VIRTUAL UNIVERSITY OF PAKISTAN

Social Statistics

SOC201

Course Handouts

GRADUP

Practice. Pass. Graduate.

thegradup.com

Lesson 01

Introduction to Statistics
Topic: 001-002

Topic 001: Introduction to Statistics

Learning Outcomes

The learning outcomes of social Statistics are given below:

- Make students understand the importance of research and statistics in psychology.
- Calculate the statistics necessary to solve problems.
- To help students find out appropriate statistical analysis based on research questions, hypotheses, and type of data/variables.
- Enable students to enter, encode and prepare data files, run analysis, and interpret output files

Chapters/Topics

- Introduction to statistics and SPSS
- Introduction to SPSS
- Descriptive statistics
- Introduction to Inferential Statistics
- Testing Hypothesis
- Computing differences between Multiple Groups using t-test, F-test
- Computing relationship between variables

Agenda

- Defining Statistics
- Why Statistics is Important?
- What are some Myths About Statistics
- Terms and Definitions: Descriptive and Inferential Statistics Sample and Population

Topic 002: Defining Statistics

The term statistics refers to a set of mathematical procedures for organizing, summarizing, and interpreting information (for example, your scores on the first assignment)

Statistics is the science of designing studies or experiments, collecting data, and analyzing data for the purpose of scientific discovery. Statistics is the science of Learning from Data.

Purposes of Statistics:

Specifically, statistics serve two general purposes:

1. Statistics are 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. Statistics help 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.

Statistics deals with:

- Collection of data,
- Organizing and summarizing the data,
- Analysis of data, and
- Making inferences, or decisions and predictions.

Step 4 is the objective of statistics in sociology, i.e., making inferences about a population based on information contained in a representative sample taken from that population. Sociological Research only makes sense when you are able to understand statistics

Topic 003: Why Statistics is Important

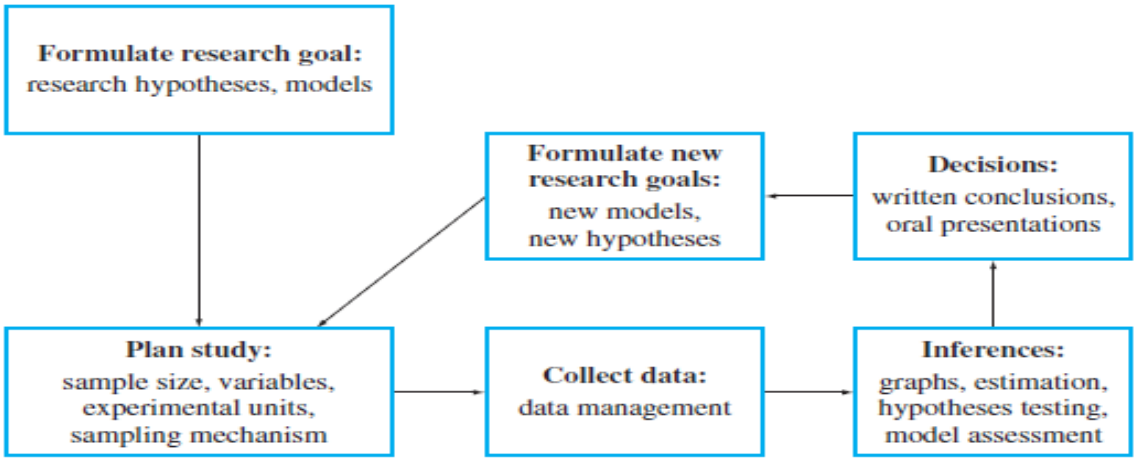
The great body of physical science, a great deal of the essential facts of financial science, and endless social and political problems are only accessible and only thinkable to those who have had a sound training in mathematical analysis, and the time may not be very remote when it will be understood that for complete initiation as an efficient citizen of one of the new great complex world, it is as necessary to be able to compute, to think in averages and maxima and minima, as it is now to be able to read and write"

H. G. Wells (1903) Mankind in the Making

Following are some of the main reasons why a sociology student needs to learn statistics.

- No Research without Statistics
- Increase Accuracy of Study
- Universal Language

1. **No Research without Statistics:**



2. **Increase Accuracy of Study:**

Statistical procedures help to ensure that the information or observations are presented and interpreted in an accurate and informative way.

3. **Universal Language**

Statistics provide researchers with standardized techniques that are recognized and understood throughout the scientific community. Thus, the statistical methods used by one researcher are familiar to other researchers, who can accurately interpret the statistical analyses with a full understanding of how the analysis was done and what the results signify.

© Virtual University of Pakistan

Lesson 02

Definition & Myth about Statistics
Topic: 004-005

Topic 004: Some Myths about Statistics

There are some myths about statistics

- Statistics is very hard.
- Statistics is math.
- Why do we teach statistics in sociology whether we are studying society?
- You lie with statistics
- If something isn't statistically significant, it is not important

Topic 005: Terms and Definitions

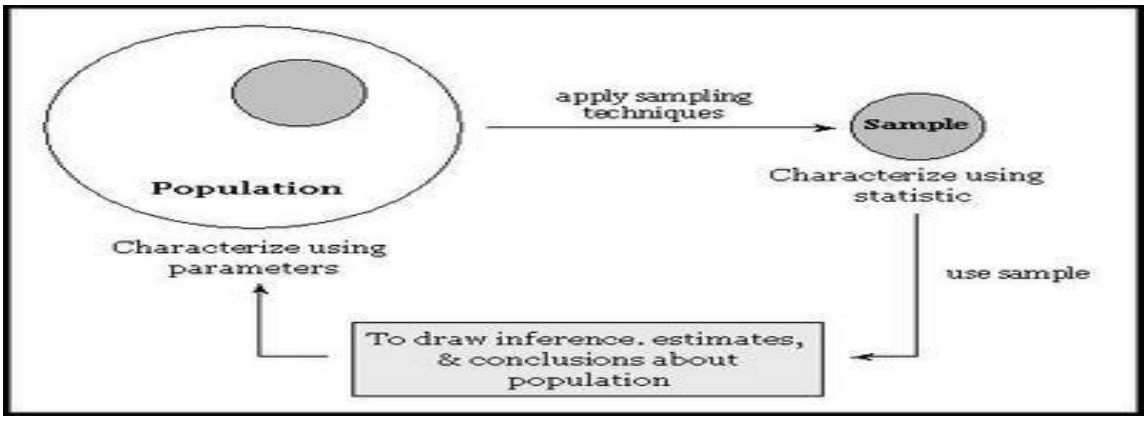
Descriptive Statistics:
It is statistical procedures that are used to summarize, organize, and simplify data.

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

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

Population is a complete set of all the individuals of interest in a particular study Population may vary in size from extremely small (all women on the planet) to very small (Women who are heads of University departments) depending on how the researcher defines it.

Sample is a set of individuals selected from the population, usually intended to represent the population in a research study.



Parameter vs Statistic

	Parameter	Statistic
Mean	μ mu	$\bar{x}$
Proportion	p	$\hat{p}$
Std. Dev.	σ sigma	s
Correlation	ρ rho	r
Slope	β beta	b

Parameter: is a value, usually, a numerical value that describes a population. A parameter is usually derived from the measurement of the individuals in the population.

Statistic: is a value, usually a numerical value that describes a sample. A statistic is usually derived from the measurement of the individuals in the sample.

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.**

Lesson 03

Scales of measurement
Topic: 006-007

Topic 06: What to Measure

The question of what to measure is easy enough to answer: to test hypotheses we need to measure variables.

Variables:

Variables are just things that can change (or vary); they might vary between people (e.g., IQ, behavior) or locations (e.g., unemployment), or even time (e.g., mood, profit, number of cancerous cells).

Independent Variable

That experimenter changes or manipulates, or controls to see its effect on the dependent variable

Dependent variable:

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.

Discrete variable:

A discrete variable consists of separate, indivisible categories. No values can exist between two neighboring categories. Discrete variables are commonly restricted to the whole, countable numbers. For example, the number of children in a family or the number of students attending class.

Continuous variable:

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. The grade point average, GPA, of any student can take an infinite number of possible values, for example in the interval 0.0 to 4.0, hence GPA is a continuous variable.

Construct some variables are simple to define, such as Height, weight, age, eye color, etc. however, some variables are that scientists/psychologists use to explain behavior. Such as Intelligence, personality, anxiety, hunger, etc.

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

Operational Definition:

It 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.

Topic 07: Scales of Measurement

- 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.

Nominal scale

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.

For example, if you were measuring the academic majors for a group of college students, the categories would be art, biology, business, chemistry, and so on. Nominal scales are not quantitative values, they are occasionally represented by numbers.

Ordinal scale

An ordinal scale consists of a set of categories that are organized in an ordered sequence. Measurements on ordinal scale rank observations in terms of size or magnitude. From an ordinal scale, you can determine whether two individuals are different and you can determine the direction of difference but cannot determine the size of the difference.

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. Other examples of ordinal scales include socioeconomic class (upper, middle, lower) and T-shirt sizes (small, medium, large).

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.

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.

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

© Virtual University of Pakistan

Lesson 04

Correlation Methods
Topic: 008-009

Topic 08: Correlation methods-I

A type of research that describes whether and to what extent there is a relationship between two or more events or factors. Correlational research involves measuring two variables and assessing the relationship between them, with no manipulation of an independent variable. For example, the relationship between class attendance and CGPA is obtained.

While correlational research cannot be used to establish causal relationships between variables, correlational research does allow researchers to achieve many other important objectives (establishing reliability and validity, providing converging evidence, describing relationships and making predictions).
The most common analysis used in correlational studies is Pearson Product Moment Correlation (*r*). In many correlation studies, researchers predict outcomes based on more than one predictor variable. Thus, they need to account for the impact of each variable. Using Mediation and Regression.

Positive relationship

A positive relationship depicts, where low (or high) scores on one variable relate to low (or high) scores on a second variable.

Negative relationship

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.

Topic 08: Correlation methods-II

There are two types of correlational designs

- Cross-sectional designs
- Longitudinal designs

Cross-Sectional Designs

One on more samples is drawn from a population, which is studied at one time on the same variables. E.g. The effect of gender on depression of people born in the 60s and 90s. (Year makes two different samples). In most cases, socio-demographic variables are used to create two or more samples.

Both samples are studied at the same time. Usually, there is no time lapse between data collection from both samples. For this reason, this design is not suitable for studying changes in behavior that are seen over time.

Longitudinal Design

Same respondents or samples are surveyed over a period of time.it Can be 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. This design is particularly used to study changes in trends across time.

Disadvantages

- Drop-out rates can be high over the years.
- Lack of control over the population.
- Direction of causality can be uncertain.

Lesson 05

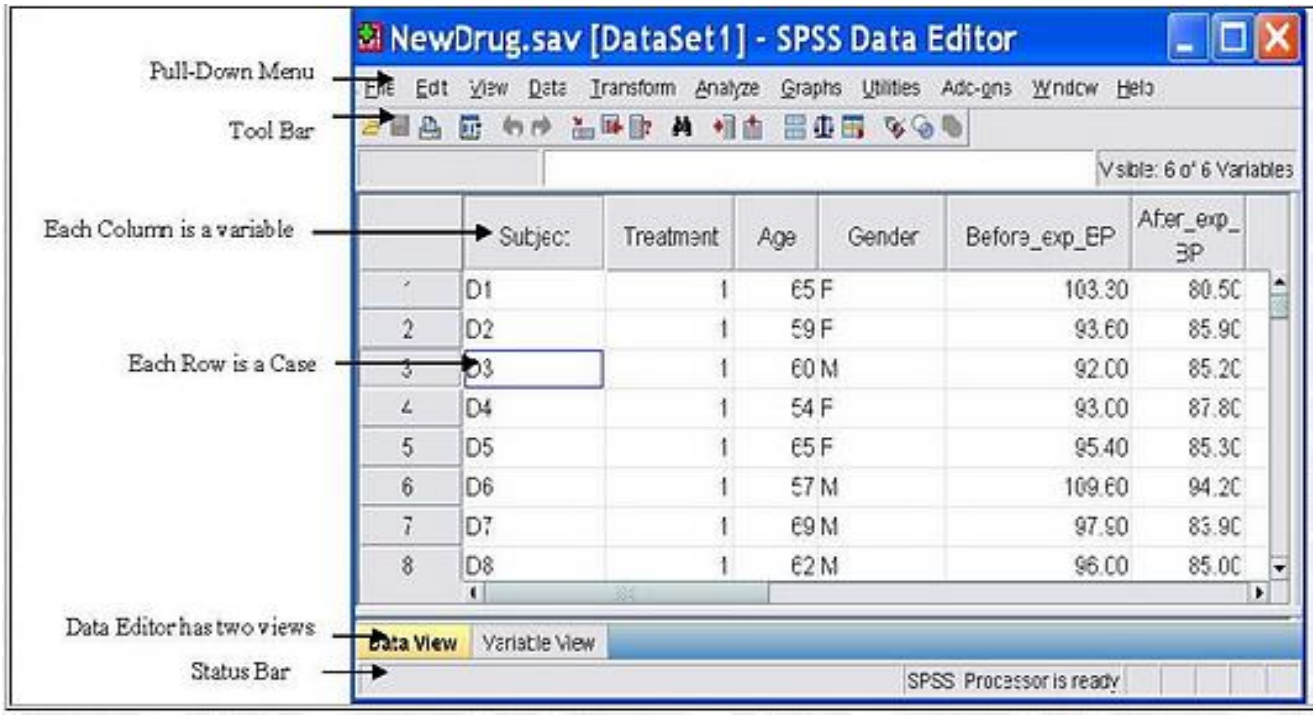
Introduction to SPSS
Topic: 010-12

Topic 10: Introduction to SPSS

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 means Statistical Package for Social Science.

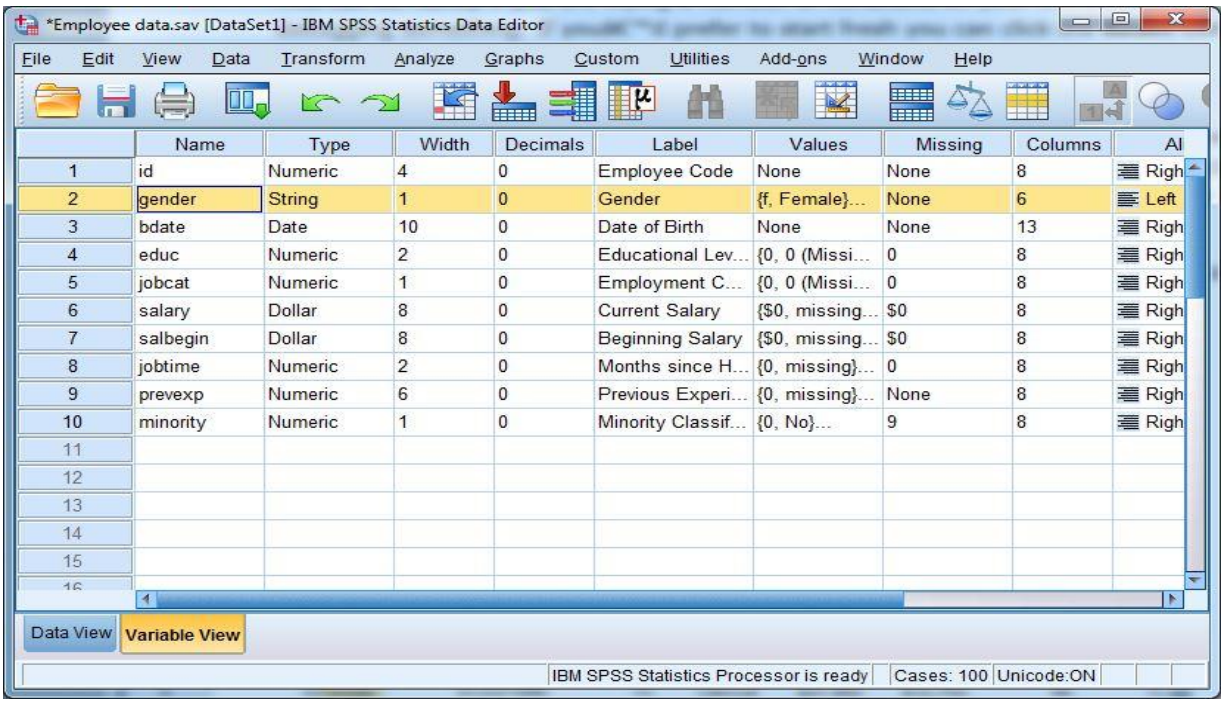
SPSS datasets always have a 2-dimensional table structure where the rows typically represent cases (such as individuals or households) and the columns or variables represent measurements (such as age, sex, or household income).

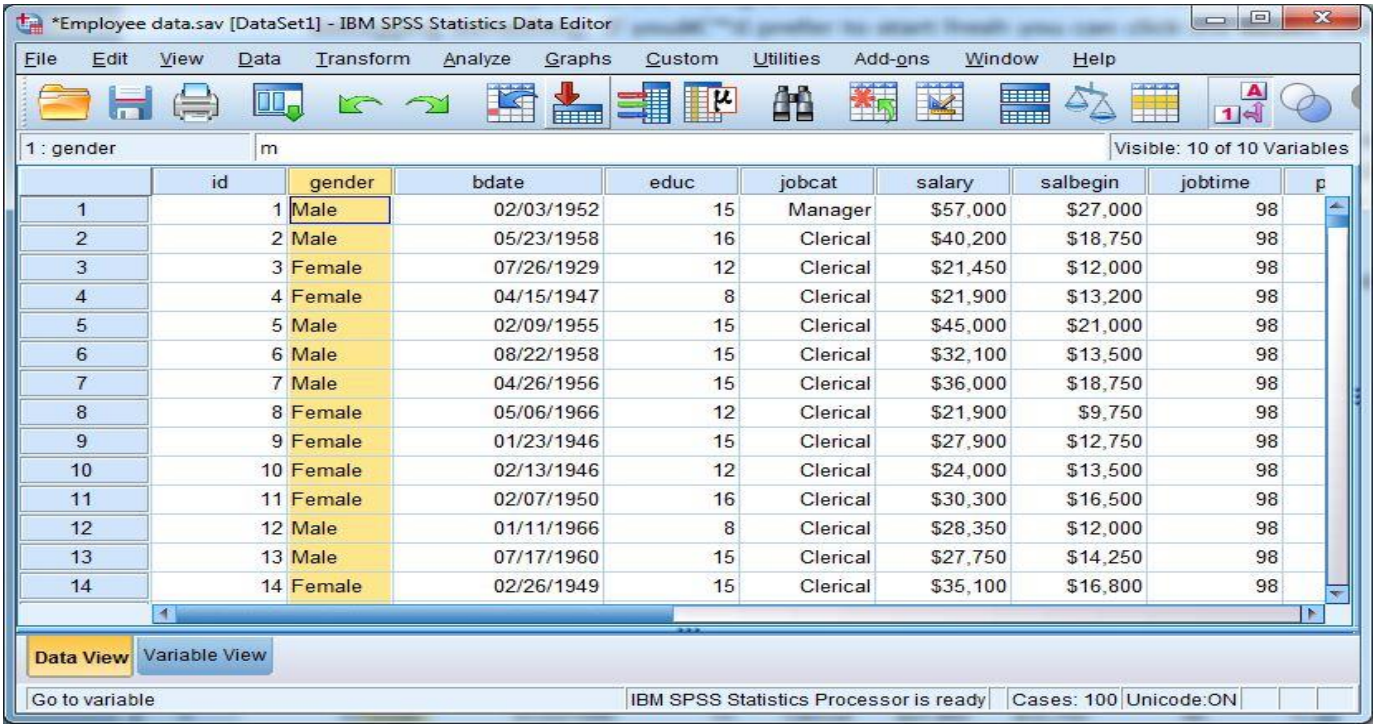




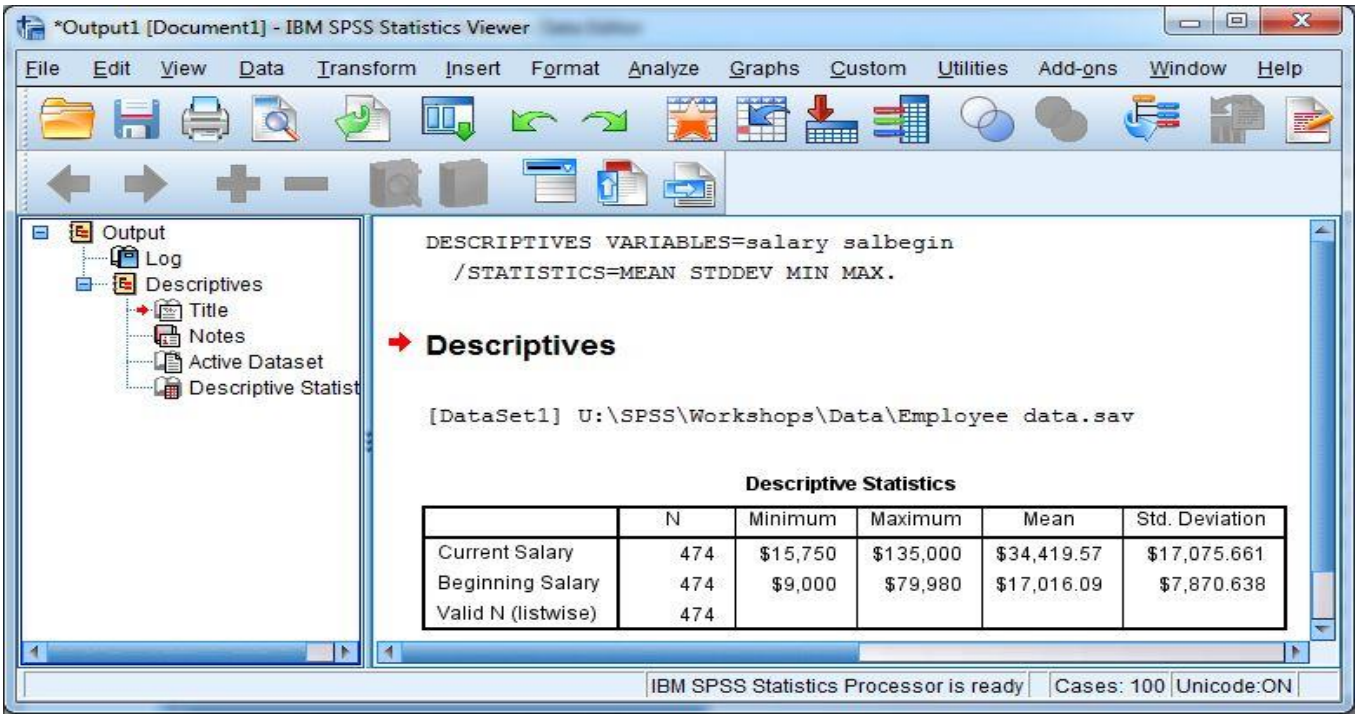
SPSS has three windows for working with data.
The Data Editor Window (.sav)

- Data view
- Variable view
- The Output Viewer Window

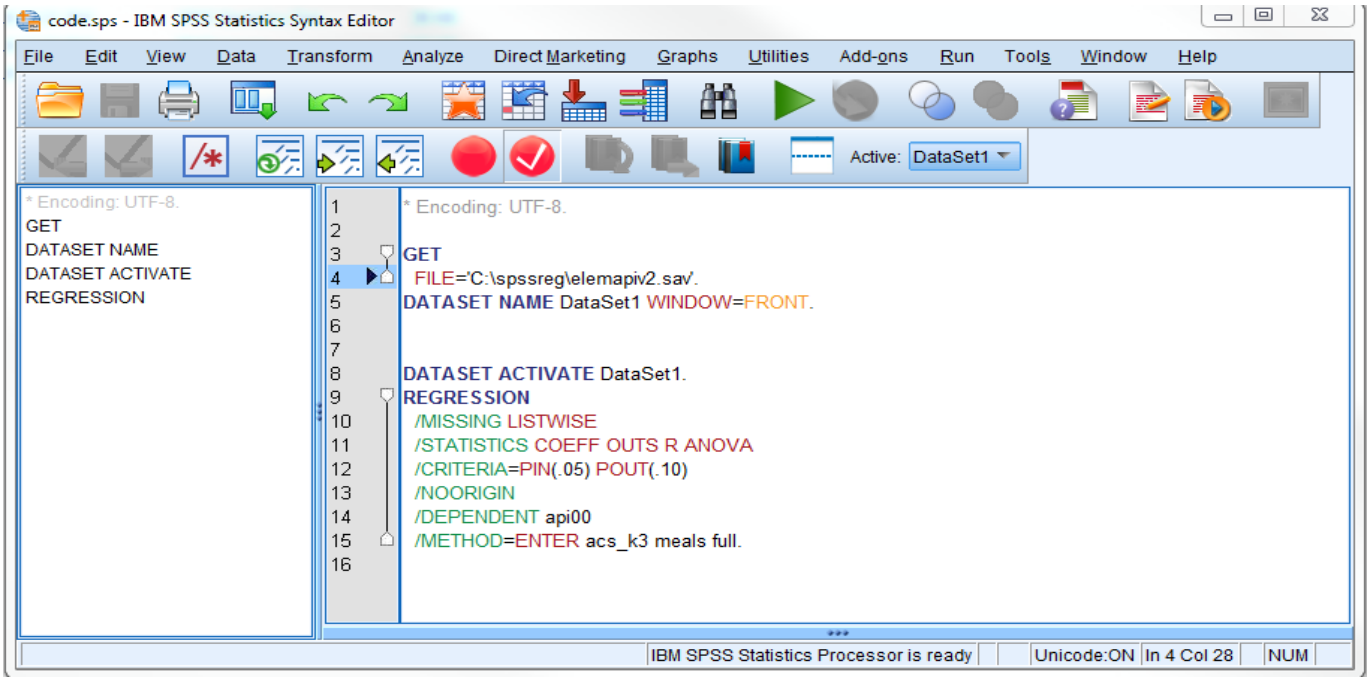


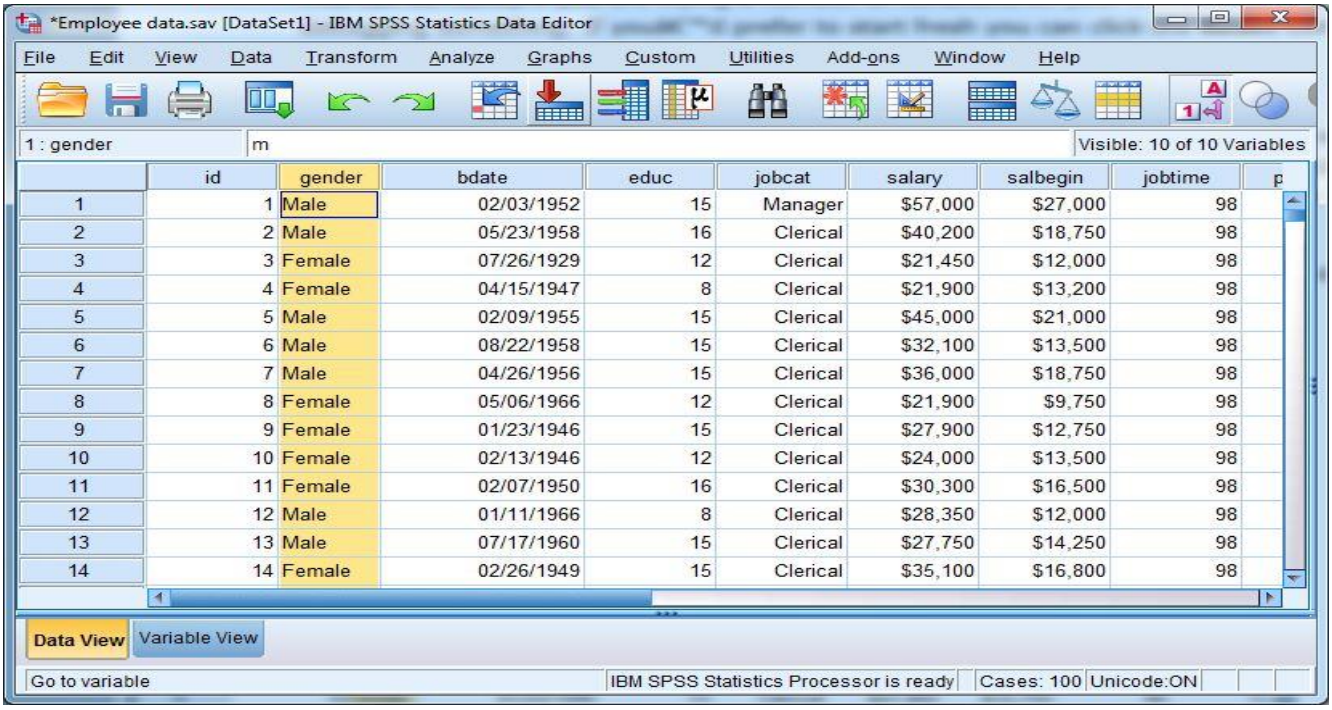


	id	gender	bdate	educ	jobcat	salary	salbegin	jobtime	p
1	1	Male	02/03/1952	15	Manager	$57,000	$27,000	98	
2	2	Male	05/23/1958	16	Clerical	$40,200	$18,750	98	
3	3	Female	07/26/1929	12	Clerical	$21,450	$12,000	98	
4	4	Female	04/15/1947	8	Clerical	$21,900	$13,200	98	
5	5	Male	02/09/1955	15	Clerical	$45,000	$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	



	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				





	id	gender	bdate	educ	jobcat	salary	salbegin	jobtime	p
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	

Topic 11: Data View in SPSS

Menu Bar Commands

The file includes all of the options you typically use in other programs, such as open, save, and 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.

Transform includes several options to change current variables. It is used for recording, computing new variables, and dealing with missing values.

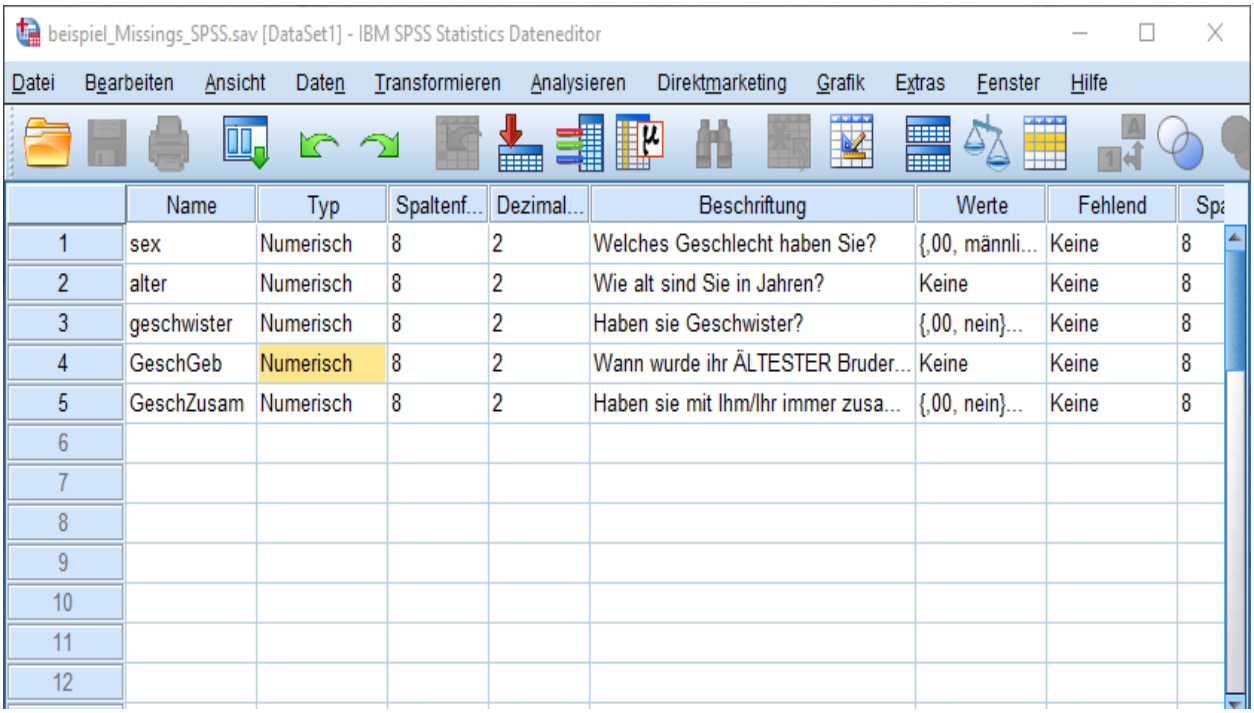
Analyze includes all of the commands to carry out statistical analyses and to calculate descriptive statistics.

Graphs include the commands to create various types of graphs including box plots, histograms, line graphs, and bar charts.

Utilities allow 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.



	Name	Typ	Spaltenf...	Dezimal...	Beschriftung	Werte	Fehlend	Spa
1	sex	Numerisch	8	2	Welches Geschlecht haben Sie?	{,00, männli...	Keine	8
2	alter	Numerisch	8	2	Wie alt sind Sie in Jahren?	Keine	Keine	8
3	geschwister	Numerisch	8	2	Haben sie Geschwister?	{,00, nein}...	Keine	8
4	GeschGeb	Numerisch	8	2	Wann wurde ihr ÄLTESTER Bruder...	Keine	Keine	8
5	GeschZusam	Numerisch	8	2	Haben sie mit Ihm/Ihr immer zusa...	{,00, nein}...	Keine	8
6								
7								
8								
9								
10								
11								
12								



This icon gives you the option to open a previously saved file



This icon allows you to save files. It will save the file you are currently working on (be it data or output).



This icon activates the Dialog box for printing whatever you are currently working on (either the data editor or the output).



Clicking on this icon will activate a list of the last 12 dialog boxes that you used.



This icon enables you to go directly to a case (a row in the data editor).



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



Icon opens a dialog box that shows you the variables in the data editor and summary information about each one.



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.



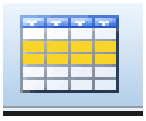
Creates a new variable to the left of the variable that is currently active.



Clicking on this icon is a shortcut to the Split File Function.



This icon shortcuts to the Weight Cases Function.



This icon is a shortcut to the Select Cases Function.

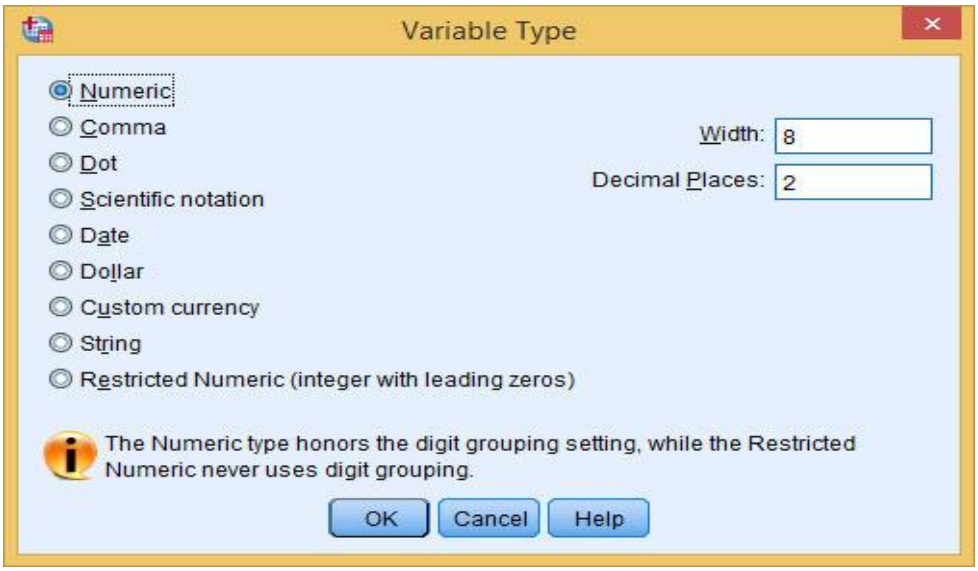


Clicking on this icon will either display or hide the value labels of any coding variables.

Topic 12: Entering Data in SPSS

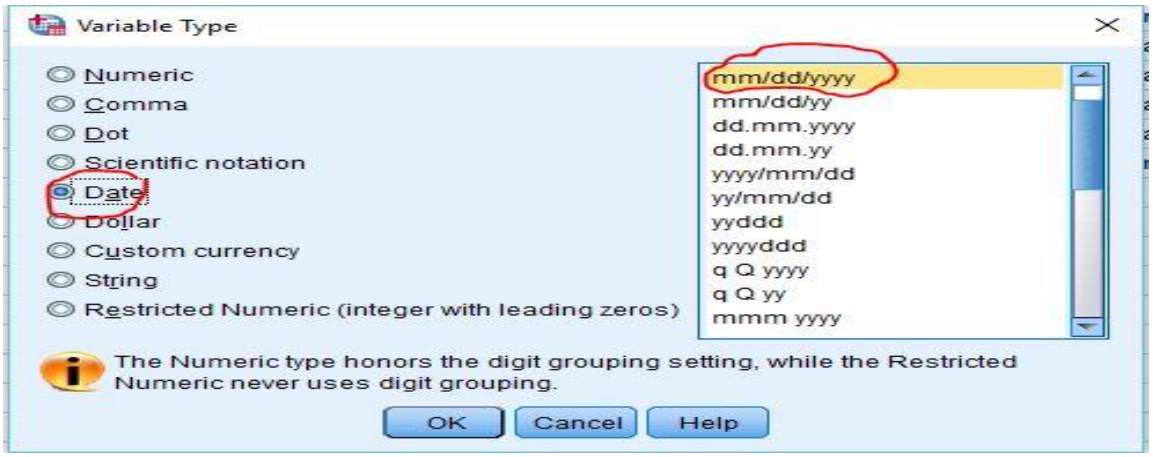
Creating a String Variable:

- Click in the first white cell in the column labeled Name.
- Type the word 'Name'.
- 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).
- Move into the column labeled TYPE. Click on it and a dialogue box will open.
- Choose the STRING option and click OK to make the variable string.
- Finally, to specify Measure and selecting either Nominal, Ordinal or Scale from the drop-down list.



