Hadhramout University
College of Medicine and Health Sciences

Department of Community Medicine

Doctorate Program of Public Health (DrPH)

Course: Biostatistics

DrPH2 4 Credit

Tutor's Guide

2023

How to Use This Guide?

The purpose of this guide is to provide instructions for tutors to supervise and manage the learning sessions with postgraduate candidates on public health. This document provides an overview of the curriculum of biostatistics with implantation details. This tutorial guide in biostatistics is a structured document that illustrates the course intending outcomes, the learning objectives of each module, and the leading activities that cover the content of the course. This tutorial guide is an implementation guide giving curriculum of biostatistics into practice. The guide is supported by training activities that facilitate the tutor to help participants how to apply different approaches in practice. This guide is designed to be used both at classroom or virtual session. So the learning process is a student-centered and the role of the tutor is facilitator.

For learning purposes; the content was divided into four modules. In each module there are three sessions, each session starts with brainstorming of multiple choice questions, and one assignment. The assignment is subject to individual assessment, at the end of each module there is a group discussion activity, in each module, the group discussion composed of three parts, tutor asks students to divide into three groups and each group work on one part of the discussion topic. This group discussion is subject to group evaluation.

The intended learning Outcomes (ILOs)

Course ILOs		Program ILOs	
يبيّن المعرفة بأساسيات الإحصاء والإحصاء المتقدم Demonstrate a knowledge base in the basic and advanced biostatistics	a1-2	A1 Demonstrate knowledge of the epidemiological, statistical and research public health related concept	A) Knowledge and understanding
يفسّر بشكل صحيح جدول Z لنتائج القياسات الجسمانية Interpret correctly the z table of anthropometric results	b1-3	B1 يحلل و يفسّر البيانات المتعلقة بالصحة العامة Be able to interpret the health problems through the scientific evidence and community diagnosis	B) Intellectual skills
يستخدم البيانات والطرق الإحصائية لتحديد وتحليل و حل مشكلة صحية use of data and statistical methods for health problem identification and resolution	**c1-3**	C1. يستخدم المهارات الوبائية والإحصائية لتحليل المشكلات الصحية Apply The epidemiological and statistical skills to analyze the health problems	**C)** Practical skills
يستخدم بشكل أفضل قواعد البيانات الإحصائية الالكترونية لإدخال و تحليل البيانات مثل برنامج الحزمة الإحصائية للعلوم الاجتماعية او برنامج ستاتا Proper use of statistical software for data entry and analysis like SPSS or STATA	**c1-4**		
يتقن مهارات استخدام الكمبيوتر والانترنت Acquire Computer and internet skills	d2-**1**	D2 Demonstrate interest for the continuous professional development يواكب التطوير المهني المستمر	D) Transferal general skills
يتقن استخدام التواصل الرقمي والتقنية الحديثة في مجال الصحة العامة Acquire skills of digital communication in public health	d2-3	D3 Communicate effectively with others بتواصل بفعالية مع الاخرين	

Module 1 Descriptive Statistics

The learning Objectives:

1. To acquire skills of classify the data distribution
2. To summarize data.
3. To acquire knowledge and skills of data presentation.

Session 1. 1: Organization and Types of Data

Brain Storming

Q1. In developing the frequency distribution table, the class interval is calculated as:

a. (maximum value -minimum value)/ No. of classes wanted
b. (maximum value - No. of classes wanted)/minimum value
c. No. of class wanted/ (maximum value -minimum value)
d. it is always 5

Answer a: maximum value -minimum value)/ No. of classes wanted

Feedback:

The first step that a mathematician does with the collected data is to organize it in the form of a frequency distribution table. A frequency distribution table is a way to organize data so that it makes the data more meaningful. A frequency distribution table is a chart that summarizes all the data under two columns - variables/categories, and their frequency. It has two or three columns. Usually, the first column lists all the outcomes as individual values or in the form of class intervals, depending upon the size of the data set. The class interval is determined by the following formula:

(maximum value -minimum value)/ No. of classes wanted

The second column includes the tally marks of each outcome. The third column lists the frequency of each outcome. Also, the second column is optional.

Q2. Weight in Kgs is

a. ordinal variable
b. Nominal variable
c. Discrete variable
d. Continuous variable

Answer d: Continuous variable

Feedback:

A variable is a characteristic that can be measured and that can assume different values. Variables may be classified into two main categories: Quantitative variables (also so called numeric) and qualitative variables (also so called categorical) and. Each category is then classified in two subcategories: nominal or ordinal for categorical variables, discrete or continuous for numeric variables.

Quantitative variable) is a quantifiable characteristic whose values are numbers. Numeric variables may be either continuous (like age in years) or discrete (like number of beds in hospital)

Qualitative variable refers to a characteristic that can't be quantifiable. Categorical variables can be either nominal (like blood group) or ordinal (like severity of the disease: mild, moderate and severe).

Q3. Temperature is ……scale

a. Ordinal scale
b. Interval scale
c. Ratio scale
d. Nominal scale

Answer b. Interval scale

Feedback:

A **ratio scale** is a quantitative scale where there is a true zero and equal intervals between neighboring points. Unlike on an interval scale, a zero on a ratio scale means there is a total absence of the variable you are measuring.. Length, area, and population are examples of ratio scales while temperature is an interval scale.

Assignment 1:

A 25 pregnant women attend the antenatal care in a certain day. Midwife registered all the personal data in records including age in years. The results are as follows:

20, 21, 23, 21, 26, 24, 20, 24, 25, 22, 22, 23, 21, 24, 21, 26, 24, 22, 21, 23, 25, 22, 21, 24, 21

a. Present these data in a frequency distribution table.
b. Which result occurs most frequently?
c. Set up a frequency distribution table including columns for the relative frequency and percentage frequency of the data.

Session 1.2 Data summary

Brain Storming

Q1. Which of the following measures are measures of central tendency, measures of dispersion, measures of relative position or not at all:

Statistical measures	Type of measures
Arithmetic mean	
Inter quartile range	
mode	
Average	
Percentile	
Standard deviation	
Confidence interval	
Histogram	
Variance	
Median	
Quartile	
Range	

Answer

Statistical measures	Type of measures
Arithmetic mean	measures of central tendency
Inter quartile range	measures of dispersion
mode	measures of central tendency
Average	measures of central tendency
Percentile	measures of relative position
Standard deviation	measures of dispersion
Confidence interval	not at all
Histogram	not at all
Variance	measures of dispersion
Median	measures of central tendency
Quartile	measures of relative position
Range	measures of dispersion

Feedback:

Histogram is a method of data presentation while confidence interval is a method of estimation. Method of data summary are three types of measures: measures of central tendency (mean, median and mode), methods of dispersion (range, variance standard deviation and interquartile range), measures of relative position (decile, quartile and percentile).

