

# Statistics guide for student and researchers

---

With SPSS illustrations

---

**Abdiasis Abdallah Jama**

**©2020**

ALL RIGHTS RESERVED. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the author.  
Published in Somalia.

This book is dedicated to my late father  
*Abdallah Jama Jibril*

## Preface

The purpose of this book is to teach students statistical techniques they need in order to write successful graduation research dissertation and also to prepare them demanding job opportunities. All job opportunities will need statistics to arrive at an effective research conclusion and to predict growth and progress. This study guide is also intended for researchers who want to update their existing knowledge on statistics as well as having understanding gap in inferential statistics to draw conclusion about population study.

The book is designed to be in study guide fashion whereby too much theory is not addressed but rather practical understanding of concepts connected with relevant examples is emphasized. Having said that the guide is not intended to cover all aspects of statistics rather it gives both students and researchers enough statistical tools they need to solve everyday research problems faced in all disciplines.

## Contents

1. <b>Preface</b> .....	4
2. Architecture of the study guide .....	8
3. <b>Introduction to descriptive statistics</b> .....	9
Levels of Measurement .....	11
Frequency distribution of ungrouped data .....	11
Frequency distribution of grouped data.....	22
Boxplot for groups variances .....	23
Review questions .....	25
4. <b>Probability basics</b> .....	27
Introduction .....	28
Probability of events.....	29
Conditional probability .....	33
Review questions .....	35
5. <b>Introduction to probability distributions</b> .....	36
Introduction to random variable .....	37
Cumulative distribution function.....	38
Expectation and standard deviation of discrete random variable .....	39
Continuous random variable .....	40
Review questions .....	42
6. <b>Common discrete probability distributions</b> .....	43
Binomial distribution .....	44
Poisson distribution.....	46
Review questions .....	48
7. <b>The normal distribution</b> .....	49
Introduction .....	50
Standard normal probabilities .....	52
Normal approximation to Binomial distribution. ....	56
Review questions .....	60
8. <b>Sampling distributions</b> .....	61
Sampling methods .....	62
Sample Statistics .....	64

## Preface

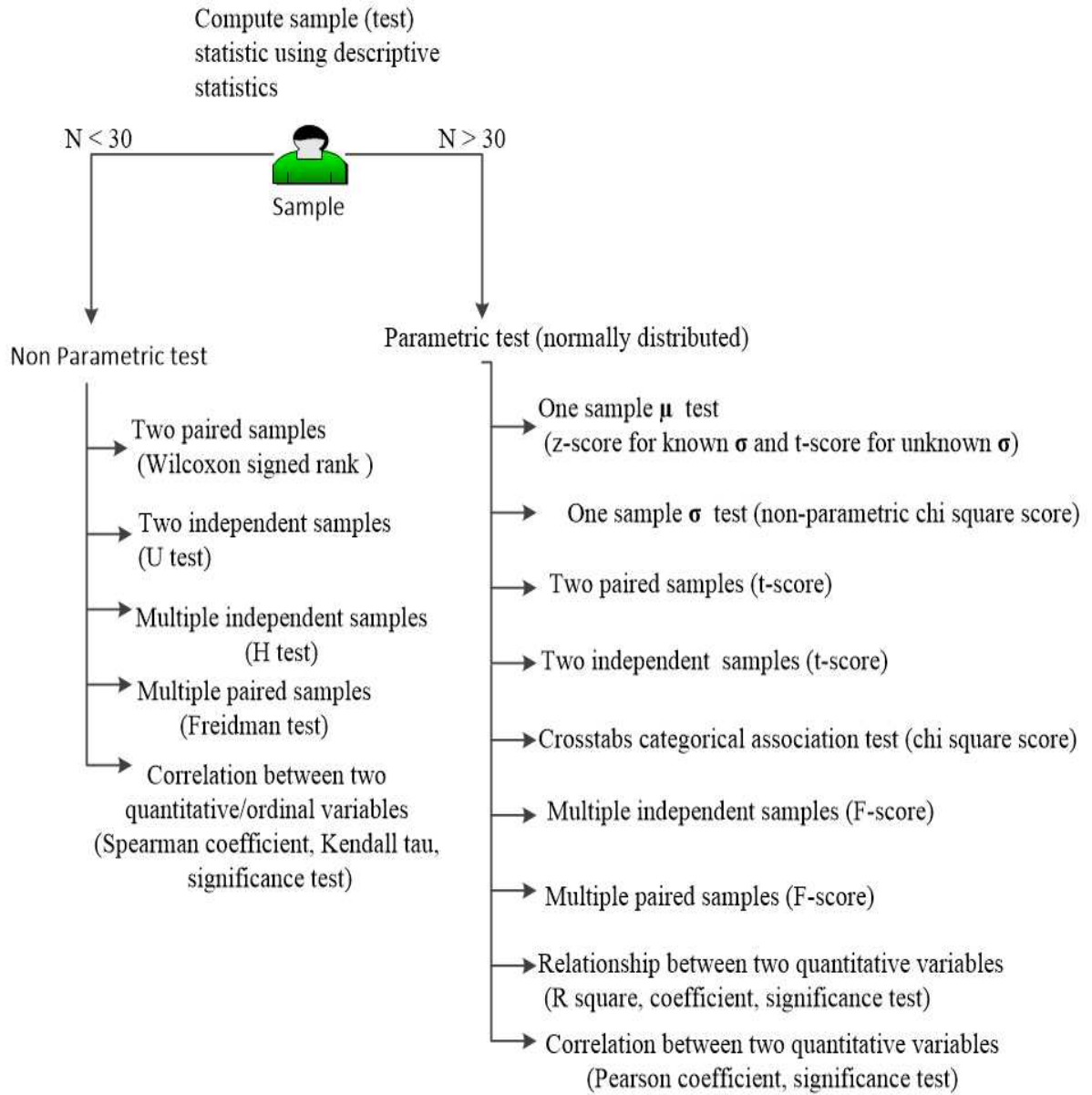
Sampling distribution statistic calculation.....	65
Central limit theorem .....	69
Interval estimation of population parameter.....	70
Confidence interval for a population mean.....	73
Confidence interval for population standard deviation .....	76
Confidence interval for population proportion .....	78
Review questions.....	80
<b>9. Hypothesis test of single sample .....</b>	<b>82</b>
Introduction .....	83
Hypothesis test for population mean .....	85
Hypothesis test for population standard deviation.....	91
Review questions.....	93
<b>10. Goodness of fit test for normality.....</b>	<b>94</b>
Chi – square goodness of fit test for normality .....	95
Kolmogorov – Smirnov (KS) goodness of fit test .....	108
Review questions.....	113
<b>11. Independent and dependent samples hypothesis test .....</b>	<b>114</b>
Dependent (paired) samples test .....	116
Independent samples t – test to compare means.....	120
Chi-square test for nominal variables independence.....	124
Review questions.....	130
<b>12. Linear regression and correlation .....</b>	<b>131</b>
Linear regression equation .....	133
Goodness of fit for the regression equation.....	139
Confidence interval for regression equation slope .....	141
Multiple linear regression.....	143
Correlation coefficient $r$ .....	147
Review questions.....	151
<b>13. One – way Analysis of variance (one – way ANOVA).....</b>	<b>152</b>
Randomized experimental design .....	153
One-way ANOVA test for three samples means equality.....	154
Fisher LSD multiple comparisons.....	162
Repeated measures one-way ANOVA .....	165

## Preface

Review questions.....	168
14. Non-parametric tests.....	170
Kruskal – Wallis test .....	171
Mann-Whitney U test.....	175
Wilcoxon signed rank test .....	178
Spearman rank correlation coefficient .....	182
Friedman test.....	184
Review questions.....	185
15. Appendix A: questionnaire design for collecting sample data .....	186
16. Appendix B: SPSS data entry.....	187
Multiple response questions .....	188
Likert scale .....	191
17. Appendix C: Guide to selecting appropriate statistical test .....	192
18. Appendix D: Statistical tables .....	194
Standard normal left tail probabilities $P(Z \leq -z)$ .....	194
Chi – square $\chi^2$ distribution table right tailed .....	196
Student’s t – distribution left tailed.....	200
F-distribution probabilities right tailed.....	201

# Architecture of the study guide

## Architecture of the study guide





# Introduction to descriptive statistics

## Chapter one

### Introduction to descriptive statistics

---

After completing this section, you should be able to

- Understand the meaning of statistics and different data levels
- Explain the importance of studying statistics
- Understand how to analysis sample data using descriptive statistics
- Appreciate how to calculate mean, variance and standard deviation of ungrouped and grouped data
- Practice descriptive statistics analysis using SPSS tool

## Introduction to descriptive statistics

Statistics is the art of collecting, summarizing, analyzing, and interpreting data. Data for statistical analysis may include financial records of an organization, population census, development of new products, and the like.

Statistics has found significant position in our daily life. Newspapers, job recruiters, managerial decisions, lottery winning, election forecasting, weather, production effectiveness all use statistical technique known as probability. Data collection and analysis are required in the following examples.

- Businesses send questionnaire to customers to determine level of service satisfaction.
- Test score of students are statistically analyzed to improve quality of education.
- To test effectiveness of new product, companies subject their new products to sample of customers. This helps the manager to decide if the new system is better than the older one.

Why write this book?

- Modern statistical data collection and analysis are carried out in computer programs which can process data much faster and provide accurate results. The focus of this book is to walk you through statistics principles with SPSS illustrations.
- To give readers comprehensive but simple text that does not flood too much mathematics and theories but rather emphasize the most important concepts and tools that you need in your day to day statistical work.

In the study of statistics it is useful to define terms like data, variable, descriptive, inference, sample, and population and so on.

The beginning chapters will mainly focus on descriptive statistics which provides information about central tendency and variation of a set of data. Frequency tables summarize the raw data, and then mean and standard deviation are calculated to account for the center of the data and its spread around the mean respectively. This analysis will depend on the scale of measurement used. To better visualize the data, graphs are plotted. For continuous data we use histograms while bar charts display discrete data.

Data is the information collected by conducting survey or experiment. Examples of data include the income salary of employees, medical records of students, and economic growth of a country.

## Introduction to descriptive statistics

### Levels of Measurement

When collecting measurement data, we use different scales.

- Nominal scale simply gives labels to data elements. There is no clear order or ranking. Examples are gender (male or female), eye color (black, blue, brown) etc.
- Ordinal scale is nominal scale plus ordering. For example if we have test grade of students we can measure them as A, B, C, and D using ordinal scale by giving grade A higher ranking than grade B and so on. A second example is asking people if they are satisfied with a TV channel program. Using ordinal scale, we have very satisfied, satisfied, not satisfied.

Each level of measurement is reported using specific statistical measure as shown below.

Level of measurement	What do we report
Ordinal data	Median and IQR
Nominal data	Mode and range
Continuous(scale) data	Mean and standard deviation

### Frequency distribution of ungrouped data

Frequency distribution is a table that summarizes values of data and the frequency of occurrence of each value. As an example consider the number of children who were born in a particular hospital for twelve months of the year as shown below

12 (Jan)	5 (Feb)	12 (Mar)	12 (Apr)	19 (May)	15 (June)
9 (July)	4 (Aug)	8 (Sep)	9 (Oct)	12 (Nov)	9 (Dec)

We can summarize this data in the form of a table with two parts: Value and frequency.

Number of children (x)	Frequency (number of months)
4	1
5	1
8	1
9	3
12	4

## Introduction to descriptive statistics

15	1
19	1

Notice that in the frequency table data is arranged from smallest to highest. When we have small data like this, we present it as ungrouped data. The frequency column represents number of months. Four months the number children born are 12 each, three months 9 children each, and so on. Which month has the highest number of children born?

To analyze this data, we can use measures like mean, median, mode, range, and standard deviation. Mean, mode, and median are called measures of central location as they locate the central value of the distribution. On the other hand range and standard deviation are called measures of dispersion as they indicate how values are dispersed in the distribution.

Using the above table, we can compute these measures as follows

- Mean is the average value of the distribution. It is the summation of the values divided by the number of values.

$$mean = \bar{x} = \frac{\sum xf}{n} = \frac{126}{12} = 10.5$$

This means on average, 10 children born per month. Note the use of the symbol  $\Sigma$  called Sigma notation which stands for  $\sum x = x_1 + x_2 + \dots + x_n$

- Median is the middle value when data is arranged in either ascending or descending order.

4, 5, 8, 9, 9, 9, 12, 12, 12, 12, 15, 19

For odd data, median is the value corresponding to  $\frac{n+1}{2}$  where n is the number of values

For even data, median is the value corresponding to average of  $\frac{n}{2}$  and  $\frac{n+1}{2}$ . Example this data is even, hence median is at  $(6 + 7) / 2 = 6.5$  position in the data which is 10.5

- The first quartile (Q1) is the 25<sup>th</sup> percentile of data left to the median value. The third quartile (Q3) is the 75<sup>th</sup> value of data right to the mean  
 Inter quartile range (IQR) = Q3 – Q1  
 From our example above median = 6.5  
 Q1 = 9 Q3 = 12 IQR = 12 – 9 = 3
- Mode is the value that repeats the most often, or the value that has the highest frequency. In this case, mode is 12 children
- Range is the difference between the highest value and the lowest value in the frequency distribution. In this example the highest value is 19 and the lowest value is 4. Thus range is 15

## Introduction to descriptive statistics

- To compute variance, we construct a table of the square of mean deviation. Let the variable be denoted by  $x$  while the mean is  $\bar{x}$ . The difference  $x - \bar{x}$  is called mean deviation.

$x$	$\bar{x}$	$x - \bar{x}$
4	10.5	-6.5
5	10.5	-5.5
8	10.5	-2.5
9	10.5	-1.5
9	10.5	-1.5
9	10.5	-1.5
12	10.5	1.5
12	10.5	1.5
12	10.5	1.5
12	10.5	1.5
15	10.5	4.5
19	10.5	8.5
		$\sum (x - \bar{x}) = 0$

As we can see from the above table, the summation of mean deviations total to zero. Some deviations are positive, while others are negative and the sum is always zero.

The formula for finding variance is  $\frac{\sum(x-\bar{x})^2}{n-1}$  the square of mean deviations

Thus we add another column to our table which will represent summation of square of deviations. This sum is always positive as the square function removes negative values.

$x$	$\bar{x}$	$x - \bar{x}$	$(x - \bar{x})^2$
4	10.5	-6.5	42.25
5	10.5	-5.5	30.25
8	10.5	-2.5	6.25

## Introduction to descriptive statistics

9	10.5	-1.5	2.25
9	10.5	-1.5	2.25
9	10.5	-1.5	2.25
12	10.5	1.5	2.25
12	10.5	1.5	2.25
12	10.5	1.5	2.25
12	10.5	1.5	2.25
15	10.5	4.5	20.25
19	10.5	8.5	72.25
		$\sum (x - \bar{x}) = 0$	$\sum (x - \bar{x})^2 = 187$

$$\text{Variance} = s^2 = \frac{\sum(x - \bar{x})^2}{n-1} = \frac{187}{11} = 17$$

Since the data is small, we call the variance sample variance.

Since variance is calculated using square of mean deviations, it will yield square units such dollar squared, exam mark squared etc. Often we are interested in linear units such kilometers, dollar, volume etc. To achieve this we take the square root of the variance.

The square root of the variance is called standard deviation.

$$\text{Standard deviation} = s = \sqrt{\text{variance}} = \sqrt{s^2} = \sqrt{17} = 4.123$$

If the reader wants physical interpretation of the meaning of standard deviation, consider the following two samples where sample one is about salary of 5 employees in public sector and the other sample is about salary of 5 employees in private sector

Private sector salary (x1000)	15	16	18	19	20	Mean = 17.6
Public sector salary (x1000)	10	13	15	20	30	Mean = 17.6

Both private mean salary and public mean salary are equal. If you look at public sector row, you will see it has large variation especially at the last values of the row. We can now say that public sector salary has large standard deviation which means there is large variation or dispersion in the public sector salary from the mean salary.

Now let us illustrate number of children example in SPSS. This first step is to create name of the variable in the **variable view** and then enter values in the **data view**. This is shown below.

## Introduction to descriptive statistics

The top screenshot shows the SPSS Data Editor window in Variable View. The variable 'number\_children' is defined with the following properties:

Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align
number_children	Numeric	8	2		None	None	8	Right

The bottom screenshot shows the SPSS Data Editor window in Data View. The data for the 'number\_children' variable is as follows:

Case	number_children
1	4.00
2	5.00
3	8.00
4	9.00
5	9.00
6	9.00
7	12.00
8	12.00
9	12.00
10	12.00
11	15.00
12	19.00
13	
14	

Next select **analyze** menu. Under analyze click **descriptive statistics** and then **frequencies** to display frequencies dialog box.

## Introduction to descriptive statistics

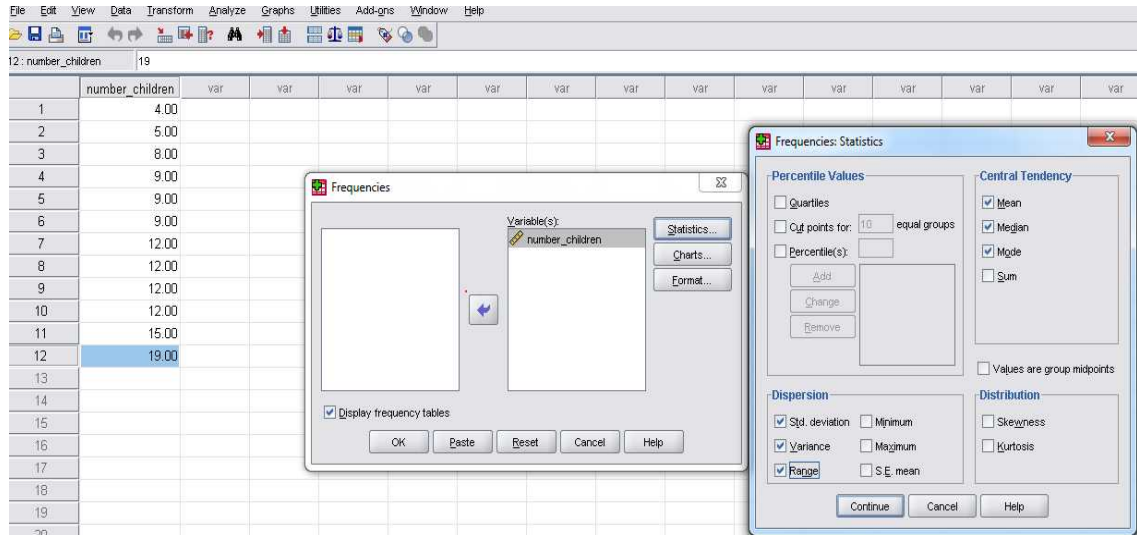
The screenshot shows the SPSS Data Editor window with a dataset named '\*Untitled1 [DataSet0]'. The data is organized into a table with 19 rows and one column labeled 'number\_children'. The values in the column are: 4.00, 5.00, 8.00, 9.00, 9.00, 9.00, 12.00, 12.00, 12.00, 12.00, 15.00, 19.00, and then blank for rows 13 through 19. The 'Analyze' menu is open, and 'Frequencies...' is selected under 'Descriptive Statistics'. Below the main window, the 'Frequencies' dialog box is open, showing 'number\_children' in the 'Variable(s):' list. The 'Display frequency tables' checkbox is checked. Buttons for 'Statistics...', 'Charts...', 'Format...', 'OK', 'Paste', 'Reset', 'Cancel', and 'Help' are visible.

Row	number_children
1	4.00
2	5.00
3	8.00
4	9.00
5	9.00
6	9.00
7	12.00
8	12.00
9	12.00
10	12.00
11	15.00
12	19.00
13	
14	
15	
16	
17	
18	
19	



## Introduction to descriptive statistics

Highlight **number\_children** and then forward it to variable(s) box using the blue arrow. Make sure **display frequency tables** is checked. Click **statistics button** and check mean, median, mode, range, variance, and standard deviation boxes. Finally click **continue** and then **ok**.



The result is shown below.