Creating a Date Variable

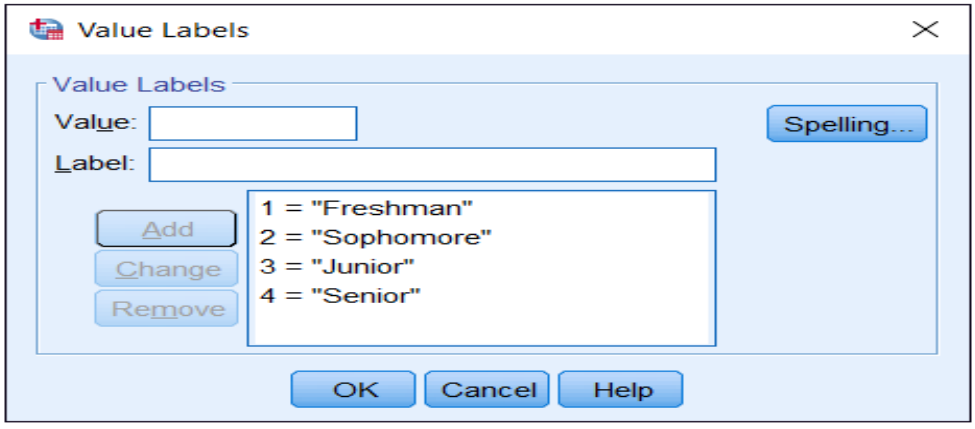
To enter date variables into SPSS we use the same procedure as with the string variable, except that we need to change the variable type.



Creating Coding Variables:

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). For coded categorical variables, the value label(s) 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).

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

Numeric variables are the easiest ones to create because SPSS assumes this format for data. E.g. Age, No. of Friends, total income etc.

- Click in the first white cell in the column labeled Name and type Age.
- Move into the column labelled TYPE. Click on it and a dialogue box will open. Choose Numeric and click ok.
- Specify the level at which a variable was measured by going to the column labeled Measure and selecting SCALE.

© Virtual University of Pakistan

Lesson 06

Descriptive Statistics
Topic: 013-14

Topic 13: Descriptive Statistics-I

Descriptive statistics are statistical procedures used to summarize, organize and simplify data. Describing and summarizing data are done in two ways.

- Graphically
- Numerically

Descriptive statistics usually include:

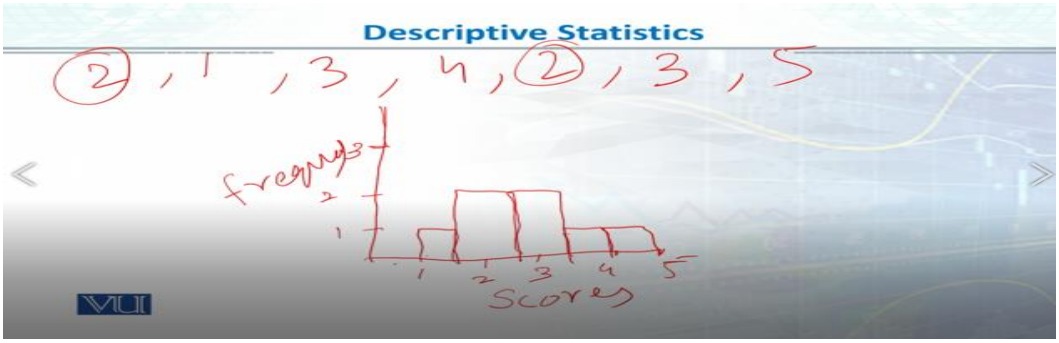
- Tables /distribution
- Central tendency
- Measure of dispersion
- Graphs/charts

The distribution of a variable:

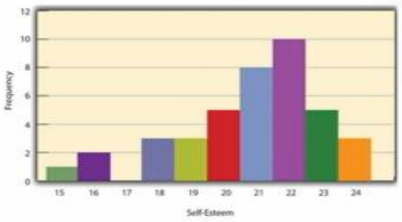
Every distribution has a variable which is the way the score is distributed across the level of that variable. in the same sample the distribution of a variable 'Sex' might be such that '44' has a score of 'male' and 56 have a score of 'female'

Topic 14: Descriptive Statistics-II

Presenting the data graph:



Descriptive Statistics



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 07

Frequency Distribution& Distribution Chart
Topic: 015-17

Topic 15: 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 a 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.

data: 3,1,1,2,5,4,4,5,3,5,3,2,3,4,3,3,4,3,2,3

X	f	p	%
5	3	.15	15%
4	4	.20	20%
3	8	.40	40%
2	3	.15	15%
1	2	.10	10%

P—proportion, the proportion of the total number of responses that fall into this category (p = f/N)
Percentages (%=p (100)) show the percentage of individuals accumulated in a category

Topic 16: Grouped Frequency Distribution

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
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

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

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

Class Limits and Grouped Frequency distribution:

When you are dealing with a continuous variable, the resulting measurement corresponds to intervals on the number rather than a single point. E.g. a score of 7 actually represents an interval by the real limits of 6.5 and 7.5

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

X	Class Limits	f
90-94	89.5-94.5	3
85-89	84.5-89.5	4
80-84	79.5-84.5	5
75-79	74.5-79.5	4
70-74	69.5-.74.5	3
65-69	64.5-69.5	1
60-64	59.5-64.5	3
55-59	54.5-59.5	1
50-54	49.5-54.5	1

Topic 17: Frequency Distribution Graphs/Charts

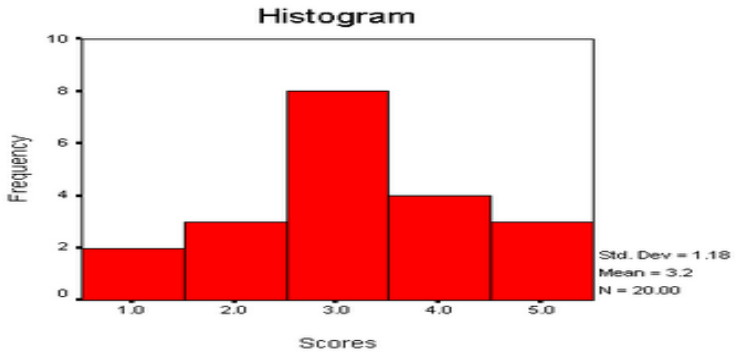
FREQUENCY DISTRIBUTION GRAPHS

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.*

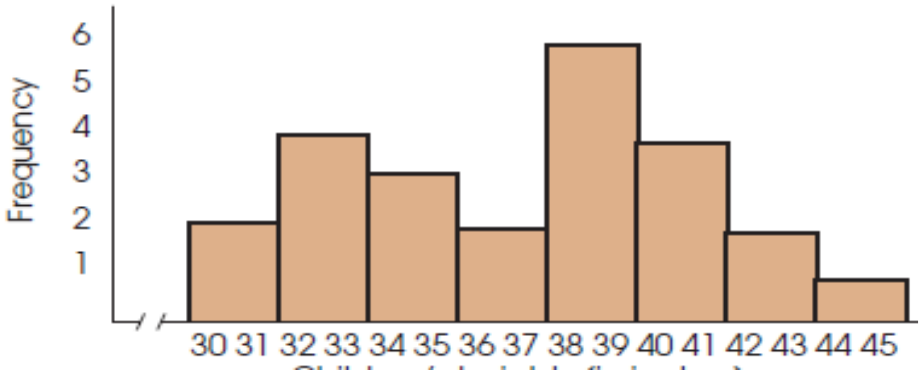
Histograms To construct a histogram, you first list the numerical scores (the categories of measurement) along the *X*-axis. Then you draw a bar above each *X* value so that

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.



Histogram from Grouped data: when data have been grouped into class intervals, you can construct a frequency distribution histogram by drawing a bar above each interval value, so that the height of the bar corresponds to the frequency of the score.

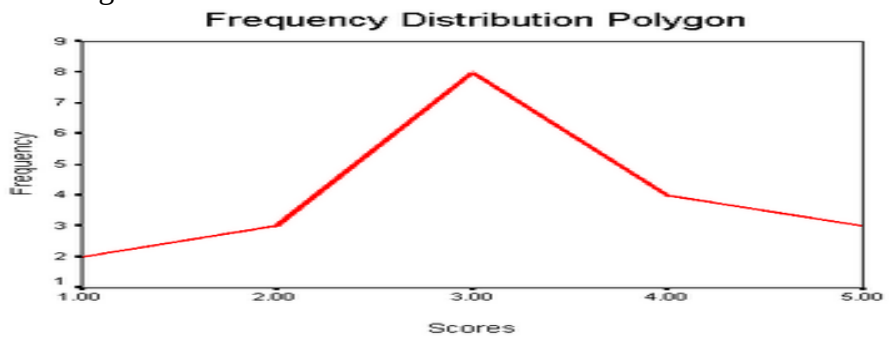


Polygons The second option for graphing a distribution of numerical scores from an interval or ratio scale of measurement is called a polygon. To construct a polygon, you begin by listing the numerical scores (the categories of measurement) along the *X*-axis. Then,

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.

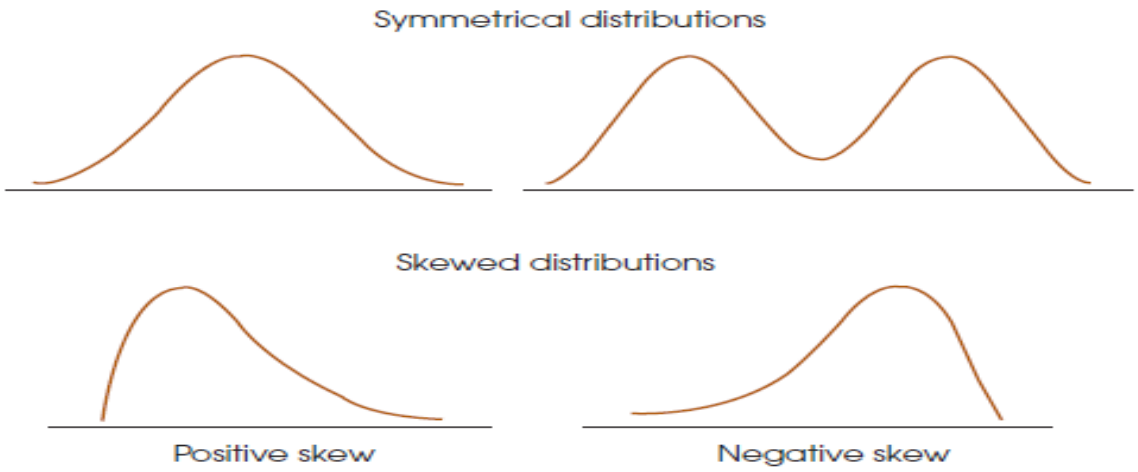
c. The graph is completed by drawing a line down to the *X*-axis (zero frequency) at each end of the range of scores. The final lines are usually drawn so that they reach the *X*-axis at a point that is one category below the lowest score on the left side and one category above the highest score on the right side.



Frequency Distribution Polygon with Grouped Data: You calculate midpoints for each class.

scores	f	Midpoints
12-13	4	12.5
10-11	5	10.5
8-9	3	8.5
6-7	3	6.5
4-5	2	4.5

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.*
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 08

Percentiles & Graphs in SPSS
Topic: 018-19

Topic 18: 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.**

Percentile

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

Suppose you have a score of 63, and exactly 70% of the class had a score of 63 or lower, then your score X= 63 has Percentile Rank = 70 and your score will be called **70th Percentile**

Calculating Percentiles and Percentile Ranks

By using frequency distributions, remember that you must take ***class limits***, as each cumulative percentage value (C %) is associated with the upper real limit of that interval

E.g. what is 95th Percentile?

What is percentile Rank for X=3.5?

X	f	cf	c%
5	1	20	100%
4	5	19	95%
3	8	14	70%
2	4	6	30%
1	2	2	10%

Percentiles and Percentile Ranks from Grouped Frequency Distributions

What if I have to calculate 50th Percentile?

X	f	cf	c%
20–24	2	20	100%
15–19	3	18	90%
10–14	3	15	75%
5–9	10	12	60%
0–4	2	2	10%

Cumulative frequencies

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		

Let's run all in SPSS

Variable: Scores on a test

2, 4, 5, 6, 5, 6, 7, 8, 9, 10, 8, 9, 11. 12. 14. 15. 17. 19. 20, 22

Data Entry >>>>> Analyze>>>> Descriptive Statistics>>> Highlight the variable and move to Variable Box>>>> Display Frequency Distribution>>> OK

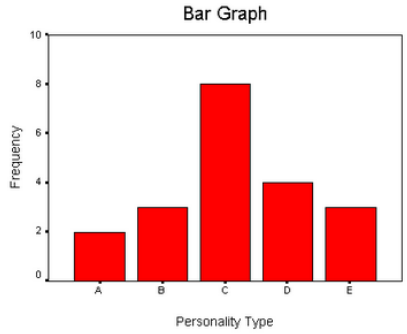
X	*f*	*cf*	*c%*
20–24	2	20	100%
15–19	3	18	90%
10–14	3	15	75%
5–9	10	12	60%
0–4	2	2	10%

Topic 19: Art of Presenting Data

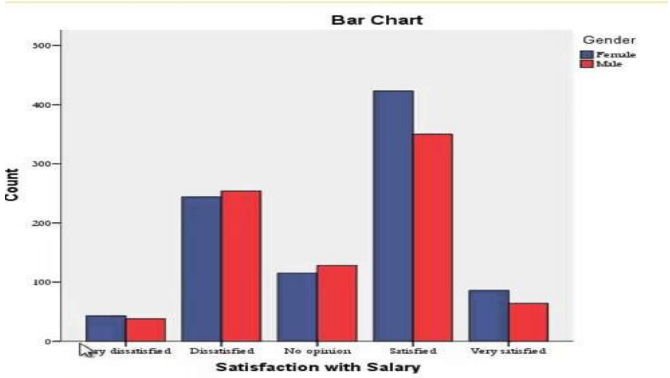
Bar graphs

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.

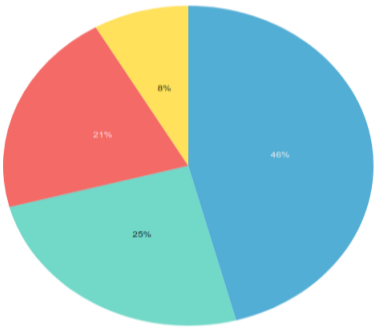


Clustered Bar Graph: can be used when you have two or more nominal or ordinal variables and want to illustrate the differences in the categories of these two variables based on some statistic. e.g. comparing satisfaction with salary in males in females



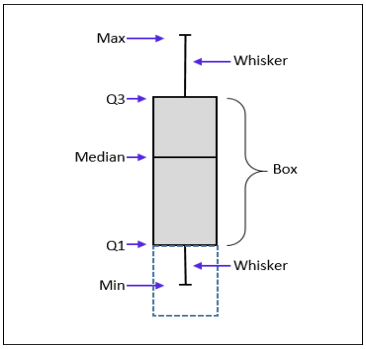
Pie Chart: It is a special chart that uses Pie slices to show the relative sizes of the data.

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 are useful when comparing distributions between many groups or datasets.



Lesson 09

Constructing Frequency Distribution and Graphs in SPSS
Topic: 020-21

Topic 20, 21: Constructing Frequency Distribution and Graphs in SPSS

Let's run all in SPSS

Frequency distribution
Percentiles
Graphs
Charts

For more details see the video of the topic to understand the topic

Lesson 10

Measures of Central Tendency

Topic: 022-24

Topic 22: Measures 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, the 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. E.g. Average yearly temperature in each province of Pakistan, Average salary of government-grade 17th employees
Favorite color of young girls in winter
The average score on SAT exam in Pakistan?
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.

Measures of Central Tendency

- Mean
- Median
- Mode

MEAN

The mean for 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:
Sample mean = M = $\Sigma x/n$

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

$$\text{overall mean} = M = \frac{\Sigma X_1 + \Sigma X_1}{n_1 + n_2}$$

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**

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

Class interval	Frequency f_i	Mid-value x_i	$f_i \times x_i$
1 – 3	12	2	24
3 – 5	22	4	88
5 – 7	27	6	162
7 – 9	19	8	152
	$\Sigma f_i = 80$		$\Sigma f_i x_i = 426$

$$\therefore \text{Mean} = \frac{\Sigma f_i x_i}{\Sigma f_i} = \frac{426}{80} = 5.325$$

Topic 23: Measures of Central Tendency II

MEDIAN
Median is the central value in the data so it divides the data into two equal parts. 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.
Median with an even number data:

When data is in even numbers, you will add the two central values of ordered data, and divide by 2
e.g. 1, 1, 4, 5, 7, 8

$$\text{median} = \frac{4+5}{2} = \frac{9}{2} = 4.5$$

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)

Scores	Class Limits	F	CF
50-60		5	
40-50		11	
30-40		10	
20-30		15	
10-20		7	
0-10		2	

Step1:
Identify Median class
= (N/2)th value
= (50/2)th value
= 25th value
Median class = 29.5 – 39.5

Step2:
Median = 29.5 + ([25 - 24]/10) x 10
= 29.5 + (1/10) x 10
= 29.5+ 1

= 30.5 Median = 31

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 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.

E.g. scores 1, 3, 4,4,5,5,5,6,7, 7

Mode: 5

E.g. Favorite fast food

Mode: McDonald

Restaurants	**F**
KFC	20
McDonald's	25
Burger King	8
Subway	12
Hardee's	15

Topic 24: Measures of Central Tendency III

When you use what:

The goal of CT is to find out one single value that best represents the entire distribution

Mean is the preferred measure as it uses every score in the distribution and is related to SD so is valuable measure in inferential statistics

But then when and why to use median and mode?

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 Mode should be used?

- For Nominal data
- Discrete variables
- When describing shape of the distribution
- Examples from real world?

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.

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.

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.

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.

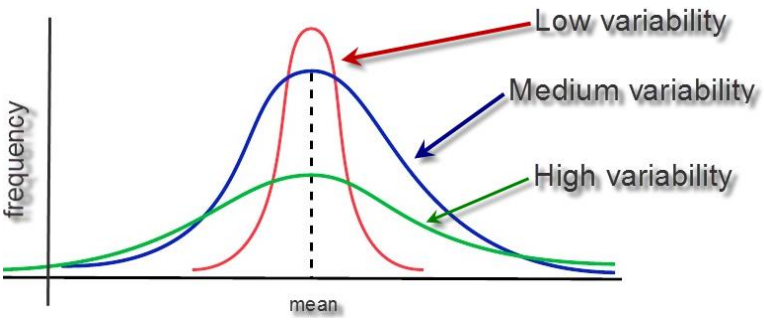
Lesson 11

Measures of Dispersion Variability
Topic: 025-26

Topic 25,26: Measures of Dispersion Variability

The measure of variability provides a single summary that describes the spread of scores within the distribution by measuring the degree to which the scores are spread out or are clustered together.

Data 1: 48, 49, 50, 51, 52
Data 2: 40, 45, 50, 55, 60
Data 3: 30, 40, 50, 60, 70
Data 4: 0, 25, 50, 75, 100



Types of Measures of Dispersion

- Range
- Inter Quartile Range
- Variance
- Standard Deviation

Range: Tells us the distance between maximum and minimum scores.

Range = max – minimum

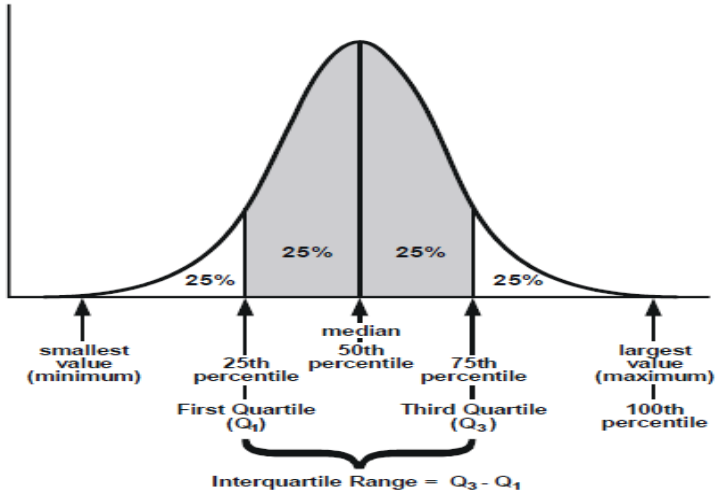
In some cases (when data is in whole numbers) we use range as an inclusive measure rather than a measure of the difference between the two scores ….then range is ….
Range = max – minimum + 1
E.g. Number of children in a family 1, 2, 3, 4, 5----- 1 & 5 are inclusive

Advantages

Quick and give a rough estimate of dispersion in the data

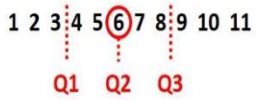
Disadvantages
Consider only extreme values and ignore all the middle values, so it's a rough and imprecise, and unreliable measure.

Inter Quartile Range:
Provides difference between the middle 50%of the cases
IQR = Q3-Q1



Provides difference between the middle 50%of the cases

$$IQR = Q3 - Q1$$
$$= 8.5 - 3.5$$
$$= 5$$



Pros:

- Easy to calculate
- Eliminates influence of extreme scores

Cons

- Discard much of the data

Variance

Deviation of each observation from the mean, and then averaging the sum of squared deviations.
Tells us whether the scores are clustered closely around the mane or widely scattered.
Deviation is distance from the mean: = X – μ (for population)
Distance/deviation of each score from its mean = x- $\bar{x}$

1. **deviation score**

Deviation = X – μ

2. **Squared deviation scores**

$(X - \mu)^2$

3. **sum of squared deviations**

$\Sigma(X - \mu)^2$

4. Variance

Variance = $(X - \mu)^2 / N$

5. Standard deviation

SD= variance

Example

X	X-x̄	$(X-\bar{x})^2$
25	10	100
20	5	25
15	0	0
10	-5	25
5	-10	100
∑=75		∑=250

For samples:

$$\text{variance} = s^2 = \frac{\Sigma (x-\bar{x})^2}{n-1}$$

$$\text{standard deviation} = s = \sqrt{s^2}$$

Calculating Formula

$$s^2 = \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n-1}$$

For populations:

$$\text{variance} = \sigma^2 = \frac{\Sigma (x-\bar{x})^2}{n}$$

$$\text{standard deviation} = \sigma = \sqrt{\sigma^2}$$

Calculating Formula

$$\sigma^2 = \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n}$$

The measure of variability provides an objective description of the differences between the Scores in distributed by measuring the degree to which the scores are spread out or are clustered together.

© Virtual University of Pakistan

Lesson 12

Measures of Variability& DS
Topic: 027-28

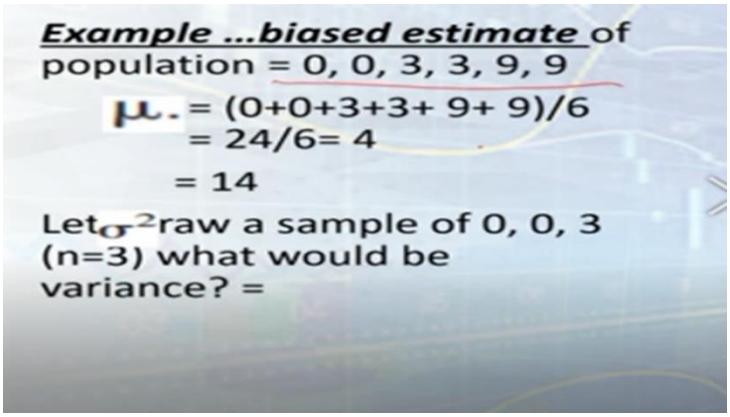
Topic 27: Measures of Variability and IS

The goal of inferential statistics is to use information from a sample and use it for the population. Samples are usually less variable than their population as extreme values in the population are less likely to be in the sample
The fact that the sample is less variable than it population means that sample variability gives a biased estimate of population variability.

Biased estimate

Sample statistics are said to be a biased **e**stimate ff, it consistently overestimates or underestimates the corresponding population parameter.
So we need to apply some corrections or adjustments which is (n-1)

Degrees of freedom:
Suppose a population has ma ean of 14 and you draw a sample n=3



Biased VS Unbiased Estimate:

To get the sample variance, dividing by n-1, is incorrect and produces biased statistics. Dividing by n-1 is correct and produces unbiased statistics.

Topic 28: Standard Deviation and Descriptive Statistics

Standard deviation summarizes the data and describes the entire distribution

Transformation of Scale:

- Adding a constant to each score does not change the standard deviation

- Multiplying each score by a constant causes the standard deviation to be multiplied by the same constant.

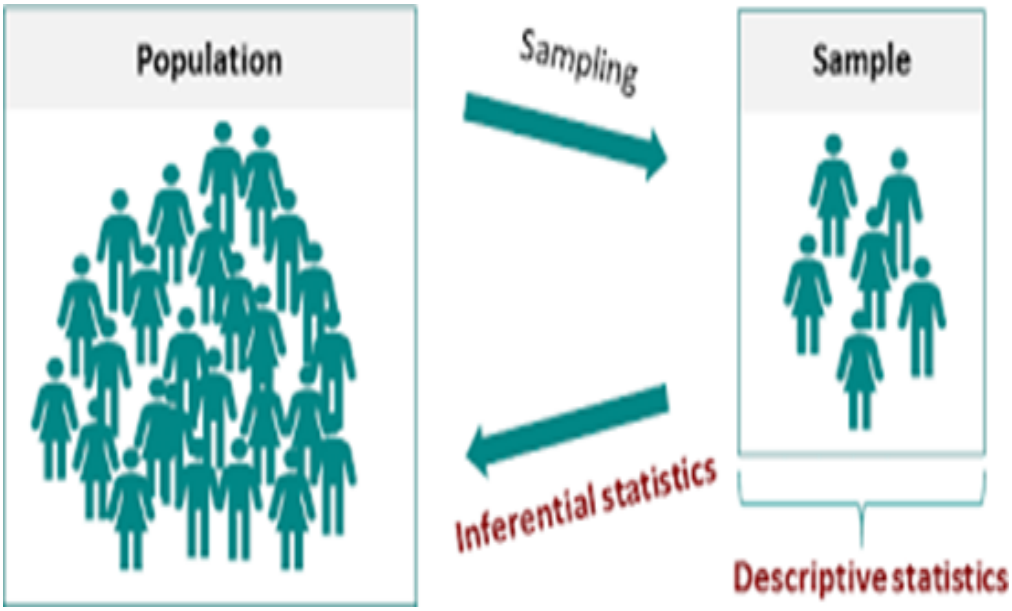
© Virtual University of Pakistan

Lesson 13

Inferential Statistic& Probability I
Topic: 29-30

Topic 029: What is 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:
Estimating population parameters.
Hypothesis testing.

- Most research uses statistical models called the Generalized Linear model and include Student's t-tests, ANOVA (Analysis of Variance), regression analysis**.**
- **Sampling error** is the naturally occurring discrepancy, or error, that exists between a sample statistic and the corresponding population parameter.

Topic 030: Introduction to Probability

A numerical description that indicates the likelihood of an event.
Probability for any specific outcome is defined as a fraction or a proportion of all the possible outcomes. e.g. If the possible outcomes are identified as 1, 2, 3, 4, 5, 6 then;

Probability of 4 = number of outcomes classified as 4, divided by total number of possible outcomes =1/6

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(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 the sample taken is right? or the 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)

Probability in Inferential Statistics

- For probability it is necessary that the outcomes be obtained without any bias in selection, which is achieved through random *sampling.*
- A **random sample** requires that each individual in the population has an equal chance of being selected.

Independent Random Sample

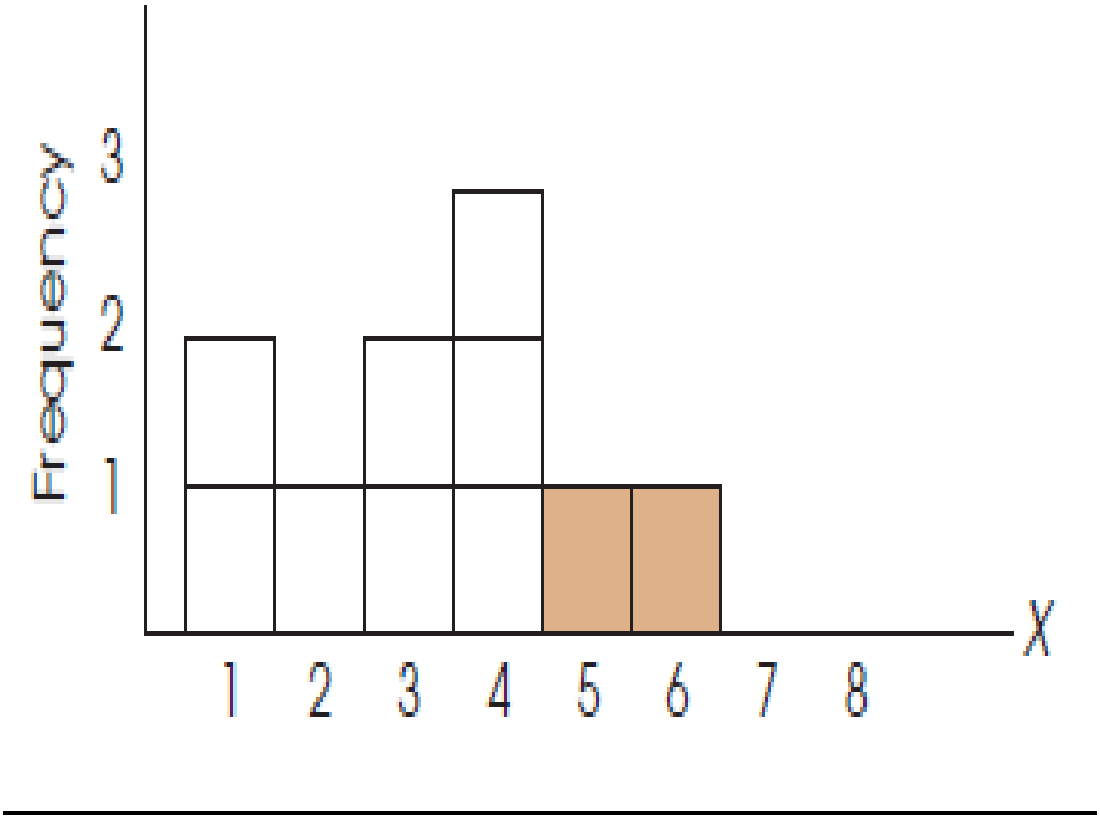
- Each individual has an equal chance of being selected in a sample, and probability remains constant from one selection to the next.
- Probabilities will change from one selection to the next if each individual is returned to the population before you make the next selection. This is called **random sampling with replacement** (e.g. cards)

Probability and Frequency Distribution

- Graphs or frequency distributions, also give you proportions.
- Proportions and probabilities are equal
- Suppose a population of sores is = 1, 1, 2, 3, 3, 4, 4, 4, 5, 6.(N=10)
- What is the probability of obtaining an individual with a score greater than 4?

$p(X>4)$=?

Suppose a population of sores is ‘
= 1, 1, 2, 3, 3, 4, 4, 4, 5, 6.(N=10)


Frequency
3
2
1
X
1
2
3
4
5
6
7
8

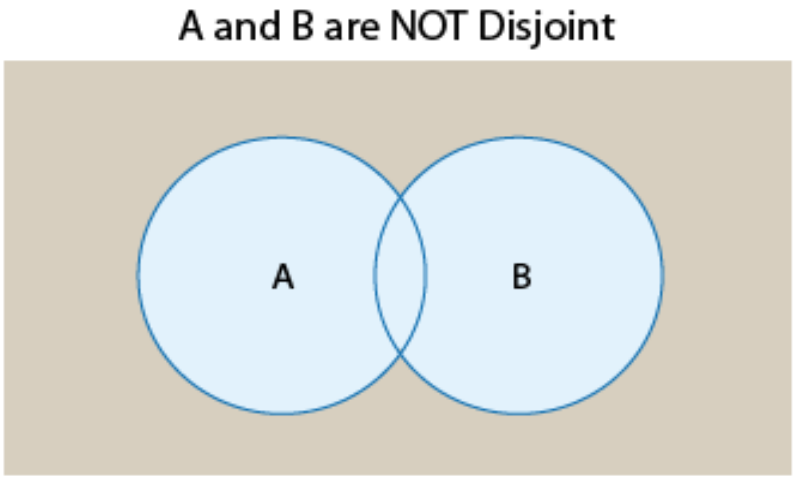
Lesson 14

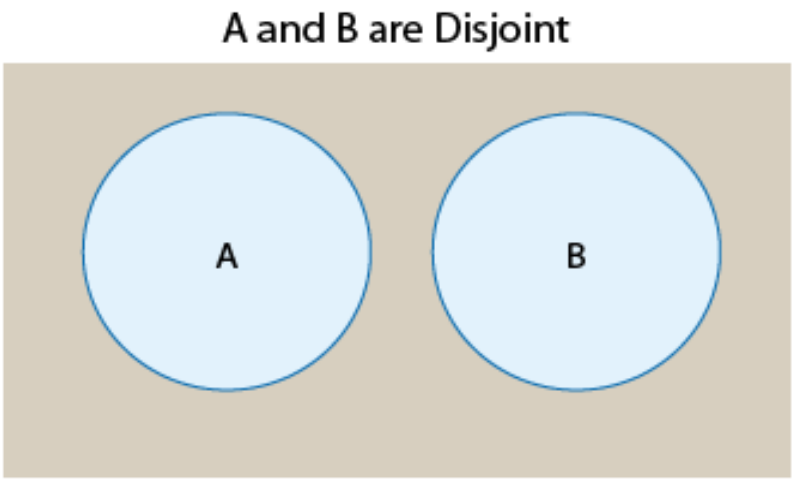
Inferential Statistic& Probability II
Topic: 031-032

Topic 031: Rules and Basic Laws of Probability

Basic Properties

- Every probability is between zero and one. If A is an event, then $0 \leq P(A) \geq 1$.
- The sum of the probabilities of all possible outcomes is 1.
- Impossible events have a probability of 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:

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

The Addition Rule

- If A and B are disjoint (Mutually Exclusive) events, then P(A or B) = P(A) + P(B).
- When dealing with probabilities, the word 'OR' will always be associated with the operation of addition; hence the name of this rule, 'The Addition Rule.'
- Example rolling 3 OR 6

The Addition Rule

If A and B are not disjoint (not mutually exclusive), then
$P(A \text{ or } B)=P(A)+P(B)-P(A \cap B)$
If Event A {a diamond card} OR
Event B {a ten card} then
$P(A) = 13/52$ and $P(B) = 4/52$.
The intersection event $A \cap B$ is the ten of diamonds hence
$P(A \cap B) = 1/52$. Therefore
$P(A \cup B) = 13/52 + 4/52 - 1/52$
$= 16/52$.