Q2. The duration of time from first exposure to HIV infection to AIDS diagnosis is called the incubation period. The incubation periods of a random sample of 7 HIV infected individuals is given below (in years):

12.0 10.5 9.5 6.3 13.5 12.5 7.2

The mean, median and standard deviation are:

a. Mean is 3.5 , median is 6.3, standard deviation is 7.2
b. Mean is 10.5 years , median is 6.3 , standard deviation is 2.71
c. Mean is 10.5 , median is 10.2 years , standard deviation is 2.71 years
d. Mean is 10.2 years , median is 10.5 years , standard deviation is 2.71 years

Answer d. Mean is 10.2 years, median is 10.5 years, standard deviation is 2.71 years

Feedback:

To calculate the mean, we just add up all 7 values, and divide by 7. In fancy statistical notation,

$$\bar{x} = (\sum x) / N$$

$\bar{x}$ = (12.0 + 10.5 + 9.5 + 6.3 + 13.5 + 12.5 + 7.2) / 7 = 71.5/7 = 10.2 years

To calculate the sample median, first rank the values from lowest to highest:

6.3 7.2 9.5 10.5 12.0 12.5 13.5

Since there are 7 values, an odd number, we can simply select the middle value, 10.5 years as a sample median. The position of the median is N/2.

To calculate the sample standard deviation! Recall the formula for SD

$$SD = \sqrt{\sum (X - \bar{X})^2 / N - 1}$$

$$SD = \sqrt{\sum(12 - 10.2)^2 + (10.5 - 10.2)^2 + (9.5 - 10.2)^2 + (6.3 - 10.2)^2 + (13.5 - 10.2)^2 + (12.5 - 10.2)^2 (7.3 - 10.2)^2 / N - 1}$$

$$SD = \sqrt{(3.24 + 0.09 + 0.49 + 15.21 + 10.89 + 5.29 + 9 / 7 - 1}$$

$$SD = \sqrt{44.21 / 6} = \sqrt{7.368} = 2.71 \text{ years}$$

Q3. If the number 6.3 in Q2 were changed to 1.5, what would happen to the sample mean, median, and standard deviation?

a. All the Sample mean, median and standard deviation will, increase
b. All the Sample mean, median and standard deviation will, increase

c. Sample mean will decrease, standard deviation will increase while, median remains 10.5 without change
d. Sample mean and median will increase while, standard deviation 2.71 remains 10.5 without change

Answer c: Sample mean will decrease, standard deviation will increase while, median remains 10.5 without change

Feedback:

Sample mean – Would decrease from 10.2 to 9.5 years, as the lowest value gets lower, pulling down the mean.

$$\bar{x} = (\sum x) / N$$

$\bar{x}$ = (12.0 + 10.5 + 9.5 + 1.5 + 13.5 + 12.5 + 7.2) / 7 = 66.7/7 = 9.5 years

Sample median – Would remain the same since the middle value is still 10.5 By replacing the 6.3 with 1.5, the rank of the 7 values is not affected.

To calculate the sample median, first rank the values from lowest to highest:

1.5 7.2 9.5 10.5 12.0 12.5 13.5

Since there are 7 values, an odd number, we can simply select the middle value, 10.5 years as a sample median. The position of the median is N/2

Sample standard deviation – Would increase. Because our minimum value has now gotten smaller, while the rest of the data points remain unchanged, the spread or variability in our data has increased; since SD is a measure of spread, it too will increase from 2.71 to 3.98 years.

$$SD = \sqrt{\sum(12 - 10.2)^2 + (10.5 - 9.5)^2 + (9.5 - 9.5)^2 + (1.5 - 9.5)^2 + (13.5 - 9.5)^2 + (12.5 - 9.5)^2 (7.3 - 9.5)^2 / N - 1}$$

$$SD = \sqrt{(1 + 0 + 0.49 + 64 + 16 + 9 + 4.84 / 7 - 1}$$

$$SD = \sqrt{95.33 / 6} = \sqrt{15.888} = 3.98 \text{ years}$$

Q3. **Regarding Standard deviation what is not correct:**

a. The square root of the variance
b. Calculated as: = $\sqrt{\sum (X - \bar{X})^2 / N - 1)}$
c. It is the deviation of an observation from the mean of the sample

d. It's value never be zero

Answer d: It's value never be zero (non-correct statement).

Feedback:
The standard deviation is the square root of the variance, so it is expressed in the same units of measurement as the original data. The symbols for standard deviation are therefore the same as the symbols for variance, but without being raised to the power of two, so the standard deviation of a population is σ and the standard deviation of a sample is S. Standard deviation is sometimes written as SD.
The gullwing formula is used to calculate variance Calculated as:

Variance = $\sum (X- \bar{X})^2 /N-1)$

A standard deviation of 0 means that all the values in the dataset are the same, and thus have no deviation from the average.

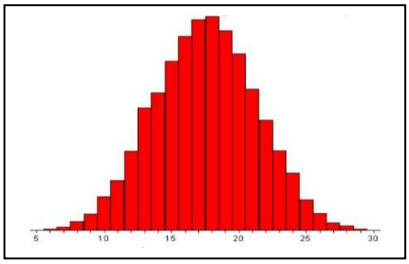
Assignment 2:
In no more than 10 PowerPoint slides present the following:

a. Define interquartile range. How to calculate it give an example
b. Give examples how to determine the percentile and quartile and what their relation with the median
c. Characteristics of the mean and median

Session 1.3 Data Presentation
Brain Storming:
Q1 - Q4

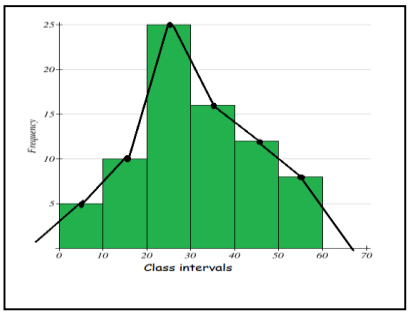
Q1. **The following graph is:**

a. Bar graph
b. Scatter diagram
c. Histogram
d. Frequency polygon



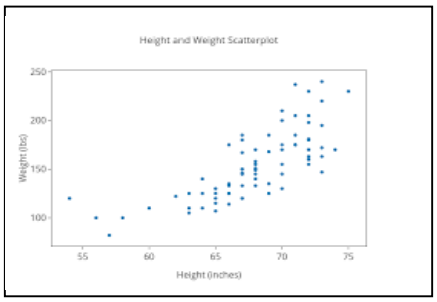
Q2. The attached graph is

a. Bar graph
b. Scatter diagram
c. Histogram
d. Frequency polygon



Q3. The attached graph is

a. Bar graph
b. Scatter diagram
c. Histogram
d. Frequency polygon



Q4. Qualitative data can be graphically represented by using a(n)

a. Bar graph
b. Scatter diagram
c. Histogram
d. Frequency polygon

Answers:

Q1. c. Histogram
Q2. d. Frequency polygon
Q3. b. Scatter diagram
Q4. a. Bar graph

Feedback:
Histogram is used to present only one continuous variable, it is an attached column, each column represents the size of the class while the height of the column represents the frequency. If the top of each column is connected with a line this graph is so called frequency polygon. If the graph presents the relative frequency, the line of the frequency polygon may be similar to the normal distribution curve if the study variable has normal distribution. Bar graph can present quantitative and qualitative variables.

Assignment 3.
Design graphs for one continuous variables, three continuous variables, one binominal variable and three categorical variables.