### ▸ Frequencies

[DataSet0]

#### Statistics

number_children		
N	Valid	12
	Missing	0
Mean		10.5000
Median		10.5000
Mode		12.00
Std. Deviation		4.12311
Variance		17.000
Range		15.00

#### number\_children

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid	4	1	8.3	8.3
	5	1	8.3	16.7
	8	1	8.3	25.0
	9	3	25.0	50.0
	12	4	33.3	83.3
	15	1	8.3	91.7
	19	1	8.3	100.0
Total	12	100.0	100.0	

## Introduction to descriptive statistics

The first result table with two columns summarizes descriptive and dispersion statistics. You can see the values of mean, median, mode, range, and standard deviation match exactly both hand calculation and the software.

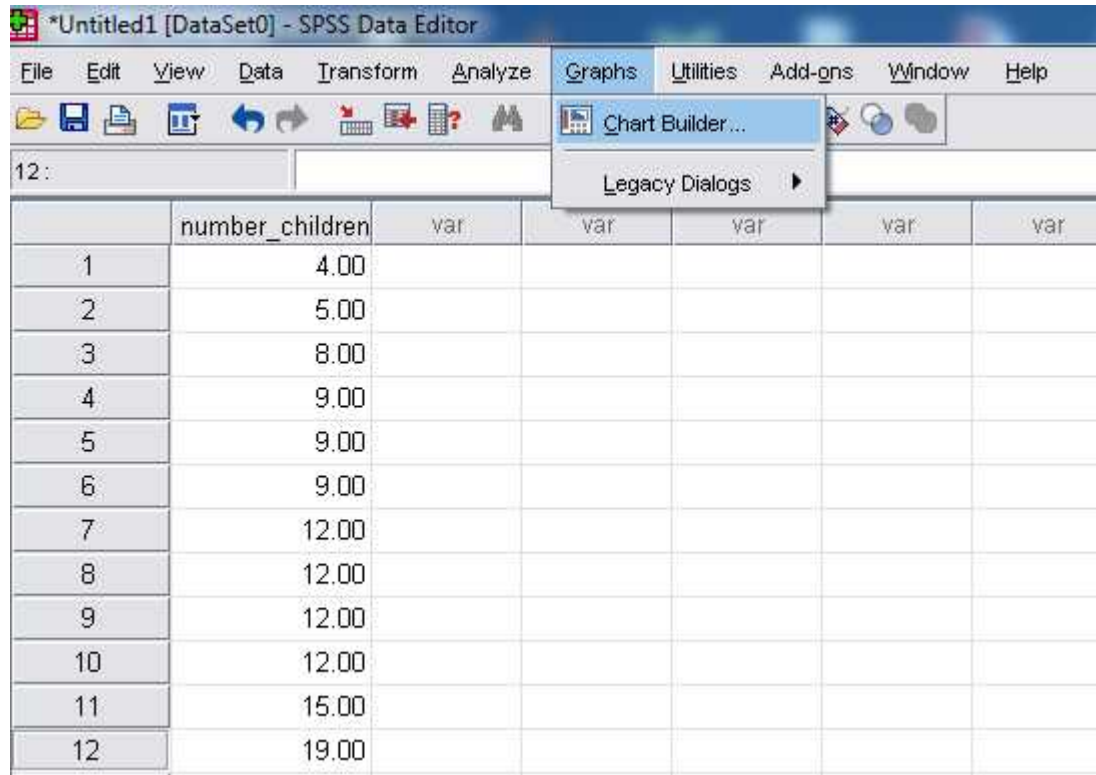
The second table is the frequency distribution of our data. The percentage column is fraction of frequency against the total (for example, first entry is 4 with frequency 1. valid percentage =  $(1/12) \times 100 = 8.3$ ).

So far we have discussed descriptive and dispersion analysis of discrete ungrouped data. The data in this example is discrete because the number of children can be counted as integer with no decimals.

The last piece remaining is graphing our data. When we have discrete data like our example, we plot it using bar chart.

In bar chart the y-axis represents frequency while the x-axis represent variable.

In data editor window, under **graphs menu** select **chart builder**.

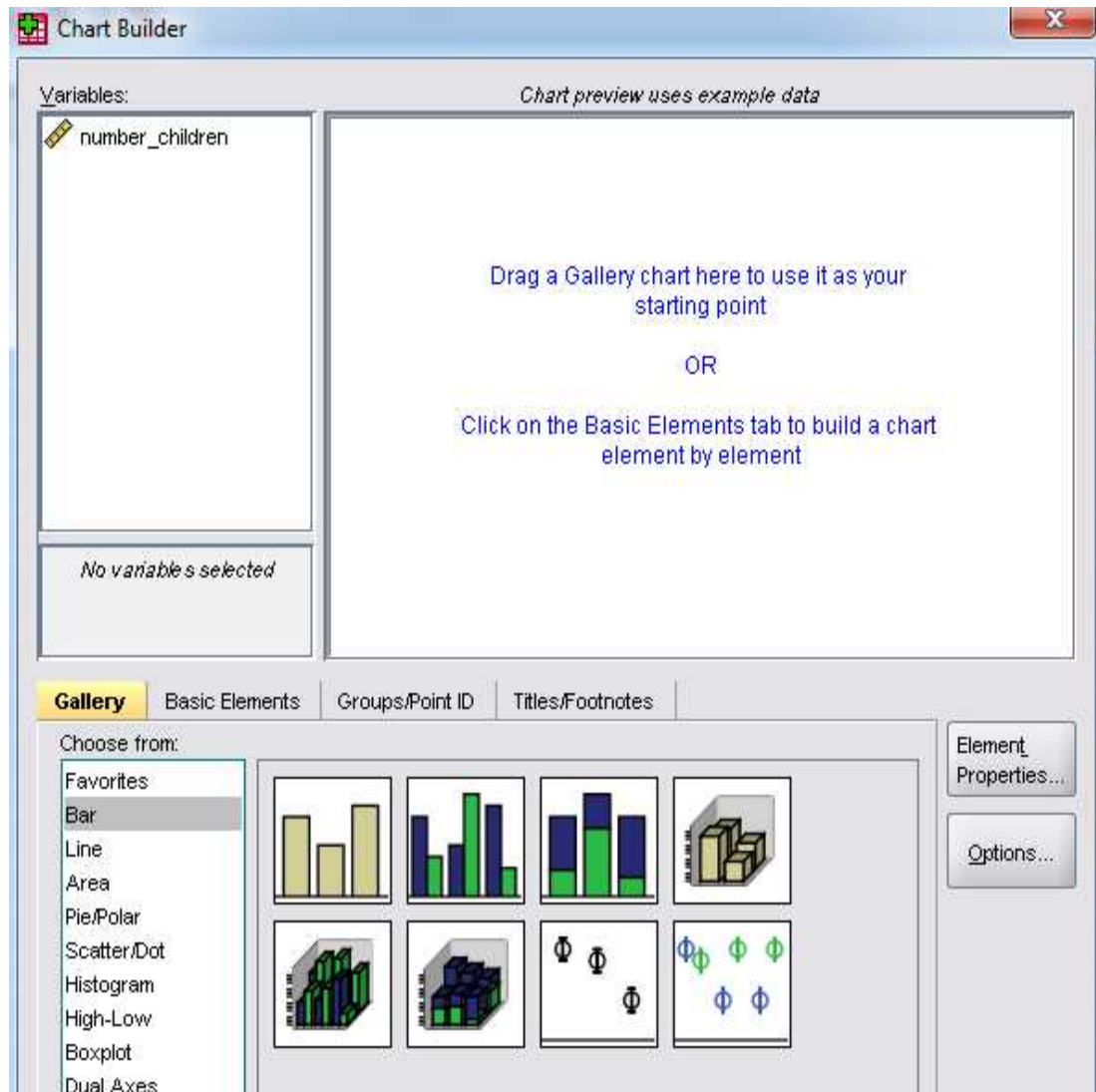


The screenshot shows the SPSS Data Editor window with the 'Charts' menu open and 'Chart Builder...' selected. The data table below is visible in the background.

	number_children	var	var	var	var	var
1	4.00					
2	5.00					
3	8.00					
4	9.00					
5	9.00					
6	9.00					
7	12.00					
8	12.00					
9	12.00					
10	12.00					
11	15.00					
12	19.00					

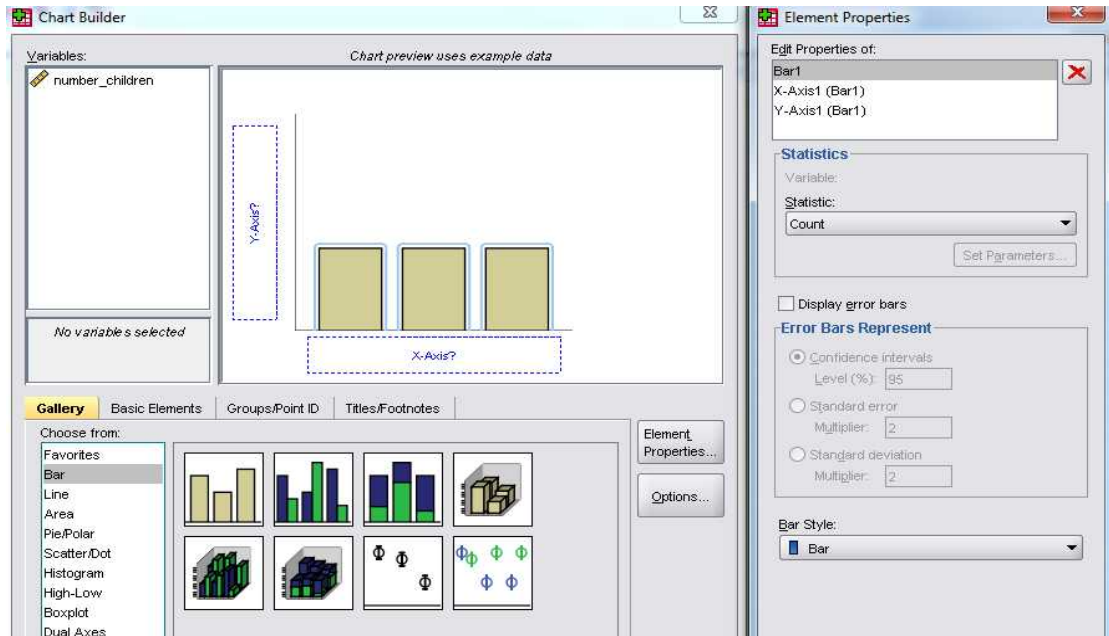
The chart Builder window is displayed

## Introduction to descriptive statistics

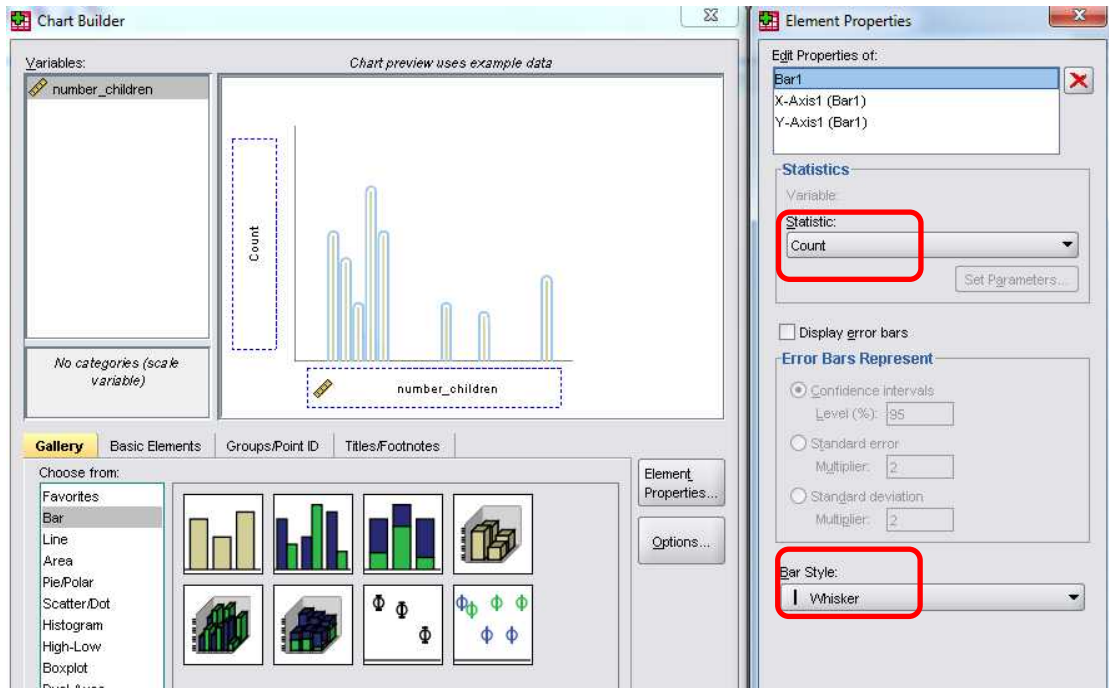


Under **Gallery** tab choose Bar. Drag and drop the first bar chart into the white area above.

## Introduction to descriptive statistics

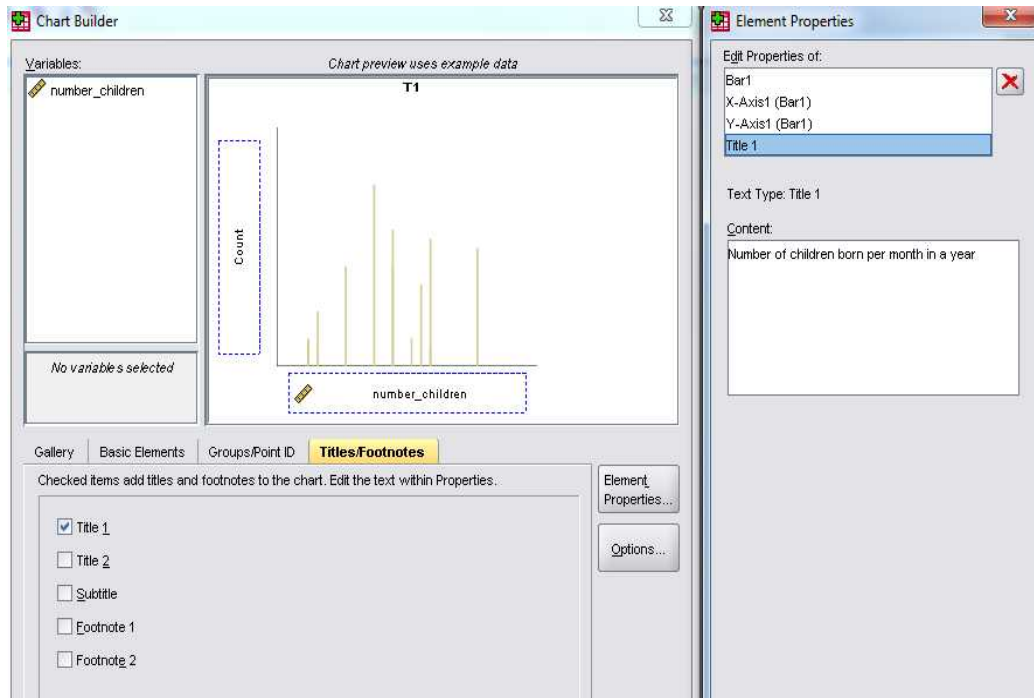


The element properties dialog box is displayed. Now right click your variable and choose scale of measurement. In this case we choose scale. Then drag your variable and drop it in the x-axis of the bar chart. In the elements properties window choose bar1 under edit properties. In the statistics group choose count statistic. Choose whisker as bar style and click apply.

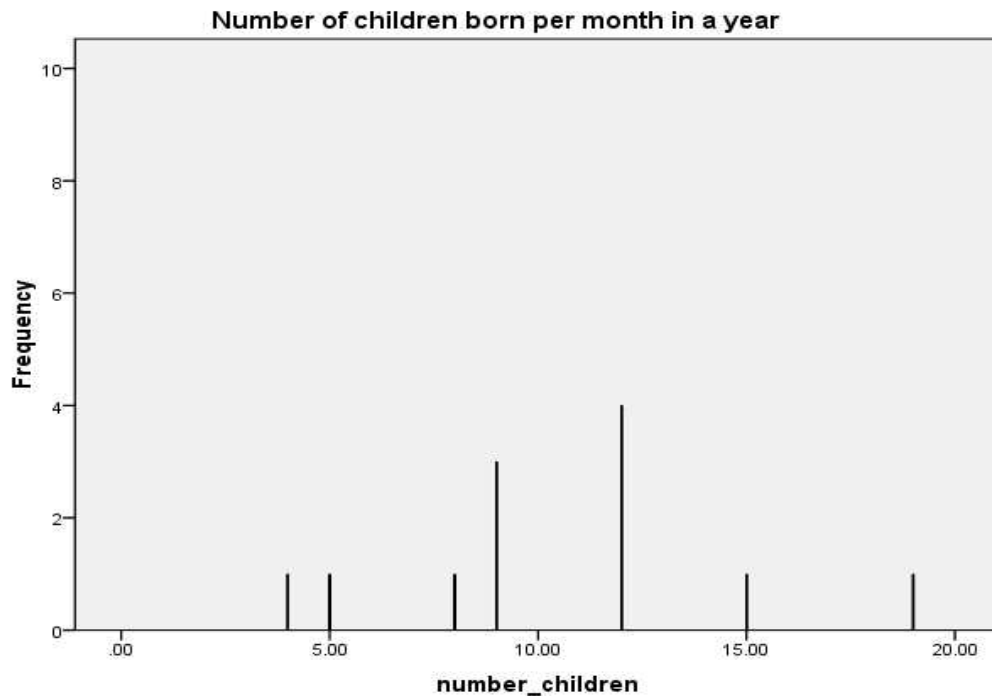


## Introduction to descriptive statistics

Label the axes. In the element properties window, set the axis label as frequency and choose an appropriate scale range. In the **title/footnotes** tab check title1. In the element properties window give title to the chart in the content box. Click apply button and then ok.



The resultant bar chart is shown in a separate window as shown below.



## Introduction to descriptive statistics

As you can see from the chart, for four months the number of children born is 12, for another three months we have 9 children. The highest number of children born for a month is 19.

### Frequency distribution of grouped data

In this situation the data is usually large and continuous. Continuous data is measured along continuous scale and assumes all values in a given interval (both integers and decimals)

As an example consider the masses of 30 students in a class to the nearest kilogram. Mass is continuous data as it can take decimal value.

55	34	39	41	51	33	49	56	35	61
59	57	44	39	38	47	43	53	60	57
45	58	60	63	46	40	50	49	51	56

Arrange this data in ascending order using SPSS. Simply right click the variable name in data view and choose sort ascending. The result is shown below.

33	34	35	38	39	39	40	41	43	44
45	46	47	49	49	50	51	51	53	55
56	56	57	57	58	59	60	60	61	63

Now that we have data arranged, we can divide it into classes. Each class has a starting point called lower class boundary and an ending point called upper class boundary.

To divide data into classes, we identify the smallest and highest value as 33 and 63 respectively.

Now we choose an appropriate class width. Normally it is taken as 5 or 10. In this example we choose 10. We construct a table and calculate the mid-value of each class. For example the mid-value of first class is  $(30+39)/2 = 34.5$

Mass (kg)	$x$	$f$	$fx$	$x^2$	$x^2f$	Frequency density
30-39	34.5	6	207	1190.25	7141.5	0.6
40-49	44.5	9	400.5	1980.25	17822.25	0.9
50-59	54.5	11	599.5	2970.25	32672.75	1.1
60-69	64.5	4	258	4160.25	16641	0.4

## Introduction to descriptive statistics

		$\sum f$ = 30	$\sum fx$ = 1465		$\sum x^2 f$ = 74277.5	
--	--	------------------	---------------------	--	---------------------------	--

$$\text{Mean} = \bar{x} = \frac{\sum fx}{\sum f} = \frac{1465}{30} = 48.83$$

Variance is given by

$$s^2 = \frac{\sum fx^2}{\sum f} - \left(\frac{\sum fx}{\sum f}\right)^2$$

$$s^2 = \frac{74277.5}{30} - (48.83)^2 = 91.54$$

Thus standard deviation is the square root of variance

$$s = \sqrt{91.54} = 9.57kg$$

To plot data on continuous scale, we use histogram. This is different from bar charts in that there is not space between histogram bars. The measured variable (in this example mass) is plotted on the x-axis. The y-axis represents frequency density.

$$\text{frequency density} = \frac{\text{frequency}}{\text{class width}}$$

For instance, frequency density of first class is  $6/10 = 0.6$

Using the same procedure we just used in bar chart, the output of SPSS gives the following histogram.