The Multiplication Rule

- probability of two separate events occurring is the product of each event
- If A and B are independent (not disjoint), then P(A and B)=P(A)P(B)
- You roll 2 dice
- Event A {6 on first die}
- Event B {6 on second die}
- P(A and B)=P(A)P(B)
- P=1/6 X 1/6 = 1/36

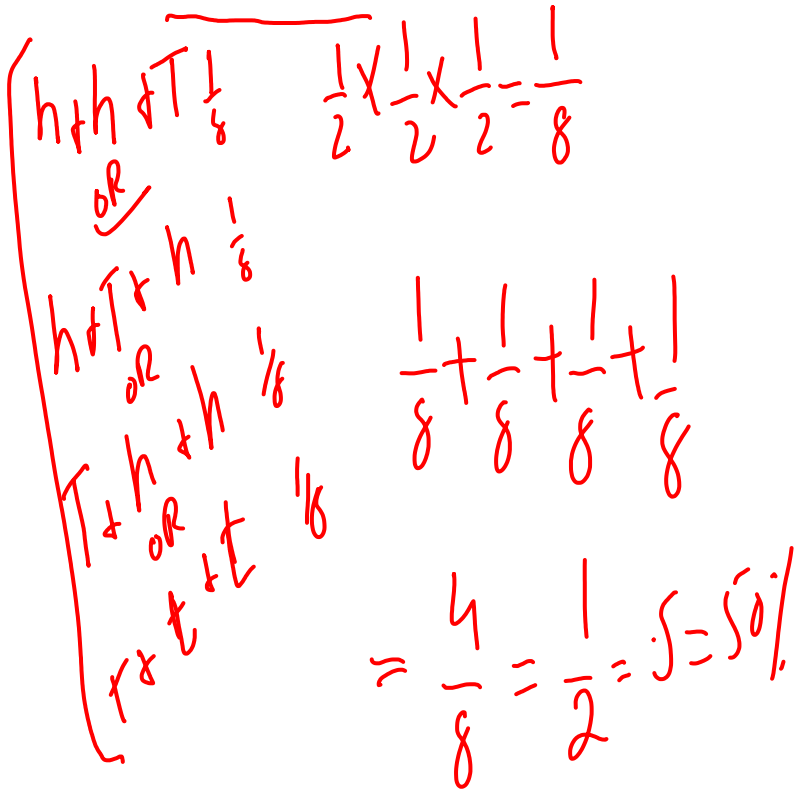
Conditional Probability /Dependent events / without replacement

- Probability of B event given that A has already occurred
- If A and B are dependent then $P(A \cap B)= P(B/A)P(A)$
- e.g. Bag contains 3 white and 3 red balls
- Event A ; Red ball
- Event B: white ball
- P(B/A) = 3/5

Topic 032: Understanding Probability

- According to Stanford University's Blood Center these are the probabilities of human blood types in the United State
- 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
- Sample Space You toss a coin 3 times,
  what is the probability of at least two heads

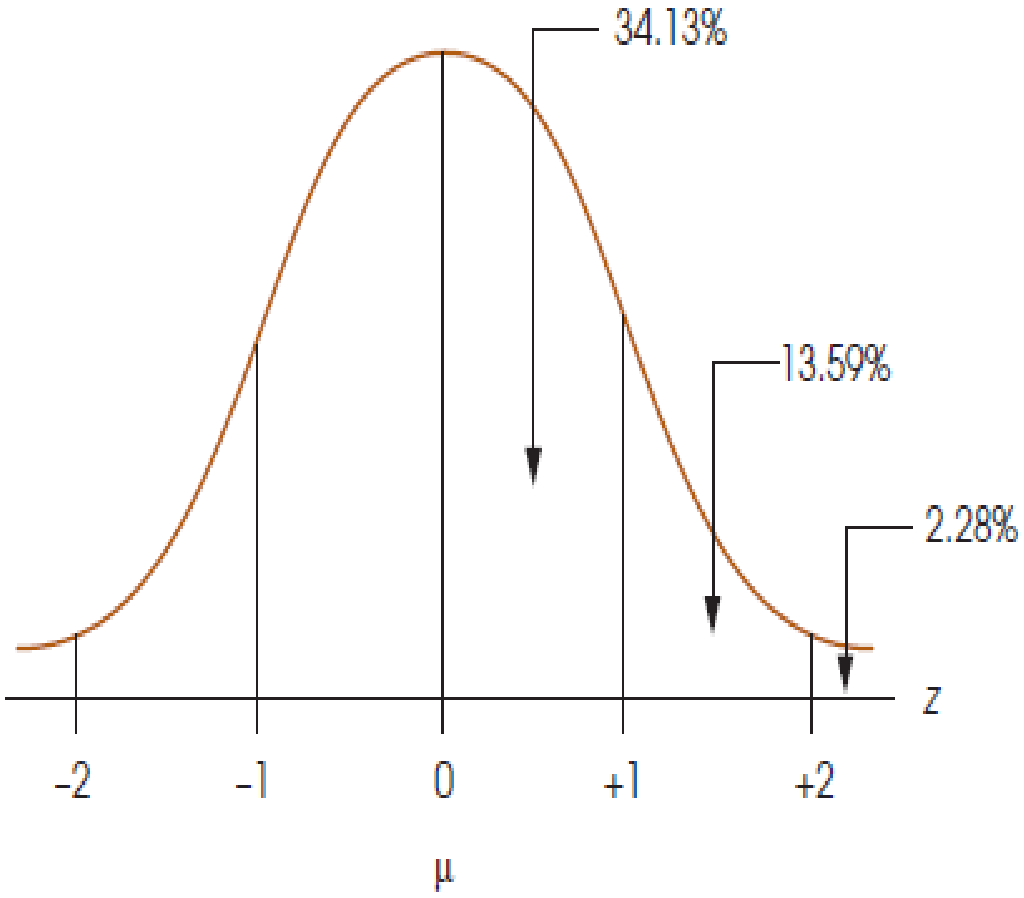


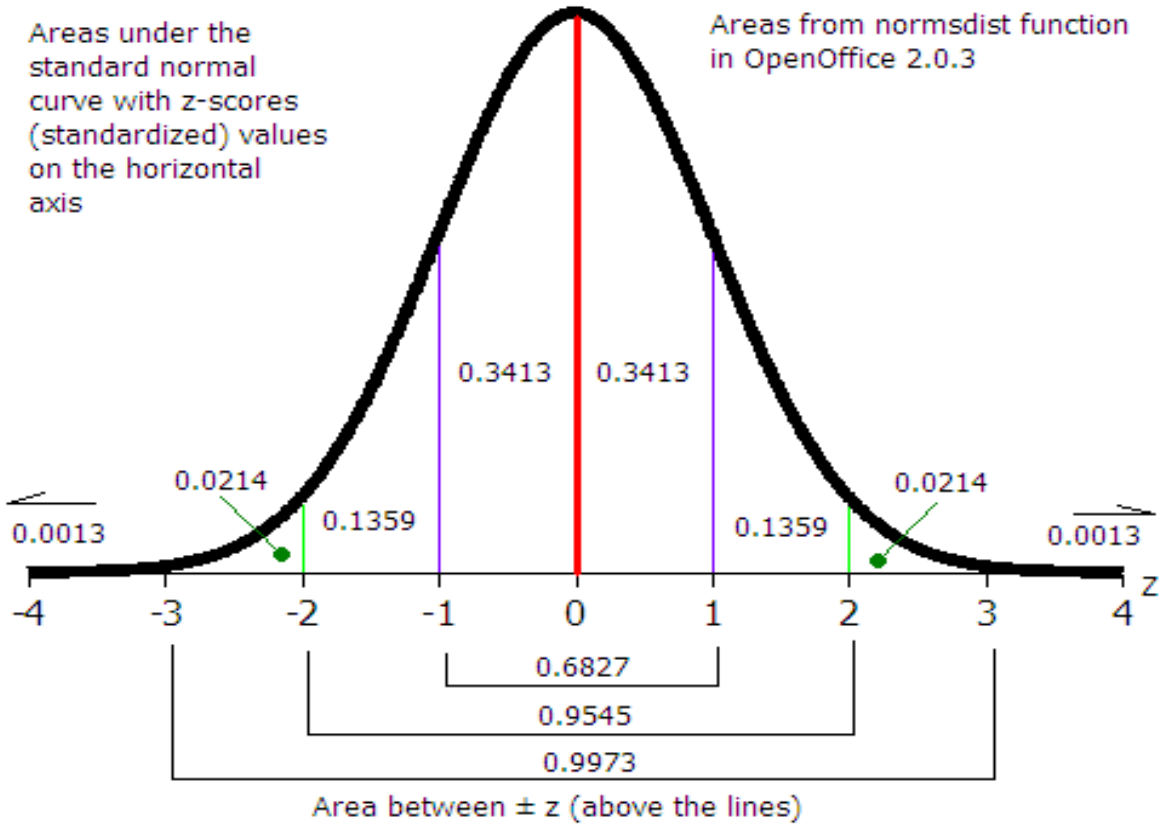
Lesson 15

Probability& Normal Distribution
Topic: 033-034

Topic 033: Probability and Normal Distribution

- **Normal Distribution** is symmetrical, with the highest frequency in the middle and frequencies tapering off as you move toward either extreme.
- The normal shape can also be described by the proportions of the area contained in each section of the distribution.
- Normal Distribution is symmetrical, with the highest frequency in the middle and frequencies tapering off as you move toward either extreme.
- The normal shape can also be described by the proportions of the area contained in each section of the distribution.





- Sections of a normal distribution are identified by using *z*-scores.
- $z = X-\mu/\sigma$
- X is sample score
- μ is population mean
- σ is population standard deviation

- When any distribution is transformed into *z*-scores, the mean becomes zero and the standard deviation becomes one.

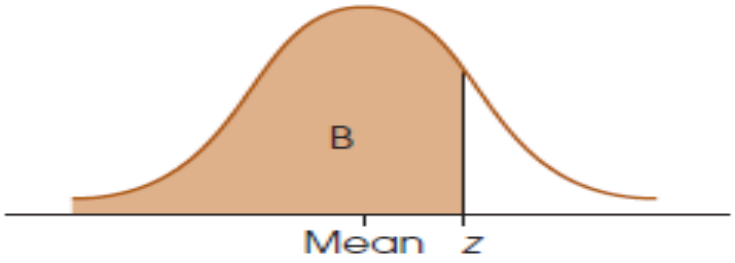
The Unit Normal Table

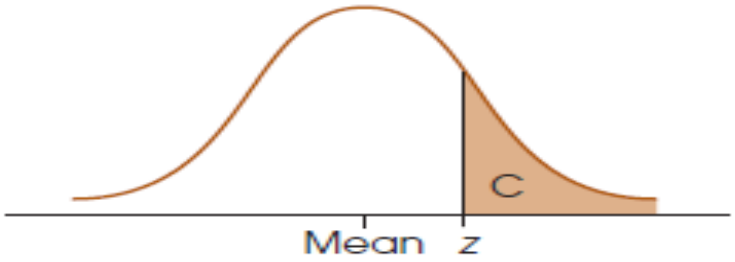
- A complete listing of *z*-scores and proportions is provided in the *unit normal table,* structured in a four-column format.
- This table lists proportions of the normal distribution for a full range of possible *z*-score values.

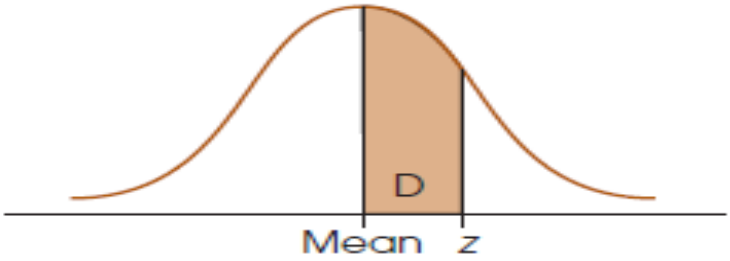
(A) z	(B) Proportion in body	(C) Proportion in tail	(D) Proportion between mean and z
0.00	.5000	.5000	.0000
0.01	.5040	.4960	.0040
0.02	.5080	.4920	.0080
0.03	.5120	.4880	.0120
0.21	.5832	.4168	.0832
0.22	.5871	.4129	.0871
0.23	.5910	.4090	.0910
0.24	.5948	.4052	.0948
0.25	.5987	.4013	.0987
0.26	.6026	.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

The Unit Normal Table

- Column A lists *z*-score values corresponding to different positions in a normal distribution.
- Column B presents the proportion in the body (the larger portion), and
- Column C presents the proportion in the tail.
- Column D, identifies the proportion of the distribution that is located between the mean and the *z*-score.

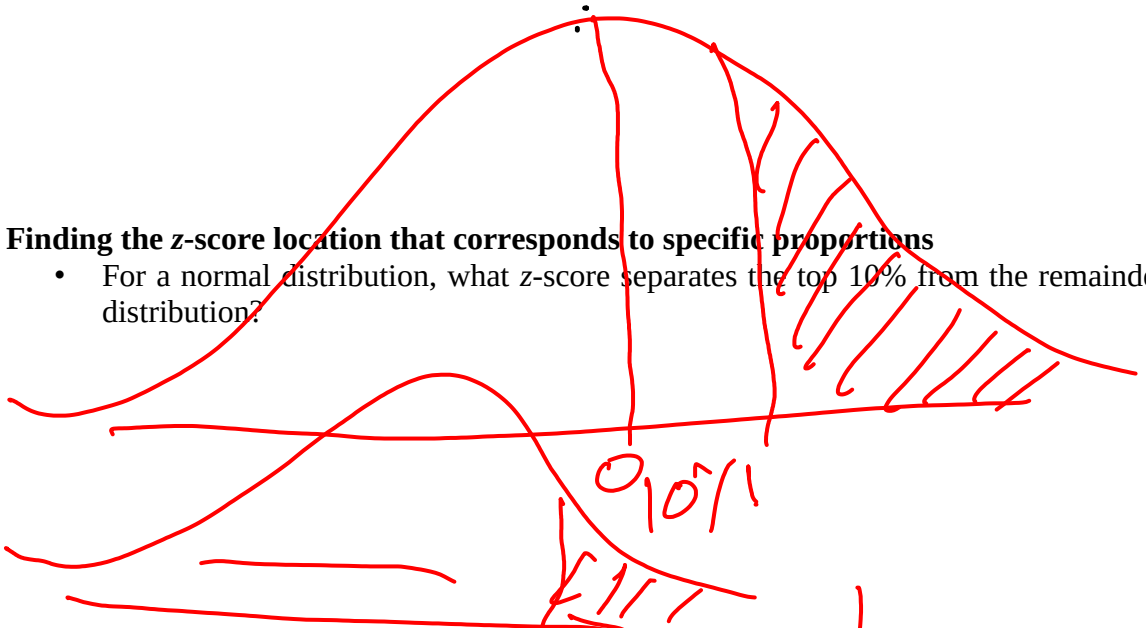






Finding proportions or probabilities for specific *z*-score values

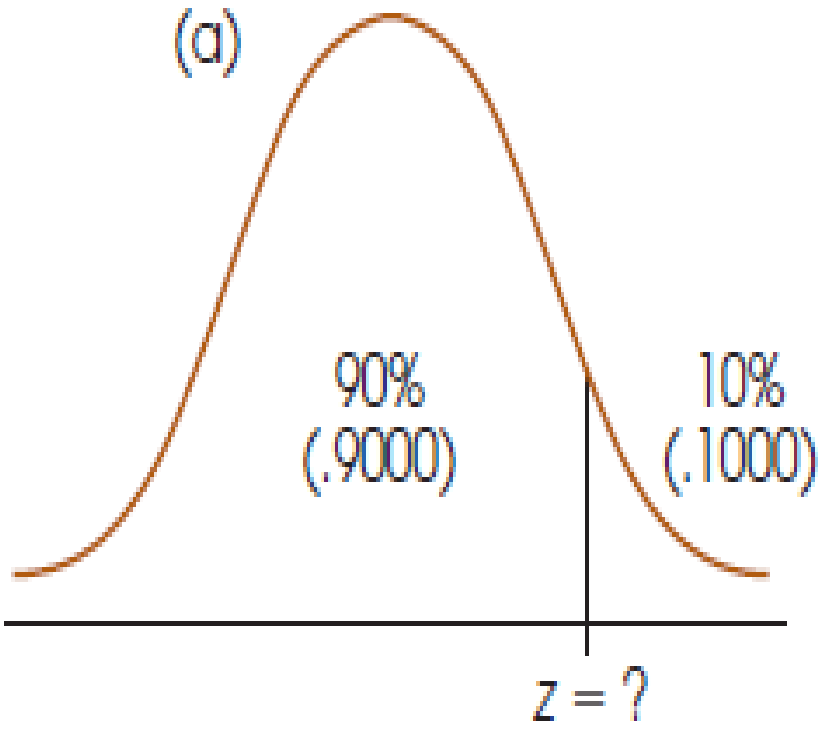
- What proportion of the normal distribution corresponds to *z*-score values greater than *z*=1.00?



Finding the *z*-score location that corresponds to specific proportions

- For a normal distribution, what *z*-score separates the top 10% from the remainder of the distribution?

For a normal distribution, what *z*-score separates the top 10% from the remainder of the distribution?

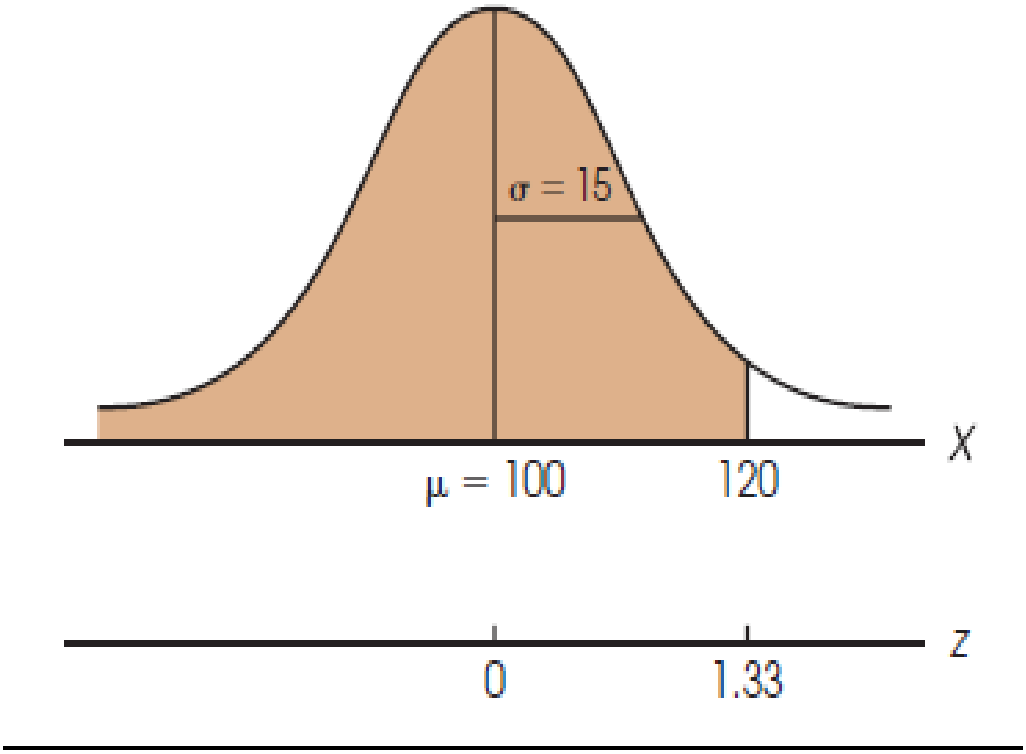


Probabilities for a Score from Normal Distribution

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.

Probabilities for a Score from Normal Distribution

- Probability of randomly selecting an individual with an IQ score less than 120?
- Transform the *X* values into *z*-scores.
- $Z=x-\mu/\sigma$
- Z=120-100/15=20/15=1.33
- Use the unit normal table to look up the proportions corresponding to the *z*-score values. = .9082. Probability of randomly selecting an individual with an IQ score less than 120?



Finding probabilities located between two scores

- The highway department conducted a study measuring driving speeds on a local section of the interstate highway. They found an average speed of μ=58 miles per hour with a standard deviation of σ=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 problem as p(55<*X*<65)=?

Finding scores corresponding to specific proportions or probabilities

We can also find the score (*X* value) corresponding to a specific proportion in the distribution. For this purpose, we begin with a specific proportion, use the unit normal table to look up the corresponding *z*-score, and then transform the *z*-score into an *X* value.

Finding scores corresponding to specific proportions or probabilities

The average reported commuting time is μ=24.3 minutes. Assuming that the distribution of commuting time is normal with a standard deviation of σ=10 minutes, how much time do you have (X=?) to spend commuting each day to be in the highest 10% (proportion=0.10)?

Area = .1003 → corresponding = 1.28

$z = X - \bar{X} / SD$

$1.28 = X - 24.3 / 10$

z=+1.28
Using formula: $z=X-\mu/\sigma$
1.28=X-24.3/10
X=1.28(10)+24.3
X=12.8+24.3=37.1

Topic 034: Understanding Probability

Probability

- For any specific outcome probability is defined as a fraction or a proportion of all the possible outcomes.
- Probability is used to predict what kind of samples are likely to be obtained from a population.

Probability
For probability, outcomes are obtained through random sampling.

Basic laws of Probability
Every probability is between zero and one. If A is an event, then $0 \leq P(A) \geq 1$.
The sum of the probabilities of all possible outcomes is 1.
Impossible events have probability zero.
Two events that cannot occur at the same time are called disjoint or mutually exclusive.

The Addition Rule

If A and B are disjoint (mutually exclusive) events, then P(A or B) = P(A) + P(B).

The Multiplication Rule

If A and B are independent, then P(A and B)=P(A)P(B)

Probability and Normal Distribution

- Normal distribution is symmetrical and the normal shape can be described by the proportions of area contained in each section of the distribution
- Sections of a normal distribution are identified by using z-scores.
- $z= X-\mu/\sigma$

Probability and Normal Distribution

We use *the unit normal table* (structured in a four-column format) to find out the relationships between *z*-score locations and proportions in a normal distribution. Scores on the Mathematics section of the SAT Reasoning Test form a normal distribution with a mean of μ 500 and a standard deviation of 100.

a. If the state college only accepts students who score in the top 60% on this test,

What is the minimum score needed for admission?

(*z* –0.25; *X* 475)

b. What is the minimum score necessary to be in the top 10% of the distribution?

1.28; *X* 628

What is the probability of selecting a score greater than 45 from a positively skewed distribution with μ 40 and 10? (Be careful.)

Lesson 16

Probability & Binomial Distribution
Topic: 035-036

Topic 035: Probability and Binomial Distribution

Binomial Distribution

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

Binomial Distribution

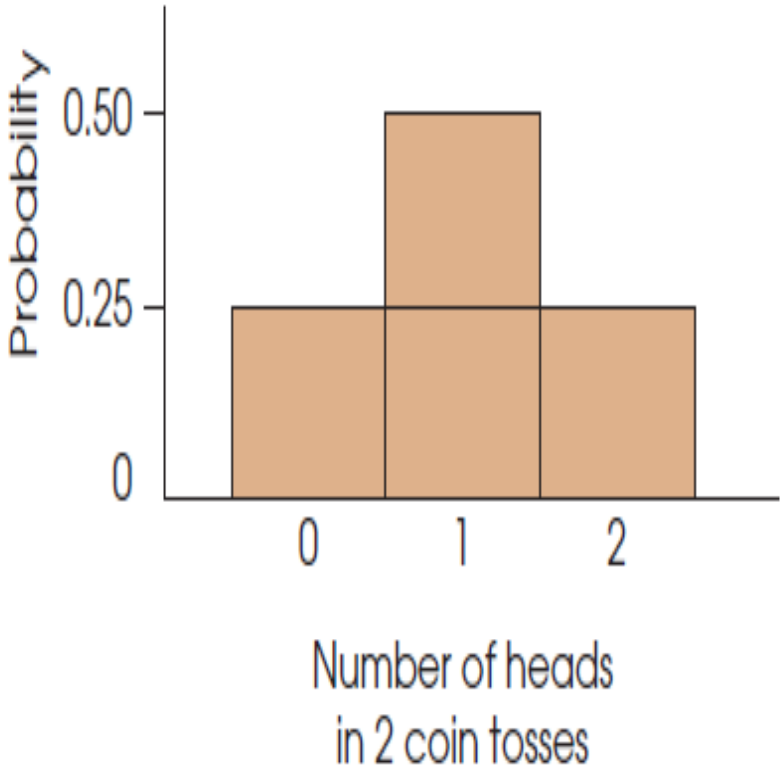
Binomial data can occur when a variable naturally exists with only two categories. For example, people can be classified as male or female, and a coin toss results in either heads or tails.

Binomial Distribution

In binomial distribution;
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 (success)*
- $q = p(B)$ = the probability of *B (failure)*
- Note that $p + q = 1.00$

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.
X can have any value from 0 to n;
0 means none of the samples in Category A, to 'n' sample in category A Number of Heads in 2 tosses of a fair coin



1st Toss	2nd Toss	
Heads	Heads	(Both heads)
Heads	Tails	(Each sequence has exactly 1 head)
Tails	Heads	
Tails	Tails	(No heads)

The binomial distribution can be used to answer probability questions.

E.g. what is the Probability of obtaining?
At least I head in 2 tosses?

Binomial Distribution

The binomial distribution tends toward a normal shape, especially when the sample size (*n*) is relatively large. Suppose you toss a coin 10 times?

The normal approximation to the binomial distribution

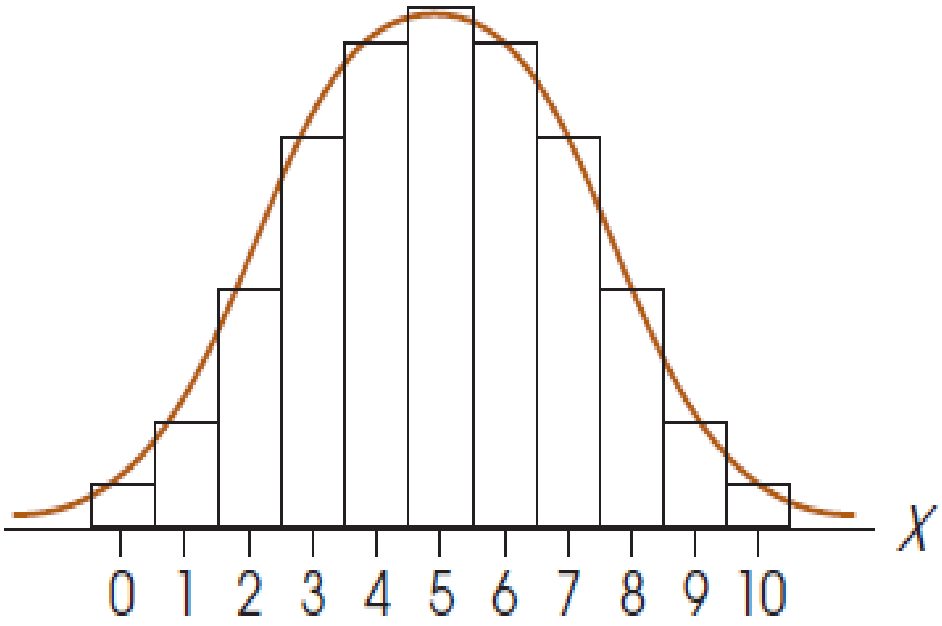
The binomial distribution is nearly perfect normal distribution when *pn* and *qn* are both equal to and greater than 10.Hence, the binomial distribution approximates a normal distribution, and however, approximation is not accurate for small samples

The normal approximation to the binomial distribution

- Mean: μ = *pn*
- Standard deviation: $\sigma = \sqrt{npq}$
- Within this normal distribution, each value of *X* has a corresponding *z*-score, $z = \frac{X-\mu}{\sigma} = \frac{\mathrm{X}-\mathrm{pn}}{\sqrt{npq}}$

The normal approximation to the binomial distribution

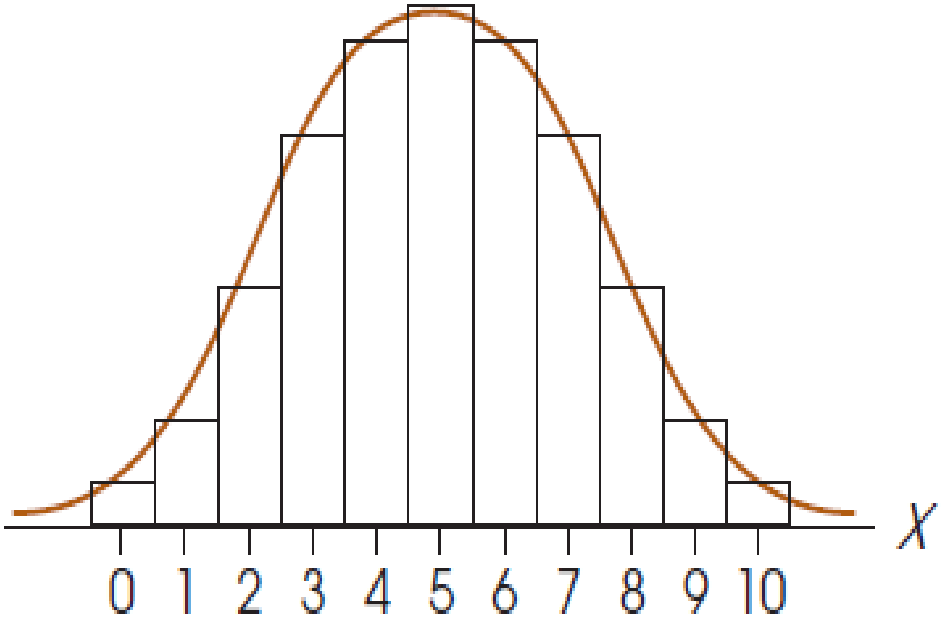
- Binomial values, such as the number of heads in a series of coin tosses, are *discrete*. The normal distribution is *continuous*.
- However, the *normal approximation* provides an extremely accurate model for computing binomial probabilities.



It is the area under the distribution that is used to find probabilities.

The normal approximation to the binomial distribution

- 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.



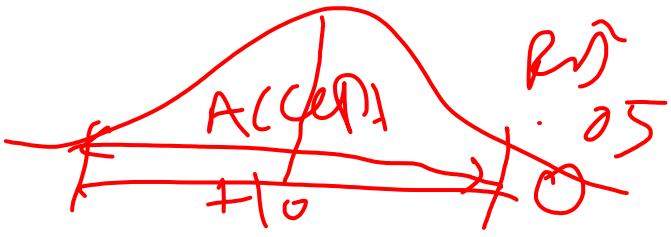
The normal approximation to the binomial distribution

- 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?

Topic 036: Using Binomial Distribution to Test Hypothesis

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

- If p > α then we do not reject the null hypothesis.
- If p< α 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

Introduction to Sampling Distribution& Hypothesis Testing
Topic: 037-038

Topic 037: Sampling Distribution

Estimating Population parameters from a sample statistic

Any statistics that are computed for the sample are not identical to the corresponding parameters for the entire population. For example, the average IQ for the sample of 25 students is not the same as the overall mean IQ for the entire population.

Sampling Error

The natural difference, discrepancy, or *error*, between sample statistics and the corresponding population parameters is called *sampling error*

Sampling Distribution

A sampling distribution is a probability distribution of a statistic obtained from a larger number of samples drawn from a specific population.

Estimating Population parameter from a sample statistic

- Think of a population and lets draw samples from it.

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

- 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: $p=1/100$

Sampling Distribution

- A **sampling distribution** is a distribution of statistics obtained by selecting all of the possible samples of a specific size from a population.
- The distribution of sample means is an example of a sampling distribution. In fact, it often is called the sampling distribution of *Means*.
- **General Characteristics of a Distribution**

- 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.

- 2, 4, 6, 8. is population

All possible samples of size 2

Sample	Scores		Sample Mean
	First	Second	(*M*)
1	2	2	2
2	2	4	3
3	2	6	4
4	2	8	5
5	4	2	3
6	4	4	4
7	4	6	5
8	4	8	6
9	6	2	4
10	6	4	5
11	6	6	6
12	6	8	7
13	8	2	5
14	8	4	6
15	8	6	7
16	8	8	8

we can use the distribution of sample means to answer probability questions about sample means. For example, if you take a sample of
n=2 scores from the original population, what is the probability of obtaining a sample mean greater than 7?
$p(M > 7)$=?



Topic 038: Hypothesis Testing

Logic of Hypothesis Testing

- Hypothesis testing is a statistical procedure that allows researchers to use sample data to draw inferences about the population of interest.
- **A hypothesis testing** is a statistical method that uses sample data to evaluate a hypothesis about a population.

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.

Logic of Hypothesis Testing

- 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

Lesson 18

Steps of Hypothesis Testing
Topic: 039-040

Topic 039: Basic Steps of Hypothesis Testing

Step 1: State the Hypothesis

- The process begins by stating a hypothesis about the unknown population.
- We state two opposing hypotheses; null and alternative hypothesis.

The Null Hypothesis

- It states that the treatment has no effect hence, the name is null.
- Identified by the symbol *H*0.

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 Alternative Hypothesis

- It states that the treatment has an effect on the dependent variable.
- Identified by (*H*1)
- In the context of an experiment, *H*1 predicts that the independent variable (treatment) *does have an effect* on the dependent variable.

The Alternative Hypothesis

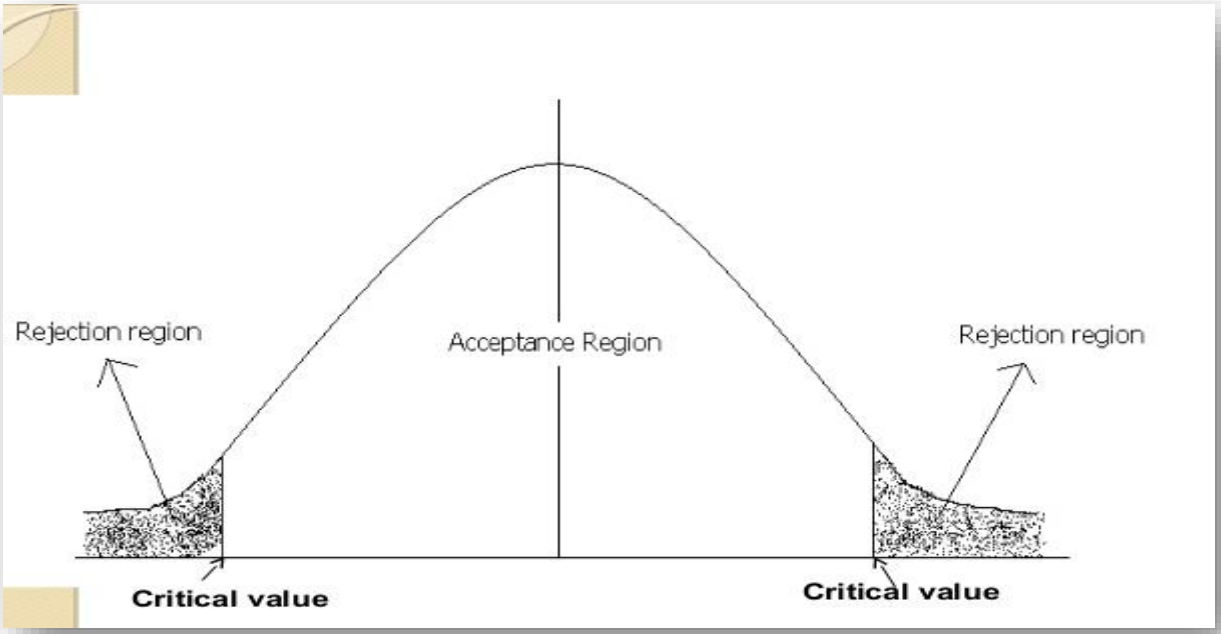
- The alternative hypothesis simply states that there will be some type of change. It does not specify whether the effect will be increased or decreased test scores.
- Sometimes, it is appropriate for the alternative hypothesis to specify the direction of the effect.

Step 2: Set the criteria for a decision

- To formalize the decision process, we use the null hypothesis to predict the kind of sample mean that ought to be obtained.

Step 2: Set the criteria for a decision

- To determine exactly which values are near the population mean and which values are very different, we examine all of the possible samples means that could be obtained if the null hypothesis is true.



Alpha Level

- To find the boundaries that separate the high-probability samples from the low-probability samples, we select a specific probability value, which is known as the *level of significance*, or the *alpha level*, for the hypothesis test.

Alpha Level

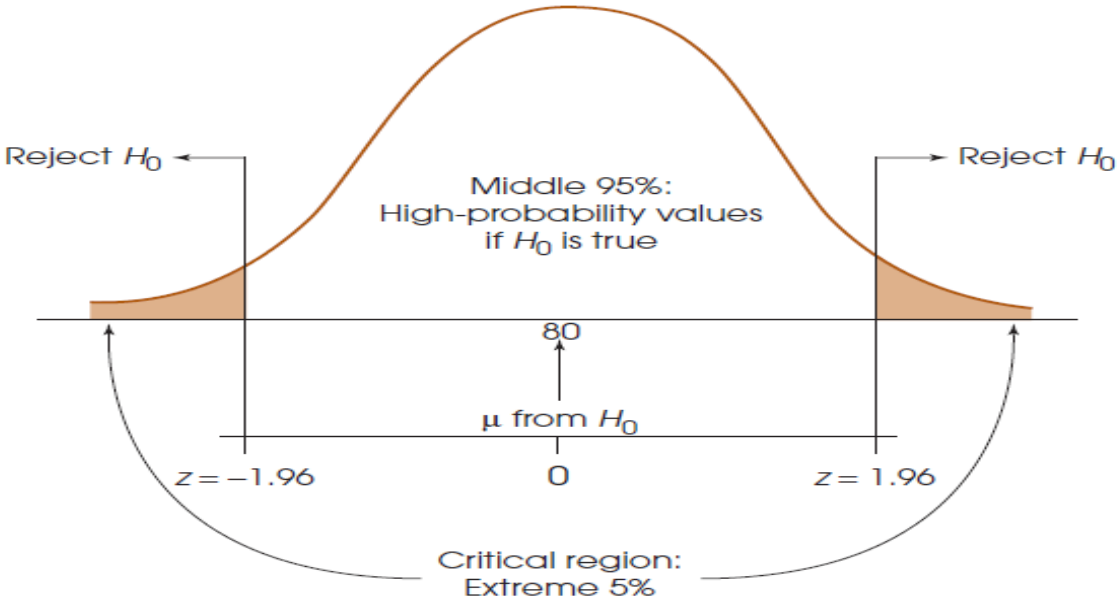
- Commonly used alpha levels are .05 (5%), .01 (1%), and .001 (0.1%).
- With $\alpha = .05$, we separate the most unlikely 5% of the sample means (the extreme values) from the most likely 95% of the sample means (the central values).

Critical Region

- It 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.

Boundaries for the Critical Region

- With $\alpha = .05$, the boundaries separate the extreme 5% from the middle 95%. Because the extreme 5% is split between two tails of the distribution, there is exactly 2.5% (or 0.0250) in each tail.
- In the unit normal table, a proportion of 0.0250 in column C (the tail) provides that the *z*-score boundary is $z = 1.96$.



Step 3: Collect data and compute sample statistics

- The raw data collected from the sample are summarized with the appropriate statistics (i.e., mean is computed)
- The most important step is to compare the data with the hypothesis.

Step 3: Collect data and compute sample statistics

- Comparison is accomplished by computing a *z*-score
- The *z*-score formula for a sample mean is
- $z = (M - \mu)/\sigma$M

Step 4: Make a decision

- While making a decision there are two possible outcomes:

1. 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*o and our decision is to *reject the null hypothesis*.

The following table gives critical values of Z for both one-tailed and two- tailed tests at various levels of significance.

level of significance	0.10	.05	.01
critical value of z for one-tailed test	1.28 or -1.28	1.645 or -1.645	2.33 0r -2.33
critical value of z for two-tailed test	1.645 or -1.645	1.96 or -1.96	2.58 or -2.58

Step 4: Make a decision

While making a decision there are two possible outcomes:

If the sample data are not in the critical region then, the sample mean is reasonably close to the population mean (in the center of the distribution). The data do not provide strong evidence that the null hypothesis is wrong, our conclusion is to *fail to reject the null hypothesis*.

Topic 040: One-Tailed and Two-Tailed Testing

One-tailed Test

In a **directional hypothesis test,** or a **one-tailed test,** the statistical hypotheses (*H*o 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.

Hypothesis for One-tailed Test

For example, a special training program is expected to *increase* student performance.

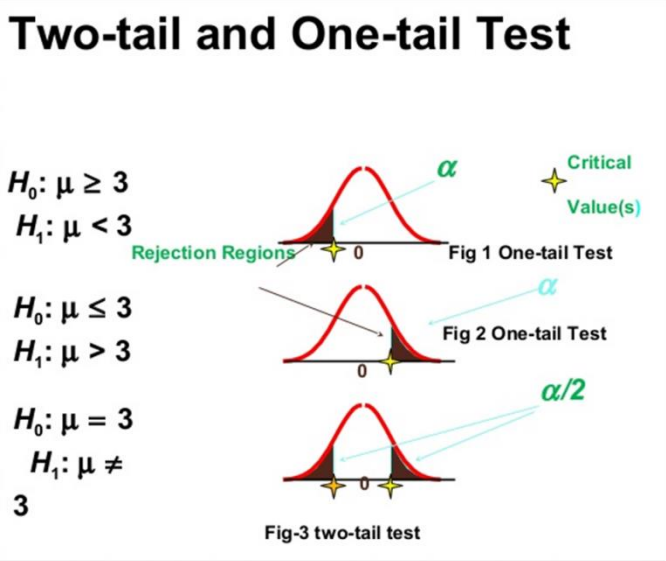
- *H*o: performance is not increased (The training does not work); $M \leq \mu$
- *H*1: performance is increased(The training works as predicted), $M > \mu$

Critical Region for One-tailed Test

The critical region is contained in one tail of the distribution in one-tailed test. Hence, a directional test is commonly called a *one-tailed* test.

