fact, some sample variances overestimate the population value and some underestimate it. However, the average of all the sample variances produces an accurate estimate of the population variance. This is the idea behind the concept of an unbiased statistic.

A sample statistic is unbiased if the average value of the statistic is equal to the population parameter. (The average value of the statistic is obtained from all the possible samples for a specific sample size, *n*.)

A sample statistic is biased if the average value of the statistic either underestimates or overestimates the corresponding population parameter.

Standard Deviation and Descriptive Statistics

Standard deviation is primarily a descriptive measure; it describes how variable, or how spread out, the scores are in a distribution. Standard deviation describes variability by measuring distance from the mean. In any distribution, some individuals are close to the mean, and others are relatively far from the mean. Standard deviation provides a measure of the typical, or standard, distance from the mean.

Transformations of Scale

Occasionally a set of scores is transformed by adding a constant to each score or by multiplying each score by a constant value. This happens, for example, when exposure to a treatment adds a fixed amount to each participant's score or when you want to change the unit of measurement (to convert from minutes to seconds, multiply each score by 60).

What happens to the standard deviation when the scores are transformed in this manner?

The easiest way to determine the effect of a transformation is to remember that the standard deviation is a measure of distance. If you select any two scores and see what happens to the distance between them, you also find out what happens to the standard deviation.

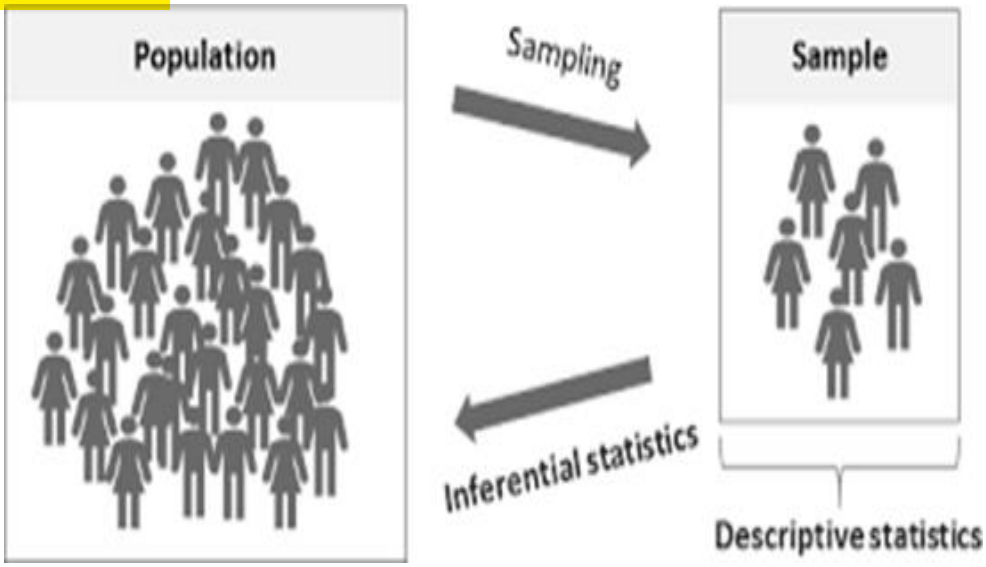
1. Adding a constant to each score does not change the standard deviation.
2. Multiplying each score by a constant causes the standard deviation to be multiplied by the same constant

Reporting the Standard Deviation

In reporting the results of a study, the researcher often provides descriptive information for both central tendency and variability. The dependent variables in psychology research are often numerical values obtained from measurements on interval or ratio scales. With numerical scores, the most common descriptive statistics are the mean (central tendency) and the standard deviation (variability), which are usually reported together. In many journals, especially those following APA style, the symbol SD is used for the sample standard deviation.

Inferential Statistics

Inferential statistics consist of techniques that allow us to study samples and then make generalizations about the populations from which they were selected.



There are *two main areas* of inferential statistics:

1. **Estimating Parameters**. This means taking a statistic from your sample data (for example the sample mean) and using it to say something about a population parameter (i.e. the population mean).
2. **Hypothesis Testing**. This is where you can use sample data to answer research questions. For example, you might be interested in knowing if a new cancer drug is effective. Or if breakfast helps children perform better in schools.

Most research uses statistical models called the *Generalized Linear model* and include Student's t-tests, ANOVA (Analysis of Variance), regression analysis. One problem with using samples, however, is that a sample provides only limited information about the population. Although samples are generally representative of their populations, a sample is not expected to give a perfectly accurate picture of the whole population. There usually is some discrepancy between a sample statistic and the corresponding population parameter. This discrepancy is called **sampling error**, and it creates the fundamental problem that inferential statistics must always address.

Lesson 13

Z-scores

Introduction to Z score

A score *by itself* does not necessarily provide much information about its position within a distribution. These original, unchanged scores that are the direct result of measurement are called *raw scores*. To make raw scores more meaningful, they are often transformed into new values that contain more information. This transformation is one purpose for *z*-scores. In particular, we transform *X* values into *z*-scores so that the resulting *z*-scores tell exactly where the original scores are located.

A second purpose for *z*-scores is to *standardize* an entire distribution. A common example of a standardized distribution is the distribution of IQ scores. Although there are several different tests for measuring IQ, the tests usually are standardized so that they have a mean of 100 and a standard deviation of 15. Because all the different tests are standardized, it is possible to understand and compare IQ scores even though they come from different tests. For example, we all understand that an IQ score of 95 is a little below average, *no matter which IQ test was used*. Similarly, an IQ of 145 is extremely high, *no matter which IQ test was used*. In general terms, the process of standardizing takes different distributions and makes them equivalent. The advantage of this process is that it is possible to compare distributions even though they may have been quite different before standardization. In summary, the process of transforming *X* values into *z*-scores serves two useful purposes:

1. Each *z*-score tells the exact location of the original *X* value within the distribution.

2. The *z*-scores form a standardized distribution that can be directly compared to other distributions that also have been transformed into *z*-scores.

One of the *primary purposes* of a *z*-score is to describe the exact location of a score within a distribution. The *z*-score accomplishes this goal by transforming each *X* value into a signed number (+ or –) so that

1. The *sign* tells whether the score is located above (+) or below (–) the mean, and

2. The *number* tells the distance between the score and the mean in terms of the number of standard deviations.

Thus, in a distribution of IQ scores with $\mu = 100$ and $\sigma = 15$, a score of $X = 130$ would be transformed into $z = +2.00$. The z value indicates that the score is located above the mean (+) by a distance of 2 standard deviations (30 points).

A z-score specifies the precise location of each X value within a distribution. The sign of the z-score (+ or –) signifies whether the score is above the mean (positive) or below the mean (negative). The numerical value of the z-score specifies the distance from the mean by counting the number of standard deviations between X and μ.

z-Scores And Location In A Distribution

The z-score definition is adequate for transforming back and forth from X values to z-scores as long as the arithmetic is easy to do in your head. For more complicated values, it is best to have an equation to help structure the calculations. Fortunately, the relationship between X values and z-scores is easily expressed in a formula. The formula for transforming scores into z-scores is

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

The numerator of the equation, $X - \mu$, is a *deviation score*; it measures the distance in points between X and μ and indicates whether X is located above or below the mean. The deviation score is then divided by _ because we want the z-score to measure distance in terms of standard deviation units. The formula performs exactly the same arithmetic that is used with the z-score definition, and it provides a structured equation to organize the calculations when the numbers are more difficult.

Computing and Interpreting Z Scores

The following examples demonstrate the use of the z-score formula.

A distribution of scores has a mean of $\mu = 100$ and a standard deviation of $\sigma = 10$. What z-score corresponds to a score of $X = 130$ in this distribution?

According to the definition, the z-score has a value of +3 because the score is located above the mean by exactly 3 standard deviations. Using the z-score formula, we obtain

$$z = \frac{X - \mu}{\sigma} = \frac{130 - 100}{10} = \frac{30}{10} = 3.00$$

The formula produces exactly the same result that is obtained using the z-score definition.

Using Z-Scores To Standardize A Distribution

It is possible to transform every X value in a distribution into a corresponding z-score. The result of this process is that the entire distribution of X values is transformed into a distribution of z-

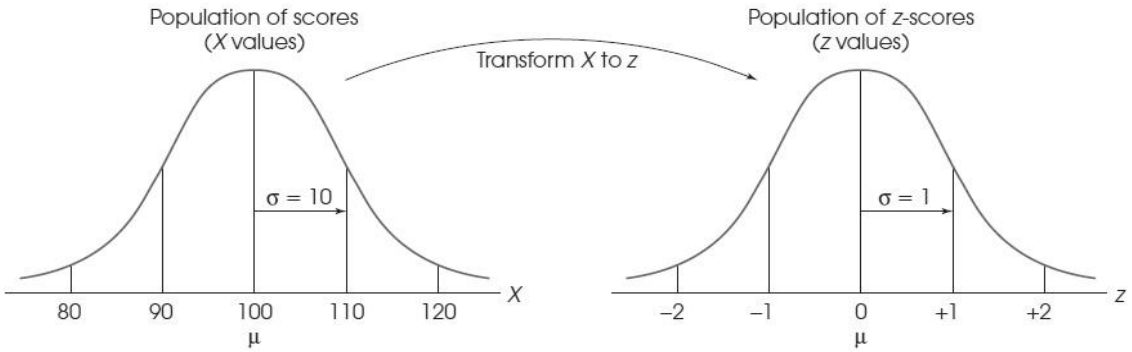
scores. The new distribution of *z*-scores has characteristics that make the *z-score transformation* a very useful tool. Specifically, if every *X* value is transformed into a *z*-score, then the distribution of *z*-scores will have the following properties:

1. Shape. The distribution of *z*-scores will have exactly the same shape as the original distribution of scores. If the original distribution is negatively skewed, for example, then the *z*-score distribution will also be negatively skewed. If the original distribution is normal, the distribution of *z*-scores will also be normal. Transforming raw scores into *z*-scores does not change anyone's position in the distribution. For example, any raw score that is above the mean by 1 standard deviation will be transformed to a *z*-score of +1.00, which is still above the mean by 1 standard deviation. Transforming a distribution from *X* values to *z* values does not move scores from one position to another; the procedure simply relabels each score. Because each individual score stays in its same position within the distribution, the overall shape of the distribution does not change.