	N	Minimum	Maximum	Mean	Std. Deviation	Variance
mass	30	33.000	63.000	4.8966E1	8.872598	78.723
Valid N (listwise)	30					

### Boxplot for groups variances

Boxplot is useful graph to compare variances of multiple samples. As an example suppose that we want to compare mean income of males and mean income of females. A sample of 60 participants were collected as shown below

<b>Male</b>									
17	27	42	19	28	31	15	45	34	30
11	27	22	29	29	31	43	21	16	12
25	35	13	27	25	32	42			
<b>Female</b>									

## Introduction to descriptive statistics

48	44	40	16	23	46	32	48	39	34
32	21	14	31	43	15	23	24	18	26
44	41	30	49	24	45	10	20	17	40
22	32	47							

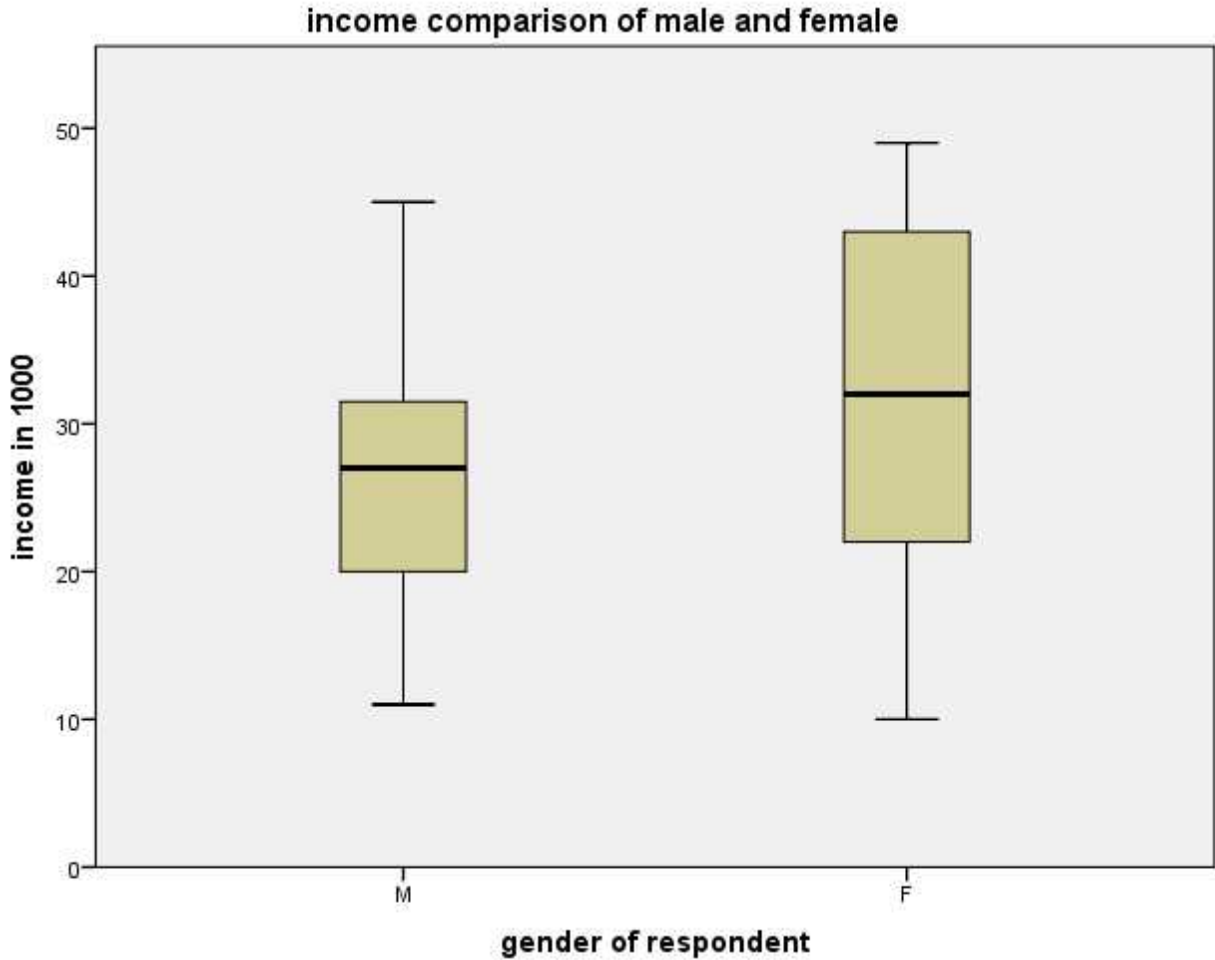
To plot the boxplot for the different groups (male and females) go to **graphs > chart builder** and under gallery choose boxplot

The screenshot shows the SPSS Chart Builder interface. On the left, the 'Variables' list contains 'instructor', 'InsRating', 'income', 'familysize', 'recoverdays', 'age', 'gen', 'test\_score', 'classSession', 'ed', and 'job'. The 'gen' variable is selected. Below the list, the 'M' and 'F' categories are visible. The 'Chart preview' window displays a boxplot titled 'T1' for 'income in 1000' across 'M' and 'F' groups. The 'Gallery' tab is active, showing 'Boxplot' selected. The 'OK' button is highlighted.

Click ok and the boxplot visualization is displayed below



## Introduction to descriptive statistics



The boxplot shows mean income of males is under 30,000 while that of female is above 30,000. The female populations has large variation from the mean. We can therefore conclude the groups have an unequal variance. This is important for some studies we will see in later chapters.

### Review questions

1. The marks obtained by 40 students in statistics test are shown below. Find the mean mark and standard deviation. Use class starting 50-59 with class width of 10.

69	54	80	77	59	66	65	81	81	88
93	57	78	71	95	91	92	86	82	89
90	83	85	75	76	62	70	81	84	87
74	72	78	84	85	79	88	95	68	91

## Introduction to descriptive statistics

2. The blood group of 10 students in a class is [A,A,AB,O,A,B,A,O,AB,A], what is the mode?
3. Select the correct answer that defines the standard deviation
  - a. The standard deviation is the square root of the mean
  - b. The standard deviation is the square root of variance
  - c. The standard deviation is the number which has the highest frequency
4. A weather student recorded temperature of a city over the past two years as shown below

16	19	33	28	34	33	31	28	25	27
28	29	30	34	28	27	21	22	23	28
28.5	23	25.6	33	34	28	29.1	30.5	23	22

Calculate mean, mode, median, standard deviation of this data in SPSS, and then plot an appropriate graph. What can you conclude from this data pattern?

## Chapter Two

### **Probability basics**

---

After completing this section, you should be able to

- Understand the meaning of probability
- Understand events, probability rules, and Venn diagram representation
- Understand the difference between dependent and independent events
- Understand conditional probability

## Probability basics

### Introduction

Probability is a quantitative measure of the likelihood of the occurrence of an event. It enables us predict the chance that something might happen on. Probability is therefore guess mathematics with certain level of confidence. Hence studying probability is very important as it will be used for the rest of chapters to study population using sample data as well as probability distributions

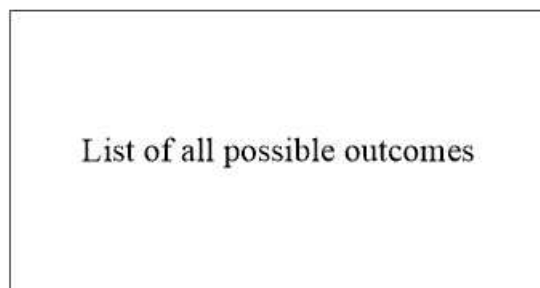
Examples of guessing something using probability is

- If someone says to you “ I am sure he will score A in the coming exam because he studied so hard” it can be translated to  $P(\text{score} = A) = 1$  where P is probability.
- If we make health care sector more private, it might probably lower health quality by 30%. This can be written as  $P(\text{health care} < \text{quality standard}) = 0.3$
- When local production of banana is doubled, our export revenue value might rise above 90% can be written as  $P(\text{export revenue} > \text{production cost}) = 0.9$

Probability values can assume the range between 0 and 1

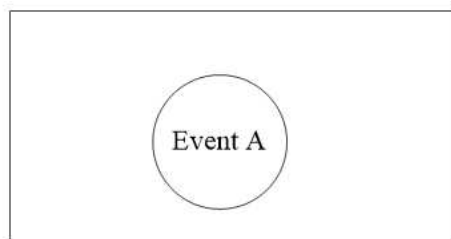
$$P(0 \leq \text{event} \leq 1)$$

In probability we often conduct an experiment about a situation. We then get sample space which contains all the possible outcomes of the experiment. Venn diagram is a useful representation to understand sample space and events



Sample space  
S

An event is a subset of the sample space. It is represented as a circle in the Venn diagram as shown below



Sample space  
S

## Probability basics

### Probability of events

Probability that event A occurs is then defined as

$$P(A) = \frac{\text{outcome in } A}{\text{total outcome in } S}$$

For example, consider tossing a fair coin. The word fair means probability of obtaining a head is the same as probability of obtaining a tail. Otherwise the coin is biased. Tossing a head will give us only two possible outcomes. Head or tail.

$$S = \{H, T\}$$

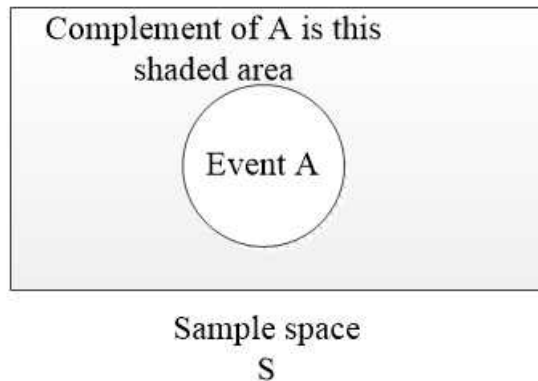
Let event A = getting a head

$$A = \{H\}$$

As can be seen there are two possible outcomes in the sample space S, while there is only one outcome in event A. let H = probability of obtaining head

$$P(H) = \frac{1}{2}$$

Remember we said that event A contains only subset of the S. The remaining subset that does not belong to A but belong to S is called complement of event A as shaded below



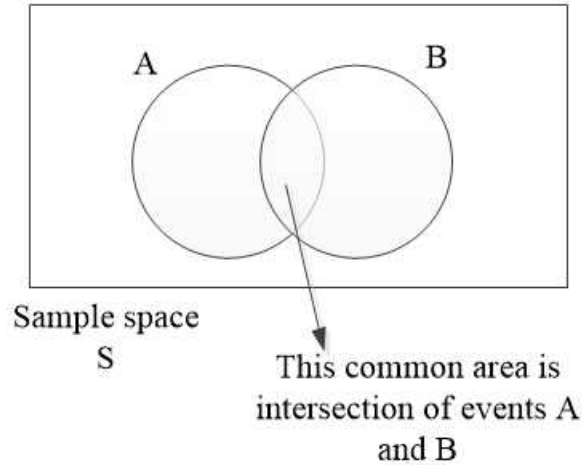
Total area of the S = 1 (probability of sample space is one)

$$P(A) + P(A^c) = 1 \text{ sum of probability of } A \text{ and its complement is always } 1$$

In most cases, we don't have only one event. Rather we are interested in multiple events as we have many questions to address in practical situations. In such case it would be helpful if we can establish some relationships between those events

Consider two events A and B that associate with the same sample space S

## Probability basics



The total shaded area that contains the two circles is called the union of events A and B, written as  $A \cup B$ . Pay attention to when forming the union, the intersection of the circles is counted twice.

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad \text{union rule}$$

A special case of the union rule is **mutually exclusive events** that cannot happen at the same time.  $P(A \cap B) = 0$  for mutually exclusive events

$$P(A \cap B) = P(A)P(B) \quad \text{interesection rule}$$

When using the intersection rule to multiply probabilities of event A and event B, the events are said to be **independent events**. This means they don't affect each other. For example, if event A is tossing a coin and event B is sampling blood group of particular school, then obtain outcome of head in event A will not be changed by event B outcome. They are independent of each other. Likewise when in another event A, the coin is tossed ten times, the outcome of obtaining head in the first toss will be independent of getting heads in the second and third tosses.

With these concepts now grasped, let us take few examples for calculating probabilities of events.

Given the following sample space  $S = \{4,6,9,7,6\}$ , calculate the probability of outcome 6 occurring.

We have five outcomes in the sample space.

Let event A = obtaining number 6 from the S

The number 6 occurs two times. Hence  $A = 2$  and  $S = 5$

$$P(A) = \frac{2}{5}$$

## Probability basics

Assume blood group of 20 students of a particular class is recorded. Hence  $S = \{A, AB, B, A, A, O, B, A, AB, O, A, O, B, AB, A, B, AB, O, A, B\}$ . What is the probability that a student selected at random has blood group B?

Let event  $B =$  student has blood group B

Total number of possible outcomes in event B is 5 while  $S = 20$

$$P(B) = \frac{5}{20}$$

What is the probability that a student selected at random has blood group A?

Let event  $A =$  student has blood group A

Total number of possible outcomes in event A is 7 while  $S = 20$

$$P(A) = \frac{7}{20}$$

Now what is the probability that a student selected at random has either blood group A or group B? Remember this is union of event A and event B. Also because the selected student cannot have both group A and group B at the same time, these are **mutually exclusive events**

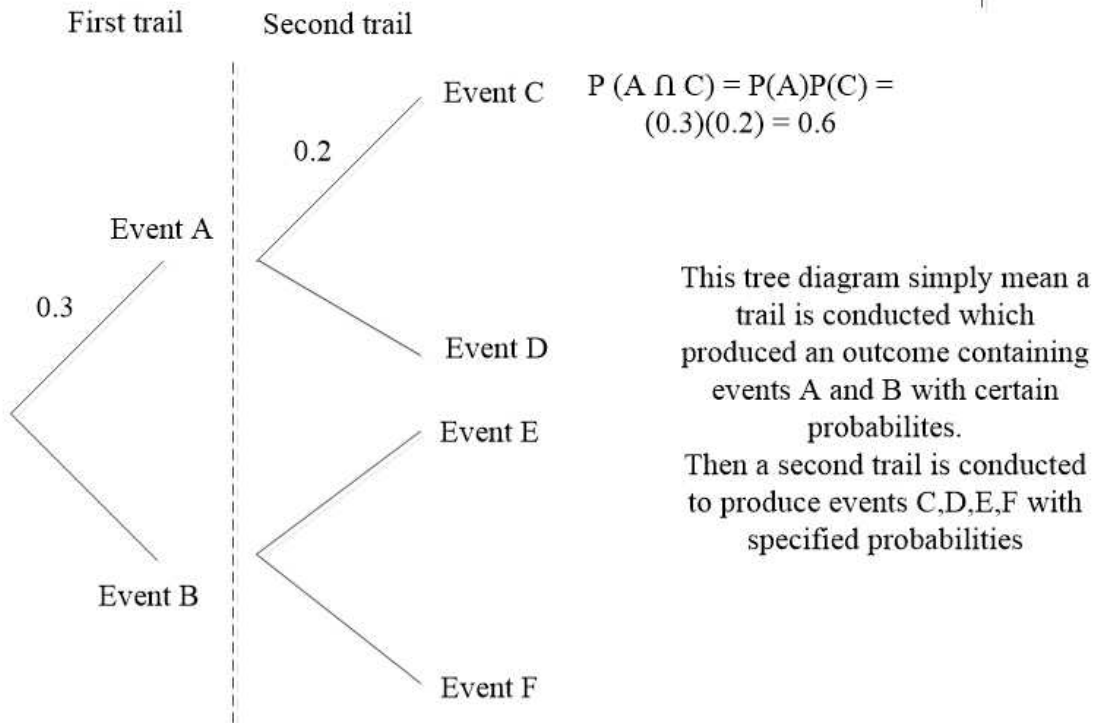
$$P(A \cup B) = \frac{7}{20} + \frac{5}{20} = \frac{12}{20}$$

Now suppose two students are selected at random. What is the probability that the first student has blood group A and the second student has blood group B? Remember because it asks event A AND event B, these are **independent events**. The first student having blood group A will not influence what the second student will have.

$$P(A \cap B) = \frac{7}{20} \cdot \frac{5}{20} = \frac{35}{400}$$

Sometimes it more easier and convenient to draw tree diagram if we have more than one experiment or trail and then compute probabilities using the branches of the tree as shown below

## Probability basics



As an example suppose an experiment was conducted about whether computer in local shop were good or defective. A sample of 10 computers were tested and found out 8 were good while only 2 were defective.

What we have here is total number of computers of 10 which is our sample space S

Let G = event that the computer is good

Let D = event that the computer is defective

Now a second test was conducted whether the computer was HP or Toshiba. It was found out that 6 were HP brand while 4 were Toshiba brand

Let H = event that the computer is HP brand

Let T = event that the computer is Toshiba brand

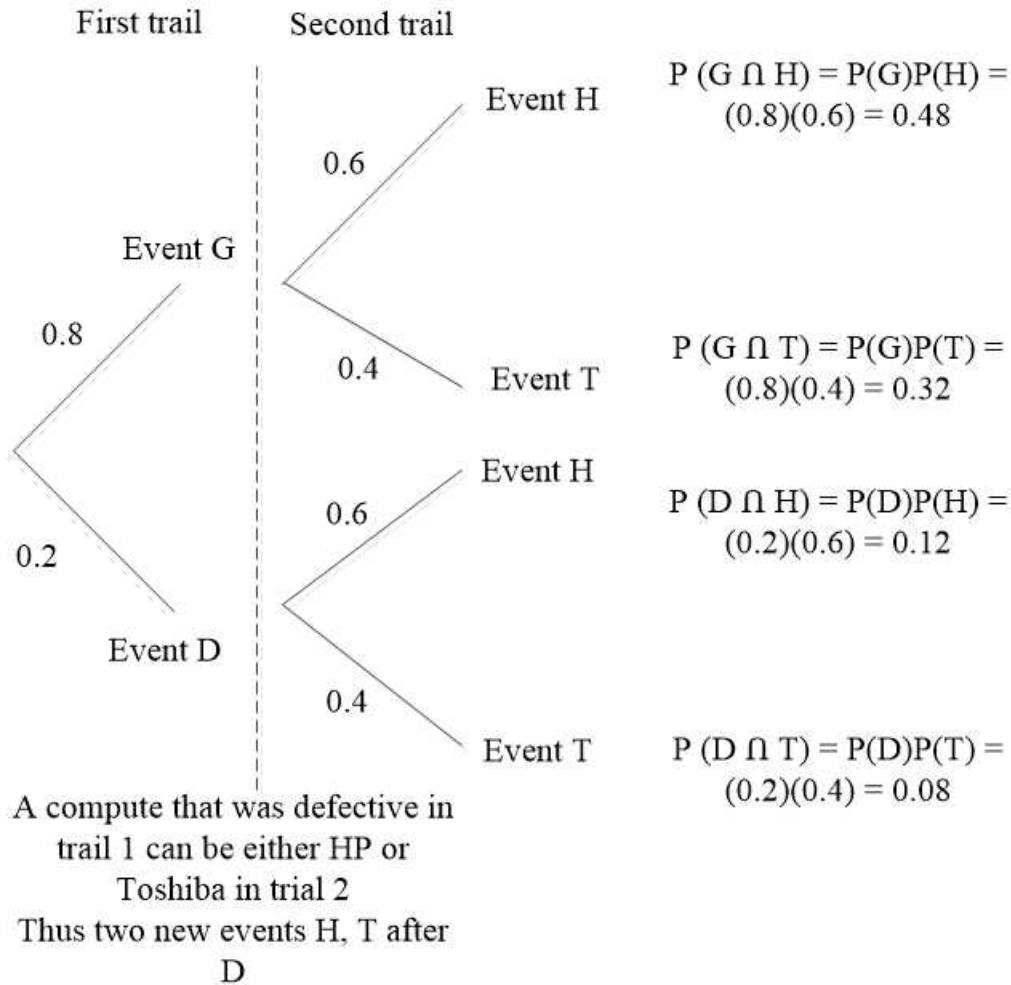
Next is to compute probability of each event as the number items in the events divided by the sample space

$$P(G) = \frac{8}{10} = 0.8 \quad P(D) = \frac{2}{10} = 0.2 \quad P(H) = \frac{6}{10} = 0.6 \quad P(T) = \frac{4}{10} = 0.4$$

Now construct the tree diagram of the two trails



## Probability basics



Let us now compute some probabilities of our interest

- What is the probability that computer selected is both defective and made of HP brand?

$$P(D \cap H) = 0.12$$

- What is the probability that a computer selected is defective and made of either HP brand or Toshiba brand

$$P(D \cap H) + P(D \cap T) = 0.12 + 0.08 = 0.2$$

### Conditional probability

Sometimes when we are finding probability of certain event, we may have knowledge that a second event has already occurred. This is the main concept in conditional probability. Probability that event A occurs given that event B has already occurred takes the following notation.

$$P(A|B) \text{ probability of event A given event B}$$

## Probability basics

Examples to understand conditional probability are

- Probability that a student selected at random from a class is doing engineering course given that he is male
- Probability that a male person selected at random is police officer given that he is older than 30 years
- Probability that GDP of a country will grow given that private sector investment has improved by 10%
- Probability that malaria prevalence will be eliminated in two years given that mosquito nets are used in every household

Let us take the last case an example to explore further our understanding of conditional probability and how to compute conditional probabilities

Suppose that we need to study malaria prevalent in normal households and hotels so that health agency can create public awareness to use mosquito nets to stay safe from the disease. Assume we have gathered the following data after conducting questionnaire study.