References:

- Anthony N. Glaser. High-yield biostatistics, epidemiology, and public health. 4th edition. Lippincott Williams & Wilkins, a Wolters Kluwer business. Printed in China 2014. https://books.google.com/books/about/High_Yield_Biostatistics_Epidemiology_an.html?id=Hb9xV2KA0m4C
- PRACTICE PROBLEMS FOR BIOSTATISTICS. https://www.math.kth.se/matstat/gru/sf1911/extraovningar/Abiostatsexercises.pdf
- Kaur P, Stoltzfus J, Yellapu V. Descriptive statistics. Int J Acad Med 2018;4:60-3. https://www.ijam-web.org/article.asp?issn=2455-5568;year=2018;volume=4;issue=1;spage=60;epage=63;aulast=Kaur
- Payum T. Graphical representation of data. PowerPoint presentation. **https://www.jncpasighat.edu.in/file/ppt/bot/graphical_rept_data.pdf**
- https://www.coconino.edu/resources/files/pdfs/academics/arts-and-sciences/college-mathematics/3rd/chapter-2-statistics-part-2.pdf Chapter 2: Statistics: Part 2

Group Discussion
Descriptive Statistics

Part 1

Theme of the discussion: Role of the mean, median and mode in the shape of the normal distribution curve.

The main question for group discussion: extreme or unusual values, may also influence the measures of central tendency and change the shape of the normal curve. Discuss with your colleagues the flowing items:

a. Types of variables can be measured by centeral tendency and dispersion?
b. The concept of positive and negative skew?

Discussion cues:

Read the following article to answer the above questions

Kaur P, Stoltzfus J, Yellapu V. Descriptive statistics. Int J Acad Med 2018;4:60-3. https://www.ijam-web.org/article.asp?issn=2455-5568;year=2018;volume=4;issue=1;spage=60;epage=63;aulast=Kaur

Part2.

Theme of the discussion: Graphic Presentation

The main question for the group discussion:

Look for the below table and Design an appropriate graph presentation of data (consider the titles of the graph, title of the x axis and y axis)

Prevalence of acute malnutrition among the sick children seeking care in health facilities by governorate, March 2022

Category of malnutrition		Lahj (n= 474)		Abyan (n= 477)		Total (N=951)		X^2	P-value
		No of children	%	No of children	%	No of children	%		
Acute malnutrition (Wasting)	MAM	84	17.7%	60	12.6%	144	15.1%	0.24	0.113
	SAM	27	5.7%	32	6.7%	59	6.2%		
	GAM	111	23.4%	92	19.3%	203	21.3%		

Discussion cues:

Take in your consideration how many variables presented in the above table. Types of variable and what graph is best present the above data? Can you read the following reference to help you design the proper graph?

- Payum T. Graphical representation of data. PowerPoint presentation. **https://www.jncpasighat.edu.in/file/ppt/bot/graphical_rept_data.pdf**

Part 3

The theme of the discussion: Measures of relative position

The main question for the discussion: what are the practical use of measures of position in public health?

Discussion cues:

Quantile is the statistical term used for the measures of relative position like quartile, decile and percentile. Define these three measures and how to determine their position. Discus with your colleague the uses of percentile and or quartile in assessing the nutritional status of the children under 5 years.

The following article may facilitate your work:

- Weir CB, Jan A. BMI Classification Percentile And Cut Off Points. 2022 Jun 27. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2023 Jan–. PMID: 31082114.

Module 2. Introduction to Inferential Statistics

The learning objectives

1. To recognize the different patterns of probability distribution.
2. To understand the concept and application of the normal distribution curve
3. To apply the z table to find the probabilities

Session 2.1 The Probability Distribution

Brain Storming

Q1. Binomial distribution is

a. discrete distribution
b. continuous distribution
c. individual distribution
d. Poisson distribution

Answer a. discrete distribution

Feedback:

Binomial Distribution:

A discrete distribution describing the results of an experiment is known as binomial distribution.

Poisson Distribution:

A discrete distribution in which the probability of the occurrence of an event within a very small time period is a very small number, the probability that two or more such events will occur within the same time interval is effectively 0, and the probability of the occurrence of the event within one-time period is independent of where that time period is.

Q2-Q5. A hospital staff comprises 132 people. 55 are males and 76 are females. 41 are medical doctors (MDs) and 90 are registered nurses (RNs). From this group, calculate the probabilities of finding match to the corresponding category of hospital staff:

Matching	Hospital staff	The probability
	Q2. A male doctor	a.40%
	Q3. A female doctor	b.18%
	Q4. A male nurse	c.28%
	Q5. A female nurse	d.13%

Answer

Matching	Hospital staff	The probability
d.13%	Q2. A male doctor	a.40%
c.28%	Q3. A female doctor	b.18%
b.18%	Q4. A male nurse	c.28%
a.40%	Q5. A female nurse	d.13%

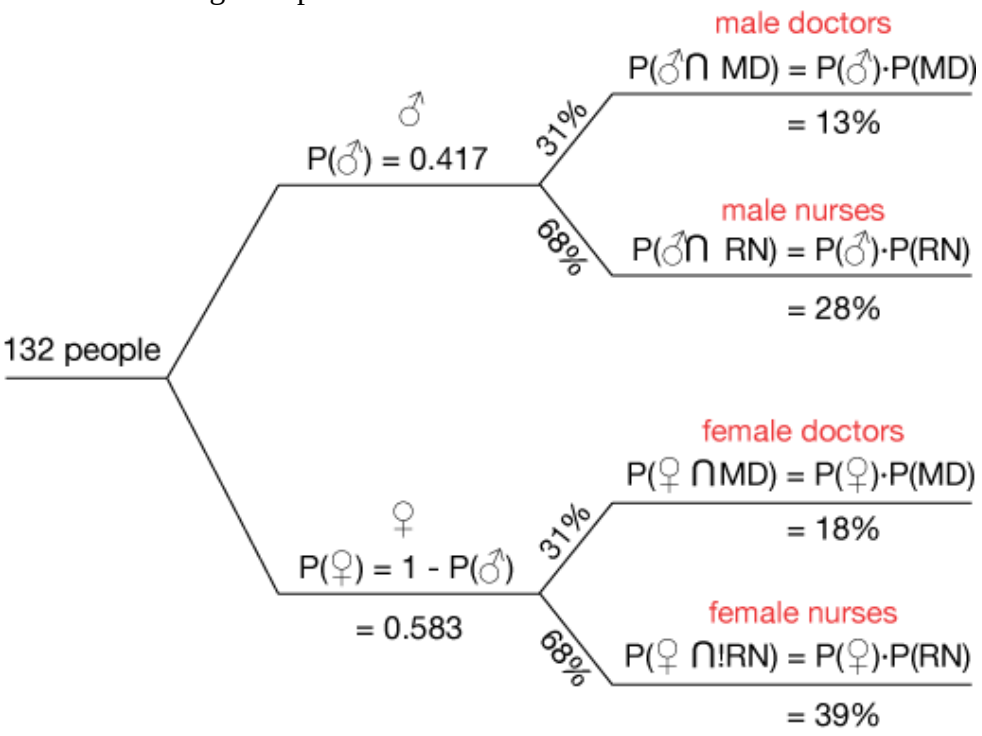
Feedback:

Being a nurse or a doctor (in this group) are mutually-exclusive outcomes, as are being male and female. These are simple AND probabilities.

a. First we have
   P(♂) =55132=0.417=55132=0.417
   P(♀) =77132=0.583=77132=0.583
   P(MD) =41132=0.311=41132=0.311
   P(RN) =90132=0.682=90132=0.682
b. Then we can have calculated probabilities a-d:

P(♂ ∩ MD) = (0.417)(0.311) = 13%
P(♂ ∩ RN) = (0.417)(0.682) = 28%
P(♀ ∩ MD) = (0.576)(0.311) = 18%
P(♀ ∩ RN) = (0.583)(0.682) = 40%

Notice that to within our rounding error here, the probabilities, which span all possible combinations of outcomes, add to 100%.
Here is a tree diagram of these probabilities, another way of looking at this situation that might help.