2. The mean. The *z*-score distribution will *always* have a mean of zero. In Figure given below, the original distribution of *X* values has a mean of 100. When this value, $X = 100$, is transformed into a *z*-score, the result is

$$z = \frac{X - \mu}{\sigma} = \frac{100 - 100}{10} = 0$$

Thus, the original population mean is transformed into a value of zero in the *z*-score distribution. The fact that the *z*-score distribution has a mean of zero makes the mean a convenient reference point.



An entire population of scores is transformed into *z*-scores. The transformation does not change the shape of the population, but the mean is transformed into a value of 0 and the standard deviation is transformed to a value of 1.

3. The Standard Deviation. The distribution of *z*-scores will *always* have a standard deviation of 1. In Figure 5.5, the original distribution of *X* values has mean 100 and standard deviation 10. In this distribution, a value of $X = 110$ is above the mean by exactly 10 points or 1 standard deviation. When $X = 110$ is transformed, it becomes $z = +1.00$, which is above the mean by exactly 1 point in the *z*-score distribution. Thus, the standard deviation corresponds to a 10-point distance in the *X* distribution and is transformed into a 1-point distance in the *z*-score distribution. The advantage of having a standard deviation of 1 is that the numerical value of a *z*-score is exactly the same as the number of standard deviations from the mean. For example, a *z*-score of $z + 1.50$ is exactly 1.50 standard deviations from the mean.

A standardized distribution is composed of scores that have been transformed to create predetermined values for μ and σ. Standardized distributions are used to make dissimilar distributions comparable.

Computing Z-Scores For a Sample

Although *z*-scores are most commonly used in the context of a population, the same principles can be used to identify individual locations within a sample. The definition of a *z*-score is the same for a sample as for a population, provided that you use the sample mean and the sample standard deviation to specify each *z*-score location. Thus, for a sample, each *X* value is transformed into a *z*-score so that

1. The sign of the *z*-score indicates whether the *X* value is above (+) or below (–) the sample mean, and

2. The numerical value of the *z*-score identifies the distance from the sample mean by measuring the number of sample standard deviations between the score (*X*) and the sample mean (*M*).

Expressed as a formula, each *X* value in a sample can be transformed into a *z*-score as follows:

$$z = \frac{X - M}{s}$$

Similarly, each Z-score can be transformed back into an X value, as follows:

$X = M + zs$

If all the scores in a sample are transformed into *z*-scores, the result is a sample of *z*-scores. The transformed distribution of *z*-scores will have the same properties that exist when a population of *X* values is transformed into *z*-scores. Specifically,

- The sample of z-scores will have the same shape as the original sample of scores.
- The sample of z-scores will have a mean of Mz = 0.
- The sample of z-scores will have a standard deviation of sz = 1.

Lesson 14

INTRODUCTION TO PROBABILITY-I

Introduction to Probability

Suppose, for example, that you are selecting a single marble from a jar that contains 50 black and 50 white marbles. (In this example, the jar of marbles is the *population* and the single marble to be selected is the *sample*.) Although you cannot guarantee the exact outcome of your sample, it is possible to talk about the potential outcomes in terms of probabilities. In this case, you have a 50-50 chance of getting either color.

Now consider another jar (population) that has 90 black and only 10 white marbles.

Again, you cannot predict the exact outcome of a sample, but now you know that the sample probably will be a black marble. By knowing the makeup of a population, we can determine the probability of obtaining specific samples. In this way, probability gives us a connection between populations and samples, and this connection is the foundation for the inferential statistics.

Probability is a huge topic that extends far beyond the limits of introductory statistics, and we do not attempt to examine it all here. Instead, we concentrate on the few concepts and definitions that are needed for an introduction to inferential statistics. We begin with a relatively simple definition of probability.

For a situation in which several different outcomes are possible, the probability for any specific outcome is defined as a fraction or a proportion of all the possible outcomes. If the possible outcomes are identified as A, B, C, D, and so on, then:

$$\text{probability of } A = \frac{\text{number of outcomes classified as } A}{\text{total number of possible outcomes}}$$

Probability values The probability of a specific outcome is expressed with a *p* (for probability) followed by the specific outcome in parentheses. Probability fractions can be expressed as either decimals or percentages

E.g., the probability of obtaining heads for a coin toss is written as:

$$p(\text{heads})=1/2=0.50=50\%$$

All probability values are contained in a range from 0 to 1

Probability in Inferential Statistics

- Relationship between sample and population is defined in terms of probability
- When we draw a sample, what is the probability that sample taken is right? or probability of making a wrong decision?
- Probability is used to predict what kind of samples are likely to be obtained from a population. (e.g.2 jars of marbles)

Random Sampling

For the preceding definition of probability to be accurate, it is necessary that the outcomes be obtained by a process called *random sampling.*

A **random sample** requires that each individual in the population has an *equal chance* of being selected.

A second requirement, necessary for many statistical formulas, states that if more than one individual is being selected, the probabilities must *stay constant* from one selection to the next. Adding this second requirement produces what is called *independent random sampling*. The term *independent* refers to the fact that the probability of selecting any particular individual is independent of those individuals who have already been selected for the sample. For example, the probability that you will be selected is constant and does not change even when other individuals are selected before you are.

An **independent random sample** requires that each individual has an equal chance of being selected and that the probability of being selected stays constant from one selection to the next if more than one individual is selected.

Probability and Frequency Distributions

The situations in which we are concerned with probability usually involve a population of scores that can be displayed in a frequency distribution graph. If you think of the graph as representing the entire population, then different proportions of the graph represent different proportions of the population. Because probabilities and proportions are equivalent, a particular proportion of the graph corresponds to a particular probability in the population. Thus, whenever a population is presented in a frequency distribution graph, it is possible to represent probabilities as proportions of the graph.

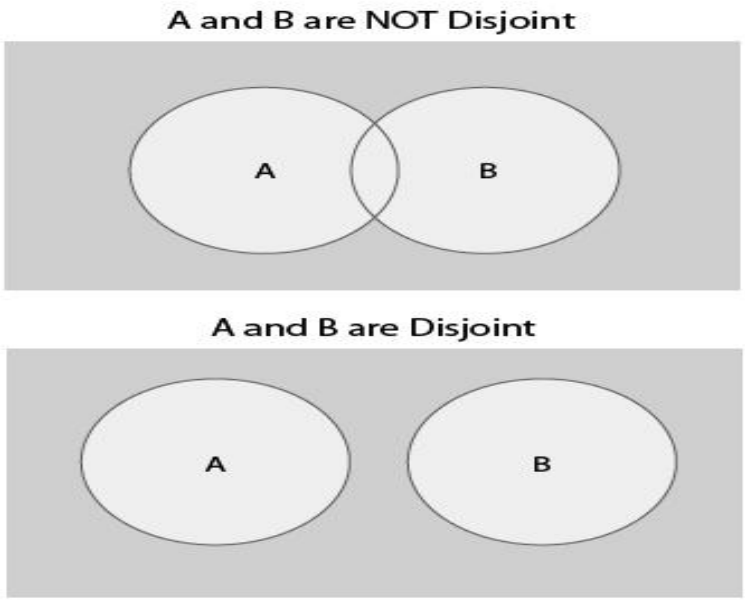
Rules and Basics Laws of Probability

Basic Properties of Probability

- Every probability is between zero and one. If A is an event, then **0≤P(A)≥ 1.**
- The sum of the probabilities of all possible outcomes is 1.
- Impossible events have probability zero. That is, if event A is impossible, then **P(A)=0.**

An example of such an event is rolling a 7 on a standard six-sided die.

- Two events that cannot occur at the same time are called *disjoint or mutually exclusive.*



Consider the following two events:

1. A-a randomly chosen person has blood type A, and
2. B-a randomly chosen person has blood type B.

We are going to assume that each person can have only one blood type. Therefore, it is impossible for the events A and B to occur together. **Events A and B are DISJOINT**

Consider the following two events:

1. A-a randomly chosen person has blood type A
2. B-a randomly chosen person is a woman.

In this case, it is possible for events A and B to occur together. **Events A and B are NOT DISJOINT.**

Basics Laws Of Probability

Additive Law of Probability-Given a set of mutually exclusive events, the probability of the occurrence of one event or another is equal to the sum of their separate probabilities.

Multiplicative Law of Probability-The probability of the joint occurrence of two or more independent events is the product of their individual probabilities.