- Proportion of household units given that they use mosquito nets is 40%
- Proportion of household units given that they don't use mosquito nets is 30%
- Proportion of hotels given that they use mosquito nets is 51%
- Proportion of hotels given that they don't use mosquito nets is 45%

Let us try to create a contingency table out of this data to summarize our data

	Does use mosquito net	Does not use mosquito net
Household	0.4	0.3
hotel	0.51	0.45

Before computing conditional probability, the following formula is used to calculate conditional probability of event A given that event B has already occurred.

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

- What is the probability that a randomly selected subject is a household given that it does use mosquito net?

Let H = event that the subject is household

Let M = event the subject uses mosquito net

$$P(\text{household}|\text{uses mosquito net}) = P(H|M) = \frac{P(H \cap M)}{P(M)} = \frac{0.4}{0.91} = 0.44$$

The value of 0.91 is obtained the adding the two values in the column "does use mosquito net"

## Probability basics

- What is the probability that a randomly selected subject is hotel given that it does not use mosquito net

Let  $h$  = event that the subject is hotel

Let  $m$  = does not use mosquito net

$$P(h|m) = \frac{P(h \cap m)}{P(m)} = \frac{0.45}{0.75} = 0.6$$

Try to calculate the probability that a subject does not use mosquito net given that is a household.

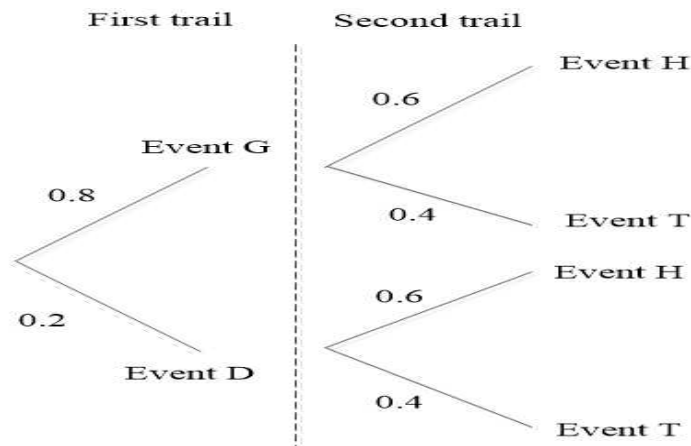
### Review questions

1. A statistic student was practicing conditional probability. He conducted study on probability of students who enrolled medicine given that there were females. He summarized his data collection in the table below

	Medicine	Other courses
Male	0.44	0.56
Female	0.56	0.44

What is the probability that a student selected has enrolled medicine given that it is female?

2. Given the following tree diagram, prove that total probability of all events is equal to 1



## Chapter Three

### **Introduction to probability distributions**

---

After completing this section, you should be able to

- Demonstrate clear understanding of what probability distribution is all about
- Understand the concept of random variable
- Differentiate between discrete random variable and continuous random variable
- Understand that a given random variable probability distribution is characterized by probability density function and cumulative distribution function
- Compute expectation and variance of probability density function

## Introduction to probability distributions

### Introduction to random variable

A random variable is a quantitative variable whose value (outcome) depends on a chance or an experiment. Its value is decided by the outcome of an experiment. Let us take flipping a coin twice as an example.

We define a random variable as  $X =$  number of tails obtained (tails obtained are random)

Then all the values of  $X$  can be listed as HH, HT, TH, and TT. In numeric terms this can be written as  $X = 0, 1, 2$ .

HH contains no tail. So  $HH = 0$ ,  $HT = 1$ ,  $TH = 1$ , and  $TT = 2$ .

As can be seen from this example,  $X = 0.5$  is not allowed as there is no half tail. Only discrete numbers are allowed. We can now call this type of random variable as discrete random variable.

Recall from the definition of probability as equal to outcome divided by total number of outcomes. Probability that a random variable  $X$  takes on a particular value  $x$  is written as  $P(X=x)$ . In this example  $X =$  number of tails obtained when a coin is flipped twice.

The probability distribution of a discrete random variable  $X$  is a listing of all possible values of  $X$  and their probabilities of occurring. The discrete probability distribution of the experiment then becomes

$$P(X = x) = \left\{ \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4} \right\}$$

In most cases, this is written as

$$P(X = x) = \begin{cases} \frac{1}{4} & \text{for } x = 0, 2 \\ \frac{1}{2} & \text{for } x = 1 \\ 0 & \text{otherwise} \end{cases}$$

This is called probability distribution function of the random variable

Let us take an example

Given the following probability distribution

$$P(X = x) = \begin{cases} a(x + 1) & \text{for } x = 0, 1 \\ ax & \text{for } x = 2, 3 \\ 0 & \text{otherwise} \end{cases}$$

Find the value of  $a$

## Introduction to probability distributions

Let us solve the problem

The first step is to draw probability distribution table to list all the values of the random variable. Just substitute  $x=0$  and  $x=1$  in the expression  $a(x+1)$  and so on

X	0	1	2	3
P(X=x)	a	2a	2a	3a

Remember from the rules of probability that sum of all probabilities of the experiment is equal to 1.

$$\sum P(X = x) = 1$$

$$a + 2a + 2a + 3a = 1$$

$$8a = 1 \text{ which gives } a = \frac{1}{8}$$

X	0	1	2	3
P(X=x)	1/8	1/4	1/4	3/8

Individual probabilities can also be calculated as

$$P(X = 1) = \frac{1}{4}$$

$$P(X \geq 1) = P(X = 1) + P(X = 2) + P(X = 3)$$

$$\frac{1}{4} + \frac{1}{4} + \frac{3}{8} = 0.875$$

### Cumulative distribution function

Cumulative distribution function is the probability distribution that the value of the random variable is less than or equal to some value  $x$ . It is denoted as  $F(x)$

$$F(x) = P(X \leq x)$$

As an example consider the following probability distribution table from the last example.

X	0	1	2	3
P(X=x)	1/8	1/4	1/4	3/8
F(x)	1/8	3/8	5/8	8/8

## Introduction to probability distributions

As can be seen the last value in the cumulative distribution is always 1. More specifically  $F(3) = 8/8 = 1$  which is the last value in  $F(x)$  row.

To reinforce understanding of cumulative distribution function, consider the following example.

$$F(x) = \frac{2x}{a} \text{ where } x = 1, 2, 3 \text{ find the value of } a$$

*the last value is 3, hence  $F(3) = 1$*

$$\frac{2(3)}{a} = 1 \text{ which gives } a = 6$$

### Expectation and standard deviation of discrete random variable

When it says expectation of random variable, it refers to mean. For the discrete case, it is computed using the following formula

$$E(x) = \sum xP(X = x)$$

Using example 1 again

x	0	1	2	3
P(X=x)	1/8	1/4	1/4	3/8

$$E(x) = 0\left(\frac{1}{8}\right) + 1\left(\frac{1}{4}\right) + 2\left(\frac{1}{4}\right) + 3\left(\frac{3}{8}\right) = \frac{15}{8}$$

Let us take one further example to reinforce understanding

Given the following distribution and  $E(x) = 2.55$  find the value of a

X	0	1	2	3
P(X=x)	0.4	a	2a	0.6

$$E(x) = \sum xP(X = x)$$

$$2.55 = 0(0.4) + 1(a) + 2(2a) + 3(0.6) \text{ gives } a = 0.15$$

## Introduction to probability distributions

Variance of the random variable is given by

$$\text{Var}(X) = E(X^2) - [E(X)]^2$$

Where  $E(X^2) = \sum x^2 P(X = x)$

Let us revisit example 1 again to demonstrate how to calculate standard deviation of the random variable

x	0	1	2	3
$x^2$	0	1	4	9
$P(X=x)$	$1/8$	$1/4$	$1/4$	$3/8$

$$E(X^2) = 0 \left(\frac{1}{8}\right) + 1 \left(\frac{1}{4}\right) + 4 \left(\frac{1}{4}\right) + 9 \left(\frac{3}{8}\right) = \frac{37}{8}$$

$$E(X) = 0 \left(\frac{1}{8}\right) + 1 \left(\frac{1}{4}\right) + 2 \left(\frac{1}{4}\right) + 3 \left(\frac{3}{8}\right) = \frac{15}{8}$$

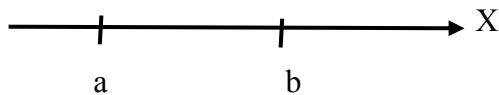
$$\text{Var}(x) = \frac{37}{8} - \left(\frac{15}{8}\right)^2 = \frac{71}{64}$$

From chapter 1, standard deviation is the square root of the variance

$$s = \sqrt{\text{Var}(X)} = \sqrt{\frac{71}{64}} = 1.053$$

### Continuous random variable

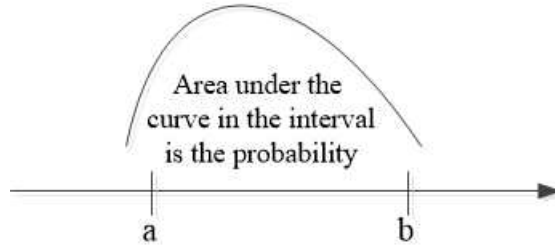
Whereas discrete random variable can take integer values, continuous random variable takes any real value in a range between a, and b. for example the temperature in a particular day can take in any value between [10, 30]. Therefore temperature distribution is a continuous random variable.



Whereas in discrete distribution we compute probability that the random variable takes a particular value as  $P(X=x)$ , continuous distribution computes probability that the random variable takes on an interval as  $P(a \leq x \leq b)$ . The area under the curve between the chosen intervals is equal to the probability.



## Introduction to probability distributions



To prove that a probability distribution is continuous, the integral of the distribution must be unity in the whole range

$$\int_a^b f(x)dx = 1$$

Expectation (mean) of the continuous random variable is given by

$$E(X) = \mu = \int_a^b xf(x)dx$$

Variance of X is given by

$$\begin{aligned} \text{Var}(X) &= \sigma^2 = E(x - \mu)^2 \\ &= \int_a^b (x - \mu)^2 f(x)dx \end{aligned}$$

### Example

To compute probabilities of continuous random variable, one needs to have basic knowledge in integral Calculus.

Give the following continuous distribution

$$f(x) = \frac{x^2}{9} \quad \text{for } 1 \leq x \leq 2$$

Find mean and standard deviation.

$$E(X) = \int_1^2 x \cdot \frac{x^2}{9} dx = \int_1^2 \frac{x^3}{9} dx = \frac{1}{9} \cdot \frac{x^4}{4} \Big|_1^2 = \frac{2^4}{9 \cdot 4} - \frac{1^4}{9 \cdot 4} = \frac{15}{36}$$

$$\text{Var}(X) = \int_1^2 \left(x - \frac{15}{36}\right)^2 \frac{x^2}{9} dx$$

## Introduction to probability distributions

$$= \frac{1}{9} \int_1^2 \left( x^2 - \frac{15}{36}x + \frac{225}{1296} \right) x^2 dx$$

$$= \frac{1}{9} \int_1^2 \left( x^4 - \frac{15}{36}x^3 + \frac{225}{1296}x^2 \right) dx$$

$$= \frac{1}{9} \left[ \frac{x^5}{5} - \frac{15}{36} \frac{x^4}{4} + \frac{225}{1296} \frac{x^3}{3} \right]_1^2$$

$$= 0.560$$

$$s = \sqrt{\text{Var}(X)} = 0.7483$$

### Review questions

- Write down the random variable X in the following cases
  - In a particular statistics class 80% of the students have passed the exam while 20% failed. A particular student was randomly selected from a class of 50 students and we want to know the probability he passed the exam.
  - The PH value of certain acidic food is claimed to be 3. A sample of 40 were tested and found acidic level of 4. What is the probability that a randomly selected food acidic level is less than 4?
- A fair coin is tossed 4 times. Is this discrete or continuous random variable? Find probability distribution table of the experiment
- Given the following random variable is given below

$$P(X = x) = \begin{cases} a & x = 1 \\ 2 & x = 3 \\ 0 & \text{otherwise} \end{cases}$$

- Is this discrete or continuous random variable? Explain
- Find the value of a
- Calculate  $P(X > 4)$
- Calculate expectation and variance

## Common discrete probability distributions

### Chapter Four

#### Common discrete probability distributions

---

After completing this section, you should be able to understand

- Binomial distribution
- Poisson distribution
- Bernoulli distribution
- Compute mean and standard deviation of these important distributions

## Common discrete probability distributions

### Binomial distribution

Suppose you want to study the number of people in a city who can speak English language. The population of the city is estimated to be 200,000 by the census board. Because it is hard to study all the population, you could take a random sample of 2000 people. For each individual of the sample, he/she can either speak English or not. This is an example of binomial random variable as it either results success (does speak English) or failure (does not speak English) with number of trails being 2000.

As another scenario, suppose you want to study population growth decline as a result of fewer marriage. If you want to know number of people who are married, you could define binomial random variable as  $X$  = number of married in 1000 people. Hence number of people married is success, while those who are not married is failure.

Let us define new random variable  $X$  as

$X$  = number of people who can speak English

A person selected from the sample can then have two answers

- Can speak English (success)
- Cannot speak English (failure)

A random variable that results in either success or failure repeated in  $n$  trials is called binomial random variable

For a random experiment to be characterized as binomial, it must satisfy all the following conditions

- The number of trails ( $n$ ) carried out is finite
- The trails are independent
- Each trail has either success or failure outcome
- Probability of success ( $p$ ) is the same for each experiment

The binomial random variable is written as

$X \sim B(n, p)$  where the symbol  $\sim$  means  $X$  is distribution of

Where  $n$  = number of trails and  $p$  = probability of success

If probability of success is  $p$ , then probability of failure is  $q = 1 - p$

What we are interested is probability of  $r$  successes in  $n$  trails  $P(X = r)$  given by

$$P(X = r) = \binom{n}{r} p^r q^{n-r} \quad r = 0, 1, 2, \dots, n$$

As an example, find  $P(X=2)$  if  $X$  is binomial distribution  $B(10, 0.5)$ . This simply asks 2 successes in 10 trials. Hence  $n = 10$  and  $r = 2$ . If probability of success  $p = 0.5$ , then probability of failure  $q = 1 - p = 1 - 0.5 = 0.5$

## Common discrete probability distributions

$$P(X = 2) = \binom{10}{2} 0.5^2 0.5^{10-2} = 0.0439$$

Where  $\binom{10}{2} = \frac{10!}{2!(10-2)!}$  and computed using factorial

Expectation and variance of binomial distribution is given by

$$E(X) = np \quad \text{Var}(X) = npq$$

From the above example of B (10, 0.5)

$$E(X) = (10)(0.5) = 5$$

$$\text{Var}(X) = (10)(0.5)(1 - 0.5) = 2.5$$

$$\text{Standard deviation } s = \sqrt{2.5} = 1.581$$

As further exercise, suppose 1000 people are sampled for marital status questionnaire. Suppose that we believe 70% of the population are married. What is that probability that 200 people are married out of 1000?

This is binomial experiment. How would you know? Well a person is either married or unmarried. We want married person which is success and we asked the same question 1000 times

In this case  $p = 0.7$ ,  $q = 0.3$ ,  $r = 200$ , and  $n = 1000$

Using the binomial distribution formula try to calculate  $P(X = 200)$

One special case of binomial is distribution is Bernoulli experiment

If binomial random variable  $X =$  number of  $r$  successes in  $n$  trials, Bernoulli random variables differs in that  $X =$  number of successes in single trial. In other words Binomial random variable is just repeated Bernoulli experiment in  $n$  times as Bernoulli can have only one trial.

As an example, suppose we flip a coin in single time (single trial). Two outcomes are produced. Head or tail. Let us assume we want to get head when flipped. Hence head is success while tail is failure. We can then label head = 1 and tail = 0

Consider we want to survey people who use particular website and those who don't use it. We sample 100 people and ask once whether they use the website. In this way we did the experiment only one time and it Bernoulli trail

Expectation and variance of Bernoulli experiment is given by

$$E(x) = p \quad \text{Var}(x) = \sqrt{pq}$$

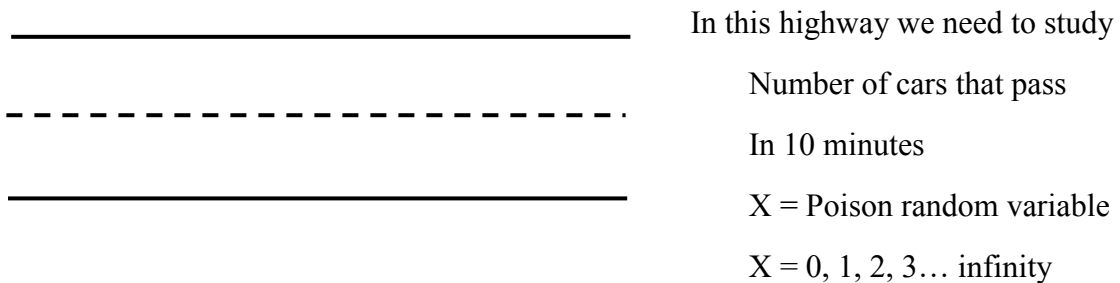
## Common discrete probability distributions

### Poisson distribution

Poisson distribution is very useful in many practical fields. For example it is used in telecommunication traffic engineering and many other social fields. Examples of Poisson random variable are

- Number of telephone calls made during busy hour
- Number of car accident in a day
- Number of cars that park in the road in one week (for example to improve parking congestion)
- Number of deaths in an hour
- Number of insect in one kilometer square of land
- Number of files downloaded from the internet in one second

In poison experiment, you select a time interval or geographic space. Then you observe number of occurrences of a given event. For example, how many observation could we get in a given time interval or space?