Viti A, Terzi A, Bertolaccini L. A practical overview on probability distributions. J Thorac Dis. 2015 Mar;7(3):E7-E10. doi: 10.3978/j.issn.2072-1439.2015.01.37. PMID: 25922757; PMCID: PMC4387424

K wak SG, Kim JH. Central limit theorem: the cornerstone of modern statistics. Korean J Anesthesiol. 2017 Apr;70(2):144-156. doi: 10.4097/kjae.2017.70.2.144. Epub 2017 Feb 21. PMID: 28367284; PMCID: PMC5370305.

Assignment 4:
Present the different types of probabilities. Use the relative frequency table to explain the probability of one continuous variable.

Session 2.2: The Normal Distribution

Brain Storming

Q1. All the following are true regarding normal distribution curve except:

a. Bill shaped
b. Symmetrical
c. Histogram
d. Present homogenous sample

Answer c: Histogram

Feedback:

The normal distribution curve indicted that the value of certain variable is normally distributed and the characteristics of the study sample and population is more or less similar regarding the study variable. It is bell shaped and the area under the curve equal one divided the curve symmetrically (0.5 at the right and 0.5 at the left). The normal distribution curve is not a histogram.

Q2. Area between one standard deviation on either side of the mean in normal distribution curve is:

a. 60%
b. 68%
c. 95%
d. 99%

Answer: b. 68%

Feedback:

Characteristics of the normal distribution curve are:

- It is bell shaped
- It is symmetrical
- The area under the curve equal 1
- Mean at the middle of the curve divide the area under the curve into two equal parts, every part = 0.5
- Mean ±1SD = 68% of the observations
- Mean ±2SD = 95% of the observations
- Mean ±3SD = 99.7% of the observations

Q3. **In normal distribution, the area under the curve is:**

a. >1
b. Equal 1
c. 0 - <1
d. 0

Answer b. equal 1

Feedback: see the feedback for Q1 and Q2.

Assignment 5.

In no more than 10 PowerPoint slides explain the importance, characteristics and uses of the normal distribution curve.

Session 2.3: Z Score

Brain Storming

Use the z table (see annexes) to answer the following questions.

Q1. The hemoglobin (Hb) level of healthy women has mean 13.5 g/dl and the standard deviation is 1.5 g/dl, what is the Z score for a woman with Hb 15g/dl?

a. 9
b. 10
c. 2
d. 1

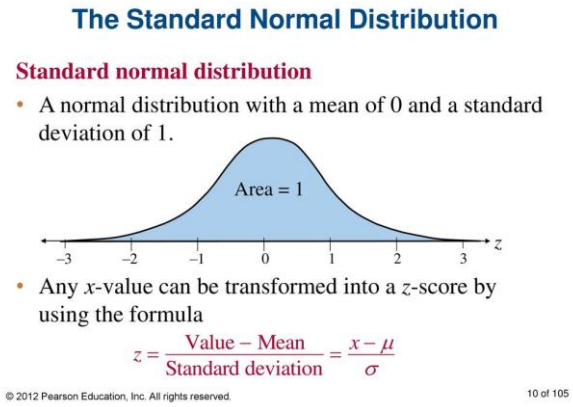
Answer d. 1

Feedback

The following formula is used to determine the value of z:

$Z = \frac{(X - \bar{X})}{SD}$ for the sample, $Z = \frac{(\bar{X} - \mu)}{\sigma}$ for population

Z = (15-13.5)/1.5 = 1.5/1.5 = 1



Q2 –Q4

Q2. The probabilities in the normal distribution table: met the z score 1.96 is:

a. 0.4750
b. 0.4505
c. 0.3997
d. 0.4893

Q3. the probabilities in the normal distribution table: met the z score 1.28 is:

a. 0.4750
b. 0.4505
c. 0.3997
d. 0.4893

Q4. The Z score of the probabilities 0.4893 from the normal distribution table is:

a. 1.96

b. 1.65
c. 1.28
d. 2.3

Answers

Q2 a 0.4750
Q3 c 0.3997
Q4 d 2.3

Feedback

Look for the z table determine the location of the value of z in the for the first column and the first raw, the corresponding cell is the probability

For the already probability, look for the cells of probability and identify the corresponding value of z in the first column and the first raw.

Q5 Q6. In a study of lead poisoning among urban adults, an epidemiologist observed that the mean blood lead level for the general population is 25µg/dl with standard deviation of 15µg/dl.

Q5. Determine probability that an individual selected at a random from urban population will have a blood lead level greater than or equal to 60 µg/d:

a. 0..0099
b. 2.33
c. 0.4901
d. 0.4901

Answer a. 0..0099

Feedback:

Z= x-u/sd = 60-25/15 = 35/15= 2.33 the probability at z core 2.33 is 0.4901

The probability that an individual selected at a random from urban population will have a blood lead level greater than or equal to 60 µg/d is 0.5000 – 0.4901 = 0..0099

Q6. Ten percent of blood pressure levels would be expected to lie above which value:

a. 1.28
b. 0.40
c. 0.10
d. 44.2 µg/d

Answer d. 44.2 µg/d

Feedback

The probability of 0.4000 correspond to z score 1.28

Using the formula of z score = x-u/sd = 1.28 = x-25/15 = 1.28*15 = x-25 = 19.2+25 = x = 44.2
Ten percent of blood pressure levels would be expected to lie above 44.2 µg/d

Assignment 6

Suppose temperature readings are normally distributed with mean 0∘ and standard deviation 1∘. Find the probabilities that a randomly selected temperature reading agrees with each of the following

a. less than −1.5∘−1.5∘
b. less than 1.23∘]
c. greater than 2.22∘
d. greater than −1.75∘−1.75∘
e. between 0.50∘ and 1.00∘
f. between −3.00∘−and −1.00∘
g. between −1.25∘ and 1.95∘
h. between −2.50∘ and 5.00∘
i. less than 3.55∘

Group Work
Probabilities and Normal Distribution

Part 1.

Theme of the group work: The probability of the continuous variable

The main question for group work: how to use the relative frequency as a key to understand the probability?