Joint and Conditional Probabilities

Two types of probabilities play an important role in discussions of probability: joint probabilities and conditional probabilities.

A ***joint probability*** is defined simply as the probability of the co-occurrence of two or more events. If those two events are independent, then the probability of their joint occurrence can be found by using the multiplicative law. If they are not independent, the probability of their joint occurrence is more complicated to compute.

A ***conditional probability*** is the probability that one event will occur, given that some other event has occurred. The probability that a person will contract AIDS, given that he or she is an intravenous drug user, is a conditional probability.

Lesson 15

INTRODUCTION TO PROBABILITY-II

Understanding Probability

According to Stanford University's Blood Center these are the probabilities of human blood types in the United State

Blood Type	O	A	B	AB
Probability	0.44	0.42	0.10	0.04

What is the probability that a randomly chosen person is a potential donor for a person with blood type A?

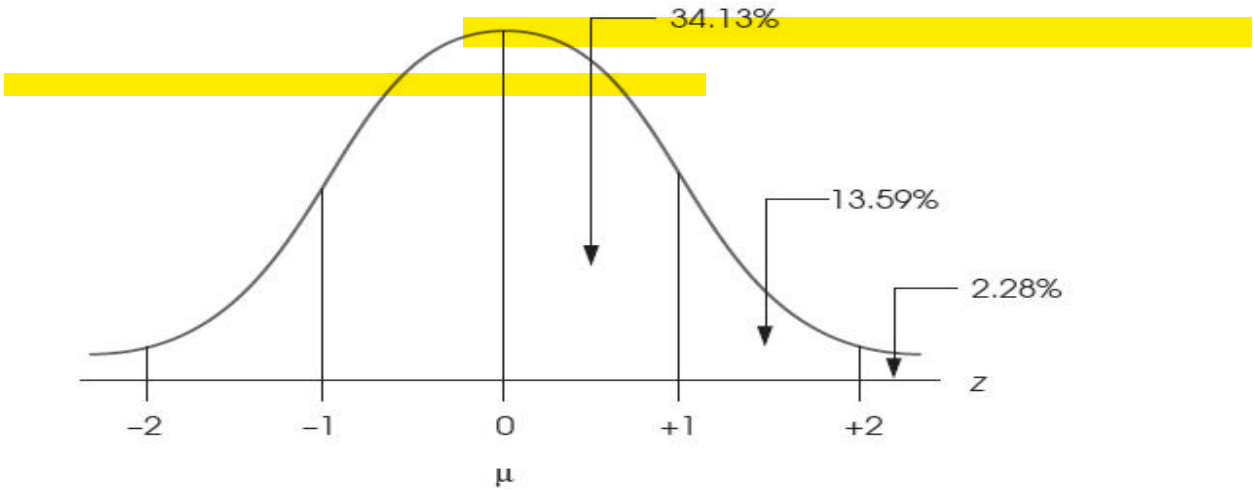
- Suppose we roll two dice. What is the probability that both dice show a 5?
- Sample Space
- You toss a coin 3 times, what is the probability of at least two heads

Probability and Normal Distribution

Note that the normal distribution is symmetrical, with the highest frequency in the middle and frequencies tapering off as you move toward either extreme. Although the exact shape for the normal distribution is defined by an equation (see Figure 6.3), the normal shape can also be described by the proportions of area contained in each section of the distribution. Statisticians often identify sections of a normal distribution by using *z*-scores. Figure 6.4 shows a normal distribution with several sections marked in *z*-score units. You should recall that *z*-scores measure positions in a distribution in terms of standard deviations from the mean. (Thus, $z = +1$ is 1 standard deviation above the mean, $z = +2$ is 2 standard deviations above the mean, and so on.) The graph shows the percentage of scores that fall in each of these sections. For example, the section between the mean ($z = 0$) and the point that is 1 standard deviation above the mean ($z = 1$) contains 34.13% of the scores. Similarly, 13.59% of the scores are located in the section between 1 and 2 standard deviations above the mean. In this way it is possible to define a normal distribution in terms of its proportions; that is, a distribution is normal if and only if it has all the right proportions.

There are two additional points to be made about the distribution shown in Figure 6.4. First, you should realize that the sections on the left side of the distribution have exactly the same areas as the corresponding sections on the right side because the normal distribution is symmetrical.

Second, because the locations in the distribution are identified by *z*-scores, the percentages shown in the figure apply to *any normal distribution* regardless of the values for the mean and the standard deviation. Remember: When any distribution is transformed into *z*-scores, the mean becomes zero and the standard deviation becomes one.

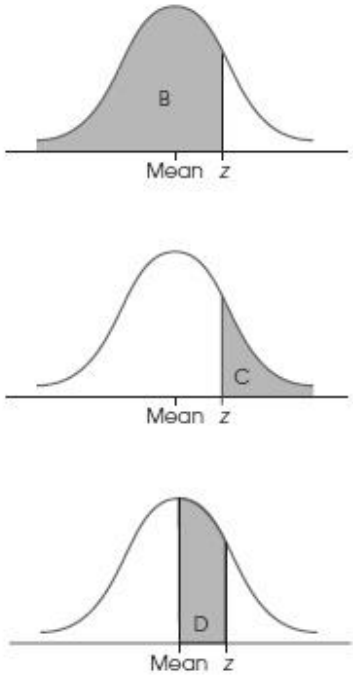


The Unit Normal Table

A more complete listing of *z*-scores and proportions is provided in the *unit normal table*. This table lists proportions of the normal distribution for a full range of possible *z*-score values. The complete unit normal table is provided in Appendix B Table B.1, and part of the table is reproduced in Figure given below. Notice that the table is structured in a four-column format. The first column (A) lists *z*-score values corresponding to different positions in a normal distribution. If you imagine a vertical line drawn through a normal distribution, then the exact location of the line can be described by one of the *z*-score values listed in column A. You should also realize that a vertical line separates the distribution into two sections: a larger section called the *body* and a smaller section called the *tail*.

Columns B and C in the table identify the proportion of the distribution in each of the two sections. Column B presents the proportion in the body (the larger portion), and column C presents the proportion in the tail. Finally, we have added a fourth column, column D, that identifies the proportion of the distribution that is located *between* the mean and the *z*-score.

(A) z	(B) Proportion in body	(C) Proportion in tail	(D) Proportion between mean and z
0.00	.5000	.5000	.0000
0.01	.5040	.4960	.0040
0.02	.5080	.4920	.0080
0.03	.5120	.4880	.0120
0.21	.5832	.4168	.0832
0.22	.5871	.4129	.0871
0.23	.5910	.4090	.0910
0.24	.5948	.4052	.0948
0.25	.5987	.4013	.0987
0.26	.6026	.3974	.1026
0.27	.6064	.3936	.1064
0.28	.6103	.3897	.1103
0.29	.6141	.3859	.1141
0.30	.6179	.3821	.1179
0.31	.6217	.3783	.1217
0.32	.6255	.3745	.1255
0.33	.6293	.3707	.1293
0.34	.6331	.3669	.1331



A portion of the unit normal table. This table lists proportions of the normal distribution corresponding to each *z*-score value. Column A of the table lists *z*-scores. Column B lists the proportion in the body of the normal distribution up to the *z*-score value. Column C lists the proportion of the normal distribution that is located in the tail of the distribution beyond the *z*-score value. Column D lists the proportion between the mean and the *z*-score value.

Probabilities and Proportions for Scores from a Normal Distribution

In most situations, however, it is necessary to find probabilities for specific *X* values. Consider the following example:

It is known that IQ scores form a normal distribution with μ=100 and σ=15. Given this information, what is the probability of randomly selecting an individual with an IQ score less than 120?

This problem is asking for a specific probability or proportion of a normal distribution.

However, before we can look up the answer in the unit normal table, we must first transform the IQ scores (*X* values) into *z*-scores. Thus, to solve this new kind of probability problem, we must add one new step to the process. Specifically, to answer probability questions about scores (*X* values) from a normal distribution, you must use the following two-step procedure:

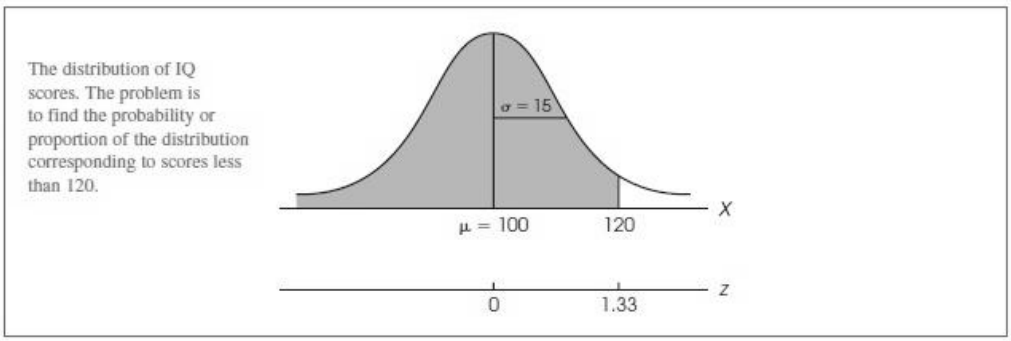
1. Transform the *X* values into *z*-scores.

2. Use the unit normal table to look up the proportions corresponding to the *z*-score values.

This process is demonstrated in the following examples. Once again, we suggest that you sketch the distribution and shade the portion you are trying to find to avoid careless mistakes.