Poisson probability distribution function if given by

$$P(X = x) = e^{-\lambda} \frac{\lambda^x}{x!}$$

Where  $e = 2.71828$  and  $\lambda = \text{mean occurrence of an event in an interval}$

From the above highway example, we assumed 10 minute interval. But what if we make the interval large like 40 hours or even a year. Then the values  $x$  of the event observations get very large and approximates to normal distribution (chapter 5)

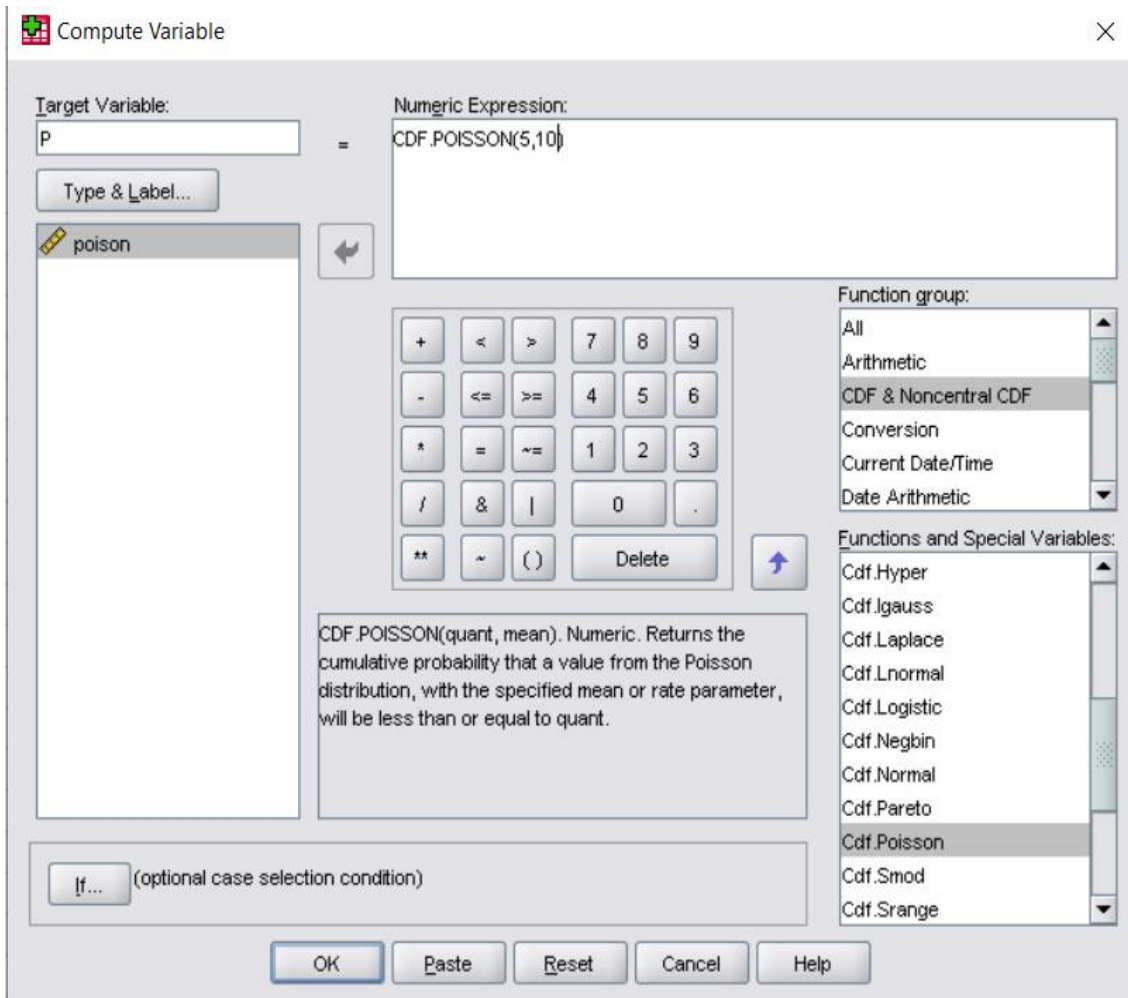
As simple example, assume we have information that mean number of cars that pass along the highway is 5 in 10 minute interval. What is the probability that 10 cars pass during this time interval

This example asks  $P(X = 10)$ , and given  $\lambda = 5$

$$P(X = 10) = e^{-5} \frac{5^{10}}{10!} = 0.0181$$

To get Poisson probability is SPSS select **transform > compute variable** as show below

## Common discrete probability distributions



There is very little probability of 1.81% of observing 10 cars pass in 10 minute interval

Expectation and variance of Poisson random variable is given by

$$E(X) = \lambda \quad \text{Var}(X) = \lambda$$

From this example,  $\lambda = 5$  and  $s = \sqrt{\lambda} = \sqrt{5} = 2.236$

## Common discrete probability distributions

The following table gives summary of the important discrete probability distribution discussed thus far.

Distribution	Random variable X	Probability function	Expectation	Variance
Binomial	X = number of successes in n trials	$P(X = r) = \binom{n}{r} p^r q^{n-r}$	$np$	$npq$
Bernoulli	X = number of successes in one trial	$P(X = x) = \begin{cases} p & x = 1 \\ 1 - p & x = 0 \\ 0 & \text{otherwise} \end{cases}$	$p$	$pq$
Poisson	X = number of event occurrences in a given interval or space	$P(X = x) = e^{-\lambda} \frac{\lambda^x}{x!}$	$\lambda$	$\lambda$

### Review questions

1. A statistics teacher writes question on the board and he thinks that randomly selected student will pass the question with 80% probability. Is this Bernoulli or Binomial experiment? What is the expected value and variance?
2. A research student wants to know approximately how many people lineup in public office per day. He believes that the mean number of people who queue at the office is 15 per day. What is the probability that 10 people queue in the first day of his observation?
3. What discrete probability distribution fits well in the following scenarios
  - a. Event that a randomly selected patient smokes or does not smoke
  - b. Event that a randomly selected student from a class of 100 does not like statistics
  - c. Event that the number of cars arrived at the university park in one month is 100



## The normal distribution

### Chapter Five

#### The normal distribution

---

After completing this section, you should be able to

- Appreciate the usefulness of the normal distribution in characterizing various phenomena in nature
- Describe the normal distribution curve and its features
- Use the standard normal distribution in solving continuous probability problems.
- Demonstrate clear understanding of how normal distribution approximates binomial distribution
- Practice in SPSS how to run the analysis of normal probabilities

## The normal distribution

### Introduction

The normal distribution is a continuous probability distribution. Recall that a continuous distribution is one in which the area under the interval assumed by the random variable gives the probability. This is in contrast to discrete distribution in which a specific value of the random variable gives probability.

The normal distribution is most popular and widely used continuous distribution. This can be attributed to the fact that not only it describes natural phenomena, but also we can use it to approximate discrete random variables such as the binomial when the number of trials get large.

Continuous variables that use normal distribution include, but not limited to

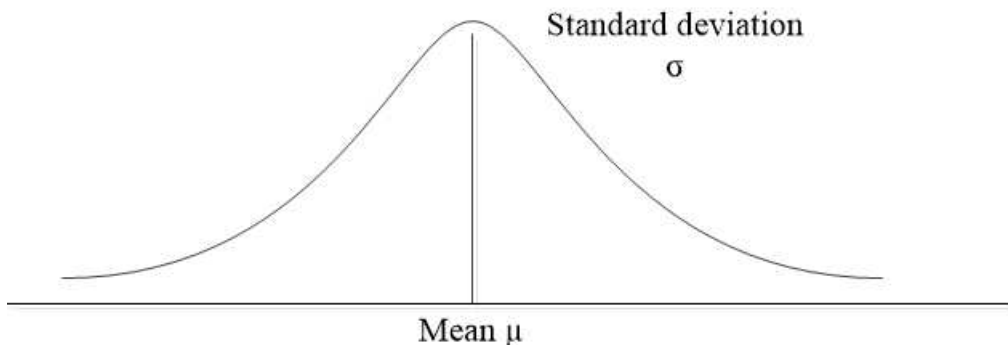
- height and weight data
- student test scores
- scientific measurements
- Business problems such as quality of new product, level of service satisfaction.

The normal variable has two parameters.

$$\mu = \text{mean}$$

$$\sigma = \text{standard deviation}$$

The distribution takes the shape of bell as illustrated below



The standard deviation tells about the width of the bell curve. Larger standard deviation will increase the width. The mean is the value the makes the curve symmetric (divides into two equal parts)

The equation of normal probability function is complicated and we omit it here. Standard normal tables are widely used to compute probability. Also in this text, we are going to use SPSS package to do most of our calculations.

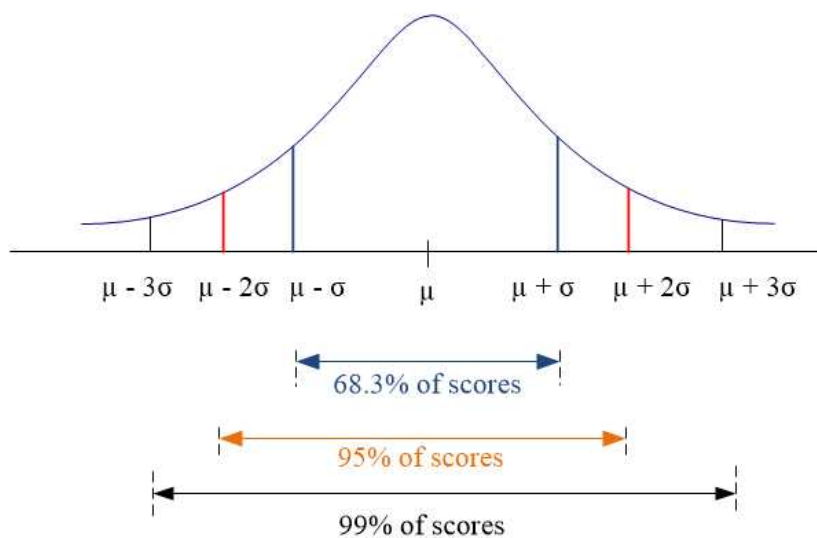
The normal curve has the following features.

## The normal distribution

- The normal curve is plotted using two parameters. The mean and the standard deviation.
- The mean, mode and the median all coincide at the highest point of the curve.
- The normal curve is symmetric with respect to the mean. Thus area to the left of mean is equal to area right of the mean
- The curve extends to infinity on either side of the mean.
- Area under the curve gives the probability of the normal distribution.

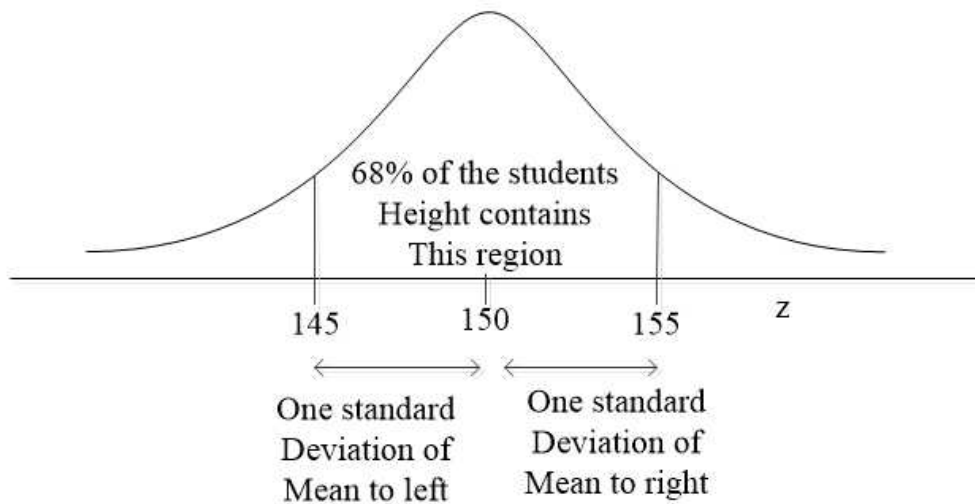
The empirical rule states that

- Approximately 68% of the distribution falls within one standard deviation of mean.
- Approximately 95% of the distribution falls within two standard deviation of mean.
- Approximately 99% of the distribution falls within three standard deviation of mean.



Let us take simple example to clarify these empirical rules. Taking 68% as an example, suppose the heights of 10 students in a class were assumed to be normally distributed with mean height of 150cm and standard deviation of 5cm as shown in the below curve

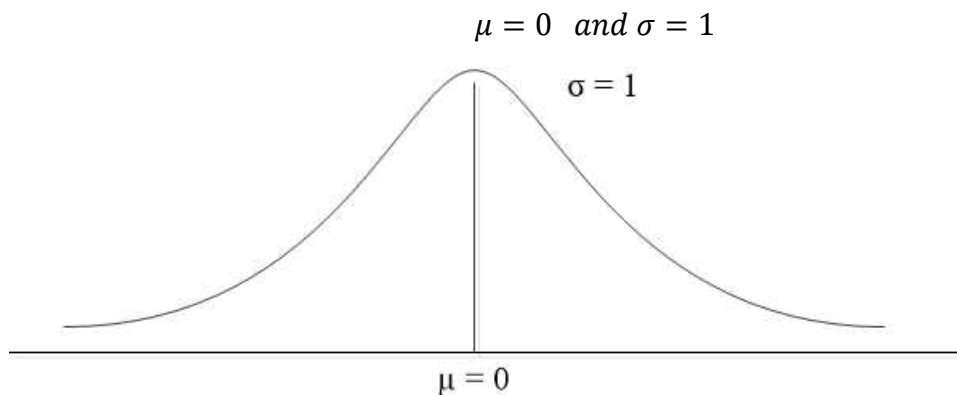
## The normal distribution



Normal distribution with different mean and standard deviation will require different tables to compute. In order to use one table for all normal probabilities we need to define standard normal distribution such that all probabilities are converted into this standard.

### Standard normal probabilities

The standard normal distribution has



For a given normal random variable value  $x$ , it is possible to convert it to standard normal variable  $z$  using the following formula

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

Once we find the  $z$ -value, we use the standard table or SPSS to find the probability. Shown below is a summary of probability cases to be encountered in normal distribution calculations.

## The normal distribution

Probability case	Example
$p(z < x)$	$p(z < 0.43) =$
$p(z > x) = 1 - p(z < x)$	$p(z > 0.43) = 1 - p(z < 0.43) =$
$p(x < z < y) = p(z < y) - p(z < x)$	$p(0.43 < z < 0.65) =$ $p(z < 0.65) - p(z < 0.43) =$
$p(z < -x) = p(z > x) = 1 - p(z < x)$	$p(z < -0.43) =$
$p(z > -x) = p(z < x)$	$p(z > -0.43) =$
$p(-x < z < y) =$ $p(z < y) - p(z < -x) =$ $p(z < y) - [1 - p(z < x)] =$ $p(z < y) + p(z < x) - 1$	$p(-0.43 < z < 0.65) =$

Consider the following example

The height of students in a class is assumed to be normally distributed with mean of 130cm and standard deviation of 50 cm.

- a) Find the probability that the height of randomly selected student is less than 140cm

Let  $x$  be the height in cm of students in a class

We are given that  $\mu = 130$  and  $\sigma = 50$

The z-values is given by

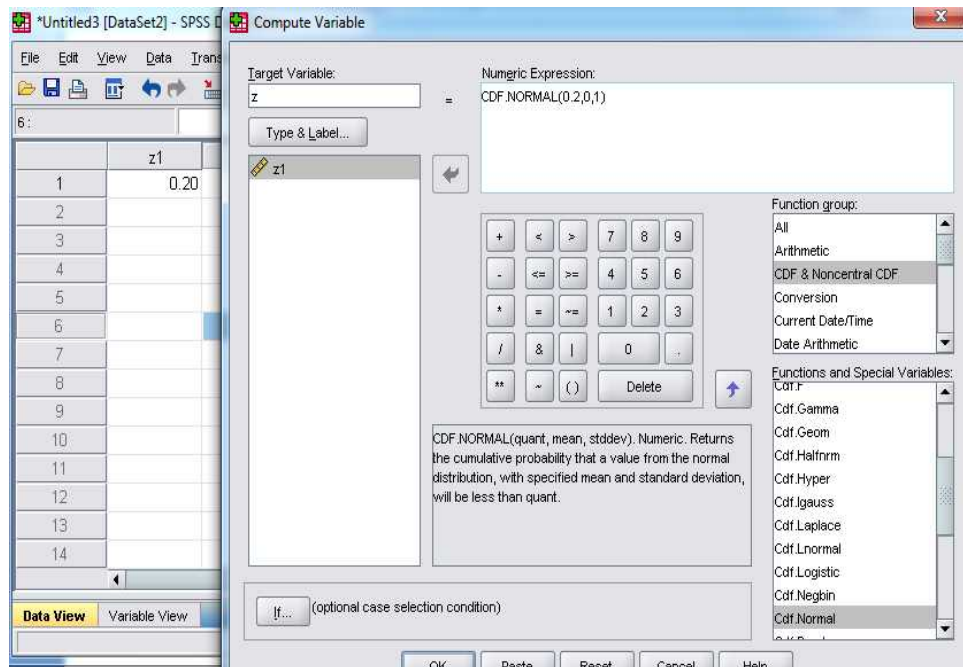
$$z = \frac{x - \mu}{\sigma} \text{ where } x = 140\text{cm}$$

$$z = \frac{140 - 130}{50} = 0.2$$

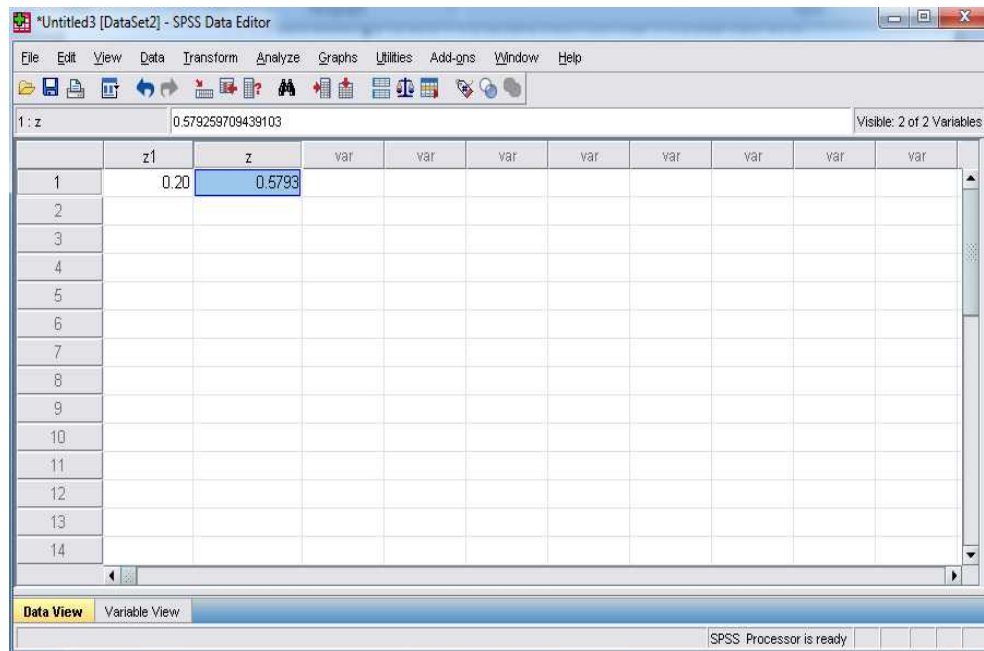
Then in probability terms this is written as  $p(z < 0.20)$

In the SPSS, create single variable in the **variable view**. Insert 0.20 in the **data view**. Go to **transform** menu and choose **compute variable**. In the function group choose **CDF & Noncentral cdf**. In the **functions and special variable** double click **cdf. Normal**. Name the **target variable**.

## The normal distribution



Click ok.



Thus  $p(z < 0.20) = 0.5793$

This means that there is 58% chance that the student height is less than 140cm

b) Find the probability that the height is one standard deviation of mean

The statement one standard deviation of mean is written as  $\mu \pm \sigma = 130 \pm 50 = 80, 180$

## The normal distribution

This gives  $p(80 < x < 180)$

$$z_1 = \frac{80 - 130}{50} = -1 \text{ and } z_2 = \frac{180 - 130}{50} = 1$$

Then using SPSS, we need to find the probability corresponding to  $p(-1 < z < 1)$

Looking at the probability case chart, this is equivalent to  $p(z < 1) + p(z < -1) - 1$

Following the same procedure in SPSS as in part a, we will get approximately 0.683

This means that there is 68.3% chance that the student height is one standard deviation of mean. This is known as the 68% empirical rule.