Group work assignment:

Look for the following table and answer the following questions:

a. calculate the relative frequency of every class
b. build a frequency polygon based in the relative frequency
c. Is the curve in the obtained relative frequency polygon indicate that the hemoglobin level in this group is normal distributed, explain your interpretation?
d. if the mean hemoglobin is 13 gr/dl and the standard deviation is 1 gr/dl, identify the area under the curve with mean ±2SD

Hemoglobin level (mg/dl) among male adolescent	# of male adolescents	%	Relative frequency
8.5 - <9	1	0.4%	
9 - <9.5	1	0.4%	
9.5 - <10	1	0.4%	
10 <10.5	2	0.7%	
10.5 - <11	3	1%	
11 <11.5	20	6.7%	
11.5- <12	21	7%	
12 - < 12.5	45	15%	
12.5 -<13	57	19%	
13 – <13.5	57	19%	
13.5 - <14	45	15%	
14 - <14.5	21	7%	
14.5 <15	19	6.3%	
15 -<15.5	4	1%	
15.5 <16	2	0.7%	
16.5 +	1	0.4%	
Total	300	100%	

Part 2.

The main theme of the group work: using the z table to identify the probabilities and the z score.

The main question of this group work: While data points are referred to as *x* in a normal distribution, they are called *z* or *z* scores in the *z* distribution. A *z* score is a **standard score** that tells you how many standard deviations away from the mean an individual value (*x*) lies:

- A positive *z* score means that your *x* value is greater than the mean.
- A negative *z* score means that your *x* value is less than the mean.
- A *z* score of zero means that your *x* value is equal to the mean.

Discuss the above paragraph to find solutions for the following exercises: You collect SAT scores from students in a new test preparation course. The data follows a normal distribution with a mean score (*M*) of 1150 and a standard deviation (*SD*) of 150. You want to find the probability that SAT scores in your sample exceed 1380.

- Find the *z* score for 1380. The *z* score tells you how many standard deviations away 1380 is from the mean.
- What is the probability to find student have a statistics score exceed 1380?
- What is the probability to find students with statistics score between 850 to 1450?
- 10% of students have lie above what score?

Part 3.
The theme of the group work: The Standard Error

- **The main question of the theme:** What means by standard error, how can you calculate it and what is the relation between standard error and standard deviation and sample size? And How to create a confidence interval with a known standard error.

Main cues to group work

Try to rich your theoretical concepts with two or three examples. Use the 95% of certainty in determining the confidence level. The following articles may facilitate your learning:

- Lee DK, In J, Lee S. Standard deviation and standard error of the mean. Korean J Anesthesiol. 2015 Jun;68(3):220-3. doi: 10.4097/kjae.2015.68.3.220. Epub 2015 May 28. PMID: 26045923; PMCID: PMC4452664.
- Hazra A. Using the confidence interval confidently. J Thorac Dis. 2017 Oct;9(10):4125-4130. doi: 10.21037/jtd.2017.09.14. PMID: 29268424; PMCID: PMC5723800.

Module 3. Testing Hypothesis

Learning objectives

1. To understand the steps for hypothesis testing
2. To test the hypothesis by using z-test, student t-test and chi square test
3. To identify the uses of the non-parametric test
4. Acquire basic concepts of the correlation, regression and multivariate analysis

Session 3.1 Inferential Statistics

Brain storming

Q1. **What is the primary purpose of inferential statistics in research?**

a. To summarize and describe data
b. To explore relationships between variables
c. To interpret qualitative data
d. **To test hypotheses and make inferences about a population**

Answer d. To test hypotheses and make inferences about a population

Q2. If null hypothesis is rejected even if it is true is

a. type I error
b. type II error
c. β error
d. μ error

Q3. Accepting null hypothesis when it is false is

a. type I error
b. type II error
c. α error
d. μ error

Answers for Q2&Q3: Q2: a. type I error. **Q3. b .** type II error

Feedback: look for the table below to understand the concept of type I and type II errors:

Decision	**Null hypothesis true**	**Null Hypothesis false**
Accept Null Hypothesis	Correct decision	Error type II
Reject null hypothesis	Error type I	Correct decision

Q4. Which of the following statistical tests is appropriate to test the difference between means of two independent samples?

a.T-test
b. ANOVA
c. Chi-square test
d. Paired-sample test

Q4. Which of the following statistical tests is appropriate to test the difference between means of two related samples?

a. T-test
b. ANOVA

c. Chi-square test
d. **Paired-sample test**

Answers to Q3 & Q4: Q3. a. T-test Q4. d. Paired-sample test

Feedback:

T-test is used to compare the two means and is used for small samples (n <30). Evaluating the t-value requires testing a null hypothesis where the means of both test samples are equal. If you perform a t-test and find the means are not equal, you reject the null hypothesis for the alternative hypothesis. You can calculate a t-value using a common t-test with the formula:

$t = (\bar{X} - \mu 0) / (s / \sqrt{n})$, where $\bar{X}$ is the sample mean, $\mu 0$ represents the population mean, s is the standard deviation of the sample and n stands for the size of the sample.

Unpaired t-test is used to compare the means of two independent groups, e.g. to compare the blood sugar of two independent groups. The data should be normally distributed and quantitative. This test is used when the SD of two means is almost the same or SD of one group is not twice greater or lesser than that of other

Paired t-test is used when one group serves as its own control, e.g. to compare the blood sugar before and after the administration of a drug.

Q5. Which of the following statistical tests is appropriate to test the relationship between two continuous variables while controlling other variables?

a. T-test
b. ANOVA
c. Chi-square test
d. **Regression Analysis**

Answer d: Regression Analysis

Q6. Which of the following statistical tests is appropriate to test the difference between more than two sample means?

a. T-test
b. ANOVA
c. Chi-square test
d. Regression Analysis

Answers for Q5 & Q6:

Q5: d. Regression Analysis Q 6 b. ANOVA

Feedback:

ANOVA test is used to compare the mean of three or more than three groups. The data should be normally distributed. One-way ANOVA is used when groups to be compared are defined by just one factor. Repeated measure ANOVA is used when groups to be compared are defined by multiple factors.

Correlation coefficient test is a parametric test; it is used to know about the linear relationship between two variables. For example, if we want to know about any linear relationship between body weight and blood pressure, correlation test will be used. Correlation only shows an association between two variables. It does not show causation. Scatter plot can be used to know about correlation between two variables. Pearson's correlation coefficient test is used for continuous variables, and Spearman's correlation coefficient is used as for categorical variables.

Regression test is a parametric test. It is used to know about the dependent relationship between two variables. We can predict the value of dependent variable, based on the value of independent variable. For example, if we draw a curve between time and plasma concentration of a drug, then we can predict a drug concentration at particular time on the basis of time plasma concentration curve. Here, time is the independent variable and plasma concentration is the dependent variable. Dependent variable is plotted on y-axis and independent variable is plotted on x-axis

References:

- Najmi A, Sadasivam B, Ray A. How to choose and interpret a statistical test? An update for budding researchers. J Family Med Prim Care. 2021 Aug;10(8):2763-2767. doi: 10.4103/jfmpc.jfmpc_433_21. Epub 2021 Aug 27. PMID: 34660402; PMCID: PMC8483143.

Assignment 7:

In no more than 10 PowerPoint slides explain the difference between Z – test and t-test, provide at least two exercises to clarify your answer.