We now answer the probability question about IQ scores that we presented earlier. Specifically, what is the probability of randomly selecting and individual with an IQ score less than 120? Restated in terms of proportions, we want to find the proportion of IQ distribution that corresponds to scores less than 120. The distribution is drawn in given Figure, and the portion we want has been shaded. The first step is to change the X values into z-scores. In particular, the score of X=120 is changed to

$$z=\frac{X-\mu}{\sigma}=\frac{120-100}{15}=\frac{20}{15}=1.33$$



The distribution of IQ scores. The problem is to find the probability or proportion of the distribution corresponding to scores less than 120.

Thus, an IQ score of $X = 120$ corresponds to a z-score of $z = 1.33$, and IQ scores less than 120 correspond to z-scores less than 1.33.

Next, look up the z-score value in the unit normal table. Because we want the proportion of the distribution in the body to the left of $X = 120$ (see Figure 6.10), the answer is in column B. Consulting the table, we see that a z-score of 1.33 corresponds to a proportion of 0.9082. The probability of randomly selecting an individual with an IQ less than 120 is $p = 0.9082$. In symbols,

$p(X <\_ 120) =\_ p(z \_< 1.33) =\_ 0.9082$ (or 90.82%)

Finding proportions/probabilities located between two scores The next example demonstrates the process of finding the probability of selecting a score that is located *between* two specific values. Although these problems can be solved using the proportions of columns B and C (body and tail), they are often easier to solve with the proportions listed in column D.

The highway department conducted a study measuring driving speeds on a local section of interstate highway. They found an average speed of μ = 58 miles per hour with a standard deviation of 10. The distribution was approximately normal.

Given this information, what proportion of the cars are traveling between 55 and 65 miles per hour? Using probability notation, we can express the problems as

p(55 < X < 65) = ?

The distribution of driving speeds is shown in figure below with the appropriate area shared. The first step is to determine the z-score corresponding to the *X* value at each end of the interval.

$$\text{For } X = 55: z = \frac{X - \mu}{\sigma} = \frac{55 - 58}{10} = \frac{-3}{10} = -0.30$$

$$\text{For } X = 65: z = \frac{X - \mu}{\sigma} = \frac{65 - 58}{10} = \frac{7}{10} = 0.70$$

Looking again at figure, we see the the proportion we are seeking can be divided into two sections: (1) the area left of the mean, and (2) the area right of the mean. The first area is the proportion between the mean and z= -0.30, and the second is the proportion between the mean and z= +0.70. using column D of the unit normal table, these two proportions are 0.1179 and 0.2580. The total proportion is obtained by adding these two sections:

p(55 < X < 65) = p(-0.30 < z < +0.70) = 0.1179 + 0.2580= 0.3759

Review of probability

- Introduction to Probability
- Rules and Basic laws of Probability
- Probability and Normal Distribution

Lesson 16

INTRODUCTION TO PROBABILITY-III

A probability distribution is a statistical function that describes all the possible values and likelihoods that a random variable can take within a given range. It is a statistical model that shows the possible outcomes of a particular event or course of action as well as the statistical likelihood of each event.

Probability and Binomial Distribution

When a variable is measured on a scale consisting of exactly two categories, the resulting data are called binomial. The term *binomial* can be loosely translated as "*two names*," referring to the two categories on the measurement scale.

Binomial data can occur when a variable naturally exists with only two categories. It is further explained by an example, people can be classified as male or female, and a coin toss results in either heads or tails. It also is common for a researcher to simplify data by collapsing the scores into two categories. For example, a psychologist may use personality scores to classify people as either high or low in aggression.

In binomial situations, the researcher often knows the probabilities associated with each of the two categories. With a balanced coin, for example, p(heads) =p(tails) =1/2. The question of interest is the number of times each category occurs in a series of trials or in a sample of individuals. For example:

- What is the probability of obtaining 15 heads in 20 tosses of a balanced coin?
- What is the probability of obtaining more than 40 introverts in a sampling of 50 college freshmen?

To answer probability questions about binomial data, we must examine the binomial distribution. To define and describe this distribution, we first introduce some notation.

1. The two categories are identified as *A* and *B*.

2. The probabilities (or proportions) associated with each category are identified as:

p* =*p*(*A*) = the probability of *A

q* =*p*(*B*) = the probability of *B

Notice that $p = q = 1.00$ because *A* and *B* are the only two possible outcomes.

3. The number of individuals or observations in the sample is identified by *n*.

4. The variable *X* refers to the number of times category *A* occurs in the sample.

Notice that *X* can have any value from 0 (none of the sample is in category *A*) to *n* (all of the sample is in category *A*).

Using the notation presented here, the binomial distribution shows the probability associated with each value of *X* from $X = 0$ to $X = n$.

The binomial distribution tends toward a normal shape, especially when the sample size (n) is relatively large. In binomial distribution, probability of success and failure in each and every trial is equal to one, p+q=1

The fact that the binomial distribution tends to be normal in shape means that we can compute probability values directly from *z*-scores and the unit normal table.

The Normal Approximation To The Binomial Distribution

Binomial distribution tends to approximate a normal distribution, particularly when n is large. To be more specific,The binomial distribution is nearly perfect normal distribution when pn and qn are both equal to or greater than 10. Under these circumstances, the binomial distribution approximates a normal distribution with the following parameters:

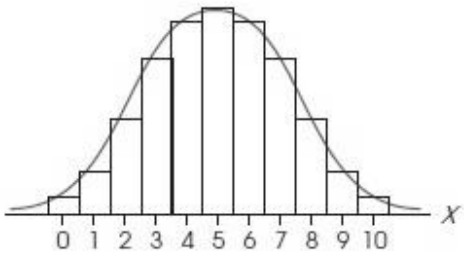
Mean: $\mu = pn$

standard deviation: $\sigma = \sqrt{npq}$

Within this normal distribution, each value of X has a corresponding z-score,

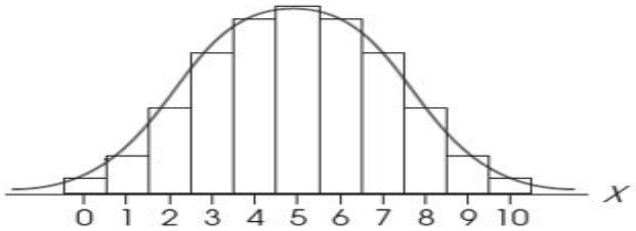
$$z=\frac{X-\mu}{\sigma}=\frac{X-pn}{\sqrt{npq}}$$

It is important to remember that the normal distribution is only an approximation of a true binomial distribution. Binomial values, such as the number of heads in a series of coin tosses, are *discrete*. The normal distribution is *continuous*. However, the *normal approximation* provides an extremely accurate model for computing binomial probabilities in many situations. Figure given below shows the difference between the true binomial distribution, the discrete histogram, and the normal curve that approximates the binomial distribution. Although the two distributions are slightly different, the area under the distributions is nearly equivalent. *Remember, it is the area under the distribution that is used to find probabilities.*



The relationship between the binomial distribution and the normal distribution. The binomial dis bution is always a discrete histogram, and the normal distribution is a continuous, smooth curve. Each *X* value is represented by a bar in the histogram or a section of the normal distribution.

In binomial distribution if the value of p is smaller or lesser than 0.5 then the binomial distribution is skewed to right. Binomial values, such as the number of heads in a series of coin tosses, are discrete. The normal distribution is continuous. However, the normal approximation provides an extremely accurate model for computing binomial probabilities. Each X value in the binomial distribution actually corresponds to a bar in the histogram. For example, if the score X=6 is represented by a bar it is bounded by real limits of 5.5 and 6.5. The actual probability of X = 6 is determined by the area contained in this bar.



Example:

To make it more understandable, it is explained through an example. In the game Rock-Paper-Scissors, the probability that both players will select the same response and tie is p=1/3 and the probability that they will pick different responses is p=2/3. If two people play 72 rounds of the game and choose their responses randomly, what is the probability that they will choose the same response (tie) more than 28 times?

Using Binomial Distribution To Test Hypothesis

In a **directional hypothesis test,** or a **one-tailed test,** the statistical hypotheses (*H*0 and *H*1) specify either an increase or a decrease in the population mean. That is, they make a statement about the direction of the effect.

Because a specific direction is expected for the treatment effect, it is possible for the researcher to perform a directional test. The first step (and the most critical step) is to state the statistical hypotheses. Remember that the null hypothesis states that there is no treatment effect and the alternative hypothesis says that there is an effect. For this example, the predicted effect is that the blueberry supplement will increase test scores. Thus, the two hypotheses would state:

H0: Test scores are not increased. (The treatment does not work.)

H1: Test scores are increased. (The treatment works as predicted.)

To express directional hypothesis in symbols, t usually is easier to begin with the alternative hypothesis (H1). Again, we know that the general population has an average test score of μ=80, and H1 states that the test scores will be increased by the blueberry supplement. Therefore, expressed in symbols, H1 states,

H1: $\mu > 80$ (With the supplement, the average score is greater than 80.)

The null hypothesis states that the supplement does not increase scores. In symbols,

Ho:$\mu \leq 80$ (With the supplement, the average score is not greater than 80.)

Note again that the two hypothesis are mutually exclusive and cover all of the possibilities.