Consider another example

Every year car racing competition is conducted. This year 60 people take part the race. Statistical analysis has shown that the time needed to complete the race is normally distributed with mean of 40 minutes and standard deviation of 15 minutes.

- a) Find the probability of completing the race in 60 minutes or less

Let  $x$  represent the time needed to complete the race.

$$\mu = 40 \text{ and } \sigma = 15$$

We are asked to compute  $p(x \leq 60)$

$$z = \frac{60 - 40}{15} = 1.333$$

*Thus in standardized probability, we have  $p(z \leq 1.333) = 0.91$*

This means there is 91% probability of completing the race in 60 minutes or less.

Try to solve this problem in SPSS as illustrated in example 1

To solve this problem using standard normal table in appendix D, start with the column 1.3, move horizontally until you find the row of .03 to get 0.9082 as illustrated below.

	0.01	0.02	0.03	0.04
3.3	0.9995	0.9995	0.9996	0.9996
3.2	0.9993	0.9994	0.9994	0.9994
3.1	0.9991	0.9991	0.9991	0.9992
3	0.9987	0.9987	0.9988	0.9988
2.9	0.9982	0.9982	0.9983	0.9984
2.8	0.9975	0.9976	0.9977	0.9977
2.7	0.9966	0.9967	0.9968	0.9969
2.6	0.9955	0.9956	0.9957	0.9959

## The normal distribution

2.5	0.9940	0.9941	0.9943	0.9945
2.4	0.9920	0.9922	0.9925	0.9927
2.3	0.9896	0.9898	0.9901	0.9904
2.2	0.9864	0.9868	0.9871	0.9875
2.1	0.9826	0.9830	0.9834	0.9838
2	0.9778	0.9783	0.9788	0.9793
1.9	0.9719	0.9726	0.9732	0.9738
1.8	0.9649	0.9656	0.9664	0.9671
1.7	0.9564	0.9573	0.9582	0.9591
1.6	0.9463	0.9474	0.9484	0.9495
1.5	0.9345	0.9357	0.9370	0.9382
1.4	0.9207	0.9222	0.9236	0.9251
1.3	0.9049	0.9066	<b>0.9082</b>	0.9099

b) How many participants will be unable to complete the race if the race period is 70 minutes?

This question simply means for a participant to succeed the race, he should arrive in 70 minutes or less. Anyone who exceeds 70 minutes will lose the race.

In probability terms we have

$$z = \frac{p(x < 70) - 40}{15} = \frac{30}{15}$$

### Normal approximation to Binomial distribution.

In the previous chapters, we have explained in detail the binomial distribution which was suitable to apply in situations in which we have number of successes in a given experiment.

If we recall, a binomial experiment can be described as:

- n identical, independent trials
- Each trial contains two outcomes. Success and a failure
- The probability of success denoted by p on a given trial is the same for all trials.

Our task in binomial distribution is to compute the probability of x success in n trials. As we have seen in chapter 4, it is time consuming to use the binomial formula as the number of trials increases. Alternatively, the normal distribution gives us an easy approximation to binomial if these simple conditions are met.

$$np \geq 5 \text{ and } nq \geq 5$$

Once we are given these conditions and the parameters n and p, we can translate them into normal distributions parameters as follows,



## The normal distribution

$$\mu = np \text{ and } \sigma = \sqrt{npq}$$

As an example to illustrate this, suppose the Drug Administration Agency decided to know how many pills of new malaria can give relieve to patients. Their past record indicate that 20% of the malaria drug prescription can give relieve to patients. They inspected a sample of 100 pills of the new drug and administered to selected patients for full recovery. Compute the probability that only 25 pieces take effect for full recovery of the patient

In this case  $n = 100, p = 0.2$ , and  $x = 25$

As a result,  $\mu = np = 100(0.2) = 20$        $\sigma = \sqrt{npq} = \sqrt{100(0.2)(0.8)} = 4$

Now we have a normal distribution with  $\mu = 20$  and  $\sigma = 4$  to approximate our binomial data.

As for any other continuous variables, the computation of normal probability requires an interval in which the variable assumes. However, in this scenario we only have a single value and this will give us zero probability.

The interval can be written as  $25 \pm 0.5$

This will give us  $p(24.5 < x < 25.5) =$

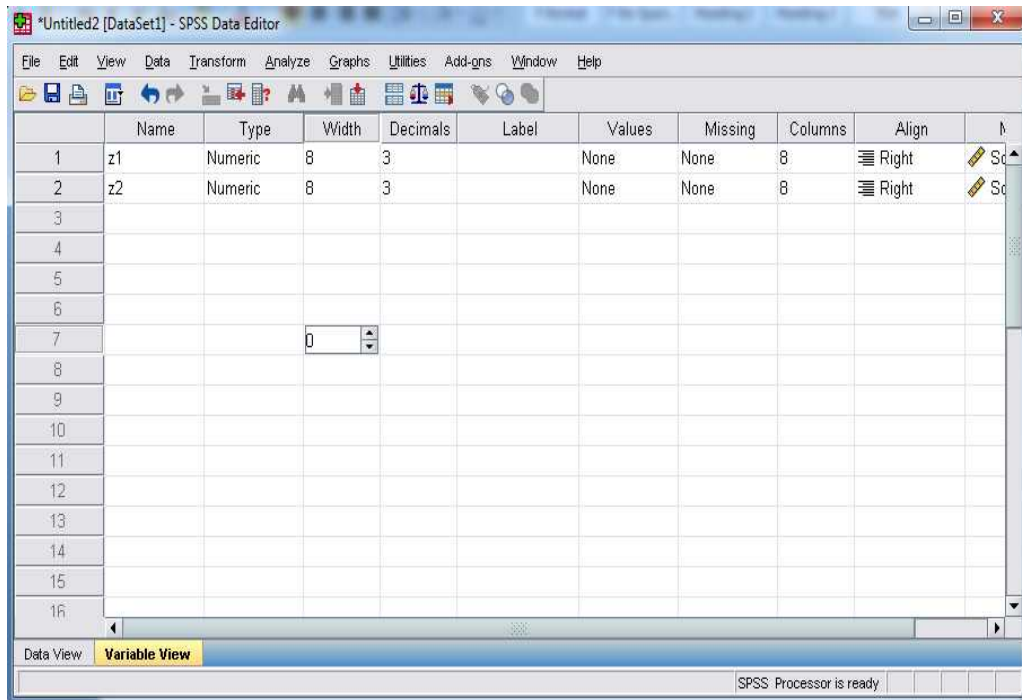
Now the z-values are  $z1 = \frac{24.5-20}{4} = 1.125$  and  $z2 = \frac{25.5-20}{4} = 1.375$

Our task is to compute  $p(1.125 < z < 1.375) = p(z < 1.375) - p(z < 1.125) = 0.0457$

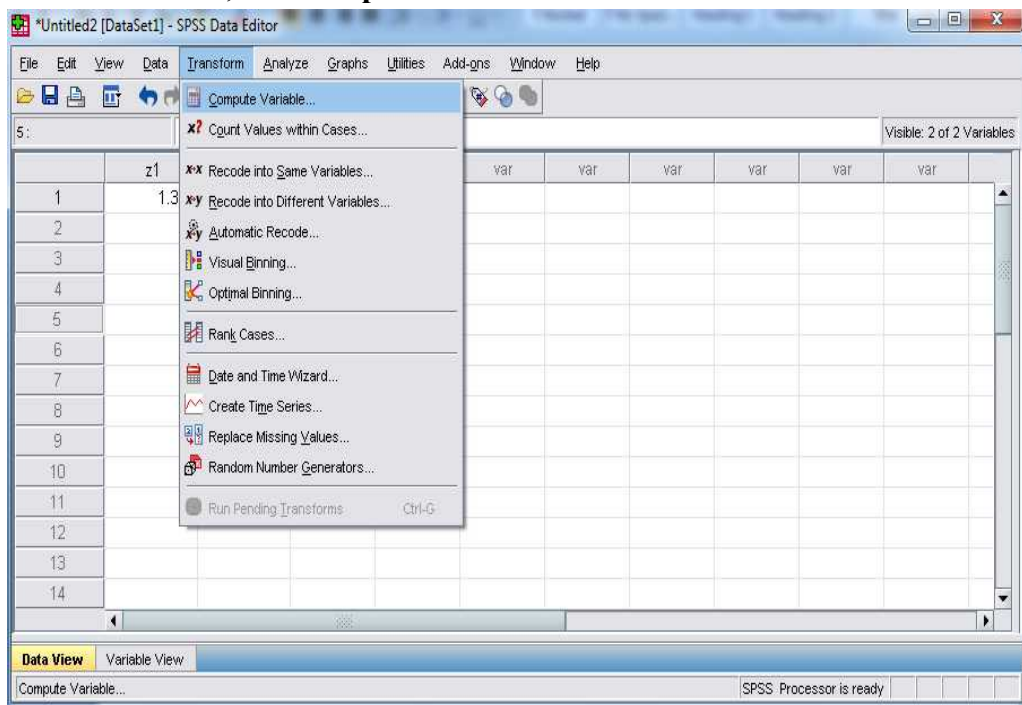
To solve this problem in SPSS, follow these steps

1. Launch the SPSS application
2. Create two variables in the variable view to represent the limits in the probability statement (say z1 for 1.375 and z2 for 1.125), then type the values in the data view.

## The normal distribution

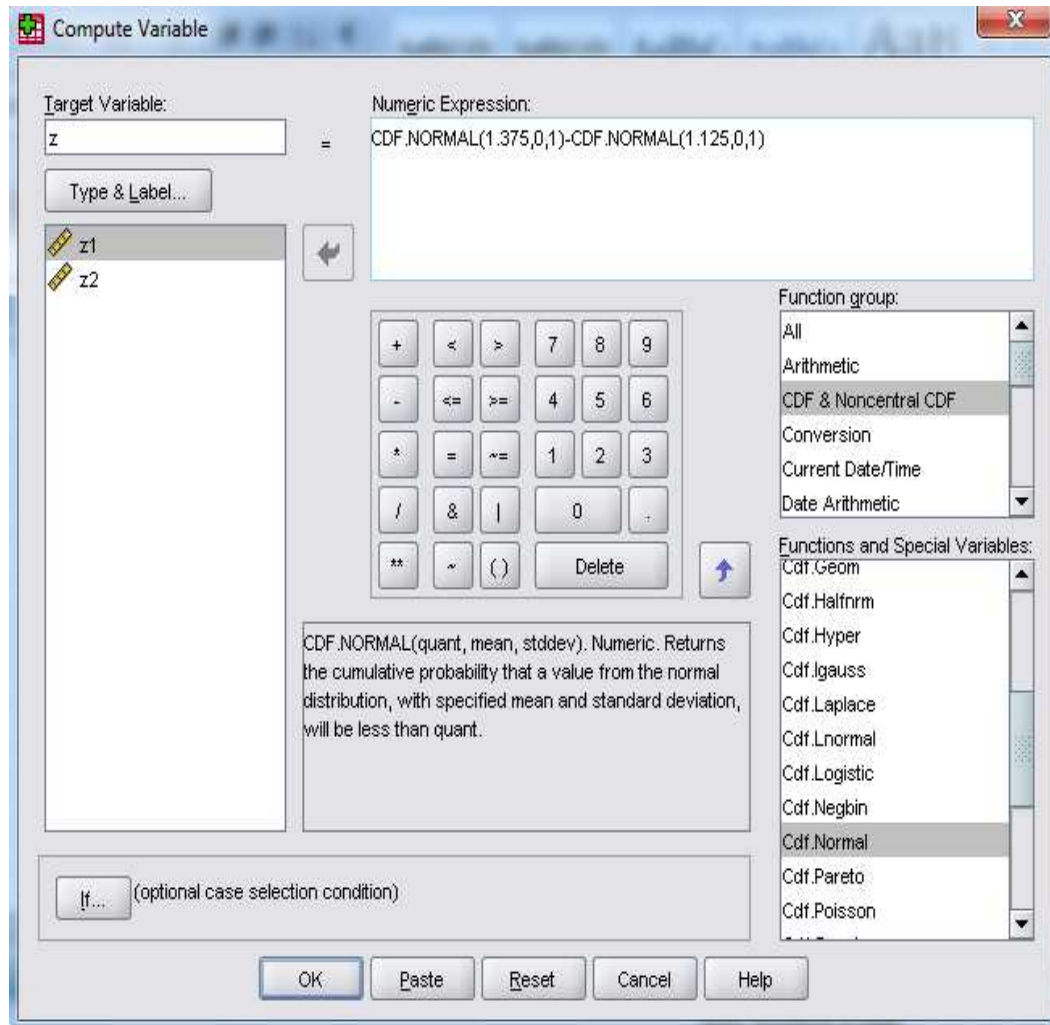


3. Go to **transform**, then **compute variable**.



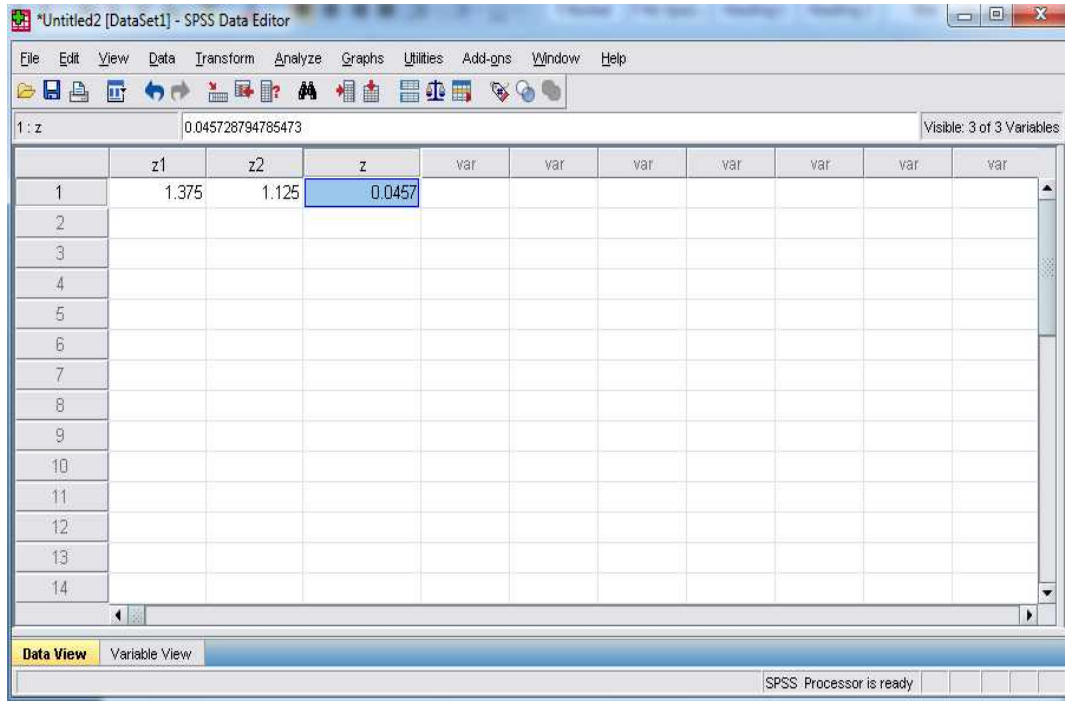
4. In the compute variable, under the function group choose **CDF & Noncentral CDF**, then under **functions and special variable** double click **cdf.normal**. Give the target variable a name.  
In the **numeric expression** box, type the z-value in the first question mark, the second is mean of 0 and third is standard deviation of 1. Then press ok.

## The normal distribution



5. The probability corresponding to  $p(z < 1.375) - p(z < 1.125)$  will then be displayed. Adjust the decimal value in the variable view to 4.

## The normal distribution



The screenshot shows the SPSS Data Editor window with a data table. The table has columns labeled z1, z2, z, and several 'var' columns. The first row contains the values 1.375, 1.125, and 0.0457. The 'z' column value 0.0457 is highlighted in blue. The status bar at the bottom indicates 'SPSS Processor is ready'.

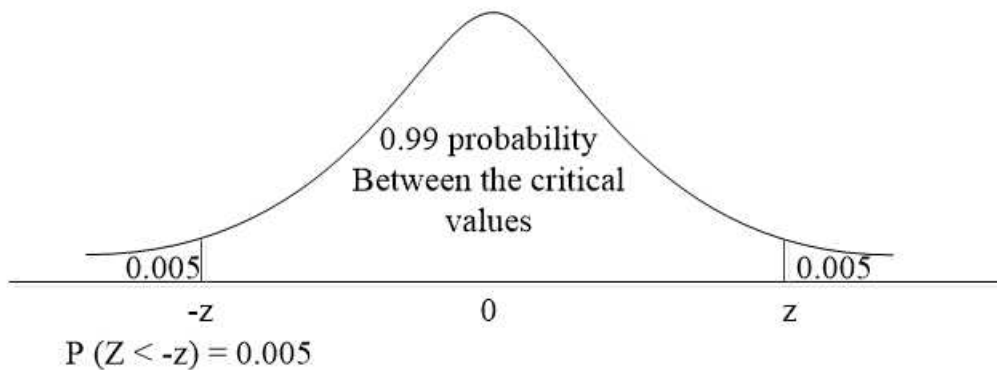
	z1	z2	z	var	var	var	var	var	var
1	1.375	1.125	0.0457						
2									
3									
4									
5									
6									
7									
8									
9									
10									
11									
12									
13									
14									

Thus the probability is 0.0457.

This means there is a very small chance of 4.5% that the new drug will take effect.

### Review questions

1. Given the following normal distribution and the area under the left tail



Find the z-value corresponding to the probability less than 0.005

2. Convert the following normal random variables into standard normal distribution
  - a. The number of people infected with Malaria in one year is 100 and normally distributed with mean of 30 per month and standard deviation of 2
  - b. The mass of 50 tuna fish is normally distributed with mean of 4 and standard deviation of 0.5

## Sampling distributions

### Chapter Six

### Sampling distributions

---

After completing this section, you should be able to

- Appreciate the usefulness of sampling in inferential statistics
- Differentiate between sample parameter and population parameter
- Understand sampling distribution
- Understand point estimates for estimating unknown population parameter
- Understand the concepts of interval estimation and confidence interval
- Compute true population parameter from sample data at particular confidence interval

## Sampling distributions

### Sampling methods

Suppose that market researcher wants to collect data on customer satisfaction on the quality of internet provided by telecom service providers. He plans his study by identifying the target population and choosing an appropriate sample to represent the population.

Population is the set of all items to be studied.

Sample is subset of population

In this study, the target population is all customers who are subscribed to the internet. This will include mobile users, fixed subscribers, organizations, and private businesses and so on.

To choose a sample, we have various types at our disposal. The choice on one type depends on the following

- Population size
- Simplicity
- nature of study

In simple random sampling, every item is given an equal chance of being selected. The only drawback for this sampling method is that it is both expensive and time consuming for larger population. In this study we can select a number of customers (respondents) from every city of the country where internet is used. To do this, we use random numbers generated by a computer program (for example SPSS)

For example, if 1000 people live in city A, then to select 100 respondents, we may use the following random numbers.

5	8	0	9	5
3	6	8	5	6
1	3	5	4	2
1	9	9	6	8
5	7	5	3	5

Let us say we need to assign three digits (ID) to each respondent. Then the first person will be 580, the second person is 953, and so on. For instant the first eight respondents will be

580, 953, 685, 613, 542, 199, 685, 753

Notice that we are reading the random numbers starting from first row to the right. People who correspond to the digits will then form the sample. When we list the sample, the list is known as sampling frame.

## Sampling distributions

But since the population is large in the study where it includes all internet users, simple random sampling will be ineffective and another method must be devised.

The second method of sampling at our disposal is called systematic sampling. In this case, the items are listed in order and starting random sample is chosen. Then after every interval, an item is selected.