Session 3.2 Chi Square and non-parametric tests

Brain Storming

Q1. What is the appropriate statistical test to test the association between two categorical variables in a research study?

a. T-test
b. ANOVA
c. Regression Analysis
d. **Chi-square test**

Answer d. Chi-square test

Feedback:

To determine whether the association between two qualitative variables is statistically significant, researchers must conduct a test of significance called the Chi-Square Test. There are five steps to conduct this test.

Q2. Which of the following statistical tests is appropriate to test the difference between medians of two groups?

e. T-test
f. ANOVA
g. **Wilcoxon rank-sum test**
h. Regression analysis

Answer c: Wilcoxon rank-sum test

Feedback:

A popular nonparametric test to compare outcomes between two independent groups is the Mann Whitney U test. The Mann Whitney U test, sometimes called the Mann Whitney Wilcoxon Test or the Wilcoxon Rank Sum Test, is used to test whether two samples are likely to derive from the same population (i.e., that the two populations have the same shape). Some investigators interpret this test as comparing the medians between the two populations.

Q3. For a r x c contingency table, the Chi-Square test has a degree of freedom (d.f)?

a. (r – 1) + (c - 1)
b. (r – 1) x (c - 1)
c. r x c
d. r – 1

Answer b. (r – 1) x (c - 1)

Feedback:

The degrees of freedom for the chi-square are calculated using the following formula: df = (r-1)(c-1) where r is the number of rows and c is the number of columns.

Assignment 8. Apply the chi square test for the following exercise

A data sample is sorted into a 3x2 contingency table based on two factors, one is severity of diseases has three sub-categories (mild, moderate and severe) and the age variable of which has two sub-categories) (≤ 5 years and > 5 years).

	≤ 5 years	> 5 years	Total
Mild	20	10	30
Moderate	15	5	20
Sever	10	20	30
Total	45	35	80

a. Find the expected number E of observations for each cell based on the assumption that the two factors are independent (that is, just use the formula $E=(R\times C)/n$.
b. Find the value of the chi-square test statistic χ^2.
c. Find the number of degrees of freedom of the chi-square test statistic.
d. Compare the calculated χ^2 with the tabulated χ^2
e. Take decision to accept or to reject the null hypothesis, interpret your conclusion based on *P*-value.

Session 3.3 Correlation, Regression and multivariate analysis

Brain Storming:

Q1. Correlation coefficient tends to lie between

a. 0 to +1
b. -1 to 0
c. -1 to +1
d. -2 to +2

Answer c: -1 to +1

Feedback:

The correlation coefficient is a statistical measure of the strength of a linear relationship between two variables. Its values can range from -1 to 1. A correlation coefficient of -1 describes a perfect negative, or inverse, correlation, with values in one series rising as those in the other decline, and vice versa.

The most common coefficient of correlation, called a Pearson correlation coefficient, measures the strength and the direction of a linear relationship between two variables. It is used for interval or ratio scale data. The other tool is the Spearman rank-order correlation, which is used for ordinal scale data.

Q2. **In the regression model (y=a+bx). Where x =2.5, y =5.5 and a=1.5. Which one of the following values is the parameter b of the model is correct?**

a. 1.75
b. 1.6
c. 2.5
d. 3.2

Answer b: 1.6

Feedback:

Linear regression is a way to model a relationship between two variables, by using the following formula: y=a+bx where x is the independent variable (plotted in x axis). b is slope in the line and a is the intercept.

Y=5.5, a= 1.5, x = 2.5

5.5 = 1.5 + bx2.5 = 5.5 – 1.5 = bx2.5 = 4=bx2.5 = 4/2.5 = b = 1.6

Q3. In a study, subjects are randomly assigned to one of three groups: control, experimental A, or experimental B. After treatment, the mean scores for the three groups are compared. The appropriate statistical test for comparing these means is:

a. The Analysis Of Variance (ANOVA)
b. The Correlation Coefficient
c. Chi Square
d. The T-Test

Answer a: The Analysis of Variance
Feedback:

- The ANOVA test allows a comparison of more than two groups at the same time to determine whether a relationship exists between them. The result of the ANOVA formula, the F statistic (also called the F-ratio), allows for the analysis of multiple groups of data to determine the variability between samples and within samples.
- The t- and z-test methods developed in the 20th century were used for statistical analysis until 1918, when Ronald Fisher created the analysis of variance method. ANOVA is also called the Fisher analysis of variance, and it is the extension of the t- and z-tests. The term became well-known in 1925, after appearing in Fisher's book, "Statistical Methods for Research Workers."
- A one-way ANOVA is used for three or more groups of data, to gain information about the relationship between the dependent and independent variables.
- A two-way ANOVA is an extension of the one-way ANOVA. With a one-way, you have one independent variable affecting a dependent variable. With a two-way ANOVA, there are two independents.

Q4. **Which analysis is portrayed by the equation:**

$$\mathbf{Y = b_O + b_1X_1 + b_2X_2 + b_3X_3 ... + b_nX_n}$$

a. Analysis of variance
b. Simple regression
c. Multivariate regression
d. T-test

Answer c. Multivariate regression

Feedback:

A simple linear regression model has a continuous outcome and one predictor, whereas a multiple or multivariable linear regression model has a continuous outcome and multiple predictors (continuous or categorical). A simple linear regression model would have the form

$$(1)\ y = \alpha + x\beta + \varepsilon$$

By contrast, a multivariable or multiple linear regression model would take the form

$$(2)\ y = \alpha + x_1\beta_1 + x_2\beta_2 + \ldots + x_k\beta_k + \varepsilon$$

Types of multivariate analysis methods

Multivariate methods can be subdivided according to different aspects. First of all, they are differentiated according to whether the aim is to discover a structure within the combination of data, or whether the data is to be checked with a certain structure. a structure The structure-determining methods include:

- Factor analysis: Reduces the structure to relevant data and individual variables. Factor studies focus on different variables, so they are further subdivided into main component analysis and correspondence analysis. For example: Which website elements have the greatest influence on purchasing behavior?
- Cluster analysis: Observations are graphically assigned to individual variable groups and classified on the basis of these. The results are clusters and segments, such as the number of buyers of a particular product, who are between 35 and 47 years old and have a high income.

Structural review procedures include, among others, the:

- Regression Analysis: Investigates the influence of two types of variables on each other. Dependent and nondependent variables are spoken of. The former are so-called explanatory variables, while the latter are explanatory variables. The first describes the actual state on the basis of data, the second explains this data by means of dependency relationships between the two variables. In practice, several changes of web page elements correspond to independent variables, while the effects on the conversion rate would be the dependent variable.
- Variance analysis: Determines the influence of several or individual variables on groups by calculating statistical averages. Here you can compare variables within a group as well as different groups, depending on where deviations are to be assumed. For example: Which groups most often click on the' Buy Now' button in your shopping cart?
- Discriminant analysis: Used in the context of variance analysis to differentiate between groups that can be described by similar or identical characteristics. For example, by which variables do different groups of buyers differ?

Assignment 9:

Explain in no more than 10 PowerPoint slides methods of multivariate analysis.