The major distinction between one-tailed and two-tailed tests is the criteria that they use for rejecting *H*0. A one-tailed test allows you to reject the null hypothesis when the difference between the sample and the population is relatively small, provided that the difference is in the specified direction. A two-tailed test, on the other hand, requires a relatively large difference independent of direction. This point is illustrated in the following example.

To test hypothesis with the binomial distribution, we must calculate probability p, of the observed event. We compare this to the level of significance α.

If $p > \alpha$ then we do not reject the null hypothesis.

If $p < \alpha$ then we accept the alternative hypothesis.

A coin is tossed 20 times (n=20) landing on head more than 6 times (X=6.5). Perform a hypothesis test at a 5% (0.05) significance level to see if the coin is biased.

Ho: the coin is not biased

H1 : the coin is biased in favor of tails

Lesson 17

SAMPLING DISTRIBUTION

A sampling distribution is a statistic that is arrived out through repeated sampling from a larger population. It describes a range of possible outcomes that of a statistic, such as the mean or mode of some variable, as it truly exists a population.

The distribution of sample means is the collection of sample means for all of the possible random samples of a particular size (n) that can be obtained from a population. Because statistics are obtained from samples, a distribution of statistics is referred to as a sampling distribution. Sample is a subset of population.

A sampling distribution is a distribution of statistics obtained by selecting all of the possible samples of a specific size from a population. Thus, the distribution of sample means is an example of a sampling distribution. In fact, it often is called the sampling distribution of M.

The Distribution Of Sample Means

It is the collection of sample means for all of the possible random samples of a particular size (n) that can be obtained from a population. The distribution of sample means contains all of the possible samples. It is necessary to have all of the possible values to compute probabilities.

For example, if the entire set contains exactly 100 samples, then the probability of obtaining any specific sample is 1 out of 100: p=1/100

General Characteristics Of a Distribution

Sampling distribution is a statistic that determines the probability of an event based on data from a small group within a large population. Its primary purpose is to establish *representative results* of small samples of a comparatively larger population.

1. Sample means should be relatively close to the population mean.
2. The pile of sample means should tend to form a normal-shaped distribution.
3. The larger the sample size, the closer the sample means should be to the population mean, μ.

Central Limit Theorem

A mathematical proposition known as the central limit theorem provides a precise description of the distribution that would be obtained if you selected every possible sample, calculated every sample mean, and constructed the distribution of the sample mean. This important and useful theorem serves as a cornerstone for much of inferential statistics. Following is the essence of the theorem. For any population with mean μ and standard deviation , the distribution of sample means for sample size n will have a mean of μ and a standard deviation of n and will approach a normal distribution as n approaches infinity.

Central limit theorem describes the distribution of sample means by identifying the three basic characteristics that describe any distribution: shape, central tendency, and variability.The central limit theorem states that if the sample size increases sampling distribution must approach normal distribution. Generally a sample size more than 30 us considered as large enough.

The Shape Of The Distribution Of Sample Means

The distribution of sample means tends to be a normal distribution. In fact, this distribution is almost perfectly normal if;

1. The population from which the samples are selected is a normal distribution.
2. The number of scores (n) in each sample is relatively large, around 30 or more.

The mean of the distribution of sample means is equal to the mean of the population of scores, μ, and is called the expected value of M.

Standard Error Of Sampling Distribution

Sampling error is the natural discrepancy, or amount of error, between a sample statistic and its corresponding population parameter.

The standard deviation (measure of variability) for the distribution of sample means is identified by the symbol M and is called the standard error of M.

The standard error serves the two purposes for the distribution of sample means.

The Standard Error Of M

1. The standard error describes the distribution of sample means. It provides a measure of how much difference is expected from one sample to another.

2. Standard error measures how well an individual sample mean represents the entire distribution.

The symbol for the standard error is σM. The σ indicates that this value is a standard deviation, and the subscript M indicates that it is the standard deviation for the distribution of sample means.

The magnitude of the standard error is determined by two factors:

(1) The size of the sample and

(2) The standard deviation of the population from which the sample is selected.

As the sample size increases, the error between the sample mean and the population mean should decrease. This rule is also known as the law of large numbers.

The law of large numbers states that the larger the sample size (n), the more probable it is that the sample mean is close to the population mean.

The Population Standard Deviation

When n=1 the standard deviation for the distribution of sample means, which is the standard error, is identical to the standard deviation for the distribution of scores.

When n=1, σM = σ (standard error = standard deviation).

Standard error = σM= σ/√n

As sample size (n) increases, the size of the standard error decreases. (Larger samples are more accurate.) When the sample consists of a single score (n=1), the standard error is the same as the standard deviation (M).

Lesson 18

CONFIDENCE INTERVAL-I

A *confidence interval* is defined as the range of values that we observe in our sample and for which we expect to find the value that accurately reflects the population. A confidence interval is an interval, or range of values, centered around a sample statistic. The logic behind a confidence interval is that a sample statistic, such as a sample mean, should be relatively near to the corresponding population parameter. Therefore, we can confidently estimate that the value of the parameter should be located in the interval.

*Confidence interva*l is an interval of values computed from sample data that is almost sure to cover the true population parameter. We are estimating population parameter from sample statistics. Point estimate will be at the center of a confidence interval. The most common level of confidence used is 95%.

For example, a confidence interval for the population mean could be calculated with data obtained from a sample and would provide an estimated range of values within which the actual population mean is believed to lie. A confidence interval often is reported in addition to the point estimate of a population parameter. We can have 90% and 99% confidence intervals. It is impossible to construct an interval in which we could be 100% confident unless we actually measure the entire population.

For an approximately normal data set, the values within one standard deviation of the mean account for about 68% of the set; while within two standard deviations account for about 95%; and within three standard deviations account for about 99.7%.

Confidence Interval For Population Proportion

Estimating population proportion from a sample proportion. The most commonly reported information that can be used to construct a confidence interval is the **margin of error.**

- To construct a 95% confidence interval for a population proportion, simply add and subtract the margin of error to the sample proportion.
- The margin of error is often reported using the symbol "±".

- The formula for a 95% confidence interval can thus be expressed as Sample proportion ± margin of error.

Constructing a Confidence Interval For a Proportion

If numerous samples are taken, the frequency curve made from proportions from the various samples will be approximately bell-shaped. The mean will be the true proportion from the population. The Standard deviation will be;

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

To be exact, we would actually add and subtract 1.96(SD) instead of 2(SD) because 95% of the values for a bell-shaped curve fall within 1.96 standard deviations of the mean. However, in most practical applications, rounding 1.96 off to 2.0 will not make much difference and this is common practice.

Applying the reasoning we used to construct the formula for a 95% confidence interval and using the information about bell-shaped curves we can construct for instance, a 68% confidence interval. Relationship between a 95 confidence interval and a 99 confidence interval from the same sample is that 99% interval will be wider. For example by simply adding and subtracting 1 standard deviation to the sample proportion instead of 2.

Lesson 19

CONFIDENCE INTERVAL- II

In previous lesson we have discussed confidence interval for population proportion as Estimating population proportion from a sample proportion. In this lesson we will discuss about confidence interval for population mean.

Confidence Interval For Population Means

We can try to estimate population mean when all we have available is a sample of measurements from the population. All we need from the sample are its mean, standard deviation, and number of observations. Sample mean do not affect the width of the confidence interval.

Normally-distributed data forms a bell shape when plotted on a graph, with the sample mean in the middle and the rest of the data distributed fairly evenly on either side of the mean. The confidence interval for data which follows a standard normal distribution is:

$$CI = \overline{X} \pm Z^* \frac{\sigma}{\sqrt{n}}$$

Where:

CI = the confidence interval

X = the population mean

Z* = the critical value of the z-distribution

σ = the population standard deviation

√n = the square root of the population size

The confidence interval for the t-distribution follows the same formula, but replaces the Z* with the t*.

In real life, you never know the true values for the population (unless you can do a complete census). Instead, we replace the population values with the values from our sample data, so the formula becomes:

$$CI = \hat{x} \pm Z^* \frac{s}{\sqrt{n}}$$

Where:

ˆx = the sample mean

s = the sample standard deviation

The Rule for Sample Means

- If numerous samples of the same size are taken, the frequency curve of means from the various samples will be approximately bell-shaped. The mean of this collection of sample means will be the same as the mean of the population. Population standard deviation/square root of sample size = σ/√n
- The standard deviation for the possible sample means is called the standard error of the mean.
- In other words, SEM = standard error = population standard deviation√n
- To construct a 95% confidence interval for a mean we can use the same reasoning we used for proportions. In 95% of all samples, the sample mean will fall within 2 standard errors of the true population mean.The values associated with a two-sided 95% confidence interval of the standard normal distribution are ± 1.96.

The formula for a 95% confidence interval for a population mean becomes:

Sample mean ± 2 standard errors

Where Standard error = σ√n

This formula should be used only if there are at least 30 observations in the sample.

To compute a 95% confidence interval for the population mean based on smaller samples, a multiplier larger than 2 is used, which is found from a "t-distribution."

Examples:

Ali got a sample of 30 graduate students and found that mean age for that sample was = 33 years with SD of 4.3 years. What is the 95% confidence interval for average age of entire university graduate students.

Wood and colleagues (1988), studied a group of 89 sedentary men for a year. 42 men were placed on a diet; the remaining 47 were put on an exercise routine. The group on a diet lost an average of 7.2 kg, with a standard deviation of 3.7 kg. The men who exercised lost an average of 4.0 kg, with a standard deviation of 3.9 kg.