Two-tailed Test

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.



Two-tailed Test

- *H*o: there is no training effect on performance; M $\leq \mu$
- *H*1: there is training effect on performance M $\neq \mu$

Lesson 19

Assumptions and Errors in Hypothesis Tests
Topic: 041-042

Topic 041: Assumptions for Hypothesis Tests

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

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.

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.

Normal sampling distribution

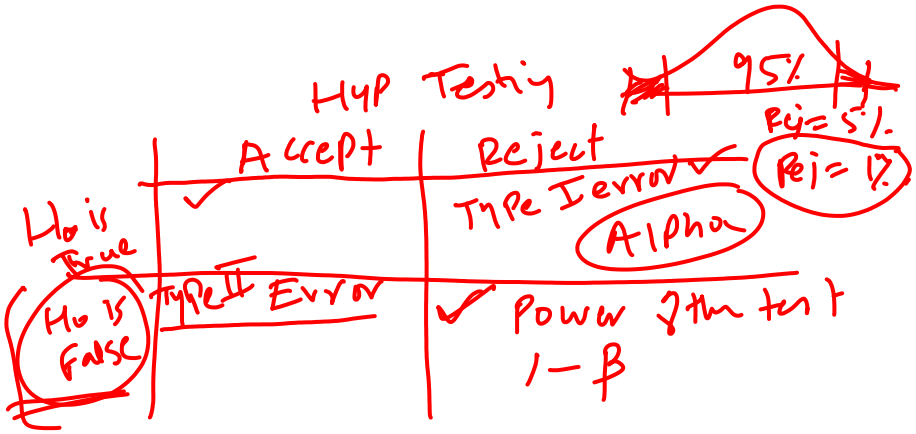
- For hypotheses with *z*-scores, the unit normal table is used to identify the critical region. This table can be used only if the distribution of sample means is normal.

Topic 042: Errors in Hypothesis Testing

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 treatment does have an effect when, in fact, it has no effect.

Type I and II Error



Type I Error

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 a variety of factors rather than a specific number.

Selecting an alpha level

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 prefer to use an alpha level such as .01 or .001 to reduce the risk.

Lesson 20

Factors Influence a Hypothesis Test
Topic: 043-044

Topic 043: An Example of Hypothesis Test

A researcher would like to investigate the effect of prenatal alcohol exposure on birth weight. 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.

State hypothesis

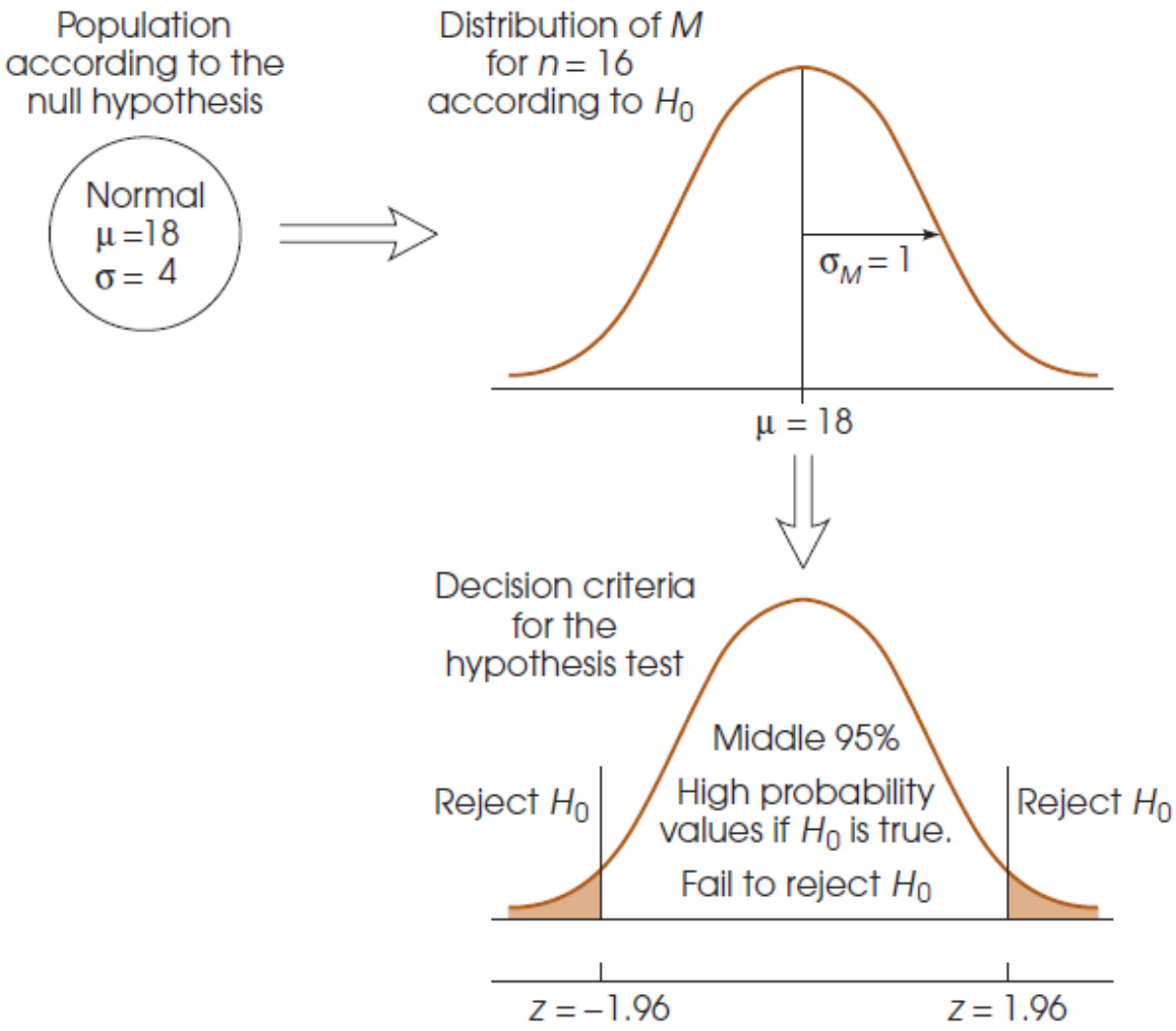
- The null hypothesis states that exposure to alcohol has no effect on birth weight
- *Ho*: μalcohol exposure =18
- The alternative hypothesis states that alcohol exposure does affect birth weight
- *H1* : μalcohol exposure ≠18

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.

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



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

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{15 - 18}{4/\sqrt{16}} = \frac{-3}{1}$$

$$z_{cal} = -3$$



Make a decision

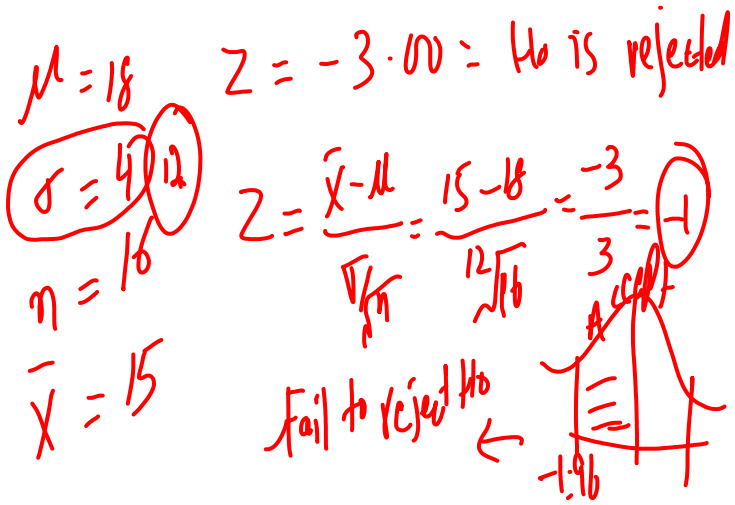
- The *z*-score computed has a value of 3.00, which is beyond the boundary of -1.96.

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

Topic 044: Factors That Influence a Hypothesis Test

1. The variability of the scores

- In a hypothesis test, higher variability can reduce the chances of finding a significant treatment effect.



2. The number of scores in the sample

- In general, increasing the number of scores can increase the chances of finding a significant treatment effect.

Lesson 21

Role of Sample Size in Statistical Significance
Topic: 045-046

Topic 045: The Role of Sample Size in Statistical Significance

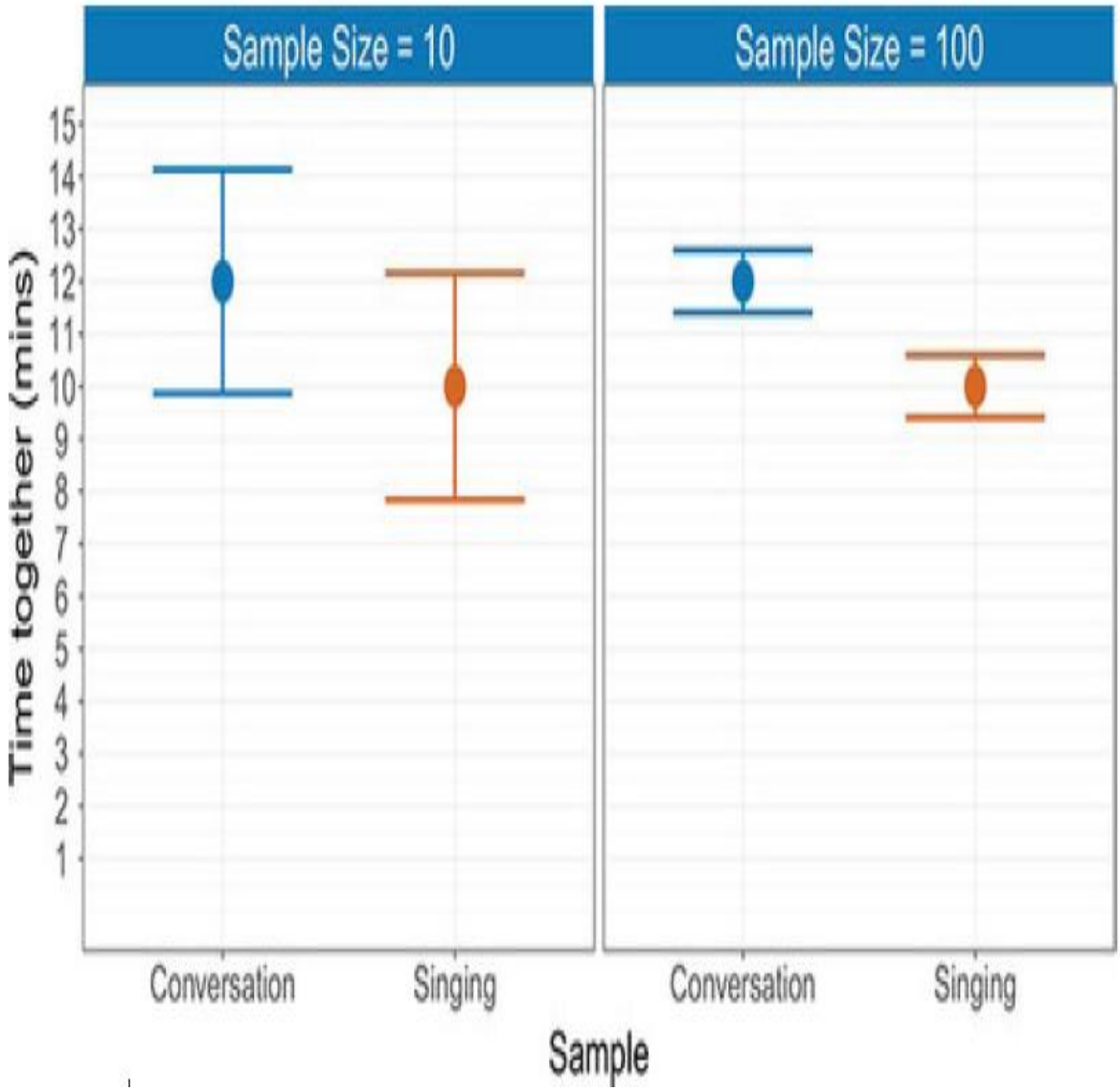
Statistical Significance

If 95% confidence intervals do not overlap then we can conclude that the means come from different populations, and, therefore, that they are significantly different.

Sample Size

- Whether the results of a study are statistically significant can depend on the sample size.
- A larger sample size tends to yield a statistically significant results
- On the other hand, if the sample size is too small, an important relationship or difference can go undetected.
- In that case, the **power** of the test is too low.
- Given that power is the ability of a test to find an effect that genuinely exists, and the effect is found by having a statistically significant result (i.e., $p < 0.05$),
- Hence, there is also a connection between the sample size and the *p*-value associated with a test statistic.
- A study got two groups of 10 heterosexual young men and got them to go up to a woman that they found attractive and either engage them in conversation (group 1) or sing them a song (group 2). Researchers measured how long it was before the woman ran away. Imagine the experiment was also repeated using 100 men in each group.
- Results showed that in both cases the singing group had a mean of 10 and a standard deviation of 3, and the conversation group had a mean of 12 and a standard deviation of 3.
- The only difference between the two experiments is that one collected 10 scores per sample and the other 100 scores per sample.

Sample Size The confidence intervals become much narrower when the samples contain 100 scores than when they contain only 10 scores.



- The sample size affects whether a difference between samples is deemed significant or not.
- In large samples, small differences can be significant, and in small samples, large differences can be non-significant.
- Even a difference of practically zero can be deemed 'significant' if the sample size is big enough.

Topic 046: Calculating Sample Size

Calculating Sample Size

- The sample size necessary to achieve a given level of power can be calculated by setting the value of α, $1 - \beta$ and effect size.

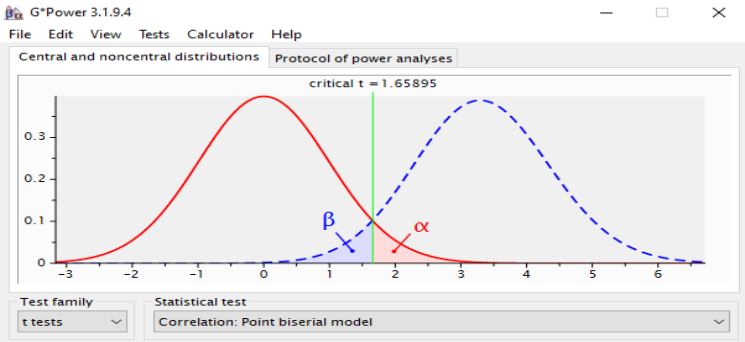
- Normally we have alpha value 0.05, power 0.8 and moderate effect size to determine the sample size.
- Using power to calculate the necessary sample size is the more common and more useful thing to do.
- The actual computations are very complicated. Therefore, we usually prefer to use a software for this purpose.

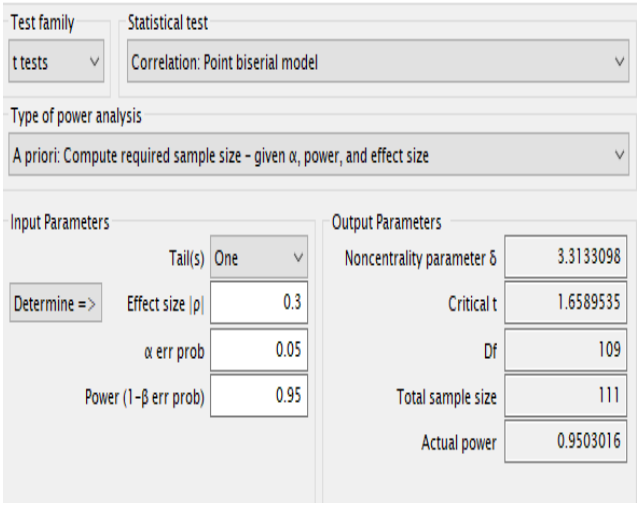
G*Power

- *G*Power* is a free-to-use software or a general power analysis programme, which is meat for determining the sample size and analyzing power in research studies.
- The program offers the ability to calculate power for a wide variety of statistical tests including t-tests, F-tests and chi-square tests among others.

To do power analysis to estimate the sample size, we need;

- Hypothesis
- Number of independent variables or number of groups
- What statistical test (one of the inferential statistics) we will be using
- Alpha level (usually .05)
- Power of the test (.80), and
- Effect size (small, moderate or large based on the test being used)




Test family
t tests
Statistical test
Correlation: Point biserial model
Type of power analysis
A priori: Compute required sample size – given α, power, and effect size
Input Parameters
Tail(s)
One
Determine =>
Effect size |ρ|
0.3
α err prob
0.05
Power (1–β err prob)
0.95
Output Parameters
Noncentrality parameter δ
3.3133098
Critical t
1.6589535
Df
109
Total sample size
111
Actual power
0.9503016

Lesson 22

Hypothesis Testing Review
Topic: 047-048

Topic 047: Hypothesis Testing Review

Hypothesis Testing
Hypothesis testing is a statistical procedure that allows researchers to use sample data to draw inferences about the population of interest.

Hypothesis testing includes the following steps;

1. **stating a hypothesis**
- Either null or alternative
- One says that an effect does not exist (**the *null hypothesis, Ho***) and the other says that an effect exists (**the *alternative hypothesis, H1***).

A researcher begins with a known population, with μ=65 and σ=15. The researcher suspects that special training in reading skills will produce a change in the scores for the individuals in the population. A sample of n=25 individuals is selected, and the treatment is given to this sample. Following treatment, the average score for this sample is M=70.

2. **set the criteria for a decision determined by;**

Alpha level (level of significance that separates the high-probability samples from the low-probability samples) and
Critical region (extreme sample values that are very unlikely to be obtained if the null hypothesis is true)

3. **data is collected data and sample statistics are computed**
- By comparing the data with the hypothesis
- Comparison is accomplished by computing a *z*-score
- $z = (M - \mu)/\sigma$M
-
4. **a decision is made**

A sample value in the critical region is very unlikely to occur if the null hypothesis is true. Therefore, the decision is to ***reject the null hypothesis*****.** If the sample data are not in the critical region then, the sample mean is reasonably close to the population mean (in the center of the distribution) conclusion is to ***fail to reject the null hypothesis*****.**

Assumptions for hypothesis testing include;

- Random selection of participants
- The values in the sample must consist of independent observations
- The value of population standard deviation (σ) is unchanged by the treatment

- Sample should be normally distributed
- In hypothesis testing 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 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.

Hypothesis test is influenced by;

- **The variability of the scores (standard deviation).** Higher variability can reduce the chances of finding a significant treatment effect.
- **The number of score in the sample.** Increasing the sample size can increase the chances of finding a significant treatment effect.

The statistical power of a test is defined as the probability that the test will reject the null hypothesis if the treatment really has an effect. In hypothesis testing, the first outcome, failing to reject *H*o when there is a real effect is a Type II error with a probability identified as $p = \beta$.
The second outcome have a probability of $1 - \beta$. Hence, rejecting *H*o when there is a real effect this and is the power of the test.

Power of the test is affected by;

- **Sample size:** 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.
- **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.

Topic 048: Hypothesis Testing Through T-Test

Goals of Inferential Statistics?

- Estimation and Testing Hypothesis

Use a sample mean (*M*) to test hypotheses about an unknown population mean (μ).

- We form Null and Alternative Hypothesi
- We use statistical Tests
- Z test
- T test
- F test

The Z Statistic

To make inferences about population we use z score test statistic.

- $z = \frac{M-\mu}{standard\ error\ (\sigma)}$

Where standard error =

$\sigma M = \sigma/\sqrt{n}$

- Z scores form a normal distribution, and we use table to find critical region for the hypothesis testing

The problem with *z*-scores

A *z*-score requires the value of the population standard deviation (or variance), which is needed to compute the standard error. In most situations, the standard deviation for the population is not known.

The t Statistic

When the variability for the population is not known, we use the **sample variability** in its place.

Sample variance as *an unbiased estimate* of the corresponding population variance is

- $s^2 = \sum(X-M)^2/n-1$

$= SS/n-1$

$=SS/df$

Sample standard deviation is then,

- $s=\sqrt{\frac{SS}{n-1}}=\sqrt{\frac{SS}{df}}$

The **estimated standard error** (*s**M***) is used as an estimate of the real standard error (σM) when the value of σ is unknown.

Estimated standard error= $SM = s/\sqrt{n} = \sqrt{\frac{s^2}{n}}$

By substituting the estimated standard error in the denominator of the *z*-score formula, the result is a new test statistic called a ***t statistic***

$t= \frac{(M-\mu)}{SM}$

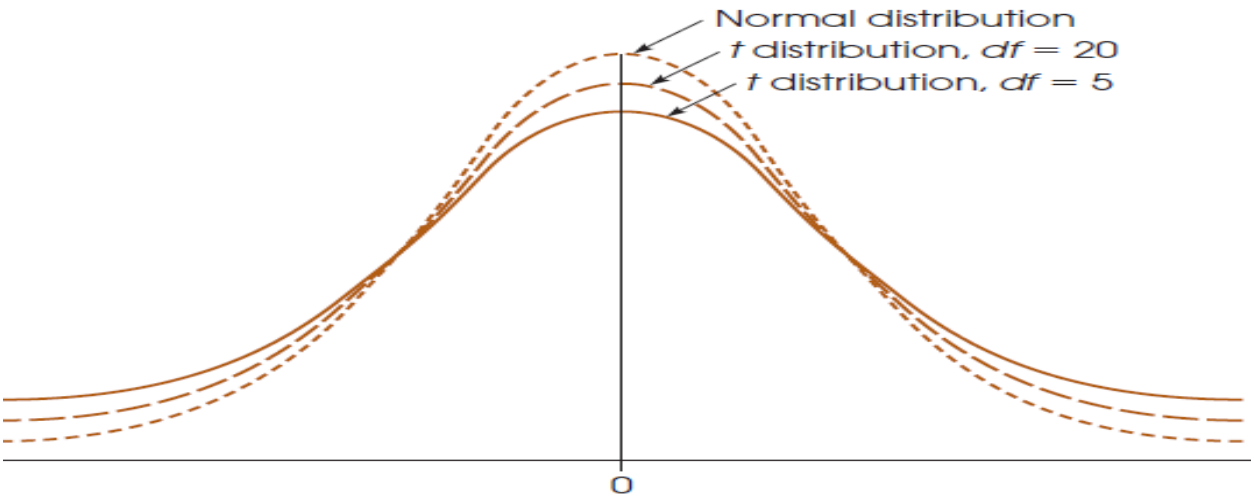
The ***t* statistic** is used to test hypotheses about an unknown population mean (μ) when the value of σ (sigma) is unknown.

Degrees of freedom and t Statistic

Degrees of freedom describe the number of scores in a sample that are independent and free to vary. Because the sample mean places a restriction on the value of one score in the sample, there are $n - 1$ degrees of freedom for a sample with n scores

The t Distribution

It is the complete set of *t* values computed for every possible random sample for a specific sample size (*n*) or a specific degrees of freedom (*df*).The *t* distribution approximates the shape of a normal distribution as n increases.



The *t* distribution has more variability than a normal *z* distribution, especially when *df* values are small. The *t* distribution tends to be flatter and more spread out, whereas the normal *z* distribution has more of a central peak.

Proportions and probabilities for t distributions

As the unit normal table is used to locate proportions associated with *z*-scores, a *t* distribution table is used to find proportions for *t* statistics.

	Proportion in One Tail					
	0.25	0.10	0.05	0.025	0.01	0.005
	Proportion in Two Tails Combined					
df	0.50	0.20	0.10	0.05	0.02	0.01
1	1.000	3.078	6.314	12.706	31.821	63.657
2	0.816	1.886	2.920	4.303	6.965	9.925
3	0.765	1.638	2.353	3.182	4.541	5.841
4	0.741	1.533	2.132	2.776	3.747	4.604
5	0.727	1.476	2.015	2.571	3.365	4.032
6	0.718	1.440	1.943	2.447	3.143	3.707

Lesson 23

Null and Alternative Hypothesis

Topic: 049-050

Topic 049: Constructing Null and Alternative Hypothesis

Hypothesis Test:

The t statistic permits hypothesis testing in situations for which a population mean is not known
The t-test does not require any prior knowledge about the population mean or the population variance.

Hypothesis Test:
The null hypothesis would state
Ho: μ=4
The alternative hypothesis would state that
H1: μ≠4

Assumptions of the t test

- The values in the sample must consist of independent observations.
- The population that is sampled must be normally distributed.

The influence of sample size and sample variance

The number of scores in the sample and the magnitude of the sample variance both have a large effect on the t statistic and thereby influence the statistical decision.
t= $((M-\mu))/SM$ where SM= $\sqrt{(s\text{^}2/n)}$
And where s=$\sqrt{(SS/(n-1))}$= $\sqrt{(SS/df)}$

The influence of sample size and sample variance

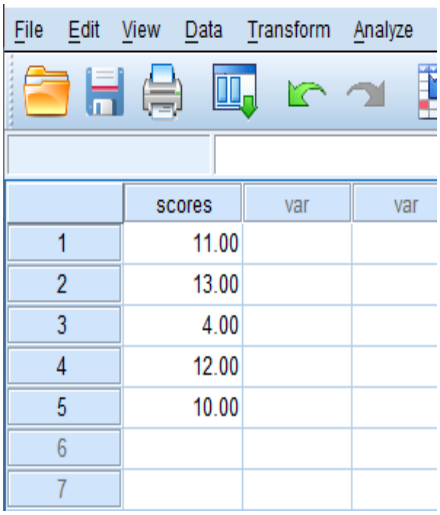
As the estimated standard error appears in the denominator of the formula, a larger value for sM produces a smaller value (closer to zero) for t.
The estimated standard error is directly related to the sample variance so that the larger the variance, the larger the error.

The influence of sample size and sample variance
The estimated standard error is inversely related to the number of scores in the sample. The larger the sample is, the smaller the error is.
If all other factors are held constant, large samples tend to produce bigger t statistics and therefore are more likely to produce significant results.

Topic 050: Doing one Sample T-Test Using SPSS

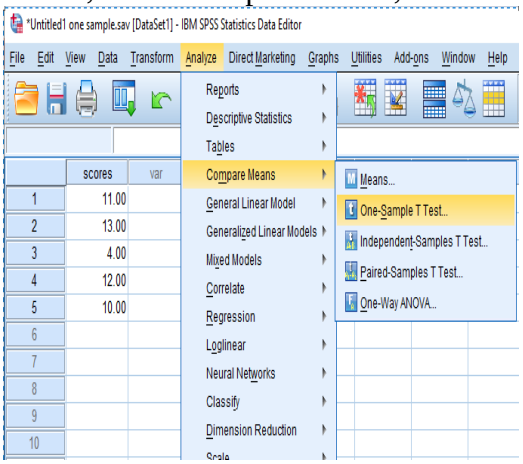
Data Entry

A random sample of n=5 participants with scores on a test (11,13,4,12,10) is obtained from a population with a mean of μ=13. A treatment is administered to the individuals in the sample. A researcher wants to examine if the sample means differ from population mean after treatment. Enter all of the scores from the sample in one column of the data editor.



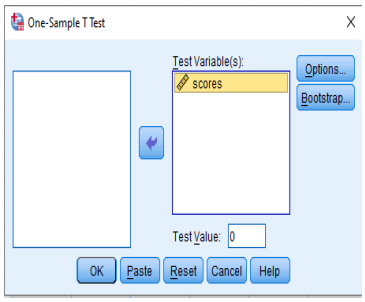
Data Analysis

1. Click Analyze on the tool bar, select Compare Means, and click on One-Sample T Test.



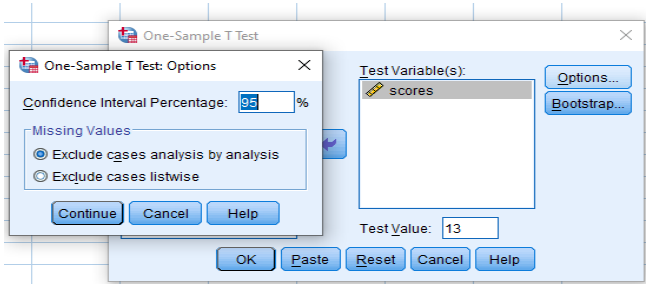
Data Analysis

2. Highlight the column label for the set of scores in the left box and click the arrow to move it into the Test Variable(s) box.
3. In the Test Value box at the bottom of the One-Sample t Test window, enter the hypothesized value for the population mean from the null hypothesis (μ=13)



Data Analysis

4. In addition to performing the hypothesis test, the program computes a confidence interval for the population mean difference. The confidence level is automatically set at 95%, but you can select Options and change the percentage.
5. Click OK.



SPSS Output

T-Test

```
[DataSet1] F:\spss\Untitled1 one sample.sav
```

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
scores	5	10.0000	3.53553	1.58114

One-Sample Test

	Test Value = 13					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
scores	-1.897	4	.131	-3.00000	-7.3899	1.3899

Lesson 24

Independent Sample T-Test& Testing Hypothesis

Topic: 051-052

Topic 051: Independent Sample T-Test

The independent-measures design

Although single sample techniques are used occasionally in real research, most research studies require the comparison of two (or more) sets of data.
For example, a social psychologist may want to compare men and women in terms of their political attitudes. The two sets of data could come from two completely separate groups of participants. This is called an independent measures research design or between-subject design.

The repeated-measures design

The two sets of data could come from the same group of participants. This is referred to as repeated-measures research design or within-subjects design.

Assumptions of independent-measures t-test

1. The observations within each sample must be independent.
2. The two populations from which the samples are selected must be normal.
3. The two populations from which the samples are selected must have equal variances also referred to as homogeneity of variance.

Topic 052: Testing Null and Alternate Hypothesis

The t statistic for an independent-sample test. The number of scores in the first sample would be identified by n_1; for the second sample, the number of scores is n_2. The sample means would be identified by M_1 and M_2. The sums of squares would be SS_1 and SS_2.

The hypothesis for an independent sample t-test

The mean for the first population is μ_1 , and the second population mean is μ_2. The difference between means is simply $\mu_1 - \mu_2$.
The null hypothesis states that there is no change, no effect, or, in this case, no difference.

- H_o: $\mu_1 - \mu_2 = 0$

The alternative hypothesis states that there is a mean difference between the two populations

- H_1: $\mu_1 - \mu_2 \neq 0$

Lesson 25

Independent Sample T-Test-I

Topic: 053-054

Topic 053: Formula and Calculation of Independent Sample T-Test Part 1,2 and 3

The formulas for an independent-measures hypothesis test .
The basic structure of the t-statistic is the same for both the independent-measures and the single-sample hypothesis tests.

$$t = \frac{sample\ statistic\ - hypothsized\ population\ mean}{estimated\ standard\ error}$$

The overall t formula

$$t = \frac{\text{sample mean difference} - \text{population mean differnce}}{\text{estimated standard error}}$$

$$t = \frac{[(M_1 - M_2) - (\mu_1 - \mu_2)]}{S(^{M}{}_{1} - {}^{M}{}_{2})}$$

The estimated standard error

For the independent-measures t formula, the standard error measures the amount of error that is expected when you use a sample mean difference ($M_1 - M_2$) to represent a population mean difference ($\mu_1 - \mu_2$). The standard error for the sample mean difference is represented by the symbol $S(M_1 - M_2)$.

Calculating the estimated standard error

1. Each of the two sample means represents it own population mean, but in each case there is some error.
   - M_1 approximates $\mu_1$with some error.
   - M_2 approximates μ_2 with some error.

Calculating the estimated standard error

For the independent-measures t statistic, we want to know the total amount of error involved in using two sample means to approximate two population means.

$$S(M_1 - M_2) = \sqrt{(s_1{}^2/n_1) + (s_2{}^2/n_2)}$$

Pooled Variance

For correcting the bias in the standard error is to combine the two sample variances into a single value called the *pooled variance*. The pooled variance is obtained by averaging or "pooling" the two sample variances using a procedure that allows the bigger sample to carry more weight in determining the final value.

For the independent-measures t statistic, there are two SS values and two df values. The values from the two samples are combined to compute what is called the pooled variance.

The pooled variance is identified by the symbol S^2p

$S^2p=(SS_1+SS_2)/df_1+df_2$

Equal samples sizes

The first sample has n=6 scores with SS=50, and the second sample has n=6 scores with SS=30. Individually, the two sample variances are

This time the first sample has n=3 scores with SS=20, and the second sample has n=9 scores with SS=48.

Estimated standard error

- Using the pooled variance in place of the individual sample variances, an unbiased measure of the standard error for a sample mean difference can be obtained as,
- $S(M_1 - M_2) = \sqrt{(Sp^2/n_1)+ (Sp^2/n_2)}$

Confidence intervals for estimating $\mu_1 - \mu_2$

- For the independent-measures *t*, we use a sample mean difference, $M_1 - M_2$, to estimate the population mean difference, $\mu_1 - \mu_2$.
- $\mu_1 - \mu_2 = M_1 - M_2 \pm t \times S(M_1 - M_2)$
- In a study of jury behavior, two samples of participants (8 each) were provided details about a trial in which the defendant was obviously guilty. Although group 2 received the same details as group 1, the second group was also told that some evidence had been withheld from the jury by the judge. For Group 1: *M*=3 and *SS*=16 and for Group 2: *M*=6 and *SS*=24
- Is there a significant difference between the two groups in their responses?

Reporting the results

The participants who were informed about the withheld evidence gave significantly longer sentences than those participants who were not informed about the withhold evidence. The mean difference was significant, t(14)=-3.53, p<.05.

If an exact probability is available from a computer analysis, it should be reported.

If a confidence interval is reported to describe effect size, it appears immediately after the results from the hypothesis test

Topic 054: Interpreting and Reporting the Result

Reporting the results

The participants who were informed about the withheld evidence gave significantly longer sentences than those participants who were not informed about the withheld evidence. The mean difference was significant, $t(14)=-3.53$, $p<.05$.
If an exact probability is available from a computer analysis, it should be reported.
If a confidence interval is reported to describe effect size, it appears immediately after the results from the hypothesis test.

Lesson 26

Independent Sample T-Test- II

Topic: 055-056

Topic 055: Independent Sample T-Test in SPSS Part-1 and Part 2

Independent Sample t-test

An independent sample t-test is used when the mean scores of two different groups of people or conditions are compared. It is also called a between-subject design. This is a **parametric test** because it requires assumptions about parameters. Parametric tests require a numerical score for each individual in the sample. The scores then are added, squared, averaged, and otherwise manipulated using basic arithmetic. This test tells whether there is a statistically significant difference in the mean scores for the two groups, for instance, whether males and females differ significantly in terms of their self-esteem levels. In statistical terms, it tests the probability that the two sets of scores (for males and females) came from the same population.

To run independent sample t-test in SPSS, we need two set of variables;

- **The Continuous /Dependent Variable**, whose mean score is compared between the two groups.
- SPSS defines it as the **Test Variable** having scale measurement.
- **The Categorical/Independent Variable** with two categories that defines which two samples will be compared in the t-test.
- In SPSS, for t-test it is defined as the **Grouping Variable.**
- The **grouping variable** must have at least two categories (groups); it may have more than two categories but a t-test can only compare two groups, so there is need to specify which two groups to compare.
- We can also use a continuous variable by specifying a cut point to create two groups (i.e., values at or above the cut point and values below the cut point).

Downs and Abwender (2002) evaluated athletes (soccer players and swimmers) to determine whether the blows to the head experienced by both groups produced long-term neurological deficits. In the study, neurological tests were administered to mature soccer players and swimmers and a researcher obtained the following data presented in the table.

- Are there any significant differences in the neurological test scores?

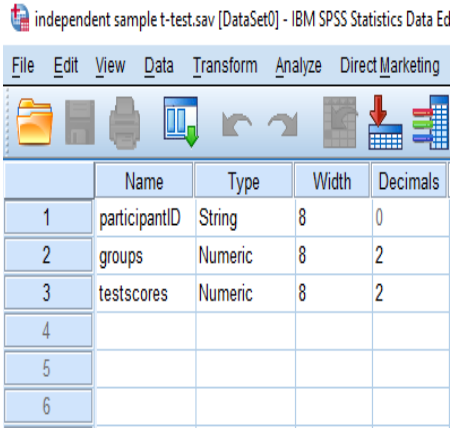
Swimmers	Soccer Players
10	7
8	4

7	9
8	3
13	7
7	
6	
12	

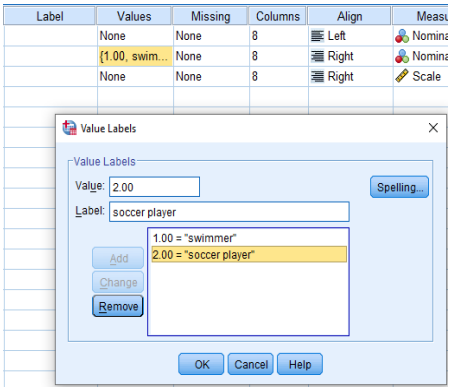
Topic 056: Running Independent Sample T-Test in SPSS

Entering Data

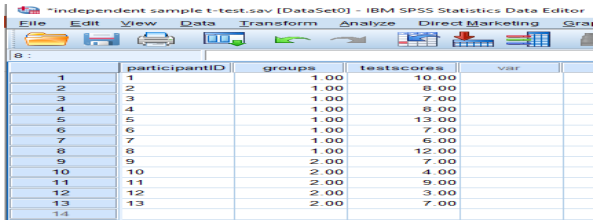
First of all, the variables must be setup in the variable view tab of SPSS data editor. Although it is not necessary but it is good practice to give each person a unique ID.



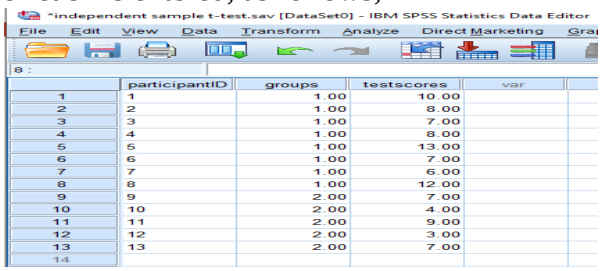
- We also need to provide codes that distinguish between the two groups (i.e., swimmers and soccer players).
- By clicking in the box of the 'values' field in the row containing the groups variable, we get a pop-up dialog that allows us to code the categorical variable. The swimmer is coded as '1' and the soccer player as '2'.