Reference:

- Rencher AC, Christensen WF. Methods of multivariate analysis. 3rd edition, 2012. A JOHN WILEY & SONS, INC., PUBLICATIoN http://ndl.ethernet.edu.et/bitstream/123456789/27185/1/Alvin%20C.%20Rencher_2012.pdf

Group Discussion

Part 1

Theme of the discussion: Use of Statistical test

The main question for the group discussion: explain how the statistical tests of significant differ according to the type of variables?

Discussion cues:

Use the following articles as a guide to your discussion with your colleagues:

- Bellolio, M.F., Serrano, L.A. & Stead, L.G. Understanding statistical tests in the medical literature: which test should I use?. *Int J Emerg Med* **1**, 197–199 (2008). https://doi.org/10.1007/s12245-008-0061-z
- Najmi A, Sadasivam B, Ray A. How to choose and interpret a statistical test? An update for budding researchers. J Family Med Prim Care. 2021 Aug;10(8):2763-2767. doi: 10.4103/jfmpc.jfmpc_433_21. Epub 2021 Aug 27. PMID: 34660402; PMCID: PMC8483143.

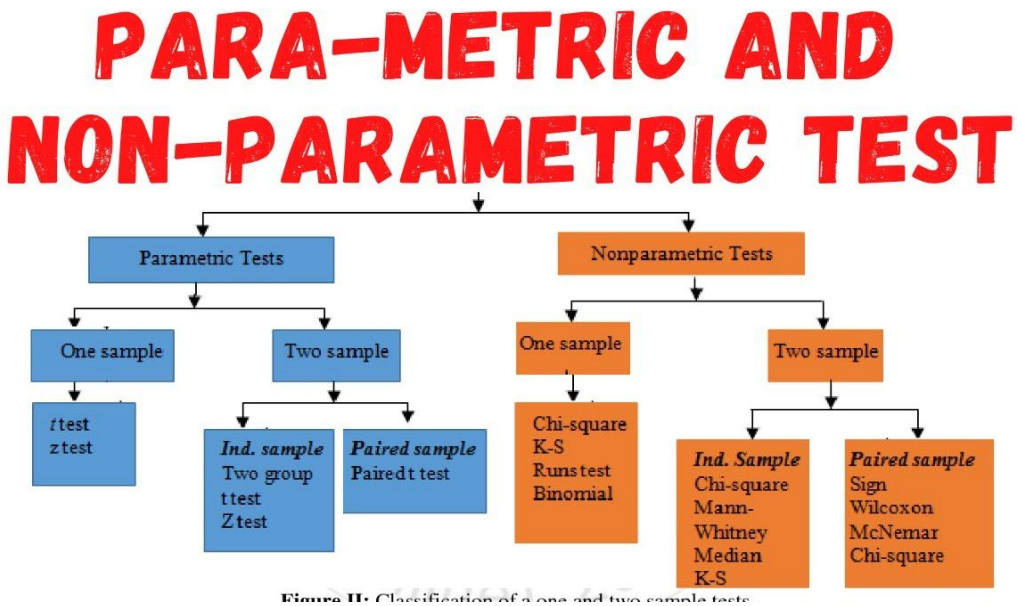
Part 2

Theme of the discussion: Non-parametric tests

The main question for the group discussion: What are the differences between parametric and non-parametric tests.

Give one exercise for using every non-parametric test

Discussion cues: Look for the below figure and add any missed non=parametric test and give an example



Read the following lectures:

- Shiek AB. Non-parametric test. https://www.narayanamedicalcollege.com/wp-content/uploads/2020/08/5.-Dr.-Sk-Ahammad-Basha_Non-Parametric-Tests-1.pdf
- Pandey R. LECTURE NOTES ON PARAMETRIC & NON-PARAMETRIC TESTS FOR SOCIAL SCIENTISTS/ PARTICIPANTS OF RESEARCH METODOLOGY WORKSHOP. https://www.lkouniv.ac.in/site/writereaddata/siteContent/202003241550010566rajeev_pandey_parametric_test.pdf

Part 3

Theme of the discussion: Logistic Regression

The main question for the group discussion: What you now about logistic regression? Concept, when use it and the interpretation of the results?

Discussion cues: use the following references to complete your task

- Sperandei S. Understanding logistic regression analysis. Biochem Med (Zagreb). 2014 Feb 15;24(1):12-8. doi: 10.11613/BM.2014.003. PMID: 24627710; PMCID: PMC3936971.
- Bewick, V., Cheek, L. & Ball, J. Statistics review 14: Logistic regression. *Crit Care* **9**, 112 (2005). https://doi.org/10.1186/cc3045
- Bzovsky, S., Phillips, M.R., Guymer, R.H. *et al.* The clinician's guide to interpreting a regression analysis. *Eye* **36**, 1715–1717 (2022). https://doi.org/10.1038/s41433-022-01949-z
- Castro HM, Ferreira JC. Linear and logistic regression models: when to use and how to interpret them? J Bras Pneumol. 2022;48(6):e20220439 https://dx.doi.org/10.36416/1806-3756/e20220439

Module 4

Use SPSS for Data Analysis (self-study)

Learning objectives:

By the end of this module, participant will be able:

1. to select the appropriate analysis method that will help you answer the question and test the hypothesis**.**
2. to run the analysis and check the output in SPSS.
3. to present and interpret the results and findings in a clear and concise way.
4. To evaluate the reliability of the results

Session 4.1
Introduction to SPSS
Brain Storming
Q1. SPSS stands for.

a. statistical package for the social sciences
b. standard package for the social sciences
c. standard package for the psychiatric sciences
d. statistical package for the public health sciences

Answer a: statistical package for the social sciences

Q2. In SPSS; The Data Editor displays the contents of the active data file. The information in the Data Editor consists of variables and cases, data will appear in data view and variable view, what of the following statement is incorrect?

a. In Data View, columns represent variables, and rows represent cases.
b. In Variable View, each row is a variable, and each column is an attribute that is associated with that variable.
c. In Data View, columns represent variables, and rows represent observations.
d. In Variable View, each row is a variable, and each column is a case.

Answer d: In Variable View, each row is a variable, and each column is an attribute that is associated with that variable (incorrect statement).

Q3. In SPSS, the variables view spreadsheet serves to define the:

a. cases
b. variables
c. data analysis
d. All of above

Answer b: Variables

Q4. Which command is required to change the existing value of variables.

a. transform
b. expression
c. compute
d. none of these

Answer a. transform

Assignment 10.