Notice that these intervals are trying to capture the true mean or average value for the population. They do not encompass the full range of weight loss that would be experienced by most individuals. Also, remember that these intervals could be wrong. Ninety-five percent of intervals constructed this way will contain the correct population mean value, but 5% will not.

Confidence Interval For Between Two Means

Instead of separately comparing the two groups from the population the efficient way is to construct a single confidence interval for the difference in the population means for the two groups or conditions.

1. Collect a large sample of observations (at least 30), independently, under each condition or from each group. Compute the mean and the standard deviation for each sample.

2. Compute the standard error of the mean (SEM) for each sample by dividing the sample standard deviation by the square root of the sample size.

3. Square the two SEMs and add them together. Then take the square root. This will give you the necessary "measure of variability," which is called the standard error of the difference in two means. In other words:

measure of variability=square root of [(SEM1)2 + (SEM1)2]

4. A 95% confidence interval for the difference in the two population means is

difference in sample means ± 2×measure of variability

or

difference in sample means ± 2×square root of [(SEM1)2 + (SEM1)2]

This method is valid only when independent measurements are taken from the two groups. For instance, if matched pairs are used and one treatment is randomly assigned to each half of the pair, the measurements would not be independent.

To construct a formula for constructing a confidence interval for the difference between two population means includes:

- A point estimate of the difference between the population means.
- The standard error of the sampling distribution of the sample means.
- The confidence level.

Reporting Confidence Interval

Confidence intervals are sometimes reported in papers, though researchers more often report the standard deviation of their estimate. If you are asked to report the confidence interval, you should include the upper and lower bounds of the confidence interval.

The first confidence interval compares the mean educational levels for the smokers and nonsmokers. The result tells us that, the average educational level for nonsmokers was 0.67 year higher than for smokers.

The interval tells us that the difference in the population is probably between 0.15 and 1.19 years of education. In other words, mothers who did not smoke were also likely to have had more education.

	Sample Means		
	0 Cigarettes	10+ Cigarettes	Difference (95% CI)
Maternal education, grades	11.57	10.89	0.67 (0.15,1.19)
Stanford-Binet (IQ), 48 mo	113.28	103.12	10.16 (5.04,15.30)
Birthweight, g	3416	3035	381.0 (167.1,594.9)

Lesson 20

HYPOTHESIS TESTING-I

Hypothesis testing is one of the most commonly used inferential procedures. In fact, most of the remainder of this book examines hypothesis testing in a variety of different situations and applications. Although the details of a hypothesis test change from one situation to another, the general process remains constant. In this chapter, we introduce the general procedure for a hypothesis test. You should notice that we use the statistical techniques that is, we combine the concepts of *z*-scores, probability, and the distribution of sample means to create a new statistical procedure known as a *hypothesis test*.

A **hypothesis test** is a statistical method that uses sample data to evaluate a hypothesis about a population. In very simple terms, the logic underlying the hypothesis-testing procedure is as follows:

- First, we state a hypothesis about a population. Usually the hypothesis concerns the value of a population parameter. For example, we might hypothesize that American adults gain an average of 7 pounds between Thanksgiving and New Year's Day each year.
- Before we select a sample, we use the hypothesis to predict the characteristics that the sample should have. For example, if we predict that the average weight gain for the population is 7 pounds, then we would predict that our sample should have a mean *around* 7 pounds. Remember: The sample should be similar to the population, but you always expect a certain amount of error.
- Next, we obtain a random sample from the population. For example, we might select a sample of $n = 200$ American adults and measure the average weight change for the sample between Thanksgiving and New Year's Day.
- Finally, we compare the obtained sample data with the prediction that was made from the hypothesis. If the sample mean is consistent with the prediction, then we conclude that the hypothesis is reasonable. But if there is a big discrepancy between the data and the prediction, then we decide that the hypothesis is wrong.

Logic of Hypothesis Testing

Hypothesis is an assumption made about a population (which is to be tested).

Before we select a sample, we use the hypothesis to predict the characteristics that the sample should have.Next, we obtain a random sample from the population. Finally, we compare the obtained sample data with the prediction that was made from the hypothesis. The researcher begins with a known population (a sample). This is the set of individuals as they exist.

The basic idea is to avoid having to reason about the real world by setting up a hypothetical world that is completely understood. The observed patterns of the data are then compared to what would be generated in the hypothetical world. If they don't match, then there is reason to doubt that the data support the hypothesis

Basic Steps For Hypothesis Testing

Step 1: State the hypothesis

As the name implies, the process of hypothesis testing begins by stating a hypothesis about the unknown population. Actually, we state two opposing hypotheses. Notice that both hypotheses are stated in terms of population parameters. The first, and most important, of the two hypotheses is called the *null hypothesis*. The null hypothesis states that the treatment has no effect. In general, the null hypothesis states that there is no change, no effect, no difference nothing happened, hence the name *null*. The null hypothesis is identified by the symbol *H*0. (The *H* stands for *hypothesis*, and the zero subscript indicates that this is the *zero-effect* hypothesis.)

For the study in Example 8.1, the null hypothesis states that the blueberry supplement has no effect on cognitive functioning for the population of adults who are more than 65 years old. In symbols, this hypothesis is:

Ho : μ(with supplement) = 80 (Even with the supplement, the mean test score is still 80.)

The **null hypothesis** (*H*0) states that in the general population there is no change, no difference, or no relationship. In the context of an experiment, *H*0 predicts that the independent variable (treatment) *has no effect* on the dependent variable (scores) for the population.

The second hypothesis is simply the opposite of the null hypothesis, and it is called the *scientific,* or *alternative, hypothesis* (*H*1). This hypothesis states that the treatment has an effect on the

dependent variable. The **alternative hypothesis** (*H*1) states that there is a change, a difference, or a relationship for the general population. In the context of an experiment, *H*1 predicts that the independent variable (treatment) *does have an effect* on the dependent variable.

For this example, the alternative hypothesis states that the supplement does have an effect on cognitive functioning for the population and will cause a change in the mean score. In symbols, the alternative hypothesis is represented as:

H1: μ(with supplement) ≠ 80 (Even with the supplement, the mean test score is different from 80.)

Notice that the alternative hypothesis simply states that there will be some type of change. It does not specify whether the effect will be increased or decreased test scores. In some circumstances, it is appropriate for the alternative hypothesis to specify the direction of the effect. For example, the researcher might hypothesize that the supplement will increase neuropsychological test scores ($\mu > 80$). This type of hypothesis results in a directional hypothesis test. For now we concentrate on non-directional tests, for which the hypotheses simply state that the treatment has no effect (*H*0) or has some effect (*H*1).

Step 2: Set the criteria for a decision

Eventually the researcher uses the data from the sample to evaluate the credibility of the null hypothesis. The data either provide support for the null hypothesis or tend to refute the null hypothesis. In particular, if there is a big discrepancy between the data and the null hypothesis, then we conclude that the null hypothesis is wrong.

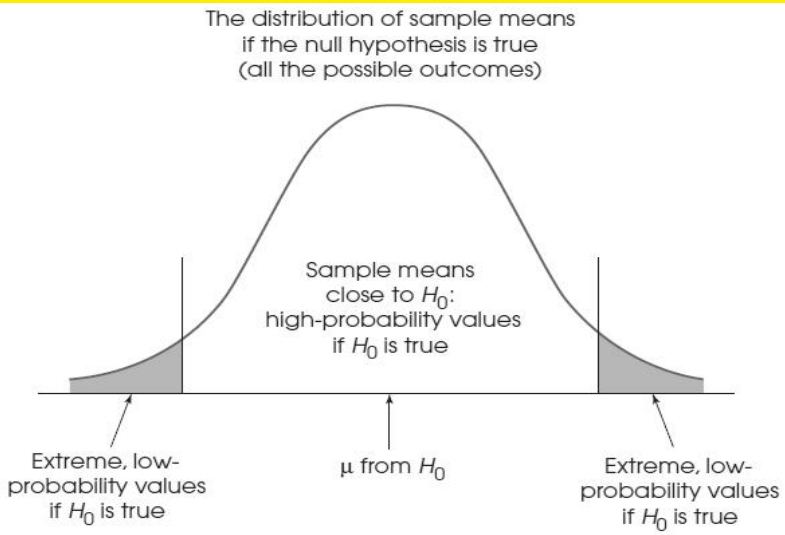
To formalize the decision process, we use the null hypothesis to predict the kind of sample mean that ought to be obtained. Specifically, we determine exactly which sample means are consistent with the null hypothesis and which sample means are at odds with the null hypothesis.

The Alpha Level

To find the boundaries that separate the high-probability samples from the low-probability samples, we must define exactly what is meant by "low" probability and "high" probability.