- The data is entered in the data view tab of the SPSS editor.
- We need to enter data for both the independent(categorical, i.e., blows to the head experienced by swimmers and soccer players)and dependent (continuous, i.e., test scores) variables.

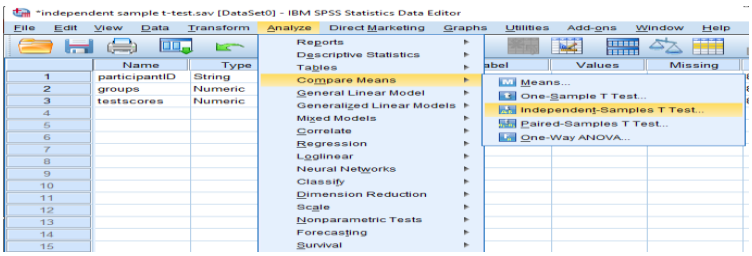


- A code is entered for the both groups as they were on (1 = swimmer, 2 = soccer player) and the test score of each is entered, as follows;

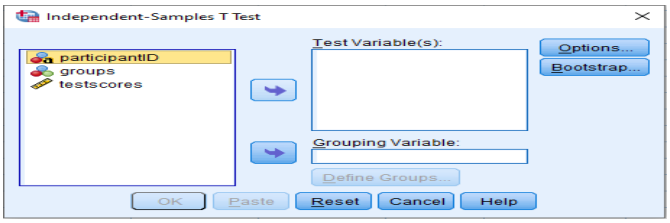


Running Analysis

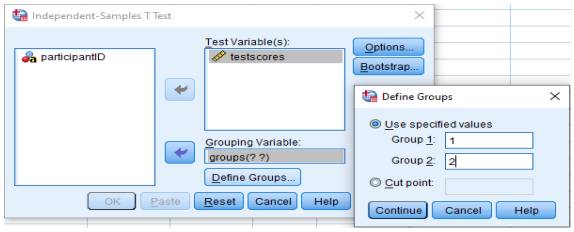
- To run independent sample t-test in SPSS, ***Click Analyze > Compare Means > Independent-Samples T Test***...on the top menu, as shown:
-



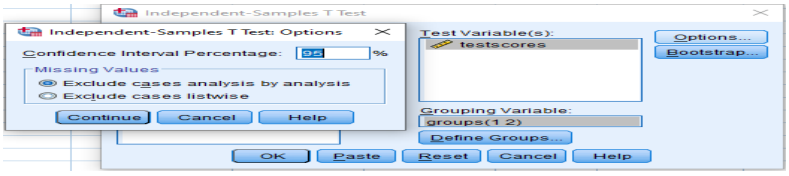
- The Independent-Samples T Test window opens where we transfer the dependent variable, **tests cores**, into the Test Variable(s): box, and the independent variable, **groups**, into the Grouping Variable: box, by highlighting the relevant variables and pressing the SPSS Right Arrow Button.
-



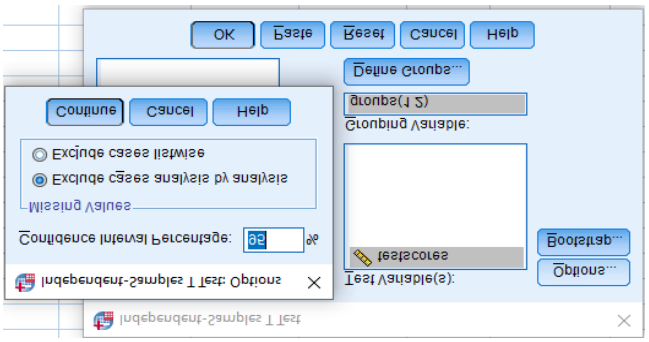
- We then need to define the groups (athletes, swimmers and soccer players).
- By clicking on the button we will be presented with the **Define Groups** dialogue box, as shown above.
- Enter **1** into the Group 1: box and enter **2** into the Group 2: box.
- Remember that we labelled the swimmer group as **1** and the soccer player group as **2**.
- Click the button **Continue**.



- The **Options** section is where we can set our desired confidence interval/level for the mean difference, and specify how SPSS should handle missing values.
- In SPSS by default confidence interval is set at 95% and we do not need to change it.



- The **Missing Values** section is not relevant if we have only specified one dependent variable; it only matters if we are entering more than one dependent (continuous numeric) variable.
- Then, click **Continue** and **OK** when finished**.**


OK
Paste
Reset
Cancel
Help
Define Groups...
Grouping Variable:
group(1 2)
Test Variable(s):
testscores
Bootstrap...
Options...
Continue
Cancel
Help
Exclude cases listwise
Exclude cases analysis by analysis
Missing Values
Confidence Interval Percentage:
95
%

Lesson 27

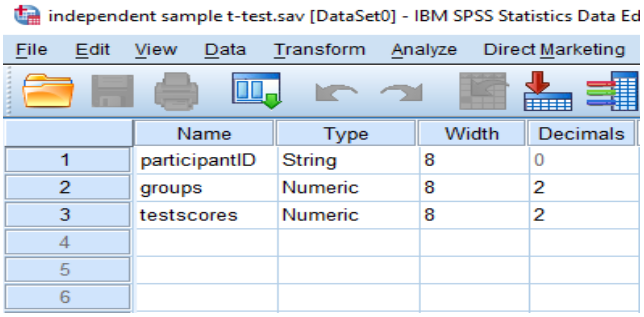
Independent Sample T-Test- III

Topic: 057-058

Topic 057:Running Independent Sample T-Test in SPSS Step by Step Tutorial Part-1 and Part 2

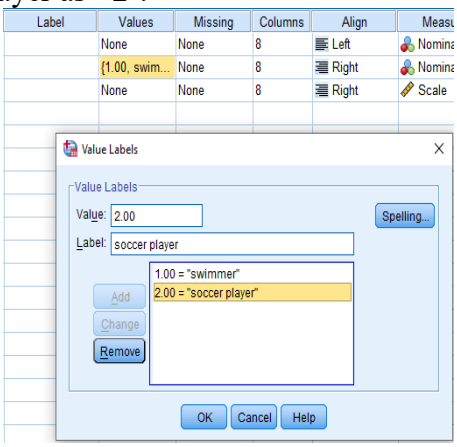
Entering Data

- First of all, the variables must be setup in the variable view tab of SPSS data editor.
- Although it is not necessary but it is good practice to give each person a unique ID.

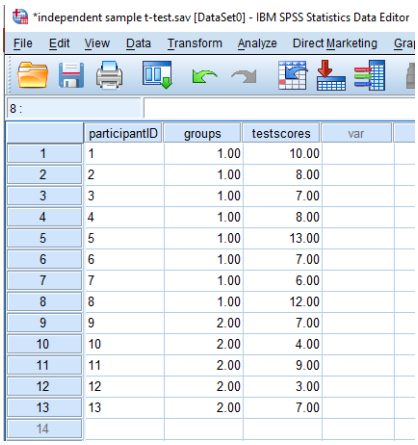


	Name	Type	Width	Decimals
1	participantID	String	8	0
2	groups	Numeric	8	2
3	testscores	Numeric	8	2
4				
5				
6				

- We also need to provide codes that distinguish between the two groups (i.e., swimmers and soccer players).
- By clicking in the box of the 'values' field in the row containing the groups variable, we get a pop-up dialog that allows us to code the categorical variable. The swimmer is coded as '1' and the soccer player as '2'.

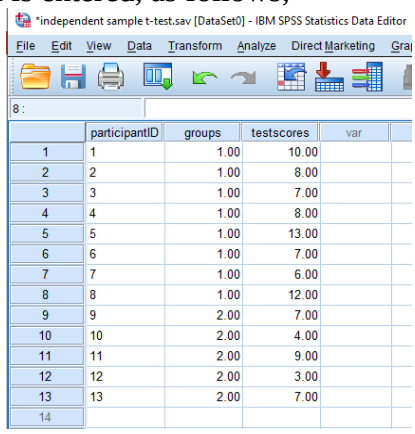


- The data is entered in the data view tab of the SPSS editor.
- We need to enter data for both the independent(categorical, i.e., blows to the head experienced by swimmers and soccer players)and dependent (continuous, i.e., test scores) variables.



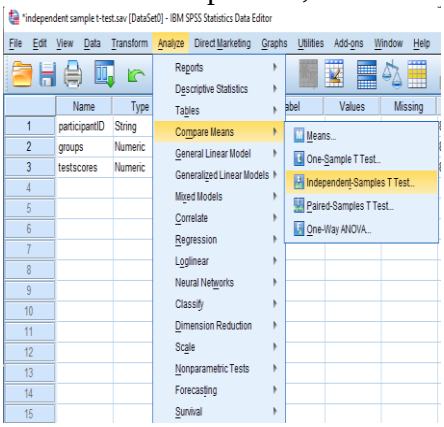
	participantID	groups	testscores	var
1	1	1.00	10.00	
2	2	1.00	8.00	
3	3	1.00	7.00	
4	4	1.00	8.00	
5	5	1.00	13.00	
6	6	1.00	7.00	
7	7	1.00	6.00	
8	8	1.00	12.00	
9	9	2.00	7.00	
10	10	2.00	4.00	
11	11	2.00	9.00	
12	12	2.00	3.00	
13	13	2.00	7.00	
14				

- A code is entered for the both groups as they were on (1 = swimmer, 2 = soccer player) and the test score of each is entered, as follows;



	participantID	groups	testscores	var
1	1	1.00	10.00	
2	2	1.00	8.00	
3	3	1.00	7.00	
4	4	1.00	8.00	
5	5	1.00	13.00	
6	6	1.00	7.00	
7	7	1.00	6.00	
8	8	1.00	12.00	
9	9	2.00	7.00	
10	10	2.00	4.00	
11	11	2.00	9.00	
12	12	2.00	3.00	
13	13	2.00	7.00	
14				

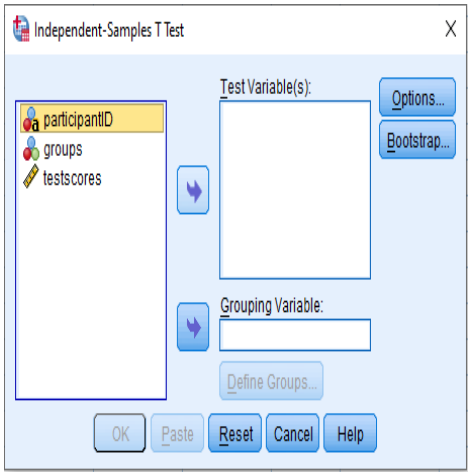
- To run independent sample t-test in SPSS, ***Click Analyze*** > ***Compare Means*** > ***Independent-Samples T Test***...on the top menu, as shown:



Running Analysis

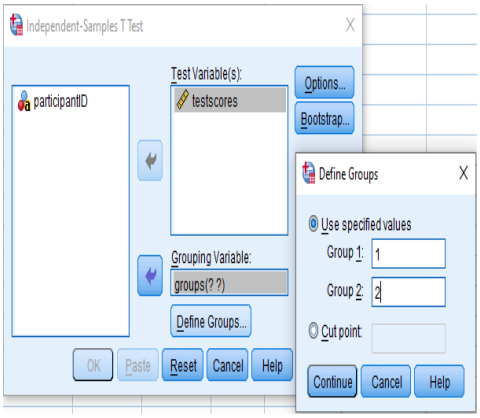
- The Independent-Samples T Test window opens where we transfer the dependent variable, **tests cores**, into the Test Variable(s): box, and the independent variable, **groups**, into the

Grouping Variable: box, by highlighting the relevant variables and pressing the SPSS Right Arrow Button.

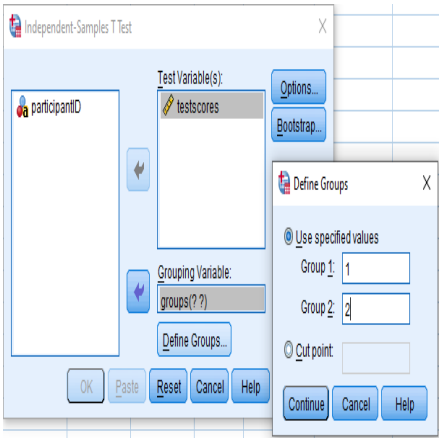


Running Analysis

- We then need to define the groups (athletes, swimmers and soccer players).
- By clicking on the button we will be presented with the **Define Groups** dialogue box, as shown above.

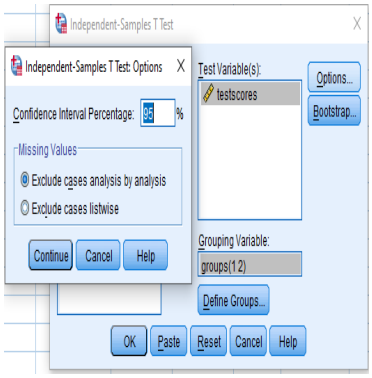


- Enter **1** into the Group 1: box and enter **2** into the Group 2: box.
- Remember that we labelled the swimmer group as **1** and the soccer player group as **2**.
- Click the button **Continue**.

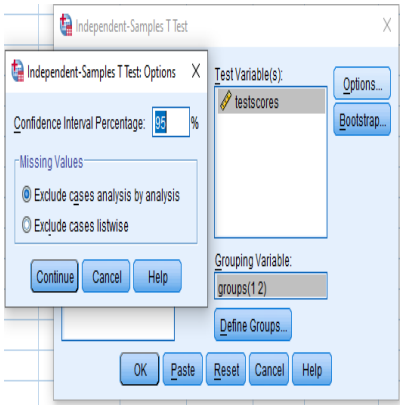


- The **Options** section is where we can set our desired confidence interval/level for the mean difference, and specify how SPSS should handle missing values.

- In SPSS by default confidence interval is set at 95% and we do not need to change it.



- The **Missing Values** section is not relevant if we have only specified one dependent variable; it only matters if we are entering more than one dependent (continuous numeric) variable.
- Then, click **Continue** and **OK** when finished**.**



Topic 058: Assumptions of Independent Sample T-Test Part-1 and Part 2

Assumption Testing

- There are some general assumptions that apply to all of the parametric tests (e.g. t-tests, analysis of variance), and additional assumptions associated with specific techniques.
- These general assumption include;

Assumption Testing
Level of Measurement

- Each of the parametric approaches assumes that the **dependent variable** is measured at the interval or ratio level; that is, using a continuous scale rather than discrete categories.

Assumption Testing

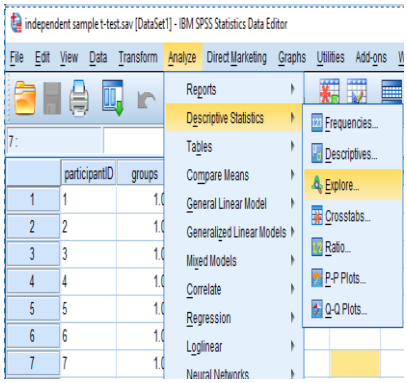
The parametric techniques assume that the scores are obtained using a random sample from the population.

Independence of Observations

- The observations that make up data must be independent of one another; that is, each observation or measurement must not be influenced by any other observation or measurement.

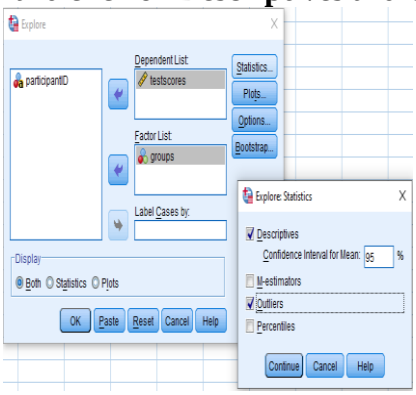
Normal Distribution

- For parametric techniques, it is assumed that the populations from which the samples are taken are normally distributed.
- The technique available in SPSS is using the **Explore** option of the **Descriptive Statistics** menu.

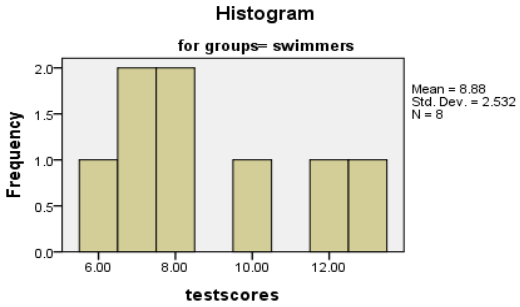


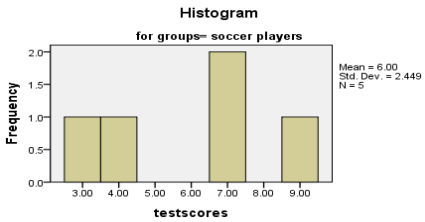
Normal Distribution

- Click on the variable(s) (test score) and move it into the **Dependent List** box.
- In the **Factor list** section add groups.
- In the **Display** section, make sure that **Both** is selected.
- Click on the **Statistics** button and click on **Descriptives** and **Outliers**. Click on **Continue**.

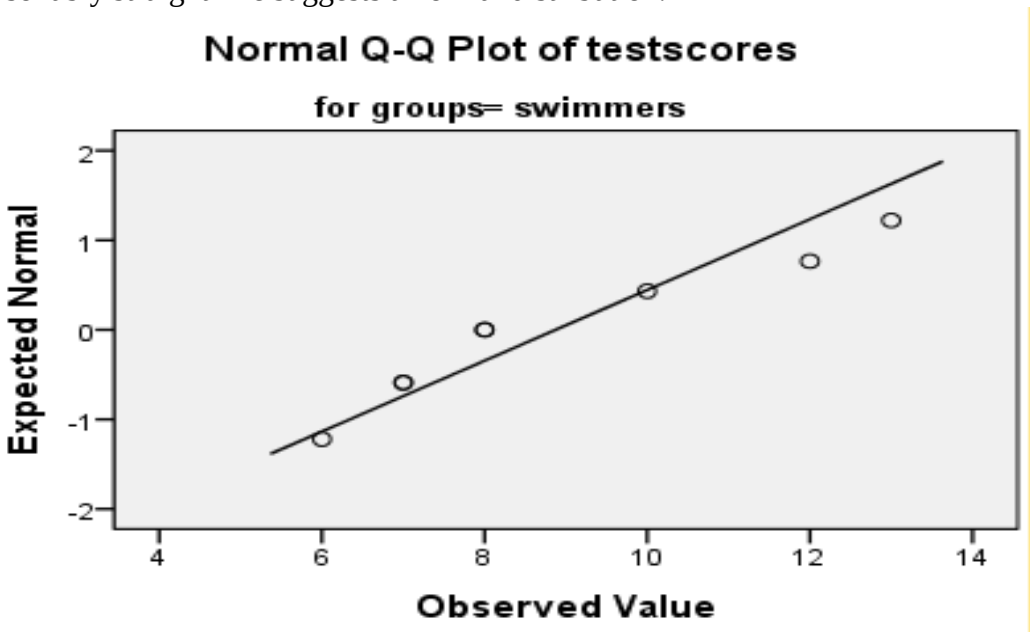


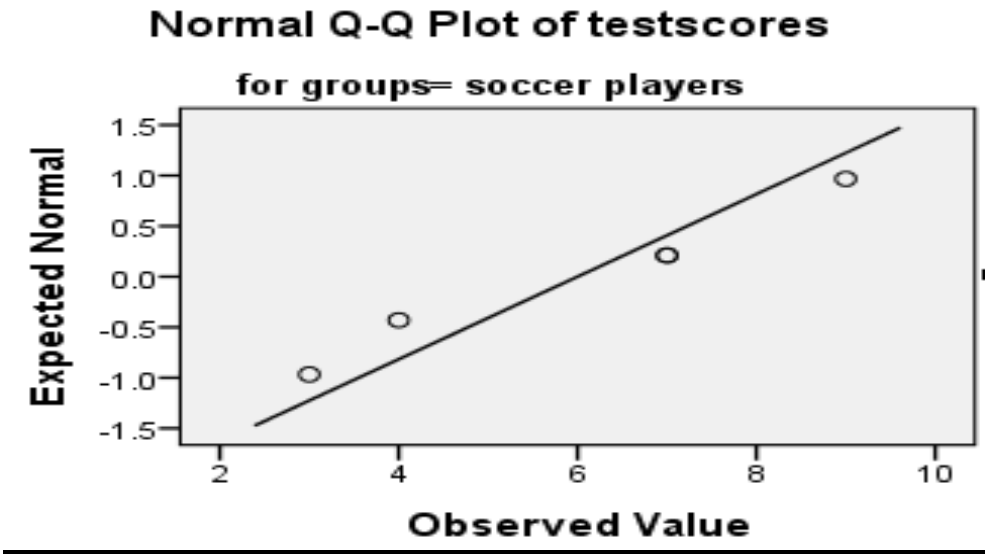
- Click on the Plots button. Under Descriptive, click on Histogram to select it.
- Click on Normality plots with tests. Click on Continue.
- Click on the Options button. In the Missing Values section, click on Exclude cases pairwise.
- The actual shape of the distribution for each group can be seen in the **Histograms**.





- This is also supported by an inspection of the normal probability plots (labelled **Normal Q-Q Plot**).
- In this plot, the observed value for each score is plotted against the expected value from the normal distribution.
- A reasonably straight line suggests a normal distribution.





- In the table labelled Tests of Normality, we are given the results of the Kolmogorov-Smirnov statistic.
- This assesses the normality of the distribution of scores. A non-significant result (Sig. value of more than .05) indicates normality.

Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	groups	Statistic	df	Sig.	Statistic	df	Sig.
testscores	swimmers	.260	8	.118	.896	8	.268
	soccer players	.258	5	.200*	.925	5	.563

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

- Other than mean and SD, descriptive statistics also provide some information concerning the distribution of scores on continuous variables (skewness and kurtosis).
- This information is needed if the variables are to be used in parametric statistical techniques (e.g. t-tests, analysis of variance).
- The **Skewness** value provides an indication of the symmetry of the distribution.
- Positive skewness values suggest that scores are clustered to the left at the low values.

- Negative skewness values indicate a clustering of scores at the high end (right-hand side of a graph).

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
testscores	13	3.00	13.00	7.7692	2.80339	.279	.616	.150	1.191
Valid N (listwise)	13								

- Kurtosis, on the other hand, provides information about the 'peakedness' of the distribution.
- Positive kurtosis values indicate that the distribution is rather peaked (clustered in the center), with long thin tails.
- Kurtosis values below 0 indicate a distribution that is relatively flat (too many cases in the extremes).

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
testscores	13	3.00	13.00	7.7692	2.80339	.279	.616	.150	1.191
Valid N (listwise)	13								

Homogeneity of Variance

- This assumption states that samples are obtained from populations of equal variances.
- This means that the variability of scores for each of the groups is similar. To test this, SPSS performs Levene's test for equality of variances as part of the t-test and analysis of variance analyses.

- The first section of the Independent Samples Test output box gives the results of Levene's test for equality of variances.
- This tests whether the variance (variation) of scores for the two groups (swimmers and soccer players) is the same. The outcome of this test determines which of the t-values that SPSS provides is the correct one to use.
- If the Sig. value for Levene's test is larger than .05 we should use the first line in the table, which refers to **Equal variances assumed**.
- If the significance level of Levene's test is p=.05 or less, this means that the variances for the two groups are not the same. Therefore, refer to second line equal **variances not assumed**.

		Levene's Test for Equality of Variances	
		F	Sig.
testscores	Equal variances assumed	.022	.886
	Equal variances not assumed		

Lesson 28

Analysis of Variance (AVOVA)

Topic: 059-060

Topic 059: Introduction to Analysis of Variance (AVOVA)

Analysis of Variance (ANOVA)

Analysis of variance (ANOVA) is a hypothesis-testing procedure that is used to evaluate mean differences between two or more treatments (or populations).
While running ANOVA we must decide between two interpretations:
1. There really are no differences between the populations (or treatments).
2. The populations (or treatments) really do have different means.
(Random and/or systematic errors).
When a researcher manipulates a variable to create the treatment conditions in an experiment, the variable is called an *independent variable.* When a researcher uses a non-manipulated variable to designate groups, the variable is called a *quasi-independent variable.*

In ANOVA, the variable (independent or quasi-independent) that designates the groups being compared is called a factor. The individual groups or treatment conditions that are used to make up a factor are called the levels of the factor. ANOVA can be used to evaluate the results from a research study that involves more than one factor. For example, a researcher may want to compare two different therapy techniques, examining their immediate effectiveness as well as the persistence of their effectiveness over time. A study that combines two factors, is called a two-factor design or a factorial design (Two way ANOVA).
A study that examines only one independent variable is called single-factor design (or One way ANOVA).

Hypotheses for (ANOVA)

The null hypothesis states that there is no treatment effect.

H_o: $\mu_1=\mu_2=\mu_3$

The alternative hypothesis states that there is at least one mean difference among the populations.
H_1: $\mu_1 \neq \mu_2 \neq \mu_3$

OR

- H_1: $\mu_1=\mu_3$ but μ_2 is different

In ANOVA we use variance to measure sample mean differences when there are two or more samples. The test statistic for ANOVA uses this fact to compute an F-*ratio* with the following structure:

$$F = \frac{\boldsymbol{Variance\ between\ treatments}}{\boldsymbol{Error\ Variance}\ (\boldsymbol{within\ group\ variance}\ )}$$

Topic 060: Logic Behind AVOVA and Its Assumptions:

The Logic of ANOVA
The first step in ANOVA is to determine the total variability for the entire set of data. (one general measure of variability for the complete experiment).
The word analysis means dividing into smaller parts. Because we are going to analyze variability, the process is called the analysis of variance.
The analysis process divides the total variability into two basic components.
1. Between-Treatments Variance
2. Within-Treatment Variance

Between-Treatments Variance
- This simply measures how much difference exists between the treatment conditions.

There are two possible explanations for these between-treatment differences:
- The differences are the result of *sampling error*.
- The differences between treatments have been caused by the *treatment effects*.

Within-Treatment Variance
- Inside each treatment condition, a set of individuals receive exactly the same treatment but their scores differ.
- The differences that exist within a treatment represent random and unsystematic differences when there are no treatment effects
- Thus, the *within-treatments variance* provides a measure of how big the differences are when H_o is true.

The F-ratio: The Test Statistic For ANOVA
- Once the total variability is analyzed into two basic components (between treatments and within treatments), we simply compare them.
- The comparison is made by computing an F-*ratio* as:

$$F = \frac{Variance\ between\ treatments}{Variance\ within\ treatments}$$

The F-ratio: The Test Statistic For ANOVA
- For ANOVA, the denominator of the *F*-ratio is called the **error term**. The error term provides a measure of the variance caused by random, unsystematic differences.
- Because ANOVA typically is used to examine data from more than two treatment conditions (and more than two samples), we need a notational system to keep track of all the individual scores and totals.
- The observations within each sample must be independent.
- The populations from which the samples are selected must be normal.
- The populations from which the samples are selected must have equal variances (homogeneity of variance).
- IV is always categorical and DV is continues.
- The assumption of homogeneity of variance is an important one.
- If a researcher suspects that it has been violated, it can be tested by **Welch** and **Brown-Forsythe tests**.

- If one of the assumptions for the independent-measures ANOVA has been violated, an alternative statistical analysis known as the **Kruskal-Wallis test** can be used.

Lesson 29

One Way ANOVA in SPSS-I

Topic: 061-062

Topic 061: One Way ANOVA in SPSS

One-Way ANOVA

One-way between-groups ANOVA is used when there is one independent (grouping) variable with three or more levels (groups) and one dependent continuous variable.For example, in comparing the effectiveness of three different teaching styles on students' math scores, you would have one factor (teaching style) with three levels (e.g. whole class, small group activities, self-paced computer activities).The dependent variable is a continuous variable (in this case, scores on a math test). One-way ANOVA tells whether there are significant differences in the mean scores on the dependent variable across the three groups.
Post-hoc tests can then be used to find out where these differences lie. To run in SPSS, we need one categorical independent variable with three or more distinct categories. This can also be a continuous variable that has been recoded to give three equal groups (e.g. age group: participants divided into three age categories, 29 and younger, between 30 and 44, 45 or above) and
One continuous dependent variable (e.g. optimism scores).
A human-factors psychologist studied three computer keyboard designs. Three samples of individuals were given material to type on a particular keyboard, and the number of errors committed by each participant was recorded. The data are shown in table.
Are these data sufficient to conclude that there are significant differences in typing performance among the three keyboard designs at an alpha level of .05?

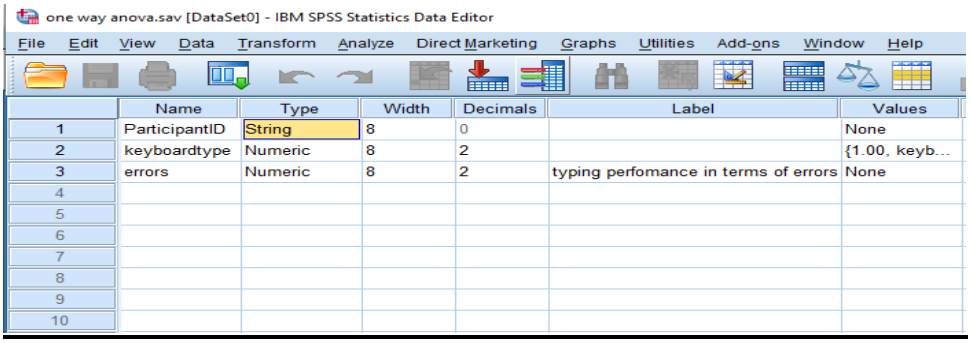
Keyboard A	Keyboard B	Keyboard C
0	6	6
4	8	5
0	5	9
1	4	4
0	2	6

Topic 062: Running One Way ANOVA in SPSS Part-1 and Part-2

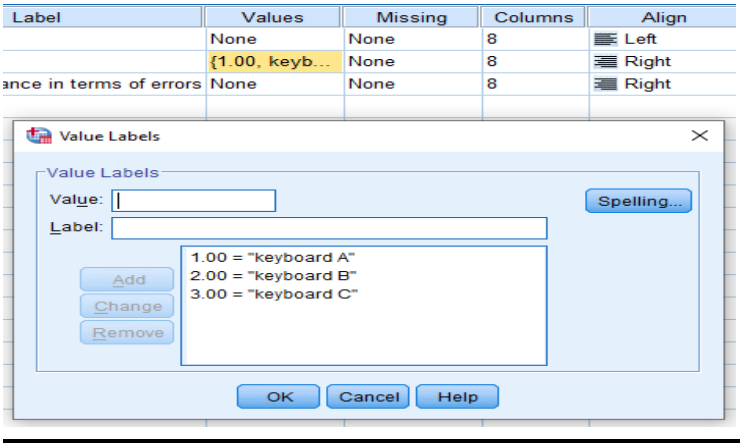
Entering Data in SPSS

We separate the groups for analysis by creating a grouping variable called Keyboard type (i.e., the independent variable.

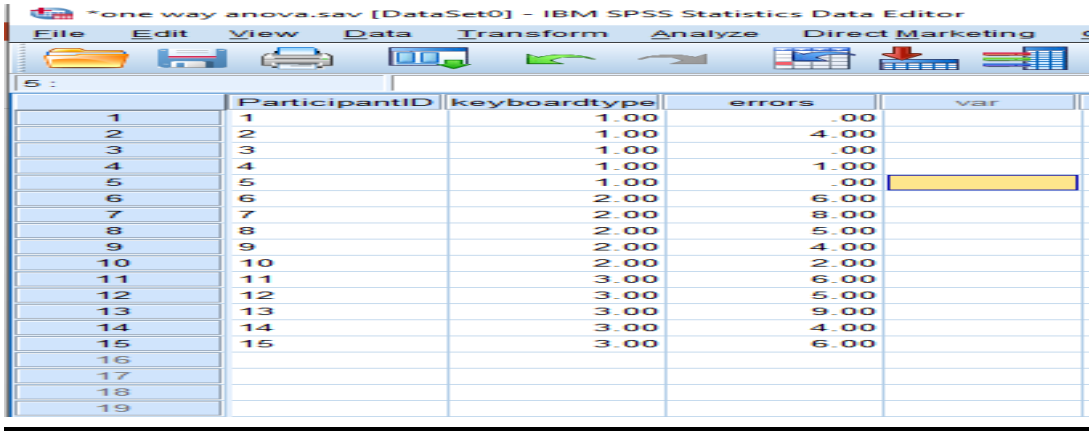
Number of errors committed is entered under the variable name errors (i.e., the dependent variable).



By clicking in the box of the 'values' field in the row containing the groups variable, we get a pop-up dialog that allows us to code the categorical variable. Three types of keyboard are coded as shown.

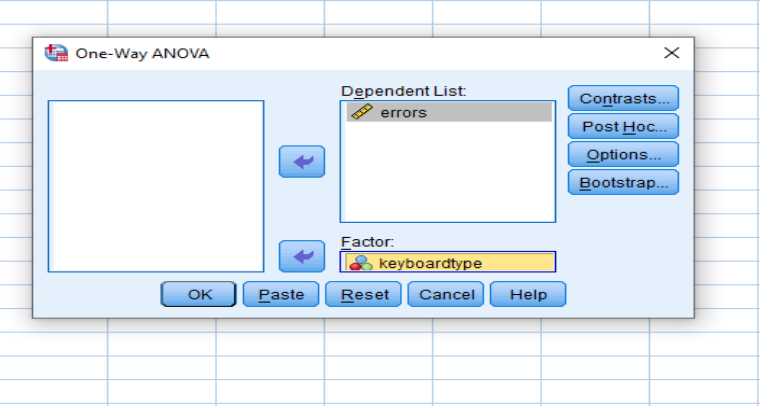


The data is entered in the data view tab of the SPSS editor. A code is entered for the categories of independent variable as (1 = keyboard A, 2 = keyboard B and 3 = keyboard C) and the number of committed errors in errors column as:

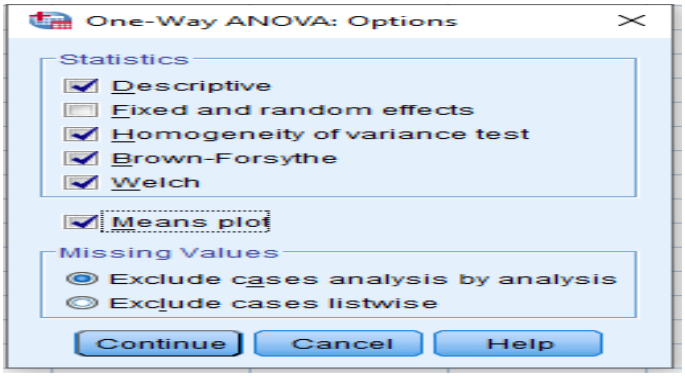


Running Analysis

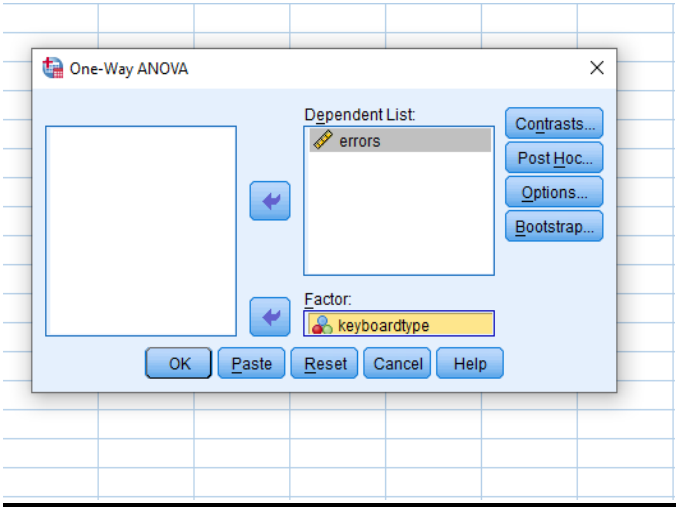
- To run one-way ANOVA in SPSS, *Click Analyze > Compare Means > One-Way ANOVA*...on the top menu.
- The One-Way ANOVA window opens.
- Click on your dependent (continuous) variable (e.g. errors). Move this into the box marked **Dependent List** by clicking on the arrow button.
- Click on your independent, categorical variable (e.g. keyboard type). Move this into the box labelled **Factor**.

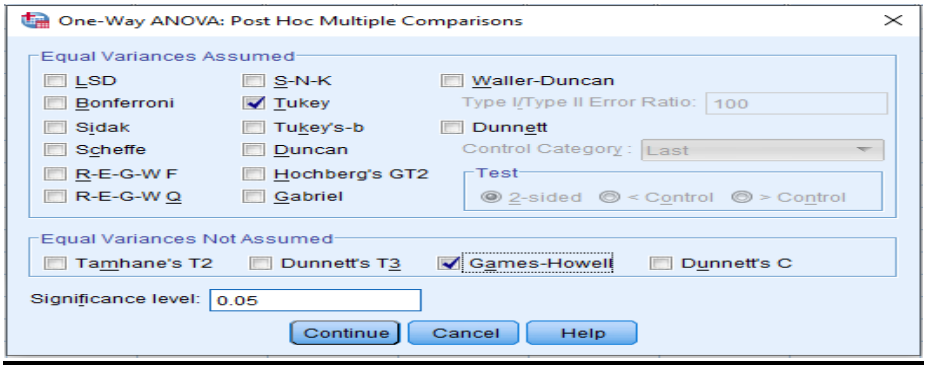


- Click the Options button and click on Descriptive, Homogeneity of variance test, Brown-Forsythe, Welch and Means Plot.
- For Missing values, make sure there is a dot in the option marked Exclude cases analysis by analysis. Click on Continue.



- Click on the button marked Post Hoc. Click on Tukey.
- Click on Games-Howell under the section Equal Variances Not Assumed
- Click on Continue and then OK.