In our example above, a suitable interval will be 10 (from 1000/100). Using our random table, we choose starting value as 5. Then our first eight respondents out of 100 in the sample is

5, 10, 15, 20, 25, 30, 35, 40, 45, 50

Now suppose we are interested to classify our target population into different layers such as mobile users, dedicated line to offices, and DSL to homes. This type of sampling is called stratified sampling. Each stratum (layer) is randomly sampled.

Suppose our target population of 1000 can be grouped into 540 mobile users, 150 dedicated lines, and 310 DSL homes. We want a sample of 50 respondents to represent the groups. How many do we need from each group?

$$\text{number of mobile users in sample} = \frac{540}{1000} \times 50 = 27$$

$$\text{number of dedicated line in sample} = \frac{150}{1000} \times 50 \cong 8$$

$$\text{number of DSL homes in sample} = \frac{310}{1000} \times 50 \cong 15$$

Thus to form a representative sample of size 50 of the whole population of internet users, we select 27 people from mobile users, 8 people from dedicated lines, and 15 people from DSL homes. Since the sample size of 50 is small, we use simple random sampling.

So far we have discussed the methods of sampling. Each one has advantage and disadvantage as compared in the following table.

	Advantage	Disadvantage
Simple random sampling	Simple and require little calculation since random numbers are generated by computer.	Expensive and cumbersome as the population size gets larger.
Systematic sampling	Better than simple random sampling for large population	The starting value chosen may correspond to person giving wrong answer

## Sampling distributions

Stratified sampling	Suitable when we have differentiable stratus.	
---------------------	---	--

### Review Exercise

1. A political party is interested to know the public rating of the current president in order to prepare for the upcoming election. They have questioned 1000 people living in different regions.
  - a) What is the sample size in this study
  - b) Explain how to use simple random sampling to select 1000 people from the population
  - c) Explain how to use systematic sampling to select 1000 people from the population
  - d) Assume the population can be grouped into different strata such as politicians, students, workers, and elders. 300 politicians, 200 students, 400 workers, and 100 elders are invited for an interview to create a representative sample of size 800. How many people from each group are needed?

### Sample Statistics

In the previous discussion we explained methods of sampling. We said that to carry out an investigation we define our target population which includes all items under consideration such as all individuals suffering from heart disease in a particular city. Since it is both costly and difficult to study every member of population, we choose sample to represent the population.

Once we choose sample and create sampling frame, we collect data either by surveying or interview. Now you are wondering what to do with the sample data. Of course we analyze it using both descriptive and inferential statistics for decision making. Now you are well familiar with descriptive statistics, it is time to study inferential statistics.

To start inferential statistics keep these two things in mind

- The characteristics of the population under study are unknown.
- We want the sample value to be close to the population value. This can be done by selecting unbiased sample, choosing larger sample size, or repeating the sampling procedure to create sampling distribution.

The population value of interest is called **parameter**

The sample value of interest is called **statistic**

Examples of population parameter include



## Sampling distributions

- Average test score of students in secondary schools.
- Variation in blood cholesterol level in patients suffering from heart disease.

Normally the population parameter (such as mean, proportion, or standard deviation) is unknown.

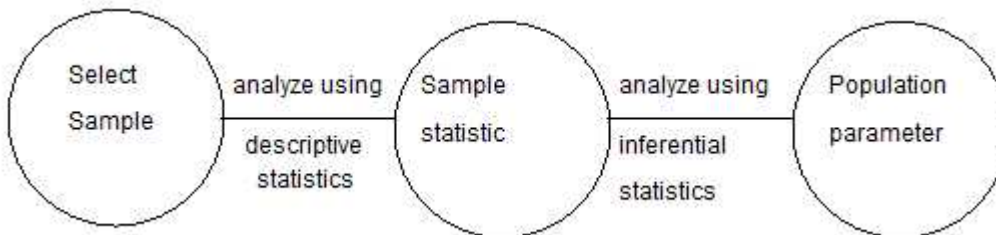
Examples of sample statistic include

- Average test score of randomly selected 500 students in secondary school
- Proportion of patients suffering from heart disease per 1000 patients.

Normally the sample statistic is known from sampling and descriptive statistics. This statistic is then used to estimate the population parameter. Two methods are used.

- Confidence interval estimation (test statistic added with margin of error to estimate unknown population parameter possible interval)
- Hypothesis test (unknown population parameter claimed and tested)

The following figure summarizes what we have said so far.



In this text our inferential analysis will be based on three population parameters, Mean, variance, and proportion. The whole discussion starts at **sampling distribution**. When we collect sample from a population we compute the sample mean, say  $x_1$ . If we draw a second sample from same population, we can compute a second sample mean, say  $x_2$ .

We continue drawing as many samples as we can from the same population and each time compute the sample mean. For the  $n$ th sample the sample mean is  $x_n$ .

The means of the different sample means form a distribution called **sampling distribution of means**.

*mean of sampling distribution,*

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_N}{N} \text{ where } N = \text{number of samples}$$

### Sampling distribution statistic calculation

Consider the following two case which highlights estimation of population parameter by sample statistic

## Sampling distributions

- Labor political party in Somalia wants to project particular candidate for presidential election. Party leaders want to know proportion of electors favoring the candidate. Sample of 1000 electors were selected across all Somali provinces. 650 out of 1000 electors preferred the candidate. Thus, an estimate of population proportion of registered electors favoring the candidate is  $650 / 1000 = 65\%$ .
- Ministry of education is considering to know information about students who don't get good university medical grades in order to enforce pre-university programs. The study sampled 500 medical students. The study produced mean GPA of 3.64, hence estimate of population mean is 3.64

Let us further extend discussion of the second case. Suppose university medical GPA as well as whether each student completed pre-university was sampled for only 15 students as given in the following table.

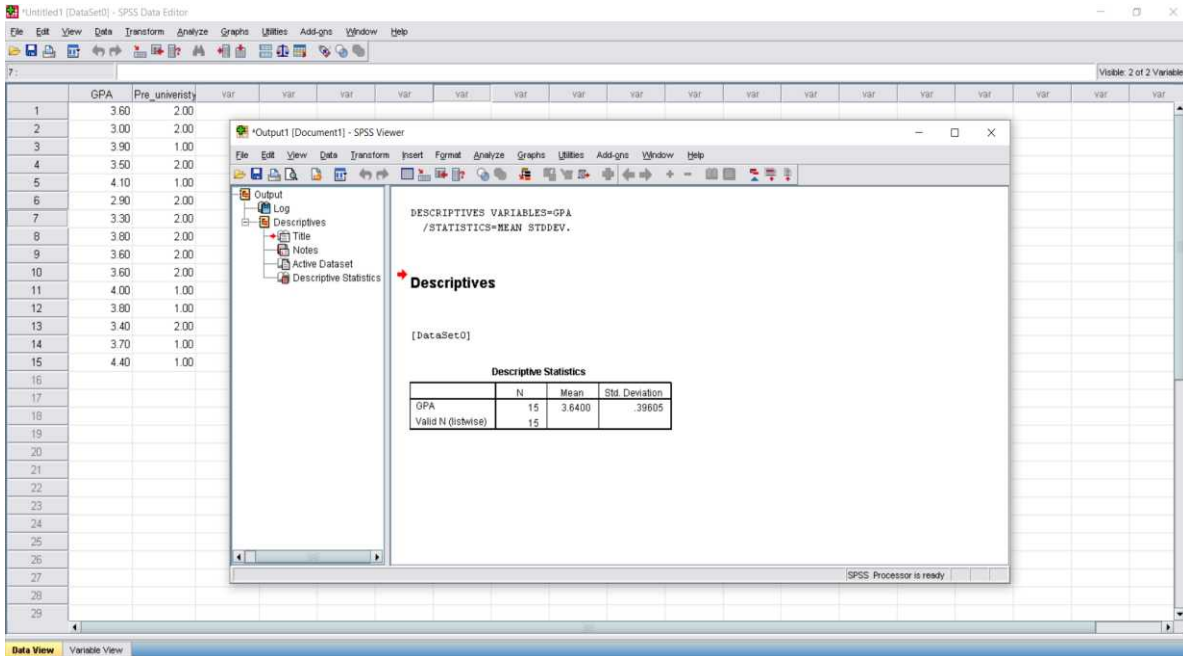
Medicine GPA	Pre-university program
3.6	NO
3.0	NO
3.9	YES
3.5	NO
4.1	YES
2.9	NO
3.3	NO
3.8	NO
3.6	NO
3.6	NO
4.0	YES
3.8	YES
3.4	NO
3.7	YES
4.4	YES

One can now compute sample statistic using simply descriptive statistics

- Mean GPA =  $\bar{x} = \frac{\sum x}{n} = \frac{54.6}{15} = 3.64$

## Sampling distributions

- Sample standard deviation =  $\sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}} = 0.396$
- Size of sample  $n = 15$
- Sampling proportion of students who took pre-university program
  - $\bar{p} = 6/15 = 0.4$



For the random sample of 15 students, we found out that

- Point estimate of population mean  $\mu$  was the sample mean  $\bar{x} = 3.64$
- Point estimate of the population proportion  $p$  was the sample proportion  $\bar{p} = 0.4$

In statistical analysis one question that may arise is what happens when you choose second, third and several other random samples and get their corresponding estimates? → We get different point estimates (different samples from the same population will give different means) as illustrate in the below table.

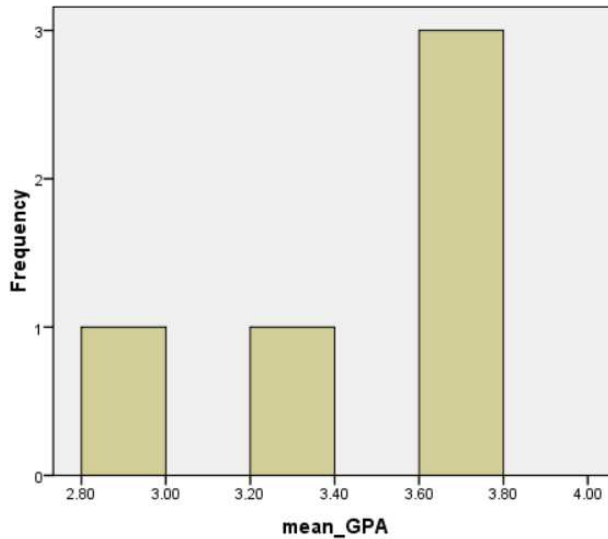
Sample	$\bar{x}$	$\bar{p}$
1	3.64	0.4
2	3.61	0.34
3	3.8	0.46

## Sampling distributions

4	2.94	0.32
5	3.4	0.33

Using descriptive statistics we can also draw frequency distribution table and histogram of this five samples.

Mean GPA	Frequency
2.5 – 3	1
3 – 3.5	1
3.5 - 4	3
	Total = 5



Recall that we said that the sample mean  $\bar{x}$  is a random variable, and its probability distribution is called the sampling distribution of  $\bar{x}$ . This brings us into an important conclusion

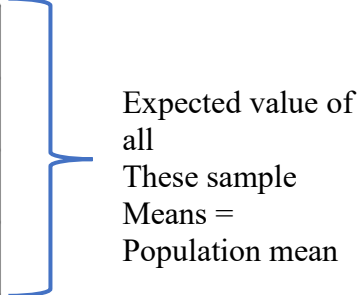
***The sampling distribution of  $\bar{x}$  is the probability distribution of all possible values of the sample mean  $\bar{x}$***

The sampling distribution of  $\bar{x}$  has

## Sampling distributions

- Expected value (mean)
- Variance and standard deviation

Expected value is the mean  $\bar{x}$  of all values that can be produced by the simple random samples. Different random samples produce different values of the sample mean  $\bar{x}$

Simple random sample #1	Sample mean $\bar{x}$ of #1	
Simple random sample #2	Sample mean $\bar{x}$ of #2	
Simple random sample #3	Sample mean $\bar{x}$ of #3	
... and so on...	... and so on...	

*With simple random sampling, the expected value of the sampling distribution*

*Of  $\bar{x}$  is equal to the population mean*

*$E(\bar{x}) = \mu$ , expected value is unbiased point estimator of the population mean*

For infinite population and to some extent for finite population (when the population size is large) the sample standard deviation is given by

$$\sigma_s = \frac{\sigma}{\sqrt{n}}$$

- Where  $n$  is sample size and  $\sigma$  is population standard deviation.
- The sampling standard deviation is also called standard error.
- As the sample size increase, the standard error of the mean decreases

### Central limit theorem

This is helpful for approximating the shape of the sampling distribution to normal distribution. Because normal distribution probabilities can be easily computed, it would be helpful if we can approximate any sampling distribution to normal. Wouldn't it?

It states that

*Choosing random samples of size  $n$  from population, the sampling distribution of the sample means take the shape of normal distribution given sample size is large enough.*

Specifically how large should the sample size be before central limit theorem is applied?

## Sampling distributions

- Sample size  $n > 30$

Using central limit theorem you could transform mean of the sampling distribution to standard normal distribution as follows

$$\begin{aligned} \text{given } \bar{x}, \text{ the standard normal is } \bar{z} &= \frac{\bar{x} - \mu}{\sigma_s} \\ &= \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \text{ test statistic of normally distributed mean} \end{aligned}$$

The sampling distribution of population proportion is the probability distribution of the all the possible values of the sample proportion

- Expectation value of  $\bar{p}$  is equal to the population proportion.
- $E(\bar{p}) = p$  where  $p$  is population proportion.
- Expected value of the sampling distribution of  $\bar{p}$  is unbiased estimator of the population  $p$
- Standard deviation of  $p$  is

$$\sigma_p = \sqrt{\frac{p(1-p)}{n}} \text{ test statistic of normally distributed proportion}$$

So far we have covered two point estimators (two sample statistic) of population parameter. However, one might ask what makes a point estimator to be called good point estimator. In other words what are the main properties of a good point estimator to model unknown population parameter? These are listed below

- *Unbiased* ... expected value of sampling distribution statistic is equal to population mean
- *Efficiency* ... the point estimator should have smaller standard error
- *Consistency* ... as the sample size increases the value of the point estimator gets closer to the population parameter

### Interval estimation of population parameter

Recall that we said in statistical analysis what we are interested is finding true value of population parameter (population mean, population standard deviation). Unfortunately in most cases this information is not available except when we have large historical data collected over many years. We said that using descriptive statistics we can compute sample statistics. Then sample mean and sample standard deviation can be used as point estimator of the unknown population parameter. This process is called inferential statistics.

## Sampling distributions

In practical sense, point estimator cannot exactly give the exact value of the population parameter, but point estimator added with some margin of error can. This process of adding and subtracting margin of error from point estimate is called *interval estimation*.

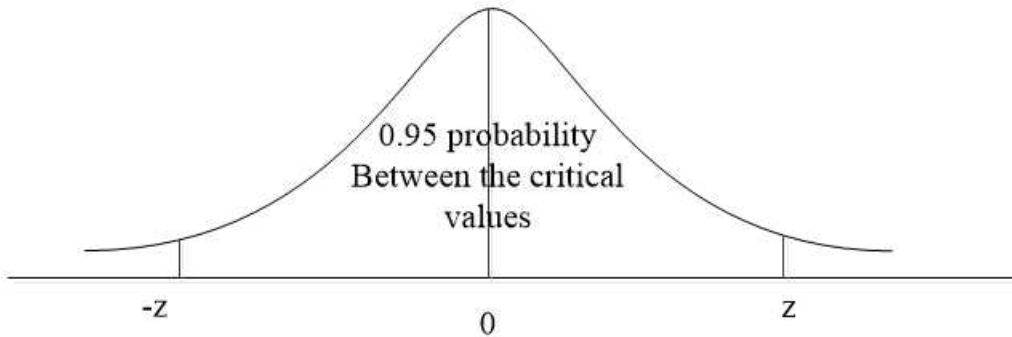
In this section, we will study

- Interval estimate of the population mean  $\bar{x} \pm \text{margin of error}$
- Interval estimate of the population proportion  $\bar{p} \pm \text{margin of error}$
- Interval estimate of the population standard deviation  $\sigma \pm \text{margin of error}$

In this chapter, we will build confidence intervals for the true population parameter given availability of the corresponding sample parameter.

Our main task is, give sample mean  $\bar{x}$  and sample standard deviation  $\sigma_s$ , can we construct an interval estimate where we will have 95% or 99% confidence that interval contains the true population mean and standard deviation respectively?. We still have 5% of doubt on this. Since the normal curve is symmetric around the mean, we have 2.5% on the right that don't contain the population mean as well as 2.5% on the left.

The confidence interval is also known as  $\alpha - \text{value}$



Using left tailed normal distribution table given in the appendix D,  $P(Z < -z) = 0.025$  will give  $z = -1.96$  as closest value to the probability of 0.025

Now the critical z values are  $-1.96 \leq z \leq 1.96$

	0.01	0.02	0.03	0.04	0.05	0.06	0.07
-3.3	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004
-3.2	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005
-3.1	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008
-3	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011
-2.9	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015

## Sampling distributions

-2.8	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021
-2.7	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028
-2.6	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038
-2.5	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051
-2.4	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068
-2.3	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089
-2.2	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116
-2.1	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150
-2	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192
-1.9	0.0281	0.0274	0.0268	0.0262	0.0256	<b>0.0250</b>	0.0244
-1.8	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307
-1.7	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384
-1.6	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475

Intuitively, the confidence interval is the interval we estimate the true population parameter to contain. For example, if we take many random samples from the same population, then at 95% probability we are certain the mean of the sampling distribution to lie in that confidence level defined by the probability of 95%.

The commonly used confidence levels with their corresponding z-values is shown below

Confidence level	z-value
90%	1.64
95%	1.96
99%	2.58

The z-value is multiplied by the standard error of the sampling distribution which is then added to and subtracted from the sample statistic to get the population parameter.

Confidence level	Confidence interval of population mean
90%	$\bar{x} \pm 1.64 \cdot \frac{\sigma}{\sqrt{n}}$
95%	$\bar{x} \pm 1.96 \cdot \frac{\sigma}{\sqrt{n}}$
99%	$\bar{x} \pm 2.58 \cdot \frac{\sigma}{\sqrt{n}}$



## Sampling distributions

--	--

### Confidence interval for a population mean

As an example, suppose that the ministry of education wants to know the average GPA of all the students in the country. This will require GPA of every student in the group to be recorded and then population mean computed.

Because the size of the population is very large, population mean is unknown and we assume population mean is normally distributed. In this case, as a statistician, you will select a sample using an appropriate sampling method, and then from the sample compute sample mean and sample standard deviation using descriptive statistics.

Suppose you selected sample size of 100 students randomly, and you computed sample mean as mean GPA of 75 with standard deviation of 30.

In this problem we then have  $n = 100, \bar{x} = 75, \sigma \approx \sigma_s = 30$

Now construct an interval estimate that the true population mean GPA will lie and you have confidence level of 95%. Let us take population standard deviation to be approximately equal to sample standard deviation of 30

Remember we said that, and always true, population mean = sample mean  $\pm$  standard error

$$\text{standard error} = z \cdot \frac{\sigma}{\sqrt{n}}$$

Where z is the critical value corresponding to the confidence level of 95% which is 1.96.

$$\text{standard error} = 1.96 \cdot \frac{30}{\sqrt{100}} = 5.88$$

This results in

$$\text{population mean GPA} = \bar{x} - z \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z \cdot \frac{\sigma}{\sqrt{n}} = 75 \pm 5.88$$

Thus at 95% confidence level, the true students mean GPA performance lies in the interval (69, 80)

To further explore this important statistical concept, suppose we now want confidence level of 97.5%

Using standard normal table in appendix D, now the critical z value is  $P(Z < -z) = 0.0125$  will give this critical value of  $z = -2.24$

	0.01	0.02	0.03	0.04	0.05	0.06
-3.3	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004

## Sampling distributions

-3.2	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006
-3.1	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008
-3	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011
-2.9	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015
-2.8	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021
-2.7	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029
-2.6	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039
-2.5	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052
-2.4	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069
-2.3	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091
-2.2	0.0136	0.0132	0.0129	<b>0.0125</b>	0.0122	0.0119

$$\text{standard error} = 2.24 \cdot \frac{30}{\sqrt{100}} = 6.72$$

This results in

$$\text{population mean GPA} = 75 \pm 6.72$$

Thus at 97.5% confidence level, the true students mean GPA lies the interval (68, 82).