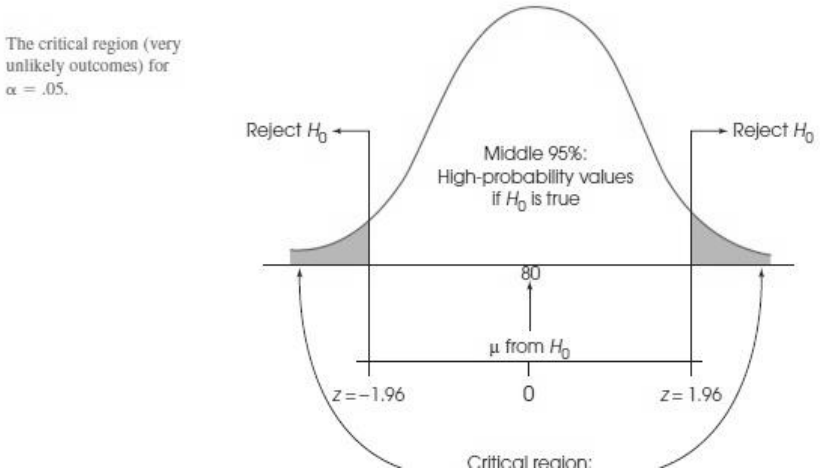
*The **alpha level,** or the **level of significance,** is a probability value that is used to define the concept of "very unlikely" in a hypothesis test.*

As in the graph given below, this is accomplished by selecting a specific probability value, which is known as the *level of significance,* or the *alpha level,* for the hypothesis test. The alpha (α) value is a small probability that is used to identify the low probability samples. By convention, commonly used alpha levels are $\alpha = .05$ (5%), $\alpha = .01$ (1%), and $\alpha = .001$ (0.1%). For example, with $\alpha = .05$, we separate the most unlikely 5% of the sample means (the extreme values) from the most likely 95% of the sample means (the central values).



The extremely unlikely values, as defined by the alpha level, make up what is called the *critical region*. These extreme values in the tails of the distribution define outcomes that are not consistent with the null hypothesis; that is, they are very unlikely to occur if the null hypothesis is true. Whenever the data from a research study produce a sample mean that is located in the critical region, we conclude that the data are not consistent with the null hypothesis, and we reject the null hypothesis.

The below graph shows the critical region. Critical region is composed of the extreme sample values that are very unlikely (as defined by the alpha level) to be obtained if the null hypothesis is true. The boundaries for the critical region are determined by the alpha level. If sample data fall in the critical region, the null hypothesis is rejected.



Step 3: Collect data and compute sample statistics

At this time, we select a sample of adults who are more than 65 years old and give each one a daily dose of the blueberry supplement. After 6 months, the neuro-psychological test is used to measure cognitive function for the sample of participants. Notice that the data are collected *after* the researcher has stated the hypotheses and established the criteria for a decision. This sequence of events helps to ensure that a researcher makes an honest, objective evaluation of the data and does not tamper with the decision criteria after the experimental outcome is known.

Next, the raw data from the sample are summarized with the appropriate statistics: For this example, the researcher would compute the sample mean. Now it is possible for the researcher to compare the sample mean (the data) with the null hypothesis. This is the heart of the hypothesis test: comparing the data with the hypothesis. The comparison is accomplished by computing a *z*-score that describes exactly where the sample mean is located relative to the hypothesized population mean from *Ho*. In step 2, we constructed the distribution of sample means that would be expected if the null hypothesis were true—that is, the entire set of sample means that could be obtained if the treatment has no effect. Now we calculate a *z*-score that identifies where our sample mean is located in this hypothesized distribution. The *z*-score formula for a sample mean is:

$$z = \frac{M - \mu}{\sigma_M}$$

In the formula, the value of the sample mean (M) is obtained from the sample data, and the value of μ is obtained from the null hypothesis. Thus, the z-score formula can be expressed in words as follows:

z = sample mean-hypothesized population mean/ standard error between M and μ.

Notice that the top of the z-score formula measures how much difference there is between the data and the hypothesis. The bottom of the formula measures the standard distance that ought to exist between a sample mean and the population mean.

Step 4: Make a decision

In the final step, the researcher uses the *z*-score value obtained in step 3 to make a decision about the null hypothesis according to the criteria established in step 2. There are two possible outcomes:

1. The sample data are located in the critical region. By definition, a sample value in the critical region is very unlikely to occur if the null hypothesis is true. Therefore, we conclude that the sample is not consistent with *H*0 and our decision is to *reject the null hypothesis*. Remember, the null hypothesis states that there is no treatment effect, so rejecting *H*0 means that we are concluding that the treatment did have an effect.

For the example we have been considering, suppose that the sample produced a mean of M = 92 after taking the supplement for 6 months. The null hypothesis states that the population mean is μ= 80 and, with n = 25 and σ=20, the standard error for the sample mean is

$$\sigma_M = \frac{\sigma}{\sqrt{n}} = \frac{20}{\sqrt{25}} = \frac{20}{5} = 4$$

Thus. a sample mean of M = 92 produces a z-score of :

$$\sigma_M = \frac{\sigma}{\sqrt{n}} = \frac{20}{\sqrt{25}} = \frac{20}{5} = 4$$

With an alpha level of a = .05, this z-score is far beyond the boundary of 1.96. Because the sample z-score is in the critical region, we reject the null hypothesis and conclude that the blueberry supplement did have an effect on cognitive functioning.

2. The second possibility is that the sample data are not in the critical region. In this case, the sample mean is reasonably close to the population mean specified in the null hypothesis (in the center of the distribution). Because the data do not provide strong evidence that the null hypothesis is wrong, our conclusion is to *fail to reject the null hypothesis*. This conclusion means that the treatment does not appear to have an effect.

For the research study examining the-blueberry supplement, suppose our sample produced a mean test score of M = 84. As before, the standard error for a sample of n = 25 is $\sigma_M = 4$, and the null hypothesis states that is μ = 80. These values produce a z-score of

$$z = \frac{M - \mu}{\sigma_M} = \frac{84 - 80}{4} = \frac{4}{4} = 1.00$$

The z-score of 1.00 is not in the critical region. Therefore, we would fail to reject the null hypothesis and conclude that the blueberry supplement does not appear to have an effect on cognitive functioning.

Types Of Hypothesis

Depending on the population distribution, you can classify the statistical hypothesis into two types:

Simple Hypothesis: A simple hypothesis specifies an exact value for the parameter. *Composite Hypothesis:* A composite hypothesis specifies a range of values.

One-Tailed And Two-Tailed Hypothesis

Hypothesis can be One-tailed and two-tailed.

In a **directional hypothesis test,** or a **one-tailed test,** the statistical hypotheses (*H*0 and *H*1) specify either an increase or a decrease in the population mean. That is, they make a statement about the direction of the effect.

Because a specific direction is expected for the treatment effect, it is possible for the researcher to perform a directional test. The first step (and the most critical step) is to state the statistical hypotheses. Remember that the null hypothesis states that there is no treatment effect and the alternative hypothesis says that there is an effect. For this example, the predicted effect is that the blueberry supplement will increase test scores. Thus, the two hypotheses would state:

Ho: Test scores are not increased. (The treatment does not work.)

H1:Test scores are increased. (The treatment works as predicted.)

To express directional hypotheses in symbols, it usually is easier to begin with the alternative hypothesis (H1). Again we know that the general population has an average test score of μ=80, and H1 states that scores will be increased by the blueberry supplement. Therefore, expressed in symbols, H1 states,

H1:μ> 80 (with the supplement, the average score is greater than 80.)

The null hypothesis states that the supplement does not increase scores. In symbols,

H0:μ ≤80 (With the supplement, the average score is not greater than 80.)

Note again that the two hypothesis are mutually exclusive and cover all of the possibilities.

Two-tailed hypothesis tests are also known as non directional and two-sided tests. It can test for effects in both directions. When a two-tailed test is performed the significance level/critical region is split between both tails of the distribution.

Ho: there is no training effect on performance; M ≤μ

H1: there is training effect on performance M ≠μ

Comparison Of One-Tailed Versus Two-Tailed Tests

The major distinction between one-tailed and two-tailed tests is the criteria that they use for rejecting *H*0. A one-tailed test allows you to reject the null hypothesis when the difference between the sample and the population is relatively small, provided that the difference is in the specified direction. A two-tailed test, on the other hand, requires a relatively large difference independent of direction.

Lesson 21

HYPOTHESIS TESTING-II

Hypothesis testing is a form of statistical inference that uses data from a sample to draw conclusions about a population parameter or a population probability distribution. First, a tentative assumption is made about the parameter or distribution. This assumption is called the null hypothesis and is denoted by H0.

In very simple terms, the logic underlying the hypothesis-testing procedure is we state a hypothesis about a population. Before we select a sample, we use the hypothesis to predict the characteristics that the sample should have. Next, we obtain a random sample from the population. Finally, we compare the obtained sample data with the prediction that was made from the hypothesis. If the sample mean is consistent with the prediction, then we conclude that the hypothesis is reasonable. But if there is a big discrepancy between the data and the prediction, then we decide that the hypothesis is wrong.

Assumptions For Hypothesis Tests

Assumptions underlying the hypothesis testing is as follows.

1. **Random sampling**

It is assumed that the participants used in the study were selected randomly.

2. **Independent observations**

The values in the sample must consist of independent observations.

Two events (or observations) are independent if the occurrence of the first event has no effect on the probability of the second event.

3. **The value of (σ) is unchanged by the treatment**

We assume that the standard deviation for the unknown population (after treatment) is the same as it was for the population before treatment.

4. **Normal sampling distribution**

For hypotheses with z-scores, the unit normal table is used to identify the critical region. This table can be used only if the distribution of sample means is normal.

Errors In Hypothesis Testing

In the framework of hypothesis tests there are two types of errors: Type I error and type II error. A type I error occurs if a true null hypothesis is rejected (a "false positive"), while a type II error occurs if a false null hypothesis is not rejected (a "false negative").