One-Way ANOVA
Dependent List:
errors
Contrasts...
Post Hoc...
Options...
Bootstrap...
Factor:
keyboardtype
OK
Paste
Reset
Cancel
Help


One-Way ANOVA: Post Hoc Multiple Comparisons
Equal Variances Assumed
LSD
Bonferroni
Sidak
Scheffe
R-E-G-W F
R-E-G-W Q
S-N-K
Tukey
Tukey's-b
Duncan
Hochberg's GT2
Gabriel
Waller-Duncan
Type I/Type II Error Ratio: 100
Dunnett
Control Category : Last
Test
2-sided
< Control
> Control
Equal Variances Not Assumed
Tamhane's T2
Dunnett's T3
Games-Howell
Dunnett's C
Significance level: 0.05
Continue
Cancel
Help

Lesson 30

One Way ANOVA in SPSS-II

Topic: 063-0

Topic: 063

Interpreting and Reporting the Output of AVOVA

ANOVA has the same general assumptions that apply to all of the parametric tests with an additional assumption of homogeneity of variances.
The general assumptions include;
1. Level of measurement (DV measured on interval or ratio scale, i.e., continuous)
The general assumptions include;
2. Random sampling
3. Independence of observations
4. Normal distribution (by assessing skewness, kurtosis, histogram, Q-Q plots, Shapiro-Wilk and Kolmogorov-Smirnov tests)

Homogeneity of Variance

- The homogeneity of variance option gives us the **Levene's test for homogeneity of variances**, which tests whether the variance in scores is the same for each of the three groups.
- Check the significance value (**Sig.**) for Levene's test. If this number is *greater* than .05, the assumption is not violated.

Test of Homogeneity of Variances

typing perfomance in terms of errors

Levene Statistic	df1	df2	Sig.
.170	2	12	.845

- If the assumption is found to be violated then the F test of the ANOVA is not robust enough to be used. In this case, there is need to consult the table in the output headed Robust Tests of Equality of Means.
- The two tests there (Welch and Brown-Forsythe) are preferable when the assumption of the homogeneity of variance is violated.

Robust Tests of Equality of Means

typing perfomance in terms of errors

	Statistic[a]	df1	df2	Sig.
Welch	9.867	2	7.919	.007
Brown-Forsythe	9.130	2	11.438	.004

a. Asymptotically F distributed.

Interpreting the Output

- The **Descriptives** table gives us the information about each group (number in each group, means, standard deviation, minimum and maximum, etc.)

Descriptives

typing perfomance in terms of errors

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
keyboard A	5	1.0000	1.73205	.77460	-1.1506	3.1506	.00	4.00
keyboard B	5	5.0000	2.23607	1.00000	2.2236	7.7764	2.00	8.00
keyboard C	5	6.0000	1.87083	.83666	3.6771	8.3229	4.00	9.00
Total	15	4.0000	2.87849	.74322	2.4059	5.5941	.00	9.00

- The **ANOVA** table gives both between-groups and within-groups sums of squares, degrees of freedom etc.
- The main thing is the **Sig.** (the p value). If the Sig. value is less than or equal to .05, there is a significant difference somewhere among the mean scores on dependent variable for the three groups.

ANOVA

typing perfomance in terms of errors

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	70.000	2	35.000	9.130	.004
Within Groups	46.000	12	3.833		
Total	116.000	14			

- In this example the overall Sig. value is .004, which is less than .05, indicating a statistically significant result somewhere among the groups with an F-ratio of 9.13.

Interpreting the Output

- The **post-hoc tests** table tells us exactly where the differences among the groups occur.
- In the column labelled **Mean Difference** look for any asterisks (*) next to the values listed. If there is an (*), this means that the two groups being compared are significantly different from one another at the p<.05 level.
- The exact significance value is given in the column labelled **Sig.**

Post Hoc Tests

Multiple Comparisons

Dependent Variable: typing perfomance in terms of errors

	(I) keyboardtype	(J) keyboardtype	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Tukey HSD	keyboard A	keyboard B	-4.00000*	1.23828	.018	-7.3036	-.6964
		keyboard C	-5.00000*	1.23828	.004	-8.3036	-1.6964
	keyboard B	keyboard A	4.00000*	1.23828	.018	.6964	7.3036
		keyboard C	-1.00000	1.23828	.706	-4.3036	2.3036
	keyboard C	keyboard A	5.00000*	1.23828	.004	1.6964	8.3036
		keyboard B	1.00000	1.23828	.706	-2.3036	4.3036
Games-Howell	keyboard A	keyboard B	-4.00000*	1.26491	.034	-7.6623	-.3377
		keyboard C	-5.00000*	1.14018	.006	-8.2620	-1.7380
	keyboard B	keyboard A	4.00000*	1.26491	.034	.3377	7.6623
		keyboard C	-1.00000	1.30384	.733	-4.7501	2.7501
	keyboard C	keyboard A	5.00000*	1.14018	.006	1.7380	8.2620
		keyboard B	1.00000	1.30384	.733	-2.7501	4.7501

*. The mean difference is significant at the 0.05 level.

- There is a significant difference in errors (typing performance) between the groups that used the keyboard A and keyboard B (p =.018), as well as between the keyboard A and keyboard C (p=.004).

However, there were no differences between the groups that used keyboard B and keyboard C (p=.706).

Post Hoc Tests

Multiple Comparisons

Dependent Variable: typing perfomance in terms of errors

	(I) keyboardtype	(J) keyboardtype	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Tukey HSD	keyboard A	keyboard B	-4.00000*	1.23828	.018	-7.3036	-.6964
		keyboard C	-5.00000*	1.23828	.004	-8.3036	-1.6964
	keyboard B	keyboard A	4.00000*	1.23828	.018	.6964	7.3036
		keyboard C	-1.00000	1.23828	.706	-4.3036	2.3036
	keyboard C	keyboard A	5.00000*	1.23828	.004	1.6964	8.3036
		keyboard B	1.00000	1.23828	.706	-2.3036	4.3036
Games-Howell	keyboard A	keyboard B	-4.00000*	1.26491	.034	-7.6623	-.3377
		keyboard C	-5.00000*	1.14018	.006	-8.2620	-1.7380
	keyboard B	keyboard A	4.00000*	1.26491	.034	.3377	7.6623
		keyboard C	-1.00000	1.30384	.733	-4.7501	2.7501
	keyboard C	keyboard A	5.00000*	1.14018	.006	1.7380	8.2620
		keyboard B	1.00000	1.30384	.733	-2.7501	4.7501

*. The mean difference is significant at the 0.05 level.

Interpreting the Output

Effect Size

- Although SPSS does not generate effect size for this analysis, it is possible to determine the effect size for this result
- The information that is needed to calculate **eta squared**, is provided in the **ANOVA** table.
- η^2 = SSbetween/SStotal
- η^2 = 70/116= 0.603

ANOVA

typing perfomance in terms of errors

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	70.000	2	35.000	9.130	.004
Within Groups	46.000	12	3.833		
Total	116.000	14			

Reporting Results in APA

A one-way between-groups analysis of variance was conducted to explore the impact of keyboard type on typing performance, as measured by the number of errors committed. Participants were divided into three groups according to their use of keyboard (Group 1: keyboard A; Group 2: keyboard B; Group 3: keyboard C). The analysis of variance indicates that there are significant differences among the three keyboard types, $F(2,12) = 9.13$, p=.004, $\eta^2 = 0.60$.

Post-hoc comparisons using the Tukey HSD test indicated that the number of errors were statistically significantly higher after using the keyboard B (M = 5.00, SD = 2.24, p=.018) and keyboard C (M = 6.00, SD = 1.87, p=.004) as compared to keyboard A (M = 1.00, SD = 1.73). There were no statistically significant differences between the keyboard B and keyboard C (p = .706).

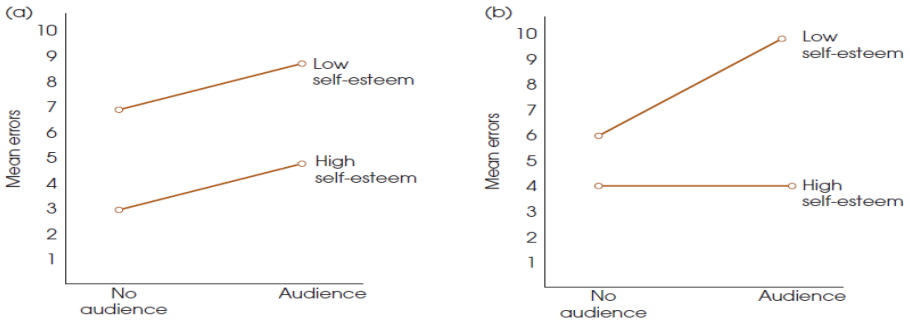
Topic 064: Introduction to Two-Way Anova

Two-Way ANOVA

- In real life, variables rarely exist or relate to other variables in isolation.
- To examine more complex situations, researchers often design research studies that include more than one independent variable.
- When a research study examines ANOVA having two factors or two independent variables, it is called Two-Way ANOVA.
- Two-way ANOVA means groups are defined by two independent variables.
- The two independent variables in a two-factor experiment are identified as factor *A* and factor *B*.
- The goal of the factorial design study is to evaluate the mean differences that may be produced by either of the factors acting independently or by the two factors acting together.
- The mean differences among the levels of one factor are referred to as the **main effect** of that factor.

	No Audience	Audience	
Low	$M = 7$	$M = 9$	$M = 8$
High	$M = 3$	$M = 5$	$M = 4$
	$M = 5$	$M = 7$	

- The real advantage of combining two factors within the same study is the ability to examine not only the main effects but the unique effects caused by an interaction.
- the concept of an interaction can be defined in terms of the pattern displayed in the graph.
- The $A \times B$ interaction typically is called the "*A* by *B*" interaction. If there is an interaction between an audience and self-esteem, it may be called the "audience by self-esteem"



interaction.

Hypotheses for Two-Way (ANOVA)

consists of three hypothesis tests, two for main effects and one for interaction

H_o: No main effect of A

H_1: There exists a main effect of factor A

H_o: No main effect of B

H_1: There exists a main effect of factor B

H_o: No interaction

H_1: There exists an interaction between factor A and factor B

Degrees of Freedom (df)

The between-treatments *df*

- $dfA = kA\text{-}1$
- $dfB = kB\text{-}1$
- $dfA \times B = dfA \times dfB$

The within-treatments *df*

- $df = N\text{-}K$

The Total df

- $df = N\text{-}1$
- We calculate *F*-ratio

Assumptions of Two-Way ANOVA

- The observations within each sample must be independent.

- The populations from which the samples are selected must be normal.
- The populations from which the samples are selected must have equal variances (homogeneity of variance).

© Virtual University of Pakistan

Lesson 31

Two-Way-Factorial (ANOVA) in SPSS

Topic: 065

Topic 065: Two-Way-Factorial (ANOVA) in SPSS

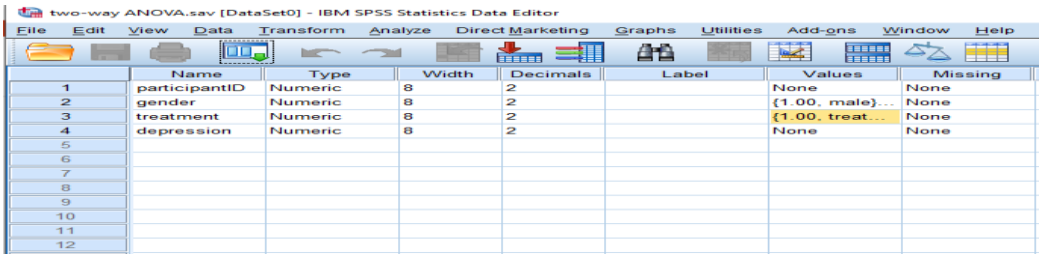
Two-Way ANOVA in SPSS

- Two-way means that there are two independent variables, and *between-groups* indicates that different people are in each of the groups.
- This technique allows us to look at the individual and joint effect of two independent variables on one dependent variable.
- Two-way ANOVA allows you to simultaneously test for the effect of each of your independent variables on the dependent variable and also identifies any interaction effect.
- To run two-way ANOVA we need three variables:
- Two categorical independent variables (e.g. sex: males/females; age group: young, middle, old)
- One continuous dependent variable (e.g. optimism).
- The following data are from a two-factor study examining the depression scores across gender and two treatment conditions. Use an ANOVA with α=.05 for all tests to evaluate the significance of the main effects and the interaction.

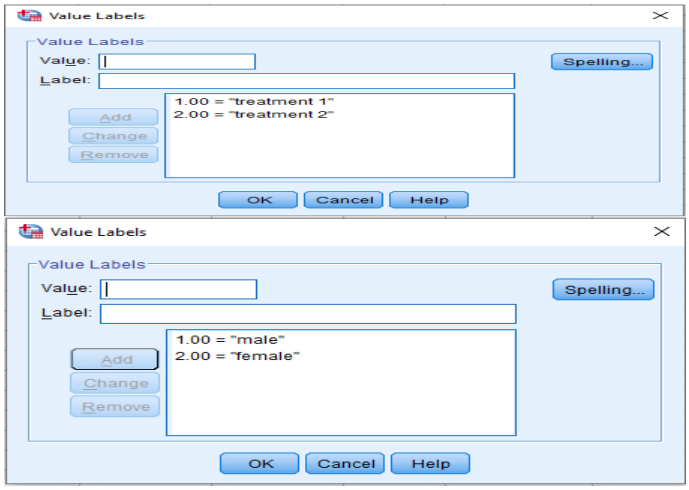
• **Factor A** • **Gender**	• **Factor B** • **Treatments**	
• **Male**	• **I**	• **II**
	• 3	• 2
	• 8	• 8
	• 9	• 7
	• 4	• 7
• **Female**	• 0	• 12
	• 0	• 6
	• 2	• 9
	• 6	• 13

Entering Data in SPSS

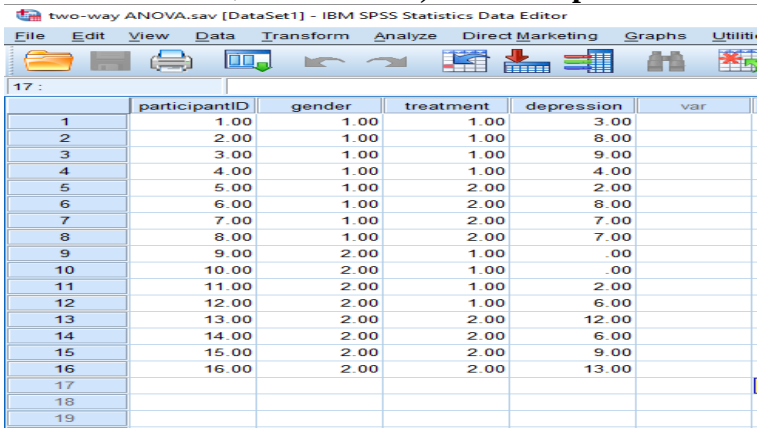
- We need to create two grouping variables in the variable view of SPSS as, **treatment and gender** (i.e., the independent variables).
- Depression scores are entered under the variable name **depression** (i.e., the dependent variable).



- In the box of the 'values' we get a pop-up dialog that allows us to code the categorical variable. Gender and Treatment conditions are coded as shown in the value labels window.

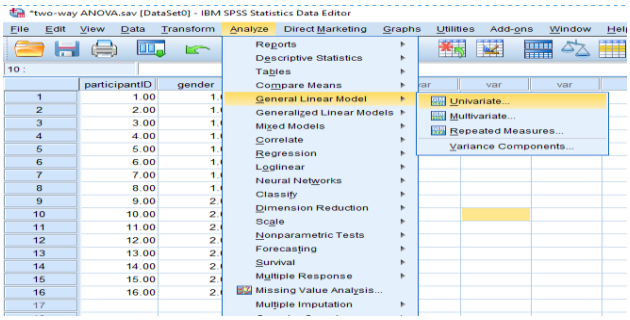


While entering data A code is entered for the categories of independent variables as (**Gender**: 1=male, 2=female; **Treatment**: 1=treatment1, 2=treatment2) and the **depression** scores in the depression column.

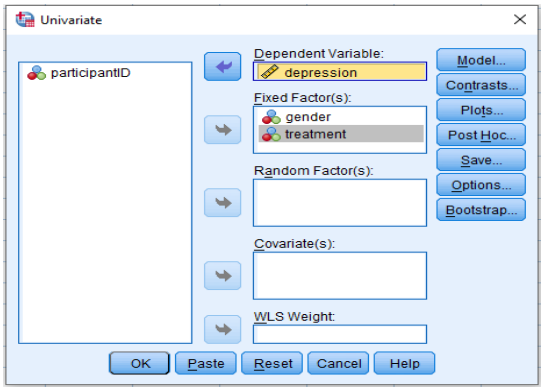


Running Analysis

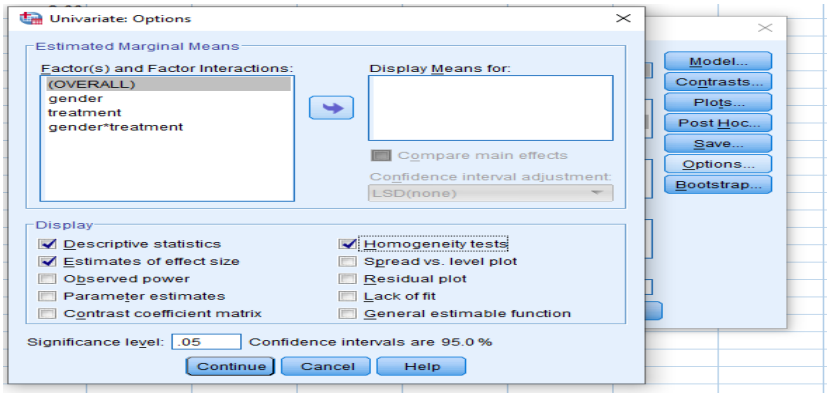
- To run one-way ANOVA in SPSS, from the menu at the top of the screen, click on **Analyze** >**General Linear Model** >**Univariate**.



- A **Univariate** window opens.
- Click on the dependent, continuous variable (e.g. depression) and move it into the box labelled **Dependent Variable**.
- Click on two independent, categorical variables (gender, treatment) and move these into the box labelled **Fixed Factors**.

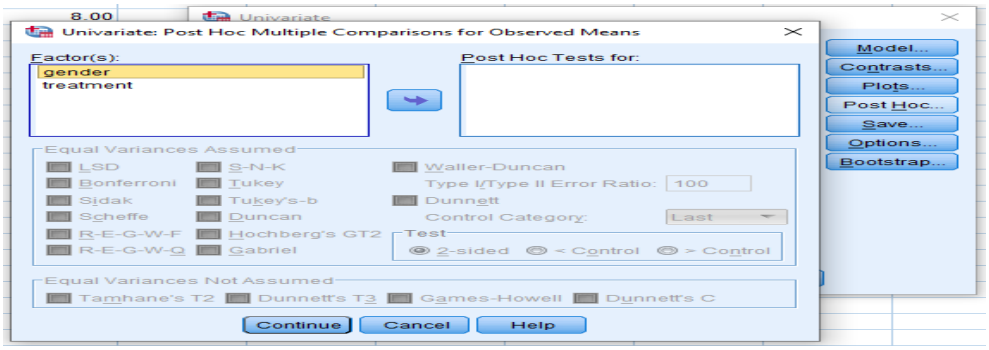


- Click on the **Options** button.
- Click on **Descriptive Statistics**, **Estimates of effect size** and **Homogeneity tests** and then **Continue** when finished.

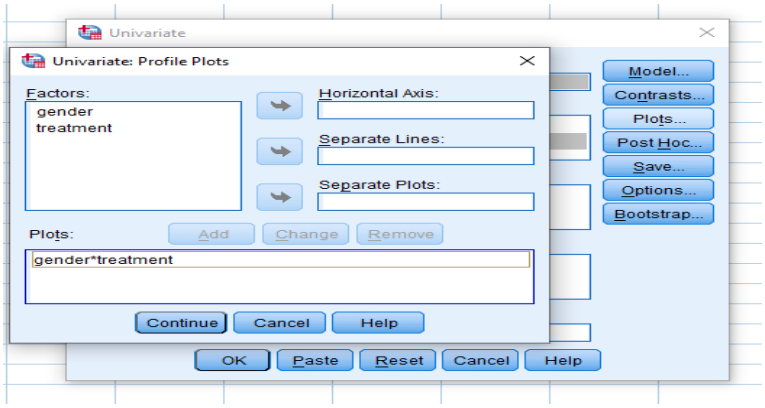


- Click on the **Post Hoc** button.
- From the **Factors** listed on the left-hand side, choose the independent variable(s) you are interested in (only if the variables have more than two levels).
- Click on the arrow button to move it into the **Post Hoc Tests for** section.

- Choose the test **Tukey** and click on **Continue**.



- Click on the **Plots** button.
- In the **Horizontal Axis** box, put the independent variable that has the most groups (e.g. treatment).
- In the box labelled **Separate Lines**, put the other independent variable (e.g. gender).
- Click on **Add**.
- In the section labelled **Plots**, you will see the two variables listed as gender*treatment
- Click on **Continue** and then **OK.**
-



Lesson 32

Interpreting the Output of Two-Way ANOVA

Topic: 066

Topic 066: Interpreting the Output of Two-Way Anova Part-1 & Part 2

Assumptions of Two-Way ANOVA

The general assumptions include;
2. Random sampling
3. Independence of observations
4. Normal distribution (by assessing skewness, kurtosis, histogram, Q-Q plots, Shapiro-Wilk and Kolmogorov-Smirnov tests).

Homogeneity of Variance

- The **Levene's test for homogeneity of variances** tests whether the variance in scores is the same for each of the groups.
- For the assumption to be fulfilled the **Sig**. value should be *greater* than .05 and therefore *non*-significant results. A significant result (Sig. value less than .05) suggests that the variance of dependent variable across the groups is not equal.

Levene's Test of Equality of Error Variances[a]

Dependent Variable: depression

F	df1	df2	Sig.
.200	3	12	.894

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept + gender + treatment + gender * treatment

Interpreting the Output

- The **Descriptive Statistics** table provides the **Mean** scores, **Std. deviations** and **N** for each subgroup.

Descriptive Statistics

Dependent Variable: depression

gender	treatment	Mean	Std. Deviation	N
male	treatment 1	6.0000	2.94392	4
	treatment 2	6.0000	2.70801	4
	Total	6.0000	2.61861	8
female	treatment 1	2.0000	2.82843	4
	treatment 2	10.0000	3.16228	4
	Total	6.0000	5.09902	8
Total	treatment 1	4.0000	3.42261	8
	treatment 2	8.0000	3.46410	8
	Total	6.0000	3.91578	16

Interpreting the Output

- The main output from two-way ANOVA is a table labelled Tests of Between-Subjects Effects.
- This table gives us the *main effects* and *interaction effects.*
- The first thing to do is to check for the possibility of an interaction effect (e.g., that the influence of treatments on depression scores depends on whether the individual is male or female).

Tests of Between-Subjects Effects

Dependent Variable: depression

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	128.000[a]	3	42.667	5.020	.018	.557
Intercept	576.000	1	576.000	67.765	.000	.850
gender	.000	1	.000	.000	1.000	.000
treatment	64.000	1	64.000	7.529	.018	.386
gender * treatment	64.000	1	64.000	7.529	.018	.386
Error	102.000	12	8.500			
Total	806.000	16				
Corrected Total	230.000	15				

a. R Squared = .557 (Adjusted R Squared = .446)

- If the interaction effect is significant, the main effects cannot be interpreted easily.
- This is because, in order to describe the influence of one of the independent variables, there is need to specify the level of the other independent variable.
- The output we need to look at is labelled gender*treatment.

Tests of Between-Subjects Effects

Dependent Variable: depression

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	128.000[a]	3	42.667	5.020	.018	.557
Intercept	576.000	1	576.000	67.765	.000	.850
gender	.000	1	.000	.000	1.000	.000
treatment	64.000	1	64.000	7.529	.018	.386
gender * treatment	64.000	1	64.000	7.529	.018	.386
Error	102.000	12	8.500			
Total	806.000	16				
Corrected Total	230.000	15				

a. R Squared = .557 (Adjusted R Squared = .446)

- In this example, the interaction effect is significant (F= 7.529, p= .018).
- This indicates that there is significant difference in the effect of treatment on depression for males and females.

Tests of Between-Subjects Effects

Dependent Variable: depression

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	128.000[a]	3	42.667	5.020	.018	.557
Intercept	576.000	1	576.000	67.765	.000	.850
gender	.000	1	.000	.000	1.000	.000
treatment	64.000	1	64.000	7.529	.018	.386
gender * treatment	64.000	1	64.000	7.529	.018	.386
Error	102.000	12	8.500			
Total	806.000	16				
Corrected Total	230.000	15				

a. R Squared = .557 (Adjusted R Squared = .446)

- To determine whether there is a **main effect** for each independent variable, in the left-hand column, find the variable (e.g., gender) and check in the column marked **Sig.** next to each variable. If the value is less than or equal to .05 there is a significant main effect for that independent variable.

Tests of Between-Subjects Effects

Dependent Variable: depression

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	128.000[a]	3	42.667	5.020	.018	.557
Intercept	576.000	1	576.000	67.765	.000	.850
gender	.000	1	.000	.000	1.000	.000
treatment	64.000	1	64.000	7.529	.018	.386
gender * treatment	64.000	1	64.000	7.529	.018	.386
Error	102.000	12	8.500			
Total	806.000	16				
Corrected Total	230.000	15				

a. R Squared = .557 (Adjusted R Squared = .446)

- In the output shown in the table, there is no significant main effect for gender (p= 1.00) but a significant main effect for treatment (p= .018).
- This means that males and females do not differ in terms of their depression scores, but there is a difference in scores for treatment I and treatment II.

Tests of Between-Subjects Effects

Dependent Variable: depression

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	128.000[a]	3	42.667	5.020	.018	.557
Intercept	576.000	1	576.000	67.765	.000	.850
gender	.000	1	.000	.000	1.000	.000
treatment	64.000	1	64.000	7.529	.018	.386
gender * treatment	64.000	1	64.000	7.529	.018	.386
Error	102.000	12	8.500			
Total	806.000	16				
Corrected Total	230.000	15				

a. R Squared = .557 (Adjusted R Squared = .446)

- The effect size for the independent variables is provided in the column labelled **Partial Eta Squared** (e.g., treatment, p= .386).

Tests of Between-Subjects Effects

Dependent Variable: depression

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	128.000[a]	3	42.667	5.020	.018	.557
Intercept	576.000	1	576.000	67.765	.000	.850
gender	.000	1	.000	.000	1.000	.000
treatment	64.000	1	64.000	7.529	.018	.386
gender * treatment	64.000	1	64.000	7.529	.018	.386
Error	102.000	12	8.500			
Total	806.000	16				
Corrected Total	230.000	15				

a. R Squared = .557 (Adjusted R Squared = .446)

- ANOVA output also provides Post-hoc tests if the post hoc option is checked in the analysis procedure since these are relevant only if we have more than two levels (groups) of an independent variable.
- However, there is no need to look at them until you find a significant main effect or interaction effect in the overall (omnibus) analysis of variance test.

Lesson 33

Correlation

Topic: 067-068

Topic 067: Introduction to Correlation

Correlation

- Correlation is a statistical technique that is used to measure and describe the relationship between two variables.
- Usually the two variables are simply observed as they exist naturally in the environment, there is no attempt to control or manipulate the variables.
- A correlation requires two scores for each individual (one score from each of the two variables).
- These scores normally are identified as *X* and *Y*.
- In the scatter plot, the values for the *X* variable are listed on the horizontal axis and the *Y* values are listed on the vertical axis.
- Each individual is then represented by a single point in the graph so that the horizontal position corresponds to the individual's *X* value and the vertical position corresponds to the *Y* value.



Although correlations have a number of different applications, a few specific examples include:

- Correlation is used for **prediction**.
- In general, the squared correlation (r^2) measures the gain in accuracy that is obtained from using the correlation for prediction.
- The squared correlation measures the proportion of variability in the data that is explained by the relationship between *X* and *Y*. It is sometimes called the ***coefficient of determination.***
- Correlation is used to demonstrate the **validity** of the test.
- Validity refers to whether or not the test measures what it claims to measure.
- Correlations are also used to determine **reliability**.
- A measurement procedure is considered reliable to the extent that it produces stable, consistent measurements.
- Furthermore, the **prediction of the theory** could be tested by determining the correlation between the two variables.
- The value of a correlation can be affected greatly by the range of scores represented in the data.

- One or two extreme data points, often called *outliers*, can have a dramatic effect on the value of a correlation.

Topic 068: Characteristics of Correlation

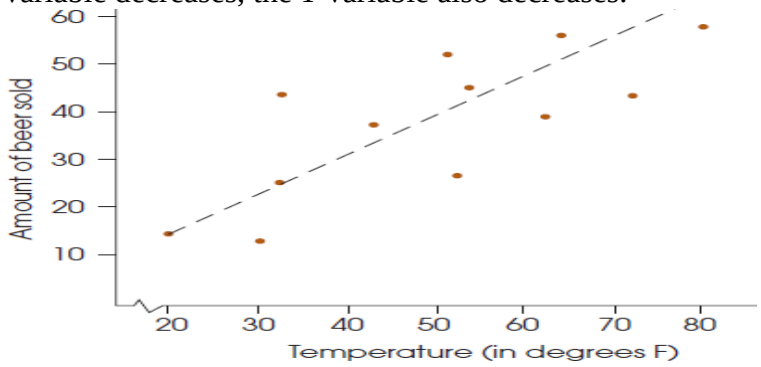
Characteristics

A correlation is a numerical value that describes and measures three characteristics of the relationship between X and Y.

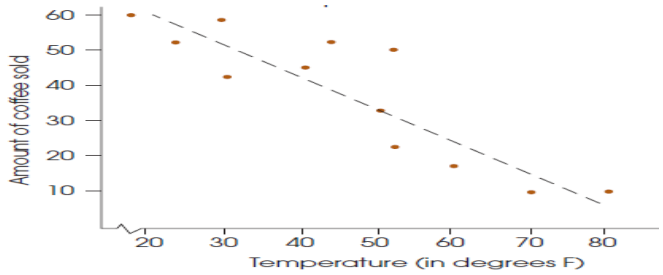
1. The Direction of the Relationship

The sign of the correlation, positive or negative, describes the direction of the relationship.

- In a **positive correlation,** the two variables tend to change in the same direction:
- As the value of the *X* variable increases from one individual to another, the *Y* variable also tends to increase;
- When the *X* variable decreases, the *Y* variable also decreases.



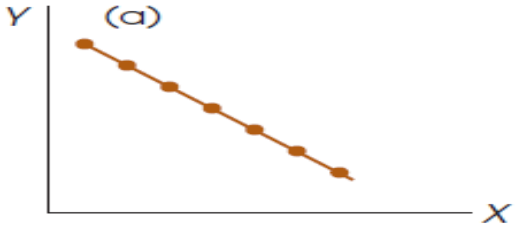
In a **negative correlation,** the two variables tend to go in opposite directions. As the *X* variable increases, the *Y* variable decreases. That is, it is an inverse relationship.



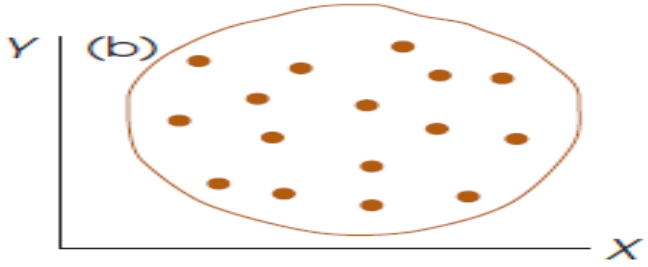
2. The Form of the Relationship

- The most common use of correlation is to measure straight-line relationships.
- The relationships tend to have a linear form; that is, the points in the scatter plot tend to cluster around a straight line.
- **3. The Strength or Consistency of the Relationship**
- The correlation measures the consistency of the relationship.

- For a linear relationship, for example, the data points could fit perfectly on a straight line. Every time *X* increases by one point, the value of *Y* also changes by a consistent and predictable amount.
- **3. The Strength or Consistency of the Relationship**
- However, relationships are usually not perfect.
- Although there may be a tendency for the value of *Y* to increase whenever *X* increases, the amount that *Y* changes is not always the same, and occasionally, *Y* decreases when *X* increases.
- **3. The Strength or Consistency of the Relationship**
- However, relationships are usually not perfect.
- Although there may be a tendency for the value of *Y* to increase whenever *X* increases, the amount that *Y* changes is not always the same, and occasionally, *Y* decreases when *X* increases.



- However, relationships are usually not perfect.
- Although there may be a tendency for the value of *Y* to increase whenever *X* increases, the amount that *Y* changes is not always the same, and occasionally, *Y* decreases when *X* increases.
- The consistency of the relationship is measured by the numerical value of the correlation.
- A **perfect correlation** always is identified by a correlation of 1.00 and indicates a perfectly consistent relationship.
- At the other extreme, a correlation of 0 indicates no consistency at all.
- For a correlation of 0, the data points are scattered randomly with no clear trend.



The sign (+ or -) and the strength of a correlation are independent.

- A correlation of 1.00 indicates a perfectly consistent relationship whether it is positive (+1.00) or negative (-1.00). Similarly, correlations of +0.80 and -0.80 are equally consistent relationships.

Correlation and Causation

- One of the most common errors in interpreting correlations is to assume that a correlation necessarily implies a cause-and-effect relationship between the two variables.
- Although there may be a causal relationship between two variables, the simple existence of a correlation does not prove it.
- To establish a cause-and-effect relationship, it is necessary to conduct a true experiment in which one variable is manipulated by a researcher and other variables are rigorously controlled.
- A study shows a positive correlation between number or churches and number of crimes in towns and cities. It is reasonable that small towns would have less crime and fewer churches and that large cities would have large values for both variables.
- Although a strong correlation exists between number of churches and crime, **the real cause of the relationship is the size of the population.**

Lesson 34

The Pearson Product Moment Correlation

Topic: 069-071

Topic 069: The Pearson Product Moment Correlation -1

The Pearson Correlation

- The most common correlation is the *Pearson correlation* (or the Pearson product-moment correlation) which measures the degree of straight-line relationship.
- In other words, it measures the degree and the direction of the linear relationship between two variables.
- The Pearson correlation is identified by the letter $\boldsymbol{r}$.

Conceptually, this correlation is computed by

$$r = \frac{degree\ to\ which\ X\ and\ Y\ vary\ together}{degree\ to\ which\ X\ and\ Y\ vary\ separately}$$

OR

$$r = \frac{covariability\ of\ X\ and\ Y}{variability\ of\ X\ and\ Y\ separately}$$

- When there is a perfect linear relationship, every change in the *X* variable is accompanied by a corresponding change in the *Y* variable.
- In this case, the covariability (*X* and *Y* together) is identical to the variability of *X* and *Y* separately, and the formula produces a correlation with a magnitude of +1.00 or –1.00.
- When there is no linear relationship, a change in the *X* variable does not correspond to any predictable change in the *Y* variable.
- In this case, there is no covariability, and the resulting correlation is zero.
- To calculate the Pearson correlation, it is necessary to know about the *sum of products* of deviations, or *SP*.
- The *SP* measures the amount of covariability between two variables.
- The *computational formula* for the sum of products of deviations is
- $SP = \sum XY - \frac{\sum X \sum Y}{n}$

Topic 070: The Pearson Product Moment Correlation-2

The Pearson Correlation

- Suppose we have n=4 pairs of scores as follows:

• X	• Y	• XY
• 1	• 3	• 3

• 2	• 6	• 12
• 4	• 4	• 16
• 5	• 7	• 35
• 12	• 20	• 66 **Total**

- In the formula for the Pearson *r*, we use *SP* to measure the covariability of *X* and *Y*.
- The variability of *X* is measured by computing *SS* for the *X* scores
- And the variability of *Y* is measured by *SS* for the *Y* scores.

The Pearson Correlation

- The final formula is,
- $r = \frac{SP}{\sqrt{SSx\ SSy}}$
- Where SS are $\Sigma(X-M)^2$ and $\Sigma(Y-M)^2$.

Topic 071: Hypothesis Testing With Pearson Correlation Part-1

Hypothesis Testing

- The Pearson correlation is generally computed for sample data.
- A sample correlation is often used to answer questions about the corresponding population correlation.
- The basic question for this hypothesis test is whether a correlation exists in the population.
- **The null hypothesis** is 'No. There is no correlation in the population,' or 'The population correlation is zero.'
- **The alternative hypothesis** is 'Yes. There is a real, nonzero correlation in the population.'
- The population correlation is traditionally represented by **ρ (the Greek letter rho).**
- Hypotheses would be stated in symbols as,
- H_o: $\rho = 0$ (There is no population correlation)
- H_1: $\rho \neq 0$ (There is real correlation)
- When there is a specific prediction about the direction of the correlation, it is possible to do a directional, or one-tailed, test. For example,
- H_o: $\rho \leq 0$ (The population correlation is not positive)
- H_1: $\rho < 0$ (The population correlation is positive)
- Samples are not expected to be identical to the populations from which they come; there is some discrepancy **(sampling error)** between a sample statistic and the corresponding population parameter.
- The hypothesis test for the Pearson correlation has degrees of freedom defined by

- ***df* = *n* – 2**
- The computations for evaluating *r* have already been completed and are summarized in **Table B.6 in Appendix B.**
- The table is based on the concept that a sample is expected to be representative of the population from which it was obtained.

	Level of Significance for One-Tailed Test			
	.05	.025	.01	.005
	Level of Significance for Two-Tailed Test			
df = *n* − 2	.10	.05	.02	.01
1	.988	.997	.9995	.9999
2	.900	.950	.980	.990
3	.805	.878	.934	.959
4	.729	.811	.882	.917
5	.669	.754	.833	.874
6	.622	.707	.789	.834
7	.582	.666	.750	.798
8	.549	.632	.716	.765
9	.521	.602	.685	.735
10	.497	.576	.658	.708
11	.476	.553	.634	.684
12	.458	.532	.612	.661
13	.441	.514	.592	.641

Example

- Calculate the Pearson correlation for these data given in the table.
- Is the correlation statistically significant? Use a two-tailed test with alpha level .05.

Person	X	Y	XY
A	0	4	0
B	2	1	2
C	8	10	80
D	6	9	54
E	4	6	24
	20	30	160 **Total**

Lesson 35

Correlation in SPSS

Topic: 072-073

Topic 072: Correlation in SPSS Part-1

Correlation in SPSS

The Pearson correlation (r) is designed for interval level (continuous) variables.
It can also be used if you have one continuous variable (e.g. scores on a measure of self-esteem) and one dichotomous variable (e.g. gender: Male/Female)-that is the point-biserial correlation. The Pearson correlation (r) is designed for **interval level (continuous)** variables. It can also be used if you have one continuous variable (e.g. scores on a measure of self-esteem) and one dichotomous variable (e.g. gender: Male/Female)-that is the point-biserial correlation. To run correlation in SPSS we need two variables: both continuous, or one continuous and the other dichotomous (two values).