Explain the different functions of the command "analysis" in SPSS

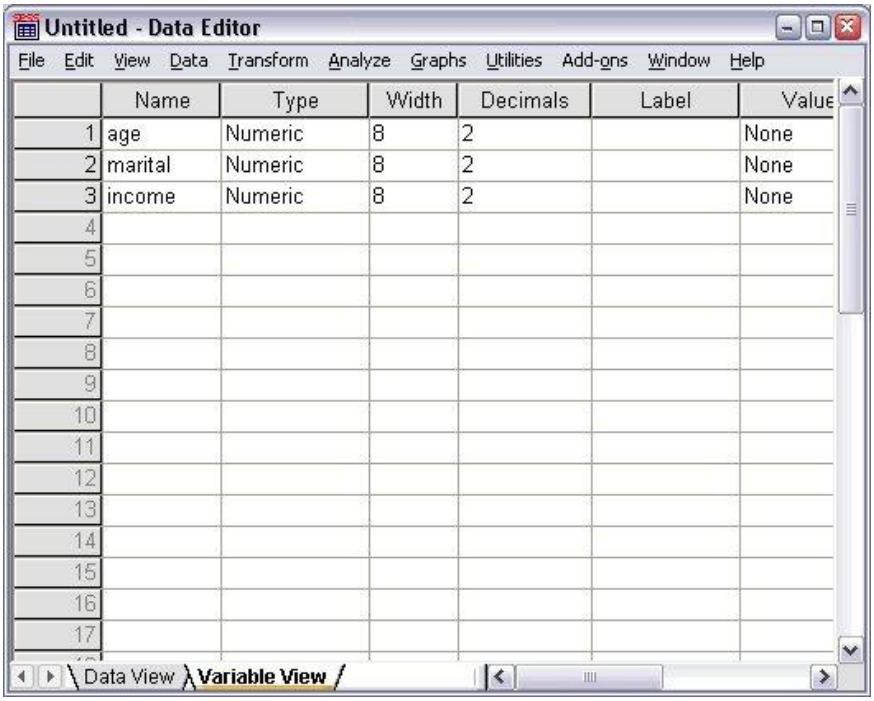
Session 4.2
Data Entry
Brain Storming

Data can be entered into the Data Editor, which may be useful for small data files or for making minor edits to larger data files.

Click the **Variable View** tab at the bottom of the Data Editor window.

You need to define the variables that will be used. In this excercise, only three variables are needed: *age*, *marital status*, and *income*.

Figure 1. Variable names in Variable View



Q1. This spreadsheet where the variable defined is:

a. Data view
b. Variable view
c. Output screen
d. All of above

Answer b. variable view

Q2. The type of the variables marital status and income are numeric for the purpose of coding, but really they are qualitative variables, so if you need to change the decimals you should change the number 2 to:

a. 0
b. 1
c. 3
d. No need to change it

Answer a. 0

Q3. In the above spreadsheet, age is defined as:

a. Continuous variable
b. Discrete variable
c. Ordinal variable
d. None of above

Answer a. Continuous variable

Assignment 11:

Referring to the above spreadsheet:

1. Write the variable name in the label section
2. Give value for each variable
3. Add four variables and determine their labels and values

Session 4.3 Data Analysis

Brain Storming

A number of tests are available to determine if the relationship between two cross tabulated variables is significant (here is the example examine the relationship between income level and PDA (personal digital assistant) ownership. One of the more common tests is **chi-square**. One of the advantages of chi-square is that it is appropriate for almost any kind of data.

1. Open the Crosstabs dialog box again.
2. Click **Statistics**.

Figure 1. Crosstabs dialog box

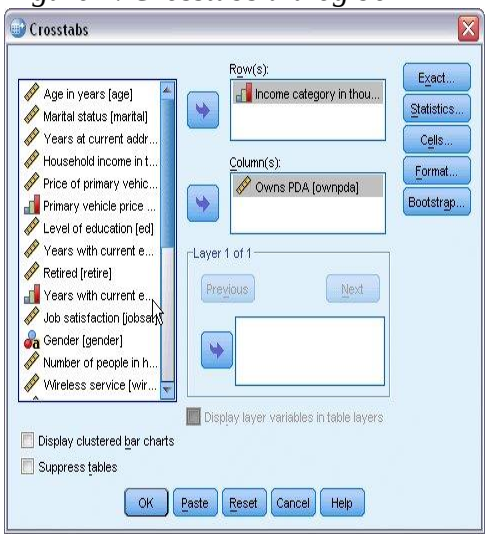
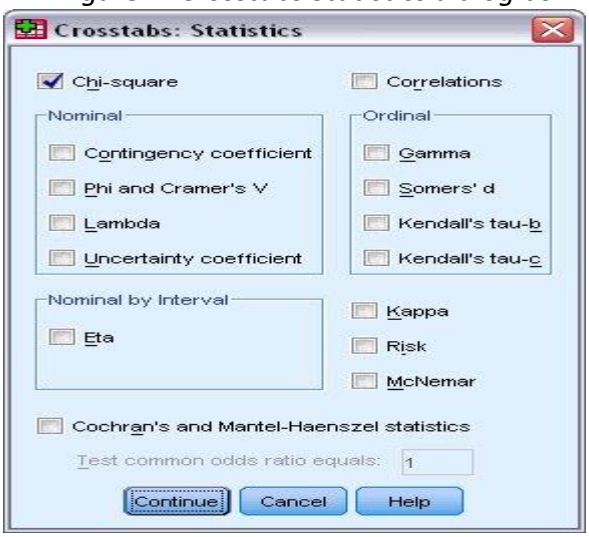


Figure 2. Crosstabs Statistics dialog box



3. Click (check) **Chi-square**.
4. Click **Continue** and then click **OK** in the main dialog box to run the procedure.

Figure 3. Chi-square statistics

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	37.677[a]	3	.000
Likelihood Ratio	37.313	3	.000
Linear-by-Linear Association	36.537	1	.000
N of Valid Cases	6400		

Q1. The degree of freedom in chi square table, give you indicators of how many rows and how many columns in the contingency table:

a. Two rows and twos columns
b. Three rows and two columns
c. Four rows and two columns
d. None of above

Answer c. Four rows and two columns

Q2. The calculated chi square is 37.677 as appear the above table, with a significant level of 0.05 and a degree of freedom of 3, this value is:

a. More than the tabulated value
b. Less than the tabulated value
c. Equal to the tabulated value
d. The tabulated value is not applicable to the chi square.

Answer a. More than the tabulated value

Q3. The *P*-value as appear in the above table indicated that:

a. The association is more likely due to chance
b. There is no significant association
c. There is true association
d. The null hypothesis is accepted.

Answer c. There is true association

Assignment 12.

Demonstrate the use of t test and ANOVA in SPSS in no more than 10 PowerPoint slides or video presentation
You can use the SPSS tutorial to help you in complete your task.

Group Work
Use SPSS in Analyzing Real Data

The facilitator asks the participants to divided into three groups, every group conduct the same three tasks but with different topic.
You and your colleagues are advised to conduct a pilot study about certain problem based in your research question

Part 1: Define the variables
Tasks

- Formulate a good research question
- Design a questionnaire
- Define the study variables in SPSS

Part 2 Data entry
Tasks

- Collect data from at least 50 subjects relevant to the study population as indicated by the research question (Field work)
- Enter data in SPSS as real data collected from the primary sources
- Examine the reliability score of the questionnaire and correct it accordingly

Part 3 Data analysis
Tasks

- Do the descriptive statistics as appropriate?
- Do univariate analysis as appropriate? (t-test, chi square test)
- Do multivariate analysis as appropriate? (ANOVA, Logistic regression

Discussion cues

- You can use the SPSS tutorial or any other educational video to facilitate your work
- You shod install SPSS program of updated version in your laptop
- Discuss with your colleagues the findings and summarize it.
- Present your findings in no more than 10 Power point slides with appropriate interpretation