- **Type I Error**

A Type I error occurs when a researcher rejects a null hypothesis that is actually true.

It means that the researcher concludes that a treatment does have an effect when, in fact, it has no effect. A Type I error is more likely to occur when a researcher unknowingly obtains an extreme, non representative sample. The alpha level determines the probability of a Type I error.

- **Type II Error**

A Type II error occurs when a researcher fails to reject a null hypothesis that is really false. It means that the hypothesis test has failed to detect a real treatment effect.

In this case the treatment does influence the sample, but the magnitude of the effect is not big enough to move the sample mean into the critical region. Because the sample is not substantially different from the original population, the statistical decision is to fail to reject the null hypothesis.

The probability of a Type II error is represented by the symbol β, the Greek letter beta and it depends on variety of factors rather than a specific number.

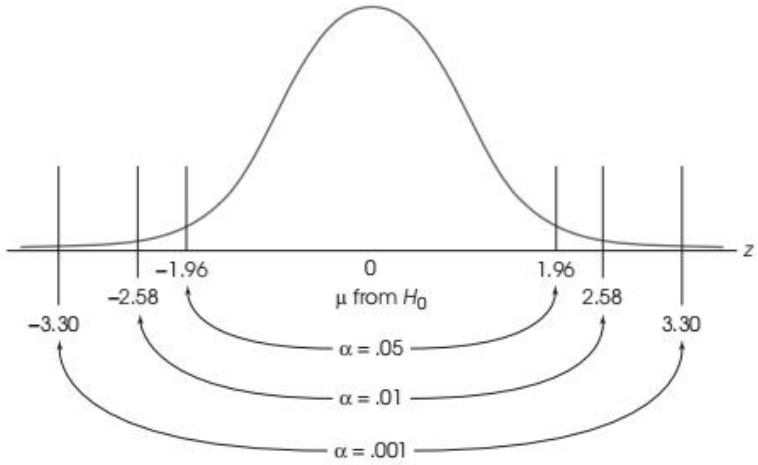
		No Effect, H_0 True	Effect Exists, H_0 False
Experimenter's Decision	Reject H_0	Type I error	Decision correct
	Retain H_0	Decision correct	Type II error

Selecting An Alpha Level

Alpha is a threshold value used to judge whether a test statistic is statistically significant. It is chosen by the researcher. Alpha represents an acceptable probability of a Type I error in a statistical test. Because alpha corresponds to a probability, it can range from 0 to 1. The alpha level for a hypothesis test serves two very important functions.

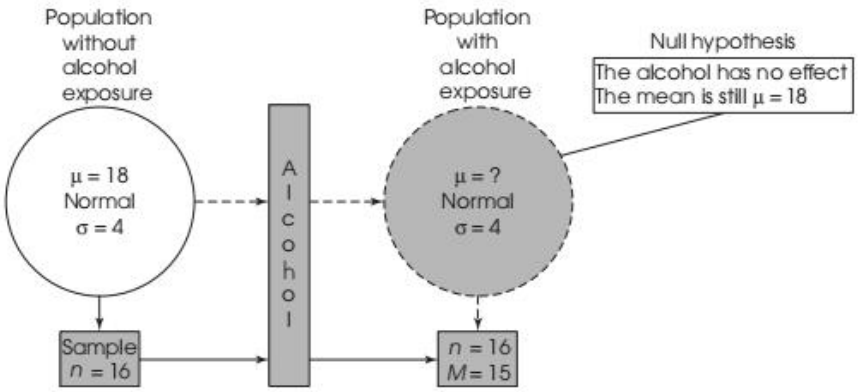
- First, the alpha level helps to determine the boundaries for the critical region
- Secondly, the alpha level determines the probability of a Type I error.

The primary concern when selecting an alpha level is to minimize the risk of a Type I error. The consequences of a Type I error can be relatively serious, therefore, many researchers and prefer to use an alpha level such as .01 or .001 to reduce the risk.



Exercise: How To Do Hypothesis Testing

For practice of students, hypothesis testing with another example is mentioned that is a researcher would like to investigate the effect of prenatal alcohol exposure on birth weight. In diagram below, a random sample of n=16 pregnant rats is obtained. The mother rats are given daily doses of alcohol. At birth, one pup is selected from each litter to produce a sample of n=16 newborn rats. The average weight for the sample is M=15 grams. The researcher would like to compare the sample with the general population of rats. It is known that regular newborn rats (not exposed to alcohol) have an average weight of μ=18 grams. The distribution of weights is normal with σ=4.



The structure of a research study to determine whether prenatal alcohol effects birth weight. A sample is selected from the original population and is exposed to alcohol. The is what would happen if the entire population were exposed to alcohol. The treated sample provides information about the unknown treated population.

1. State hypothesis

The null hypothesis states that exposure to alcohol has no effect on birth weight

Ho: μalcohol exposure =18

The alternative hypothesis states that alcohol exposure does affect birth weight

H1 : μ alcohol exposure ≠18

2. Select alpha level

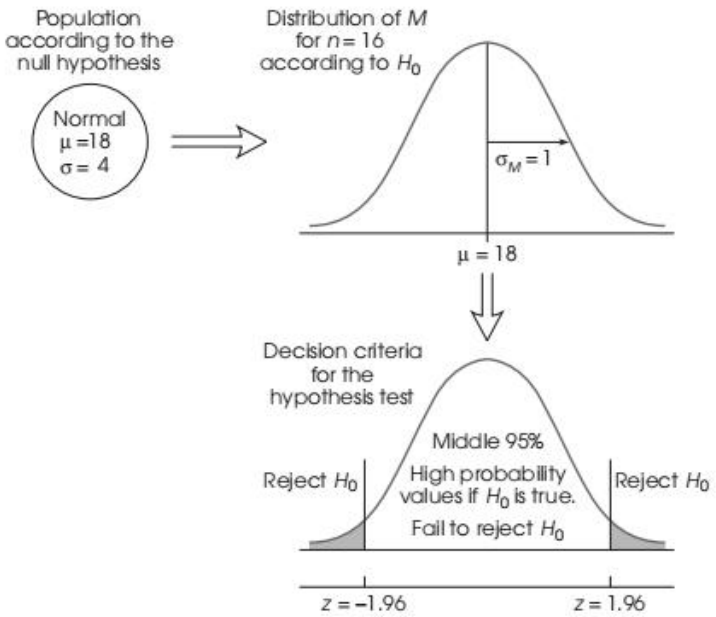
An alpha level of .05 will be used

That is, we are taking a 5% risk of committing a Type I error.

3. ***Set the decision criteria by locating the critical region***

Begin with the population of scores

Construct the distribution of sample means for the sample size



Use z-scores to separate the extreme outcomes (as defined by the alpha level) from the high-probability outcomes.

4. Collect the data, and compute the test statistic

Researcher would select one newborn pup from each of the n=16 mothers that received alcohol during pregnancy. The birth weight is recorded for each pup and the sample mean is computed; M= 15grams

The sample mean is then converted to a z-score

5. Make a decision

The z-score computed has a value of 3.00, which is beyond the boundary of -1.96.

Therefore, the sample mean is located in the critical region.

Lesson 22

Hypothesis Testing-III

The final decision in a hypothesis test is determined by the value obtained for the z-score statistic. If the z-score is large enough to be in the critical region, then we reject the null hypothesis and conclude that there is a significant treatment effect. Otherwise, we fail to reject H0 and conclude that the treatment does not have a significant effect.

Factors Influencing Hypothesis Test

The most obvious factor influencing the size of the z-score is the difference between the sample mean and the hypothesized population mean from H0. A big mean difference indicates that the treated sample is noticeably different from the untreated population and usually supports a conclusion that the treatment effect is significant. In addition to the mean difference, however, there are other factors that help determine whether the z-score is large enough to reject H0.

In this section we examine two factors that can influence the outcome of a hypothesis test.

1. The variability of the scores, which is measured by either the standard deviation or the variance. The variability influences the size of the standard error in the denominator of the z-score.
2. The number of scores in the sample. This value also influences the size of the standard error in the denominator.

Statistical Power Of The Test

The power of a test is defined as the probability that the test will reject the null hypothesis if the treatment really has an effect.

The power of a statistical test is the probability that the test will correctly reject a false null hypothesis. That is, power is the probability that the test will identify a treatment effect if one really exists.

There are only two possible outcomes for a hypothesis test: either fail to reject Ho or reject Ho.

- The first outcome, failing to reject Ho when there is a real effect is a Type II error with a probability identified as p= β.
- The second outcome have a probability of 1 – β. Hence, rejecting Ho when there is a real effect this and is the power of the test.

Factors Affecting Power Of The Test

There are many factors that can influence the power of a test. Such as power is to detect the effect when it is actually present, power is to avoid/lower the probability of committing type II error. Beta is the probability of Type II error. Beta is equal to proportion of alternate distribution that falls below the critical value. Factors that can effect the power of a test is mentioned below:

- **Sample size**

When the sample size is reduced, power decreases to less than 50%. In general, a larger sample produces greater power for a hypothesis test.

- **Alpha level**

Reducing the alpha level for a hypothesis test also reduces the power of the test. For example, lowering from .05 to .01 lowers the power of the hypothesis test.

- **One-tailed versus two-tailed test**

Changing from a regular two-tailed test to a one-tailed test increases the power of the hypothesis test.