Assumptions

- **The scale of measurement** for the variables should be interval or ratio (continuous).
- One exception to this is if we have one dichotomous independent variable (with only two values e.g., gender) and one continuous dependent variable, we can still run correlation.
- Each subject must provide a score on both variable X and variable Y **(related pairs).**
- Both pieces of information must be from the same subject.
- The observations that make up the data must be **independent** of one another. That is, each observation or measurement must not be influenced by any other observation or measurement.
- Scores on each variable should be **normally distributed.**
- The relationship between the two variables should be **linear**.
- This means that when we look at a scatterplot of scores there should be see a straight line (roughly), not a curve.

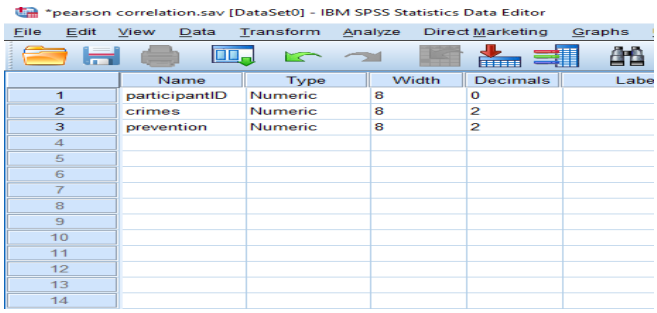
Example

- A researcher records the annual number of serious crimes and the amount spent on crime prevention. The data are given in table.
- Compute the Pearson correlation to measure the degree of relationship between the two variables. Is the correlation statistically significant? Use a two-tailed test with alpha .01.

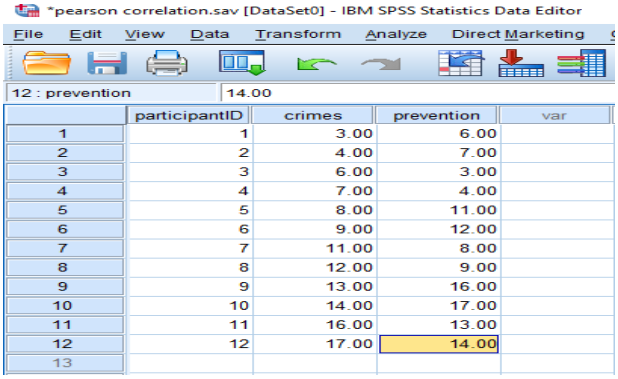
Number of crimes	Amount spent for prevention
3	6
4	7
6	3
7	4
8	11
9	12
11	8
12	9
13	16
14	17
16	13
17	14

Data Entry

- First of all, we need to create two variables on scale measurement namely crime (independent variable) and prevention (dependent variable) in variable view.



- The data are entered into two columns in the data editor, one for the *X* values (variable crimes) and one for the *Y* values (Variable prevention), with the two scores for each individual in the same row.



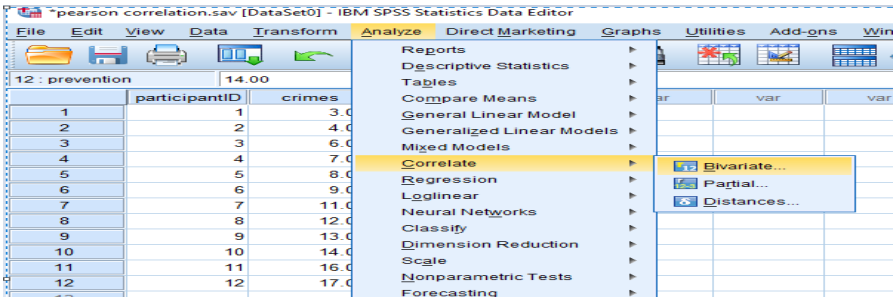
	participantID	crimes	prevention	var
1	1	3.00	6.00	
2	2	4.00	7.00	
3	3	6.00	3.00	
4	4	7.00	4.00	
5	5	8.00	11.00	
6	6	9.00	12.00	
7	7	11.00	8.00	
8	8	12.00	9.00	
9	9	13.00	16.00	
10	10	14.00	17.00	
11	11	16.00	13.00	
12	12	17.00	14.00	
13				

Topic 073: Running Correlation in SPSS

Running Analysis

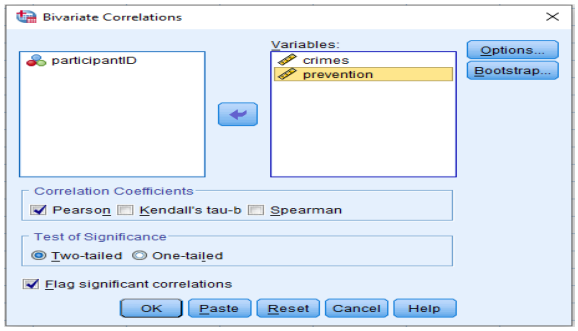
- From the menu at the top of the screen, click on **Analyze**, then select **Correlate**, then **Bivariate**.

- From the menu at the top of the screen, click on **Analyze**, then select **Correlate**, then **Bivariate**.



- Select two variables and move them into the box marked **Variables** (e.g., crimes and prevention).

- In the **Correlation Coefficients** section, the **Pearson** box is the default option.



- Click on the **Options** button. For **Missing Values**, click on the **Exclude cases pairwise** box.

- Under **Options**, we can also obtain means and standard deviations if we wish.

- Click on **Continue** and then on **OK**



Interpreting Results

- SPSS provides a table giving the correlation coefficients between each pair of variables listed, the significance level and the number of cases.

Correlations

		crimes	prevention
crimes	Pearson Correlation	1	.765**
	Sig. (2-tailed)		.004
	N	12	12
prevention	Pearson Correlation	.765**	1
	Sig. (2-tailed)	.004	
	N	12	12

**. Correlation is significant at the 0.01 level (2-tailed).

- In the example given here, the Pearson correlation coefficient (r=.77, p<.01) indicating a significant positive correlation between number of crimes and the amount spent on prevention.

- The more number of crimes people commit the more time is spent to prevent these crimes.

Correlations

		crimes	prevention
crimes	Pearson Correlation	1	.765**
	Sig. (2-tailed)		.004
	N	12	12
prevention	Pearson Correlation	.765**	1
	Sig. (2-tailed)	.004	
	N	12	12

**. Correlation is significant at the 0.01 level (2-tailed).

- The size of the value of the correlation coefficient ranges from –1 to 1. This value indicates the strength of the relationship between your two variables.
- Cohen (1988) suggests the following guidelines;
- Small r=.10 to .29
- Medium r=.30 to .49
- Large r=.50 to 1.0

Correlations

		crimes	prevention
crimes	Pearson Correlation	1	.765**
	Sig. (2-tailed)		.004
	N	12	12
prevention	Pearson Correlation	.765**	1
	Sig. (2-tailed)	.004	
	N	12	12

**. Correlation is significant at the 0.01 level (2-tailed).

Reporting Results

The relationship between number of crimes and amount spent on prevention was investigated using Pearson product-moment correlation coefficient. Preliminary analyses were performed to ensure no violation of the assumptions. There was a strong significant positive correlation between the two variables, $r = .77$, $n = 12$ $p < .01$, with more number of crimes associated with more time spent on prevention.

© Virtual University of Pakistan

Lesson 36

Other Types of Correlations
Topic: 074-75

Topic 74,75: Other Types of Correlations

Partial Correlation

Sometimes a researcher may suspect that the relationship between two variables is being distorted by the influence of a third variable. A statistical technique, known as *partial correlation,* allows a researcher to measure the relationship between two variables while eliminating the influence of a third variable.
In a situation with three variables, X, Y, and Z, it is possible to compute three individual Pearson correlations:

1. rXY measuring the correlation between X and Y
2. rXZ measuring the correlation between X and Z
3. rYZ measuring the correlation between Y and Z

The statistical significance of a partial correlation is determined using the same procedure as is used to evaluate a regular Pearson correlation. Specifically, the partial correlation is compared with the critical values listed in **Table B6.**
For a partial correlation, however, we must use ***df*=*n* – 3 instead of the *n* – 2** value that is used for the Pearson correlation. A significant correlation means that it is very unlikely ($p < \alpha$) that the sample correlation would occur without a corresponding relationship in the population.

The Spearman Correlation

When the Pearson correlation formula is used with **data from an ordinal scale (ranks),** the result is called the *Spearman correlation*. It is identified by the symbol ***rS*** to differentiate it from the Pearson correlation. The Spearman correlation is used to measure the relationship between *X* and *Y* when both variables are measured on ordinal scales.

Additionally, it can be used as a valuable alternative to the Pearson correlation, even when the original raw scores are on an interval or a ratio scale. The Spearman correlation measures consistency, rather than form. When two variables are consistently related, their ranks are linearly related. For example, a perfectly consistent positive relationship means that every time the *X* variable increases, the *Y* variable also increases.
When there **is a consistently one-directional relationship between two variables, the relationship is said to be *monotonic***. Thus, the Spearman correlation measures the degree of a monotonic relationship between two variables.

The Point-Biserial Correlation

The point-Biserial correlation is used to measure the relationship between two variables in a situations in which one variable consists of regular, numerical scores, but the second variable has only two values.

A variable with only two values is called a *dichotomous variable* or a *binomial variable*. For example,

- Male versus female
- College graduate versus not a college graduate
- Older than 30 years versus younger than 30 years

How to calculate point-based correlation:

To compute the point-biserial correlation, the dichotomous variable is first converted to numerical values by assigning a value of zero (0) to one category and a value of one (1) to the other category.

Then the regular Pearson correlation formula is used with the converted data

- Squaring the value of the point-biserial correlation **produces r^2** which is exactly the value of r^2 we obtain by measuring the effect size of the independent sample t-test.
- As with most correlations, the *strength* of the relationship is best described by the value of r^2.
- r^2 is the coefficient of determination, which measures how much of the variability in one variable is predicted or determined by the association with the second variable.

The Phi-Coefficient

When both variables (X and Y) measured for each individual are dichotomous, the correlation between the two variables is called the *phi-coefficient.* It is represented by the symbol (ϕ).

A two-step procedure is followed:

1. Convert each of the dichotomous variables to numerical values by assigning a 0 to one category and a 1 to the other category for each of the variables.
2. Use the regular Pearson formula with the converted scores.

The phi-coefficient can be used to assess the relationship between two dichotomous variables. The more common statistical procedure is a chi-square statistic.

Lesson 37

Regression-I
Topic: 076-77

Topic 76: Introduction to Regression

Regression

The line drawn in the scatters plot defines the relationship between two variables. Regression is a Statistical technique that draws a best-fit line to define the relationship between variables. It establishes a precise, one-to-one relationship between each *X* value and a corresponding *Y* value.

The line can be used for prediction. It establishes a precise, one-to one relationship between each X value (SAT score) and a corresponding Y value (GPA). The statistical technique for finding the best-fitting straight line for a set of data is called regression. And the resulting straight line is called the **regression line.**

A *linear relationship* between two variables *X* and *Y* can be expressed by the equation

- ***Y= bX + a***
- Where *a* and *b* are fixed constants

In the general linear equation, **the value of *b* is called the *slope***. The slope determines how much the *Y* variable changes when *X* is increased by 1 point. **The value of *a* in the general equation is called the Y-*intercept*** because it determines the value **of *Y* when *X* 0.**

Topic 77: Least Square Method and Linear Equation

To determine how well a line fits the data points, there is a need to define mathematically the distance between the line and each data point. For every *X* value in the data, the linear equation determines a *Y* value on the line. This value is the predicted *Y* and is called $\hat{Y}$ ("*Y* hat"). The distance between the predicted value and the actual *Y* value in the data is determined by distance = $Y - \hat{Y}$

Because some of the distances are positive and some are negative, we need to square each distance to obtain a uniformly positive measure of error. Total squared error = $\Sigma(Y - \hat{Y})^2$. We can define the *best-fitting* line as the one that has the smallest total squared error. The resulting line is commonly called the ***least-squared-error solution or least squared method*.**

A final linear equation becomes

- $\hat{Y} = bX + a$
- For each value of *X* in the data, this equation determines the point on the line ($\hat{Y}$) *that gives the best prediction of Y*.

The value of b is determined by

- b= SP/SSx
- Where *SP* the sum of products and SSx is the sum of squares for the *X* scores.

The value of the constant *a* is determined by

- a= My-bMx

Example

X	Y
2	3
6	11
0	6
4	6
7	12
5	7
5	10
3	9

In general, there is some error between the predicted *Y* values and the actual data. Although the amount of error varies from point to point, on average the errors are directly related to the magnitude of the correlation. With a correlation near 1.00 (or 1.00), the data points generally are clustered close to the line and the error is small. As the correlation gets nearer to zero, the points move away from the line and the magnitude of the error increases.

Lesson 38

Regression-II

Topic: 078-79

Topic 78: Standard Error of Estimate

A regression equation, by itself, allows to make predictions, but it does not provide any information about the accuracy of the predictions. To measure the precision of the regression, it is customary to compute a *standard error of the estimate.*
The **standard error of estimate** gives a measure of the standard distance between the predicted *Y* values on the regression line and the actual *Y* values in the data. The standard error of the estimate is very much like a standard deviation both provide a measure of standard distance.

How to calculate the standard error of estimate

To calculate the standard error of estimate, we need the sum of squared deviations (*SS*). This sum of squares is commonly called ***SSresidual*** **because** it is based on the remaining distance between the actual *Y* scores and the predicted values.

***SS*residual** = $\Sigma(\boldsymbol{Y} - \hat{Y})^2$

The obtained *SS* value is then divided by its degrees of freedom to obtain a measure of variance.

SSresidual = SS/df

The degrees of freedom for the standard error of estimate are *df* = *n*-2.
The final equation is

$$Standard\ error\ of\ estimate = \frac{\sqrt{SS}}{\sqrt{df}} = \frac{\sqrt{\Sigma(Y - \hat{Y})^2}}{\sqrt{n-2}}$$

Since the standard error of estimate provides a measure of how accurately the regression equation predicts the *Y* values. In this case, the standard distance between the actual data points and the regression line is measured by a standard error of an estimated.

Coefficient of determination

Squaring the correlation provides a measure of the accuracy of prediction. The squared correlation, r^2, is called **the coefficient of determination.** Because it determines what proportion of the variability in *Y* is predicted by the relationship with *X*. Because of r^2 measures the predicted portion of the variability in the *Y* scores, **we can use the expression (1-r^2) to measure the unpredicted portion.**

Predicted variability= *SS*regression =r^2SSY

Unpredicted variability = *SS*residual $(1 - r^2)SSY$

The standard error of estimate can also be computed as

$$Standard\ error\ of\ estimate = \frac{\sqrt{SS}}{\sqrt{df}} = \frac{\sqrt{(1 - r^2)SSY}}{\sqrt{n-2}}$$

Topic 79: Assumptions of Regression

Assumptions

There are the following assumption of regression:

- Data should be linear.
- For any two observations, the residual terms should be uncorrelated (i.e., independent). This eventuality is sometimes described as a lack of autocorrelation. The assumption of autocorrelation can be tested with the Durbin–Watson test, which tests for serial correlations between errors. Specifically, it tests whether adjacent residuals are correlated. The test statistic varies between 0 and 4, with a value of 2 meaning that the residuals are uncorrelated.
- At each level of the predictor variable(s), the variance of the residual terms should be constant. This assumption means that the residuals at each level of the predictor(s) should have the same variance (homoscedasticity). When the variances are very unequal there is said to be heteroscedasticity.
- Normally distributed errors: It can be helpful if the residuals in the model are random, normally distributed variables with a mean of 0. This assumption means that the differences between the predicted and observed data are most frequently zero or very close to zero, and that differences much greater than zero happen only occasionally.
- All predictor variables must be quantitative (continuous) or categorical (with two categories), and the outcome variable must be quantitative (continuous).
- No perfect multicollinearity: If the model has more than one predictor then there should be no perfect linear relationship between two or more of the predictors. So, the predictor variables should not correlate too highly. To identify multicollinearity we need to scan the correlation matrix for predictor variables that correlate very highly (values of r above 0.80 or 0.90).

SPSS Statistics also compute the variance inflation factor (VIF), which indicates whether a predictor has a strong linear relationship with the other predictor(s).SPSS also gives the tolerance statistic, which is the reciprocal of VIF, (1/VIF).

Some general guidelines for interpreting the VIF include:

- If the largest VIF is greater than 10 (or the tolerance is below 0.1) then this indicates a serious problem.
- If the average VIF is substantially greater than 1 then the regression may be biased
- Tolerance below 0.2 indicates a potential problem

© Virtual University of Pakistan

Lesson 39

Multiple & Simple Linear Regression through SPSS
Topic: 80-81

Topic 80: Multiple Regression

Multiple regression is a family of techniques that can be used to explore the relationship between one continuous dependent variable and a number of independent variables or predictors. It allows a more sophisticated exploration of the interrelationship among a set of variables.

We cannot just throw variables into a multiple regression. However, we should have a sound theoretical or conceptual reason for the analysis and, in particular, the order of variables entering the model. There are a number of different types of multiple regression analyses that can be used, depending on the nature of the question. **The three main types of multiple regression analyses are:**

Multiple linear regression

Multiple linear regression is also called standard or simultaneous regression. All the independent (or predictor) variables are entered into the **model simultaneously**. Each independent variable is evaluated in terms of its predictive power, over and above that offered by all the other independent variables.

Hierarchical multiple regression

Hierarchical multiple regression (also called sequential regression). Variables or sets of variables are entered in **steps (or blocks**), with each independent variable being assessed in terms of what it adds to the prediction of the dependent variable after the previous variables have been controlled for.

Stepwise multiple regression

The researcher provides a list of independent variables and then allows the program to select which variables it will enter and in which order they go into the **equation,** based on a set of statistical criteria. Stepwise multiple regression selects the variable which is **statistically stronger.**

There are three different steps of this approach:

- Forward selection
- Backward deletion
- Stepwise regression

The forward selection method is often used to provide an initial screening of the variables when a large group of variables exists. SPSS selects the variable that has the highest R-Squared. At each step, it selects the predictor variable that increases R-Squared the most. Finally, it stops adding variables when none of the remaining variables are significant.

The backward deletion procedure begins with a model that includes all the independent variables. It deletes one variable at a time by determining whether the least significant variable currently in the model can be removed because its p-value is less than the user-specified or default value.

Stepwise regression is a modification of the forward selection so that after each step in which a variable was added, all variables in the model are checked to see if their significance has been reduced below the specified tolerance level. If a non-significant variable is found, it is removed from the model.

Topic 81: Simple Linear Regression through SPSS

Linear regression

Linear regression is the next step up after correlation. It is used when we want to predict the value of one variable based on the value of another variable.

Dependent variable

The variable we want to predict is called the dependent variable (or sometimes, the outcome variable).

Independent variable

The variable we are using to predict the other variable's value is called the independent variable (or sometimes, the predictor variable).

For simple linear regression dependent variable should be measured at the continuous level (i.e., it is either interval or ratio variable). The independent variable could be categorical (having only two categories) or continuous. **If the dependent variable is categorical**, that is having two levels, the appropriate procedure would be a **logistic regression (non-parametric test) instead of linear regression.**

Example

For the following set of data for income (DV) and weight (IV). To simplify the weight variable, the women are classified into five categories that measure actual weight relative to height, from 1=thinnest to 5=heaviest. Find the linear regression. Does the regression account for a significant portion of the variance? Use alpha level .05 to evaluate the *F*-ratio.

Weight	Income

1	125
2	78
4	49
3	63
5	35
2	84
5	38
3	51
1	93
4	44

Data Entry

For the predictor variable (weight), enter the values in one column and for the outcome variable (income), enter the values in a second column of the SPSS data editor as shown.

Lesson 40

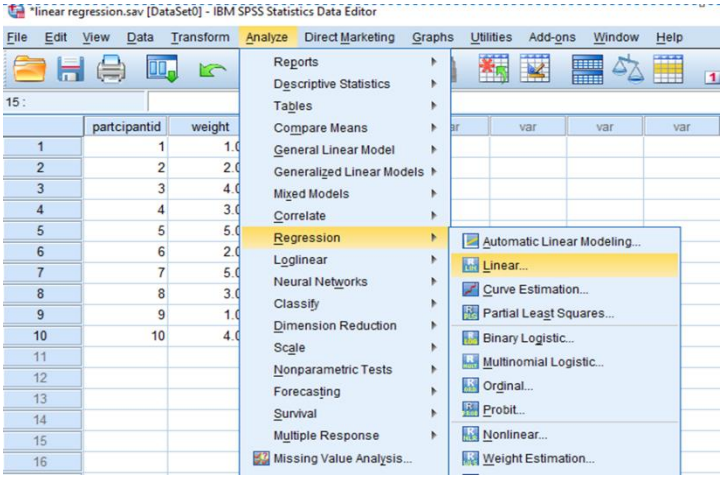
Interpreting Results in Simple Linear Regression
Topic: 82-83

Topic 82: Simple Linear Regression through SPSS

Running Analysis:

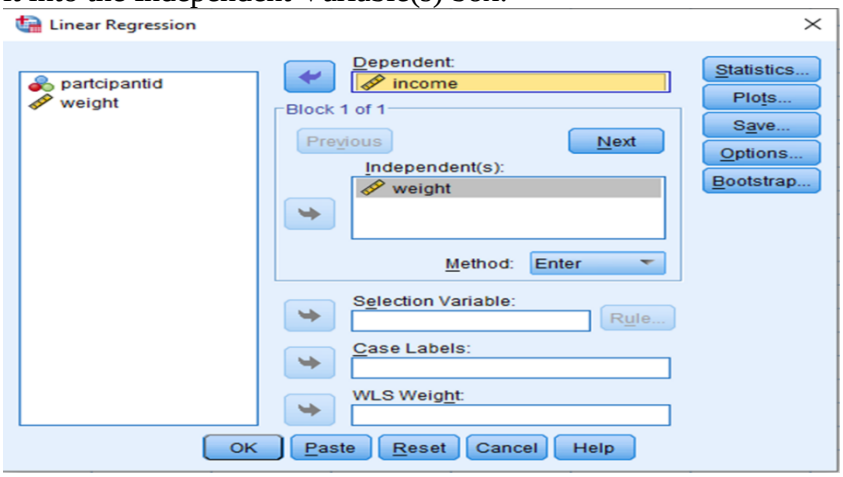
There are the following steps to analyze our data in SPSS by using simple linear regression.

1. Click Analyze on the toolbar, select Regression and click on Linear.



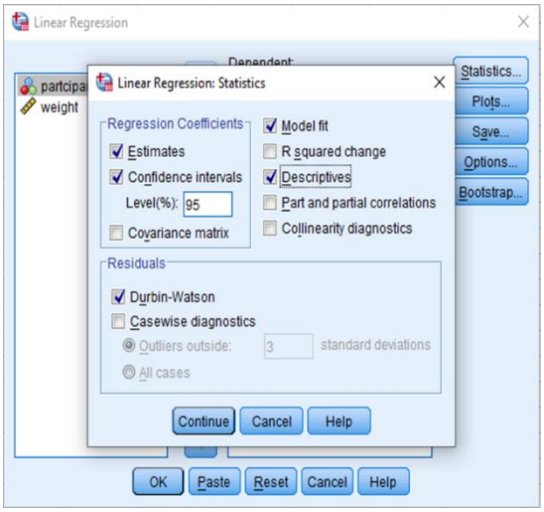
2. In the left-hand box, highlight the column label for the outcome variable, then click the arrow to move the column label into the Dependent Variable box.

For the predictor variable, highlight the column label for the weight and click the arrow to move it into the Independent Variable(s) box.



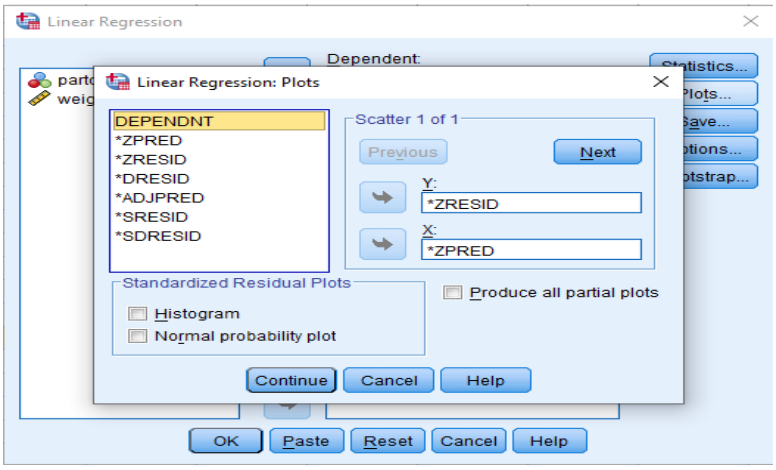
3. From the Statistics section, check the Estimates, Confidence intervals, and Descriptive and model fit boxes.

To assess the assumption of independence of observations (no autocorrelation) check the Durbin-Watson box.
Click Continue.



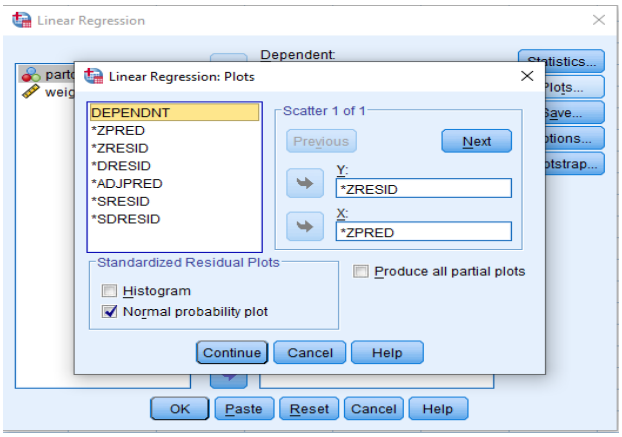
4. In Plots section, move the ZRESID (standardized residuals) and ZPRED (standardized predicted values) in Y and X boxes respectively.

Purpose of scatter plot:
The scatter plot will assess the homoscedasticity in the data.



5. In the section headed Standardized Residual Plots, tick the Normal probability plot option **to assess the normality of the data.**

Click on Continue and OK.



Topic 83: Interpreting Results in Simple Linear Regression through SPSS

Interpreting Results:

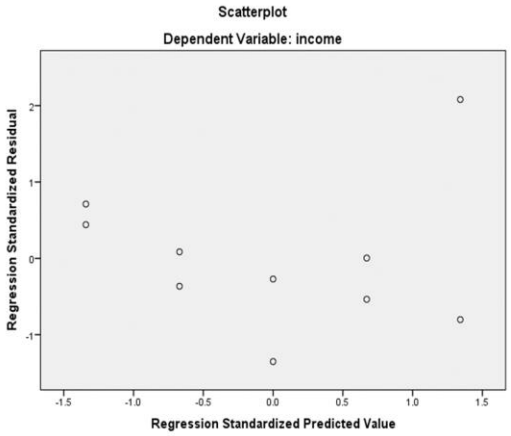
There are four main tables in the output field:

1. Descriptive
2. Model summary
3. ANOVA
4. Coefficient

Descriptive table;

In the **Scatterplot** of the standardized residuals, we are hoping that the residuals will be roughly rectangular distributed, with most of the scores concentrated in the center (along with the 0 points).

- There should not be a curvilinear pattern in the data.
- Deviations from a centralized rectangle suggest some violation of the assumptions.



- In the Normal P-P Plot, we check whether the data points lie in a reasonably straight diagonal line from bottom left to top right.
- This would suggest no major deviations from normality.

Model Summary box

Now we need to look at the **Model Summary** box. This table provides R and R square values. The R-value represents the correlation and is (R=.931) which indicates a high correlation. **The R square value tells us how much of the variance in the DV** (income) is explained by the IV (weight).

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.931[a]	.867	.851	11.10152	1.663

a. Predictors: (Constant), weight

b. Dependent Variable: income

The above table also gives the Durbin-Watson value which is 1.6 and is under the acceptable range. This indicates that there is no autocorrelation and the assumption is satisfied.

ANOVA table

The next table is the ANOVA table, which reports **how well the regression equation fits the data (i.e., predicts the dependent variable).**
This tests the null hypothesis that R in the population equals 0.
The model in this example reaches statistical significance (F=52.29, p<.001).

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6444.050	1	6444.050	52.287	.000[b]
	Residual	985.950	8	123.244		
	Total	7430.000	9			

a. Dependent Variable: income

b. Predictors: (Constant), weight

Coefficients table

The Coefficients table provides the necessary information to predict income from the weight.
The weight is significantly negatively predicting income (B=-17.95, p<.001).
This means that for every 1-unit increase in weight, the income will decrease by 17.95 units.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	119.850	8.233		14.557	.000
	weight	-17.950	2.482	-.931	-7.231	.000

a. Dependent Variable: income

Simple linear regression was run to assess whether weight predicts income in women**. Preliminary analyses were conducted to ensure no violation of the assumptions of normality, linearity, and homoscedasticity**. The total variance explained by the model as a whole was 86.7%, F (1, 8) = 52.28, p < .001. Weight was found to be significant negative predictor of income (B=-17.95, p<.001).

Lesson 41

Non-Parametric Tests
Topic: 84-85

Topic 84: Non-Parametric Tests

Often, researchers are confronted with experimental situations that do not conform to the requirements of parametric tests. In these situations, it may not be appropriate to use a parametric test.

The main reasons to apply the nonparametric test include:

- The underlying data do not meet the assumptions about the population sample
- The population sample size is too small
- The analyzed data is ordinal or nominal

Nonparametric statistics include

- Nonparametric descriptive statistics
- Statistical models, inference, and statistical tests.
- The model structure of nonparametric models is not specified a priori but is instead determined from data.
- **Nonparametric test usually involves qualitative data or Nominal / Ordinal data**
- Most of the time even quantitative data is converted to nominal/ordinal data for use with nonparametric tests
- For example, numerical scores measuring self-esteem can be converted to categories like high, medium, and low self-esteem, etc.

Topic 85: Chi-Square Test

The most obvious difference between the chi-square tests and the other hypothesis tests we have considered (t and ANOVA) is the nature of the data. For chi-square, the **data are in frequencies rather than numerical scores.**

In general terms, this chi-square test uses the proportions obtained for sample data to test hypotheses about the corresponding proportions in the population.

Assumptions of Chi-Square Test:

1. **Independence of observations**

One consequence of independent observations is that each observed frequency is generated by a different individual. A chi-square test would be inappropriate if a person could produce responses that can be classified in more than one category.

2. **Less than 5.**

A chi-square test should not be performed when the expected frequency of any cell is **less than 5.** The chi-square statistic can be distorted when f_e is very small.

3. Chi-square statistic involves two tests: the chi-square test for goodness of fit and the chi-square test for independence.

© Virtual University of Pakistan

Lesson 42

Chi-Square Test-I
Topic: 86-87

Topic 86: Chi-Square Test for Goodness of Fit

Definition:
The chi-square test for goodness of fit uses sample data to test hypotheses about the proportions of population distribution. The test determines how well the obtained sample proportions fit the population proportions specified by the null hypothesis. The null hypothesis specifies the proportion (or percentage) of the population in each category.
For example, a hypothesis might state that 50% of all lawyers are men and 50% are women.

Generally, the null hypothesis states that there is no relationship among the different categories.

- H_o states that the population is divided equally among the categories.
- The alternative hypothesis says that the population is not divided equally.

The data for a chi-square test is very simple. There is no need to calculate a sample mean or SS; we just select a sample of *n* individuals and count how many are in each category. The resulting values are **called observed frequencies (f_o).**

The **general goal of the chi-square test for goodness** of fit is to compare the data (the observed frequencies) with the null hypothesis.
The problem is to determine how well the data fit the distribution specified in H_o—hence the name goodness of fit.

For example, the following data represent observed frequencies for a sample of n=40 college students. The students were classified into three categories based on the number of times they reported exercising each week.

No exercise	1 time a week	More than once a week
15	19	6

1. The first step in the chi-square test is to construct a **hypothetical sample** that represents how the sample distribution would look if it were in perfect agreement with the proportions stated in the null hypothesis.

Category A	Category B	Category C
25%	50%	25%

2. To find the exact frequency expected for each category, multiply the sample size (n) by the proportion (or percentage) from the null hypothesis. For this example, we would expect:

- 25% of 40=0.25(40)=10
- 20% of 40=0.50(40)=20
- 25% of 40=0.25(40)=10

Category A	Category B	Category C
25%	20%	25%

The frequency values predicted from the null hypothesis are called **expected frequencies.**
The symbol for expected frequency is $\mathbf{f_e}$ and is computed as
$\mathbf{f_e = pn}$
Where p is the **proportion** stated in the null hypothesis and n is **the sample size.**

Topic 87: Chi-Square Test for Goodness of Fit II

The chi-square statistic simply measures how well the data (f_o) fit the hypothesis (f_e).
The symbol for the chi-square statistic **is χ^2.**
The formula for the chi-square statistic is $\chi^2=\Sigma (f_o- f_e)^2/f_e$

For the goodness-of-fit test, **the degrees of freedom are determined by**
df = C – 1
Where C is the number of categories.
To determine whether a particular chi-square value is significantly large, we need to consult the table entitled The Chi-Square Distribution (Appendix B).

Example
A psychologist selected an abstract painting that had no obvious top or bottom. Hangers were placed on the painting so that it could be hung with any one of the four sides at the top. The painting was shown to a sample of n=50 participants, and each was asked to hang the painting in the orientation that looked correct. The following data indicate how many people chose each of the four sides to be placed at the top:

Top up (Correct)	Bottom up	Left side up	Right side up
18	**17**	**7**	**8**

H_o: In the general population, there is no preference for any specific orientation.
Thus, the four possible orientations are selected equally often, and the population distribution has the following proportions:
H_1: In the general population, one or more of the orientations is preferred over the others.

Top up (Correct)	Bottom up	Left side up	Right side up

25%	25%	25%	25%

Difference between Chi-Square and Single Sample t-test:

In particular, both tests are intended to use the data from a single sample to test hypotheses about a single population. If the sample data consist of numerical scores (from an interval or ratio scale), it is appropriate to compute a sample mean and use a t-test to evaluate a hypothesis about the population mean.
On the other hand, if the individuals in the sample are classified into non-numerical categories (on a nominal or ordinal scale), then the researcher would use a chi-square test to evaluate a hypothesis about the population proportions.

Lesson 43

Chi-Square Test-II

Topic: 88-89

Topic 88: Chi-Square Test for Independence

The chi-square test for independence uses the frequency data from a sample to evaluate the relationship between two variables in the population. Each individual in the sample is classified **on both of the two variables**, creating a two-dimensional frequency-distribution matrix.
The frequency distribution for the sample is then used to test hypotheses about the corresponding frequency distribution for the population.

Chi-Square Test for Goodness of Fit	Chi-Square Test for Independence
The chi-square test for goodness of fit uses sample data to test hypotheses about the proportions of population distribution	The chi-square test for independence uses the frequency data from a sample to evaluate the relationship between two variables in the population.

Usually, the data for this test are presented in the form of a matrix, where the rows correspond to the categories of one variable and the columns correspond to the categories of the second variable.

For example, a group of students could be classified in terms of personality (introvert, extrovert) and in terms of color preference (red, yellow, green, or blue).

	Red	Yellow	Green	Blue	
Introvert	10	3	15	22	50
Extrovert	90	17	25	18	150
	100	20	40	40	$n = 200$

The null hypothesis for the chi-square test for independence states that the two variables being measured are independent.

- **Null hypothesis H_o:** For the general population of students, there is no relationship between color preference and personality.
- **The alternative hypothesis,** H_1, states that there is a relationship between the two variables.

Topic 89: Calculating Chi-Square Test for Independence

The chi-square test for independence uses the same basic logic that was used for the goodness-of-fit test.

Observed frequencies

The frequencies in the sample distribution are called observed frequencies and are identified by the symbol f_o

Expected frequencies

The expected frequencies define an ideal hypothetical distribution that is in perfect agreement with the null hypothesis. Once the expected frequencies are obtained, we compute a chi-square statistic to determine how well the data (f_o) fit the null hypothesis (f_e).

The column totals for the matrix describe the overall distribution of color preferences.
For these data, 100 people selected red as their preferred color. Because the total sample consists of 200 people, the proportion selecting red is 100 out of 200, or 50%.

	Red	Yellow	Green	Blue	
Introvert	10	3	15	22	50
Extrovert	90	17	25	18	150
	100	20	40	40	$n = 200$

The complete set of color preference proportions is as follows:

100 out of 200=50% prefer red
20 out of 200=10% prefer yellow
40 out of 200=20% prefer green
40 out of 200=20% prefer blue

According to the null hypothesis, both personality groups should have the same proportions for color preferences.
Beginning with the sample of 50 introverts in the top row, we obtain the expected frequencies of
f_e=pn

- f_e(red)=50% of 50=.50(50)=25
- f_e(yellow)=10% of 50=.10(50)=5
- f_e(green)=20% of 50=.20(50)=10
- f_e(blue)=20% of 50=.20(50)=10

Using exactly the same proportions for the sample **of n 150** extroverts in the bottom row, we obtain the expected frequencies of
f_e=pn

- f_e(red)=50% of 150=.50(150)=75
- f_e(yellow)=10% of 150=.10(150)=15

- f_e(green)=20% of 150=.20(150)=30
- f_e(blue)=20% of 150=.20(150)=30

- Although expected frequencies are derived directly from the **null hypothesis** and the sample characteristics, there is a simple formula that determines f_e for any cell in the frequency distribution matrix:
  $f_e = f_c f_r / n$

- The chi-square test for independence uses exactly the same chi-square formula as the test for goodness of fit:
  $\chi^2 = \Sigma (f_o - f_e)^2 / f_e$

- The total number of f_e values that we can freely choose is (R – 1)(C – 1), and the degrees of freedom for the chi-square test of independence are given by the formula
  df=(R – 1)(C – 1)

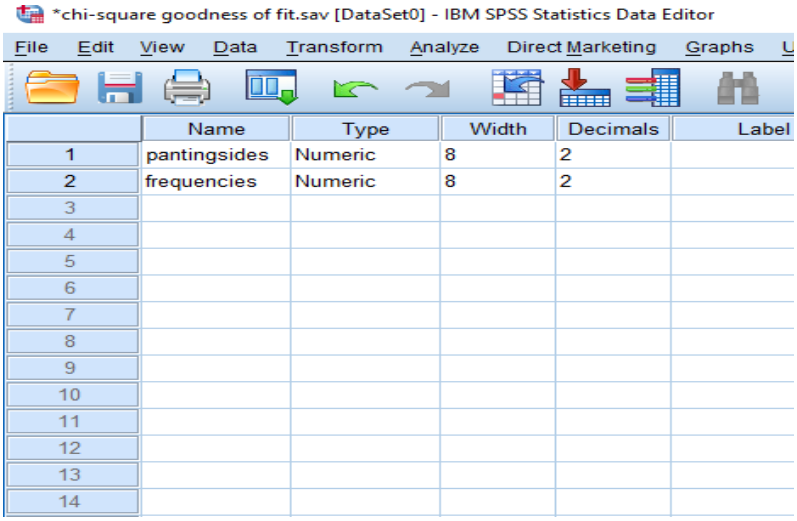
Lesson 44

Chi-Square Test-III
Topic: 90-91

Topic 90: Chi-Square Test for Goodness of Fit in SPSS

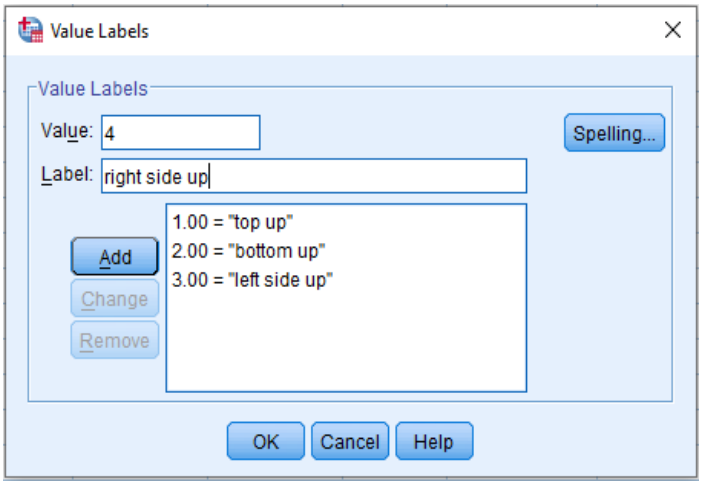
Entering Data

There are **two methods** of entering data into SPSS Statistics in order to run a chi-square goodness-of-fit test.
The first step is to create two variables. One categorical or ordinal variable having levels (e.g., pantingsides).

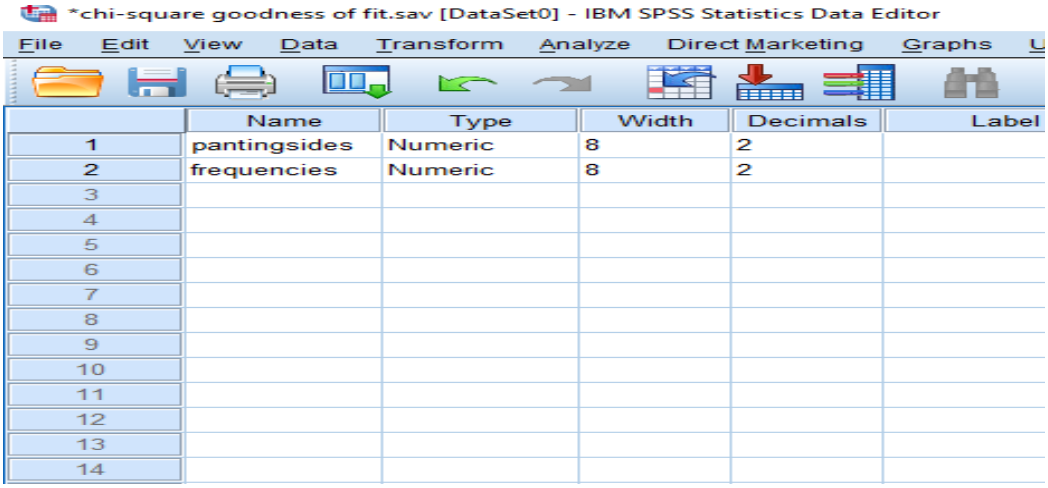


	Name	Type	Width	Decimals	Label
1	pantingsides	Numeric	8	2	
2	frequencies	Numeric	8	2	
3					
4					
5					
6					
7					
8					
9					
10					
11					
12					
13					
14					

In Value Labels section, assign codes to the levels of the variable as shown.



If the frequency data has already been summated for the various categories, we need to create a second column that contains the respective frequency counts.

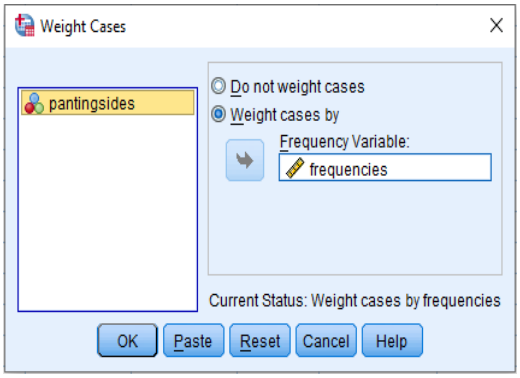


If data are entered in this way, we cannot run the chi-square goodness-of-fit test without first "weighting" our cases.

This is a procedure that tells SPSS Statistics that we have summated the frequencies:

- Click Data in the top menu and go to Weight Cases.
- Check the option Weight cases by and drag the frequencies and variable into the Frequency Variable box.
- Click Ok.

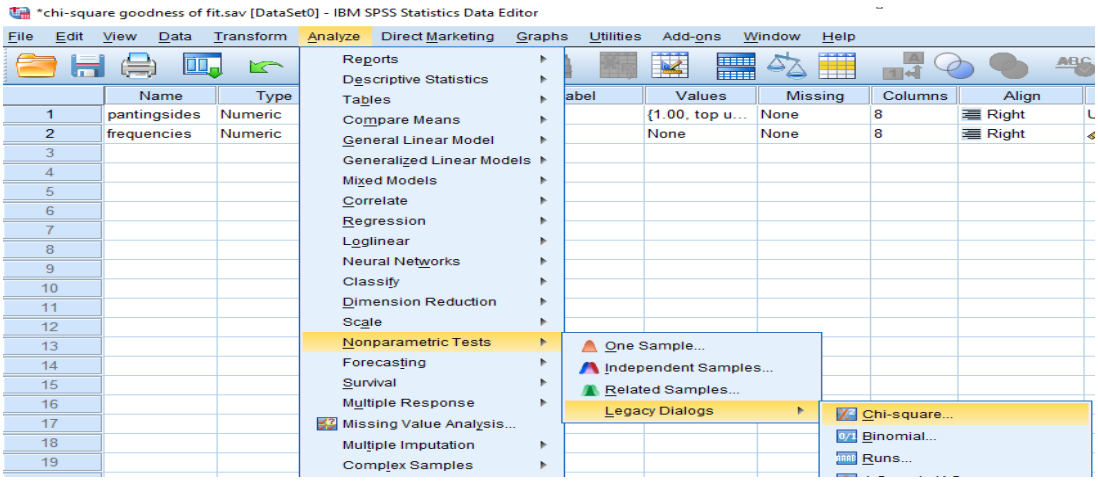
Data>weight cases>frequencies variable>OK



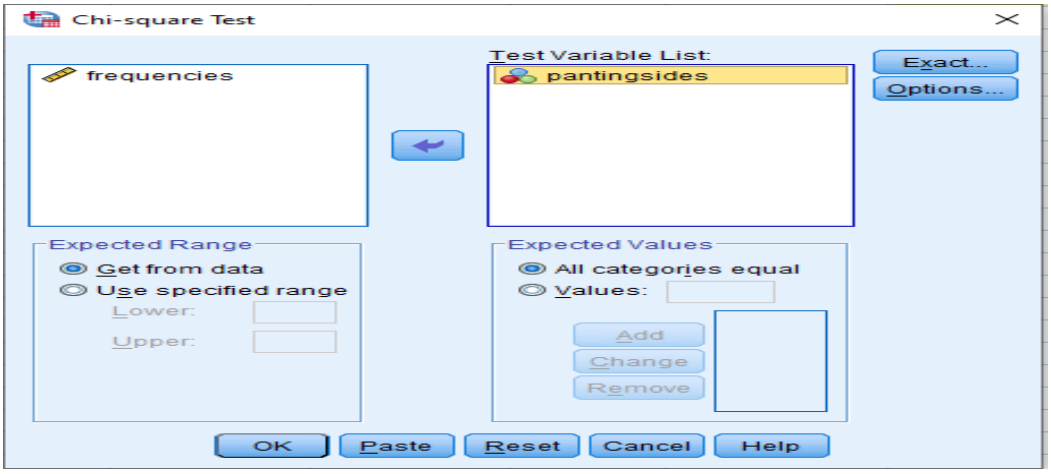
Alternatively, we may have the data in raw form (i.e., we have not summated the frequencies). In this case, we do not need a second column as SPSS Statistics can calculate the frequencies of occurrence of each category itself. This would mean that, in this example, there will be 50 rows of data.

When we select frequencies then we move to another step by following order:

Click Analyze > Nonparametric Tests > Legacy Dialogs > Chi-square



- There will appear the Chi-square Test dialogue box.
- Transfer the pantingsides variable into the Test Variable List box.
- Keep the All categories equal option selected in the Expected Values area as we are assuming equal proportions for each category.
- Click Ok.



Interpreting Results:

- The table pantingsides provides the observed frequencies for each side, as well as the expected frequencies which are the frequencies expected if the null hypothesis is true.
- The difference between the observed and expected frequencies is provided in the **Residual column.**

pantingsides

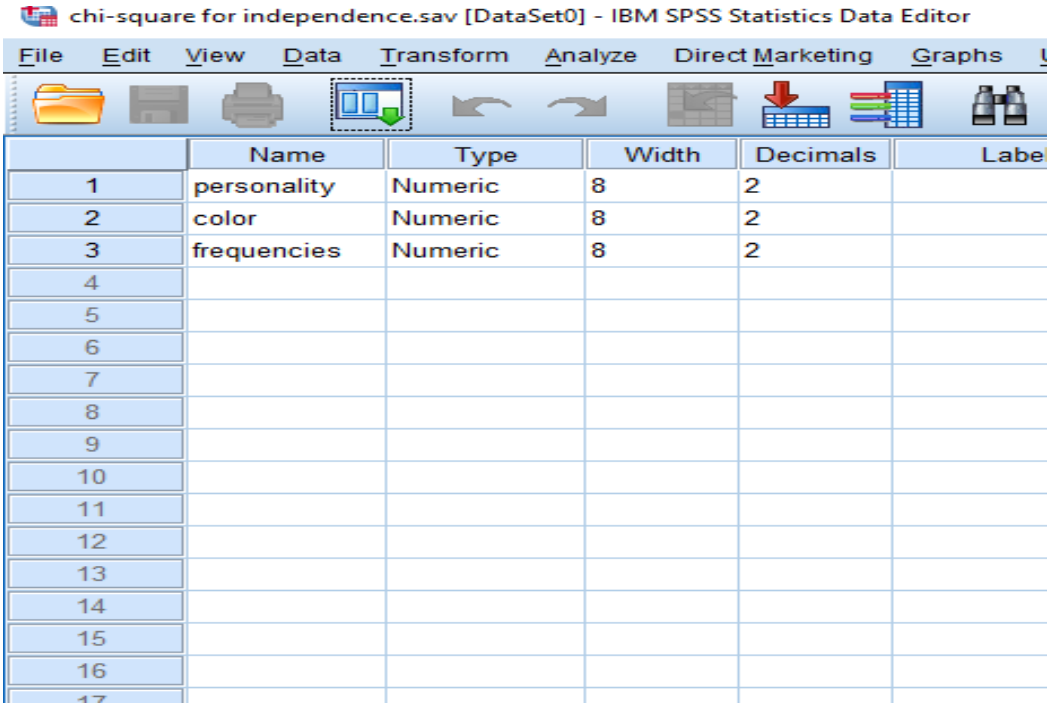
	Observed N	Expected N	Residual
top up	18	12.5	5.5
bottom up	17	12.5	4.5
left side up	7	12.5	-5.5
right side up	8	12.5	-4.5
Total	50		

- The table Test Statistics gives the actual result of the chi-square goodness-of-fit test.
- **The results indicate that our test statistic is statistically significant: $\chi 2$ (3) =8.08, p <.05.**
- Therefore, we can reject the null hypothesis and conclude that there are statistically significant differences in the preference of the type of panting side

Topic 91: Chi-Square Test for Independence in SPSS

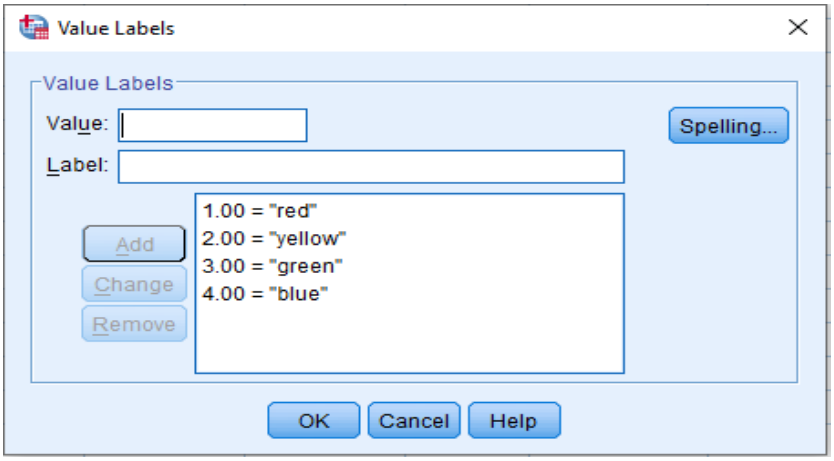
Entering Data

In the SPSS data editor**, create three variables**, one for frequencies and two categorical variables (e.g., personality and color).

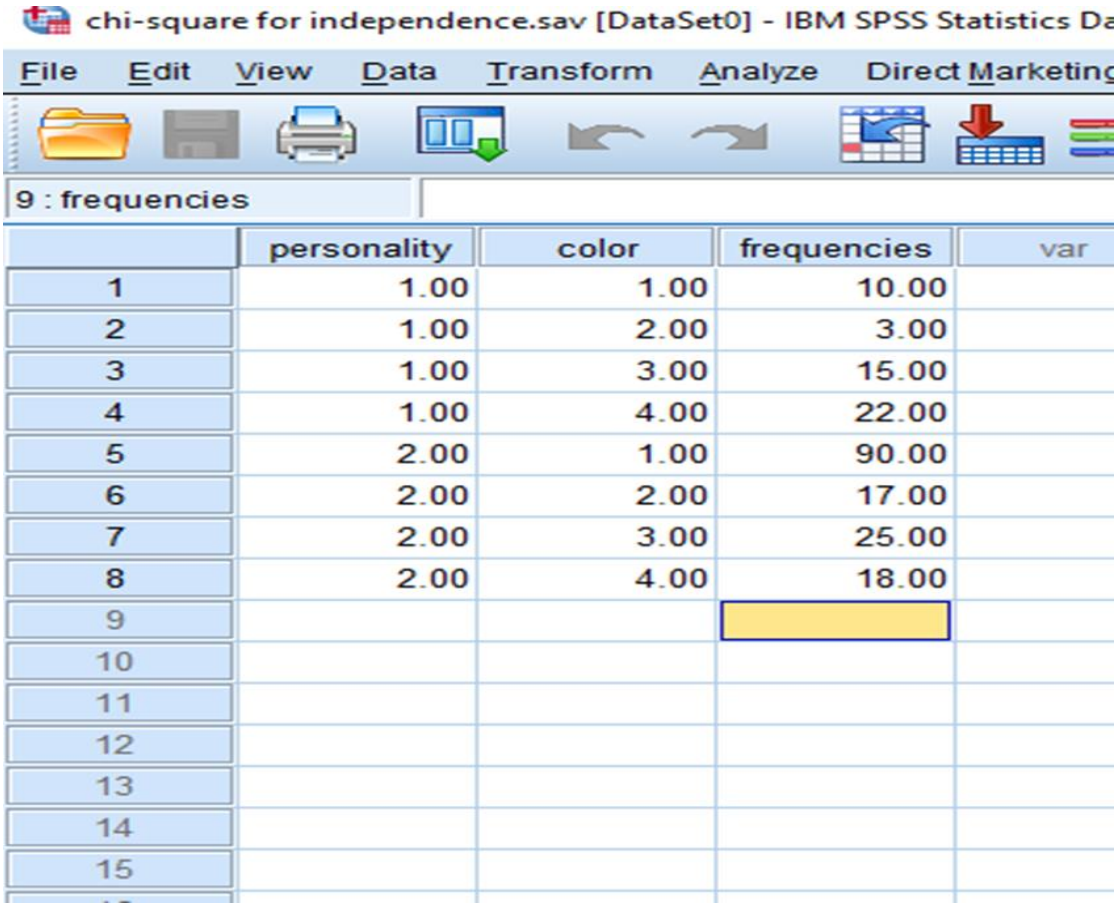


	Name	Type	Width	Decimals	Label
1	personality	Numeric	8	2	
2	color	Numeric	8	2	
3	frequencies	Numeric	8	2	
4					
5					
6					
7					
8					
9					
10					
11					
12					
13					
14					
15					
16					

In Value Labels section, assign codes to the levels of the variables.



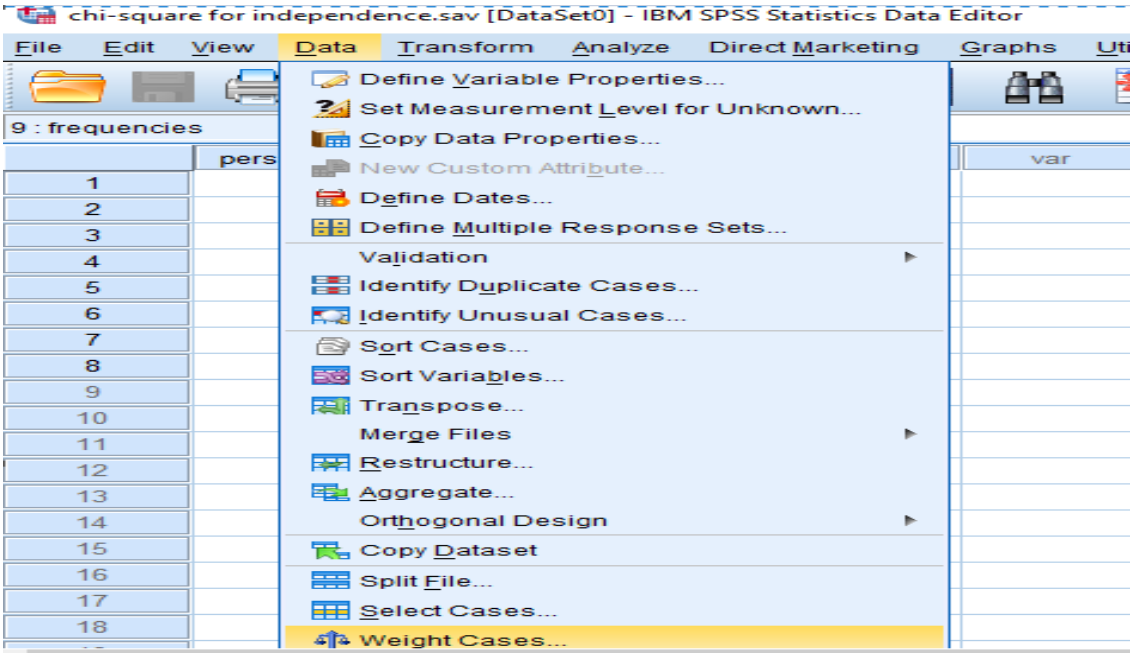
- In a first column, enter a code (e.g., 1 and 2) that identifies the row corresponding to each observed frequency.
- In the second column, enter a code (e.g., 1, 2, 3 and 4) that identifies the column corresponding to each observed frequency.
- Enter the complete set of observed frequencies in a column named frequencies.



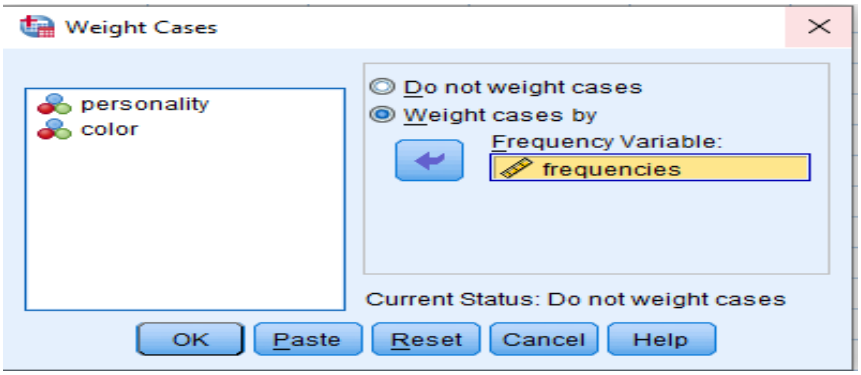
	personality	color	frequencies	var
1	1.00	1.00	10.00	
2	1.00	2.00	3.00	
3	1.00	3.00	15.00	
4	1.00	4.00	22.00	
5	2.00	1.00	90.00	
6	2.00	2.00	17.00	
7	2.00	3.00	25.00	
8	2.00	4.00	18.00	
9				
10				
11				
12				
13				
14				
15				

Running Analysis

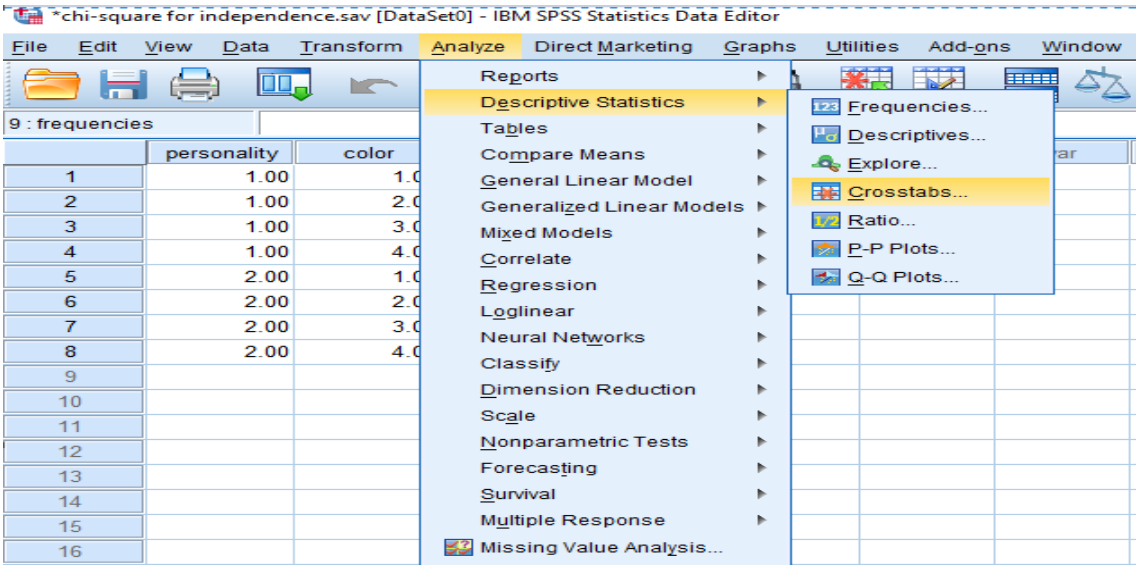
We need to Weight our cases to tell the SPSS Statistics that we have summated the categories. Click **Data** in the top menu and go to **Weight Cases.**



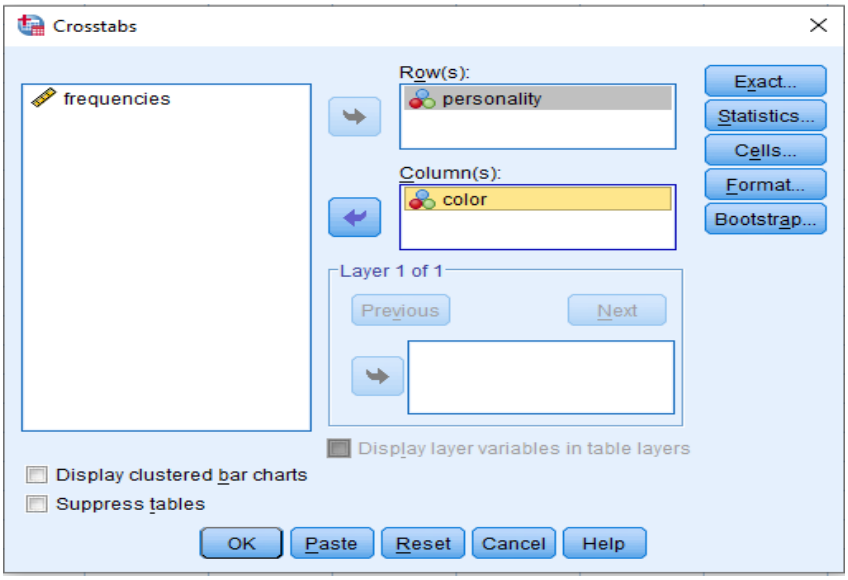
Click on the circle Weight cases by and drag the frequencies and variable into the Frequency Variable box.
Click Ok.



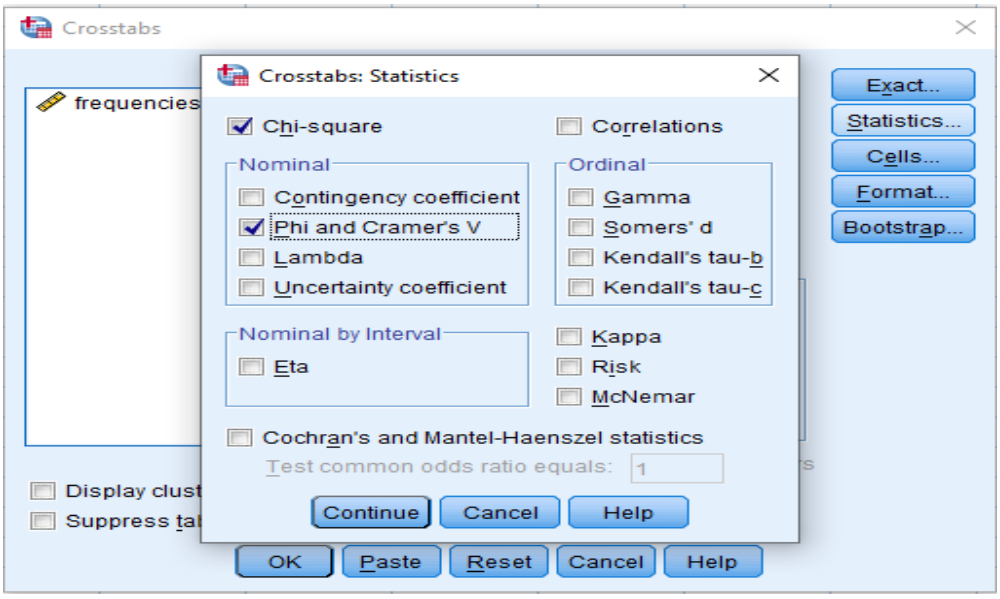
Click Analyze on the tool bar at the top of the page, select Descriptive Statistics, and click on Crosstabs.



- Highlight the label for the column containing the rows (personality) and move it into the Rows box by clicking on the arrow.
- Highlight the label for the column containing the columns (color) and move it into the Columns box by clicking on the arrow.



Click on the Statistics button. Tick Chi-square and Phi and Cramer's V. Click on Continue. Click OK.



Interpreting Results

The first thing we should check is whether we have violated one of the assumptions of chi-square concerning the 'minimum expected cell frequency', which should be 5 or greater.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	35.600[a]	3	.000
Likelihood Ratio	35.033	3	.000
Linear-by-Linear Association	34.476	1	.000
N of Valid Cases	200		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 5.00.

The main value that we are interested in is the Pearson Chi-Square value, which is presented in the Chi-Square Tests.

In the example, the chi-square value is $\chi 2(3)=35.6$, with an associated significance level of $p<.001$, presented in the column labelled Asymptotic Significance (2-sided).when asymptotic significance will be less than five it indicate that our results are significant.
This indicates that our result is significant and there appears to be an association between personality type and color preference.

For 2 by 2 tables the most commonly used effect size is the phi coefficient, which is a correlation coefficient and can range from 0 to 1, with higher values indicating a stronger association between the two variables.

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	.422	.000
	Cramer's V	.422	.000
N of Valid Cases		200	

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

For tables larger than 2 by 2 the value to report is **Cramer's V,** which takes into account the degrees of freedom.

In this example the Cramer's V value (shown in the table Symmetric Measures) is .422 indicating a large effect size.

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	.422	.000
	Cramer's V	.422	.000
N of Valid Cases		200	

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

A chi-square test for independence indicated significant association between personality types and color preferences, $\chi 2$ (3, n = 200)= 35.6, p<.001, Cramer's V = .422.

Lesson 45

The Spearman Correlation
Topic: 93-93

Topic 92: The Spearman Correlation

Definition of Spearman correlation

When the Pearson correlation formula is from a used with data ordinal scale (ranks), the result is called the Spearman correlation also called **Spearman rank-order correlation**. The Spearman correlation is used to measure the relationship between X and Y when both variables are measured on ordinal scales. It is also used when a researcher wants to measure the consistency of a relationship between X and Y. When two variables are consistently related, their ranks are **linearly related.**

Relationship between Pearson's and Spearman's

Pearson's r and Spearman's rho are very similar. They are both correlation coefficients, interpreted in the same way.

Difference between Pearson's and Spearman's

Pearson's r is used when our data meet the assumptions for a parametric test.
Spearman's rho (ρ) is used when the data do not conform to these assumptions.

Assumptions of Spearman's correlations:

- In Spearman, one or more variables are ratings given by participants (e.g., the attractiveness of a person), or they have to put pictures in rank order of preference. In these cases, data might not be normally distributed.
- When we have small participant numbers and are uncertain as to whether we meet the assumptions for Pearson's r, it is advised to use Spearman's rho.
- Spearman's rho transforms the original scores into ranks before performing further calculations. This test **is less sensitive to bias due to the effect of outliers**
- This can be used to reduce the weight of outliers (large distances get treated as a one-rank difference). It does not require assumption of normality.

Topic 93: The Spearman Correlation in SPSS

Example

The following data were gathered from nine people who were asked to rate the attractiveness of a target person, and then rate themselves (my attractiveness) on a 10-point scale from 1 (awful) to 10 (wonderful).
Use alpha level .05.

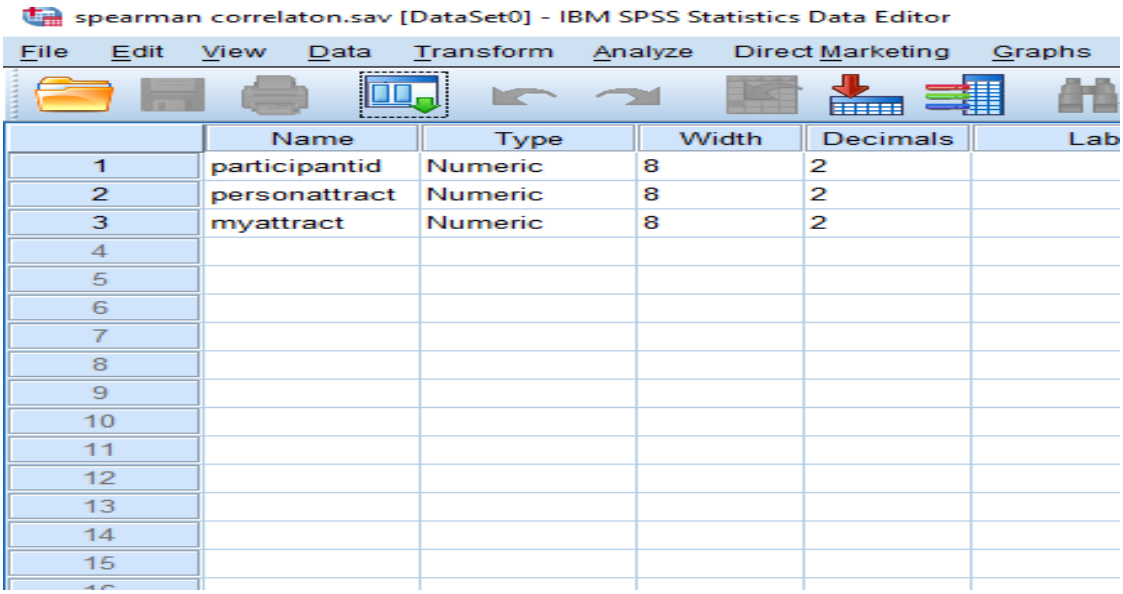
Attractiveness of another person	My attractiveness

7	9
5	4
5	5
8	9
9	9
6	8
1	3
2	5
8	9

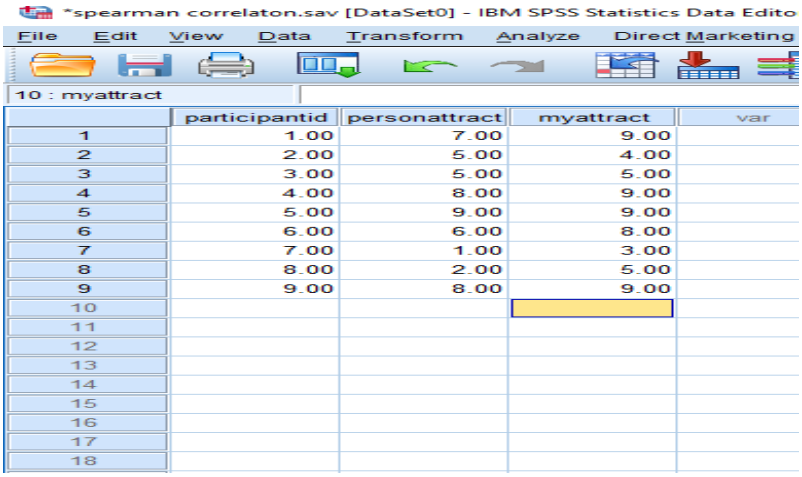
Hypotheses:

H_o: There is no correlation between the X (person attractiveness) and the Y (self-attractiveness).
H_1: There is a correlation between the X and the Y.

Entering Data
Create two variables (i.e., personattract and myattract) in the SPSS data editor as shown.

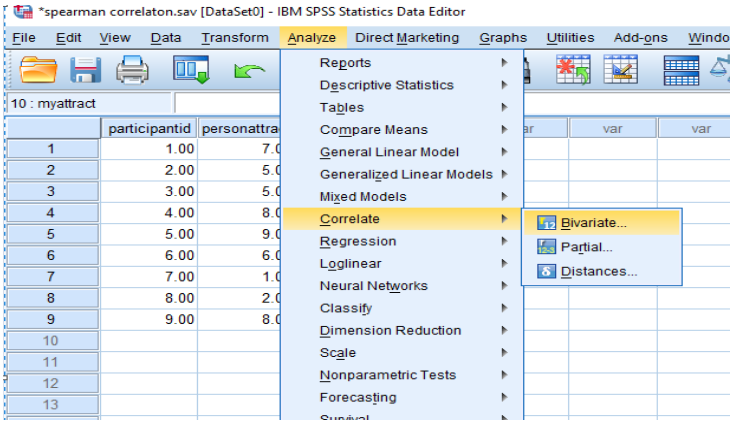


Enter data for personattract and myattract in two columns respectively.

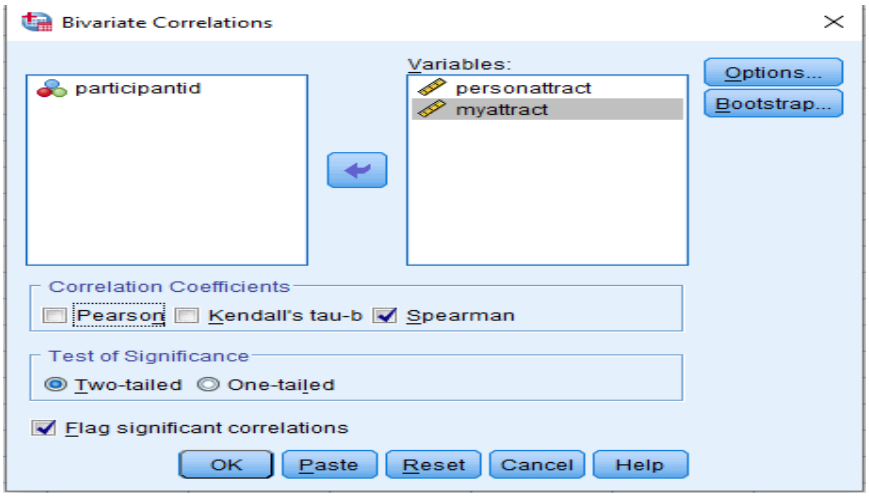


Running Analysis

In the top menu, go to Analyze > Correlate > Bivariate.



- This brings us to the bivariate correlations dialogue box.
- We need to uncheck the default Pearson option and select Spearman.
- Move the variables of interest from the left-hand side to the variable list on the right.
- Click OK.



Interpreting Results:

As in the matrix from Pearson's r, we only need to focus on one half of the matrix.
The results indicate that there is a significant positive correlation between the ratings given to a target person and oneself (ρ=.92, p<.001).

Nonparametric Correlations

```
[DataSet0] C:\Users\Peaceful Tears\Desktop\spearman correlaton.sav
```

Correlations

			personattract	myattract
Spearman's rho	personattract	Correlation Coefficient	1.000	.921**
		Sig. (2-tailed)	.	.000
		N	9	9
	myattract	Correlation Coefficient	.921**	1.000
		Sig. (2-tailed)	.000	.
		N	9	9

**. Correlation is significant at the 0.01 level (2-tailed).

Reporting Results:

The relationship between person attractiveness and self-attractiveness was investigated using the Spearman correlation coefficient. There was a strong, positive correlation between the two variables, ρ = .92, n = 9, p < .001.