As can be seen as the confidence level increases, the interval widens and the standard error increases.

Let us revisit the same problem but now assume we don't know population standard deviation ( $\sigma$  unknown). When the sample size is very small (like less than 30) and  $\sigma$  unknown, instead of using the normal distribution, we can use the t – distribution. The t – distribution uses degrees of freedom and it has similar curve as the normal distribution

Degree of freedom =  $n - 1$  where  $n$  is sample size and

$$\text{population} \sim t$$

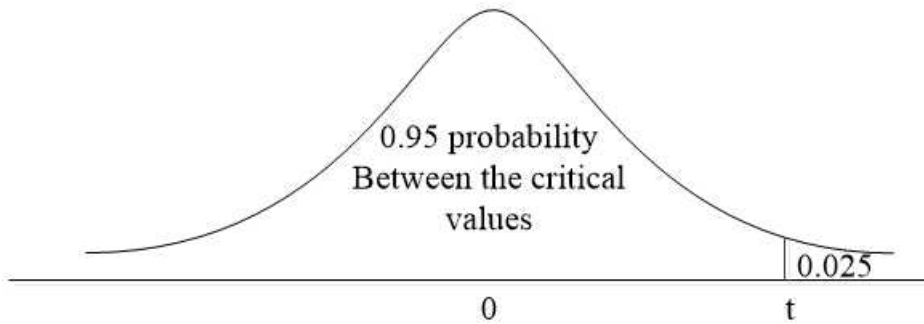
Revisit the same example and find 95% using t – distribution when  $n = 20$ ,  $\sigma_s = 30$   $df = 19$

$$\text{standard error} = z_t \cdot \frac{\sigma_s}{\sqrt{n}}$$

## Sampling distributions

Whereas the z-multiplier for normal distribution was fixed for a given confidence level (for example 1.96 for 95%), the t-multiplier,  $z_t$  depends on the degrees of freedom of the t-distribution.

Now instead of z-table, we will use the t – table shown below to find the critical value corresponding to confidence probability 95%



As show from the t – table,  $\alpha$  – value of 0.025 and  $df = 19$  intersect at 2.093 highlighted

df	0.01	0.025
1	-31.821	-12.706
2	-6.965	-4.303
3	-4.541	-3.182
4	-3.747	-2.776
5	-3.365	-2.571
6	-3.143	-2.447
7	-2.998	-2.365
8	-2.896	-2.306
9	-2.821	-2.262
10	-2.764	-2.228
11	-2.718	-2.201
12	-2.681	-2.179
13	-2.650	-2.160
14	-2.624	-2.145
15	-2.602	-2.131
16	-2.583	-2.120

## Sampling distributions

17	-2.567	-2.110
18	-2.552	-2.101
19	-2.539	-2.093
20	-2.528	-2.086

$$\text{standard error} = 2.093 \cdot \frac{30}{\sqrt{20}} = 14.04$$

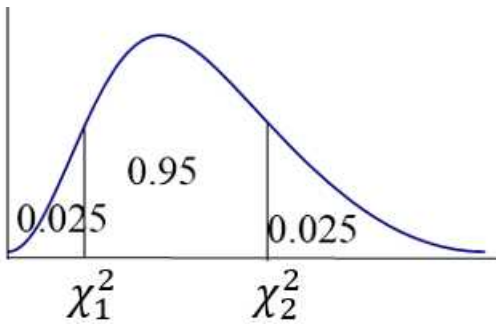
Thus at 95% confidence level, the true students mean GPA performance lies in the interval (61, 89)

To summarize when finding interval estimate for the true population mean use z-values in case population standard deviation is known and the sample size is large. Use t-values when the population standard deviation is unknown and the sample size is very small.

### Confidence interval for population standard deviation

Now let us focus on confidence interval for population standard deviation. A different distribution called Chi – squared test will be used.

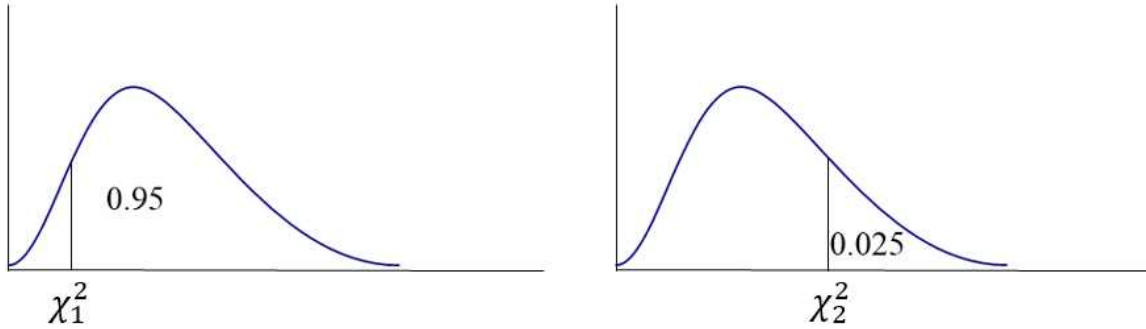
A Chi – squared test curve will look like follows. It is not symmetric like normal curve. The confidence interval limits using Chi-square is shown below



$$\text{population} \sim \chi^2$$

We can break down the above curve into two to easily use the Chi-square table in appendix D

## Sampling distributions



For Chi – squared test, the sample statistic is given by

$$\frac{(n - 1)\sigma_s^2}{\sigma^2} \text{ sample statistic}$$

Revisiting again our last example, construct confidence interval that will contain true population (students) standard deviation at confidence level (or  $\alpha$  – value) of 95%. Remember 95% will leave 2.5% probability both to the left tail and right tail.

Also we had  $n = 100$  and  $\sigma_s = 30$  and  $df = 99$  ( $n - 1 = 100 - 1 = 99$ )

$$\frac{(100 - 1)30^2}{\sigma^2}$$

Let us now, using Chi – square table in appendix D, find the critical values corresponding to right tail probability of 0.025. The left tail probability is 0.025 while right tail is  $95 + 0.025$

$$P(\chi_1^2 \leq \chi^2 \leq \chi_2^2) = 0.025$$

Using Chi – square table below will give the critical values

$$\chi_1^2 = 128.42$$

$$P(77.05 \leq \chi^2 \leq 128.42) \text{ with } 95\% \text{ probability}$$

	0.01	0.025	0.05	0.10	0.90	0.95	0.99
<b>90</b>	124.12	118.14	113.15	107.57	73.29	69.13	61.75
<b>91</b>	125.29	119.28	114.27	108.66	74.20	70.00	62.58
<b>92</b>	126.46	120.43	115.39	109.76	75.10	70.88	63.41
<b>93</b>	127.63	121.57	116.51	110.85	76.01	71.76	64.24
<b>94</b>	128.80	122.72	117.63	111.94	76.91	72.64	65.07
<b>95</b>	129.97	123.86	118.75	113.04	77.82	73.52	65.90
<b>96</b>	131.14	125.00	119.87	114.13	78.73	74.40	66.73

## Sampling distributions

<b>97</b>	132.31	126.14	120.99	115.22	79.63	75.28	67.56
<b>98</b>	133.48	127.28	122.11	116.32	80.54	76.16	68.40
<b>99</b>	134.64	<b>128.42</b>	123.23	117.41	81.45	<b>77.05</b>	69.23
<b>100</b>	135.80	129.56	124.34	118.50	82.36	77.92	70.06

Now substitute our test statistic between 77.05 and 128.42 at confidence level of 95%

$$77.05 \leq \frac{(100 - 1)30^2}{\sigma^2} \leq 128.42$$

Now isolate the unknown population standard deviation from the inequality by inverting everything

$$1156 \geq \sigma^2 \geq 694$$

$$694 \leq \sigma^2 \leq 1156$$

$$26.34 \leq \sigma \leq 34$$

This conclusion mean given a sample standard deviation of 30, we can believe that the true standard deviation of the whole population lies in the interval (26.34, 34) with confidence probability of 95%

### Confidence interval for population proportion

Suppose that a researcher wants to know whether an increase in road car accidents is as a result of drivers not using safety driver seat belt. A random sample of 150 drivers taken shows that only 60 drivers use seat belt while driving. Thus 90 drivers don't use seat belt.

In this case sample proportion of drivers who don't use seat belt is

$$\bar{p} = \frac{90}{150} = 0.6$$

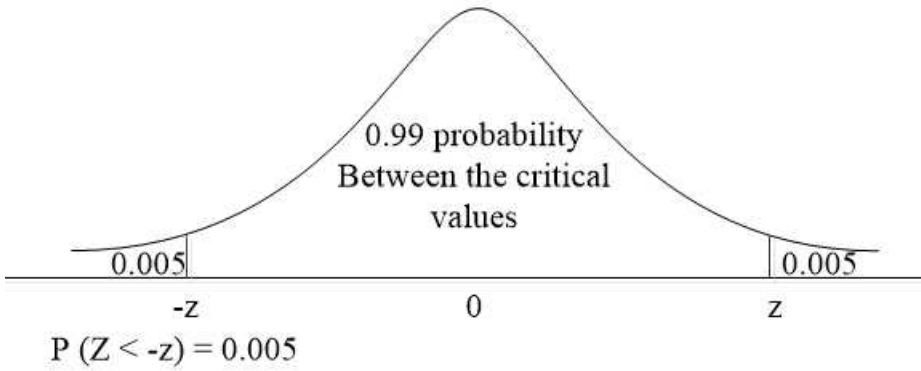
This is a single point estimate for the true population proportion. A single point estimator is not enough. Hence construct a confidence interval of 99% that will contain the true population proportion.

The test statistic for population proportion is given by the following confidence interval

$$\bar{p} - z \cdot \sqrt{\frac{\bar{p}(1 - \bar{p})}{n}} \leq P \leq \bar{p} + z \cdot \sqrt{\frac{\bar{p}(1 - \bar{p})}{n}}$$

Using central limit theorem we can assume the population is normally distributed as  $n > 30$

## Sampling distributions



From the normal distribution in appendix D, the closest z value is 2.58 as shown below

	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.3	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	<b>0.0049</b>	0.0048
-2.4	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183

Substituting values into the test statistic

$$0.6 - 2.58 \sqrt{\frac{0.6(1 - 0.6)}{90}} \leq P \leq 0.6 + 2.58 \sqrt{\frac{0.6(1 - 0.6)}{90}}$$

## Sampling distributions

$$0.4667 \leq P \leq 0.733$$

This result concludes that given probability of 95% we are confident that the proportion of drives who don't use safety seat belt lies between 47% and 73%

Population parameter	Distribution	Confidence interval
Mean ( $\mu$ )	<ul style="list-style-type: none"> <li>t-distribution when <math>\sigma</math> is unknown and sample size is very small</li> <li>z-distribution when <math>\sigma</math> is taken as the sample deviation and sample is very large</li> </ul>	<ul style="list-style-type: none"> <li><math>\bar{x} \pm t \cdot \frac{\sigma_s}{\sqrt{n}}</math> df = n - 1</li> <li><math>\bar{x} \pm z \cdot \frac{\sigma}{\sqrt{n}}</math></li> </ul> <p>Normally distributed</p>
Standard deviation ( $\sigma$ )	Use Chi-square distribution	$\chi^2 \leq \frac{(n-1)\sigma_s^2}{\sigma^2} \leq \chi^2$
Proportion (P)	Use z-distribution using central limit theorem	$\bar{p} \pm z \cdot \sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \sim z(0,1)$

### Review questions

- The government TV is reviewing a study about whether a large proportion of the population prefer watching their program compared with other private TV programs. A sample of 1000 people is taken and 650 claimed they prefer watching program from government TV channel than other privately owned TV channels. Construct 95% confidence interval for the true proportion of the whole population (all people in the country)
- The quality control agency wants to know whether a particular drug imported into the country meets certain standard. A sample of 100 drug of particular type is tested in the lab and the ibuprofen content is assumed to be normally distributed with standard deviation of 20mg. Construct 99% confidence interval for the true population (all drugs) standard deviation
- An independent research organization wants to study if physical exercise can reduce heart diseases. If they want to find true population mean number of people who do exercise per day at confidence level of 98%, what sample size would they need assuming mean of the population is normally distributed?
- Choose the correct answer that defines interval estimation of confidence interval
  - Interval estimation tests hypothesis of a claimed true population parameter
  - Interval estimation provides probability of normal distribution
  - Interval estimation provides true population parameter with a given confidence level from sample statistic
  - Interval estimation is comparing means of two samples



## Sampling distributions

5. From historical data, we have population standard deviation as 4 and if we collect sample size of 100, what is the standard error of the sampling distribution?

## Hypothesis test of single sample

### Chapter Seven

#### Hypothesis test of single sample

---

After completing this section, you should be able to

- Make a claim about population parameter
- Collect data about the population parameter and check if your claim is valid or not
- Understand null and alternative hypothesis
- Understand type I and type II errors and minimize them
- Demonstrate clear understanding on how to do hypothesis tests on population mean, standard deviation and proportion

## Hypothesis test of single sample

### Introduction

Hypothesis tests are everywhere. Everyday people make claim on certain aspects of research area. Most cases people base their claims on educational guess and researchers collect statistical data to prove their claims. Examples of claims are listed below.

- The mean number of females who underwent FGM (female genital mutilation) is 15 per day in this month alone
- The proportion of voters who support a particular candidate for an election is 70% and thus the candidate has higher public rating
- Exam scores of students has improved in recent year with standard deviation of 35 as a result of curriculum improvement

Each of those claims that we want to prove if it is true or false is called NULL hypothesis. The null hypothesis is the assumption we made about value of population parameter. It is the “status quo” that we don’t want to change unless there is sufficient evident available to reject it.

The null hypothesis is given the following notation  $H_0$  and will be used in this text. From the above claim examples, our null hypothesis can be written as follows

$$H_0 = 15 \text{ for population mean}$$

$$H_0 = 0.7 \text{ for population proportion}$$

$$H_0 = 35 \text{ for population standard deviation}$$

There are combinational cases the null hypothesis might experience base whether it is true or false and also whether it is rejected or accepted as summarized below

- $H_0$  accepted and true (normal case)
- $H_0$  accepted and false (type II error)
- $H_0$  rejected and true (type I error)
- $H_0$  rejected and false (normal case)

In best case, probability of type I error occurring must be minimized as we prefer not changing  $H_0$  unless sufficient evidence is available to reject it.

To perform hypothesis testing of rejecting the null hypothesis, a second hypothesis must be established against the null hypothesis. This second claim is called alternative hypothesis and it is an assumption that says the null hypothesis is false and has to be rejected.

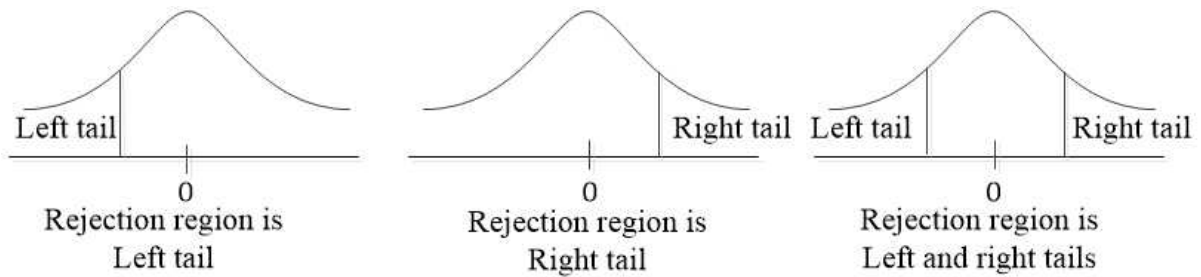
The alternative hypothesis takes the following notation and test cases, using the first claim of the above examples

$$H_1 < 15 \text{ left tailed test}$$

$$H_1 > 15 \text{ right tailed test}$$

$$H_1 \neq 15 \text{ two – tailed test}$$

## Hypothesis test of single sample



The probability of making type I error when  $H_0$  is true is called  $\alpha$  – value or significance level. An appropriate significance value is selected to minimize type I error. The significance level specifies the small area we want to reject null hypothesis if the statistic falls in.

Now  $H_0$  and  $H_1$  defined, they are always against each other when doing hypothesis testing and only one of them is true.

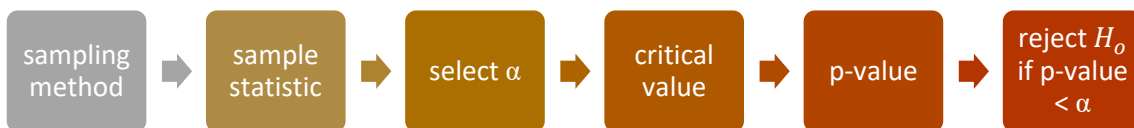
$$H_0 = 15 \text{ null hypothesis}$$

$$H_1 \neq 15 \text{ alternative hypothesis}$$

For two – tailed test.

Again using this example, if we make type I error it means we rejected the mean females who underwent FGM is 15 per day when that claim is actually true.

In this chapter we will also see how to test  $H_0$  using probability called p – value which if large, will support the evidence of the null hypothesis claim. A different method used than the p – value is called rejection region.



The general procedure for hypothesis is summarized by the following diagram considering normal distribution, but will also apply for other distributions.

## Hypothesis test of single sample

Calculate test statistic	P-value approach	Rejection region approach
$\frac{\bar{x} - \mu}{\sigma_s / \sqrt{n}} \sim Z(0, 1)$ <div style="text-align: center; margin-top: 20px;">↓</div>	<p style="text-align: center;">Given <math>\alpha</math> Obtain z test statistic</p> <p style="text-align: center;">P-value is <math>P(z &gt; \text{test statistic})</math></p> <p style="text-align: center;">If p-value <math>&gt; \alpha</math> Accept <math>H_0</math></p>	<div style="text-align: center;"> </div> <p style="text-align: center;"> <math>H_0: \mu = \text{value}</math>  <math>H_1: \mu \neq \text{value for two-tailed test}</math> </p> <p style="text-align: center;">Z-value should be outside in the rejection region</p> <p style="text-align: center;">Accept <math>H_0</math></p>

### Hypothesis test for population mean

As an example, suppose the government has published mean number of females who underwent FGM as 15 per day as a part of new law restricting FGM practice as well as for public awareness. To test their claim a random sample of female subjects were asked if they underwent FGM. Because of the sensitivity of the topic and cultural boundaries of the society making individuals reveal such topic very hard, we were able to get small sample of 30 females. The sample mean was found to be 20 and standard deviation 10. This is an example of hypothesis test of population mean.

Test the claim that the mean is different than 15 at the 5% level of significance

First step is to write down null and alternative hypothesis statements

$$H_0: \mu = 15$$

$$H_1: \mu \neq 15$$

We are given

$$\alpha = 0.05, n = 30, \bar{x} = 20, \sigma_s = 10$$

Since our sample is very small, we will use the student's t – distribution with degrees of freedom of  $n - 1 = 30 - 1 = 29$

Our test statistic for the t – distribution will then be

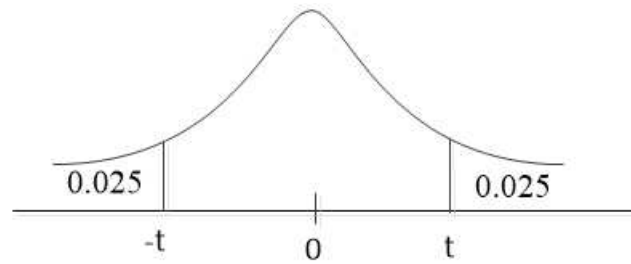
$$\frac{\bar{x} - \mu}{\sigma_s / \sqrt{n}} \sim t(29)$$

From this test statistic we can write down what we observed based on our sample

## Hypothesis test of single sample

$$\frac{20 - 15}{10/\sqrt{30}} = 2.7386$$

Because we want to test that the population mean is not 15 (different than 15) we have two-tailed test and we will reject both left and right tails with significance level of 0.05 (5%)



Rejection region is  
Left and right tails

No find from the t – table, the critical value corresponding to probability value of 0.025 with degrees of freedom of 29 as shown below

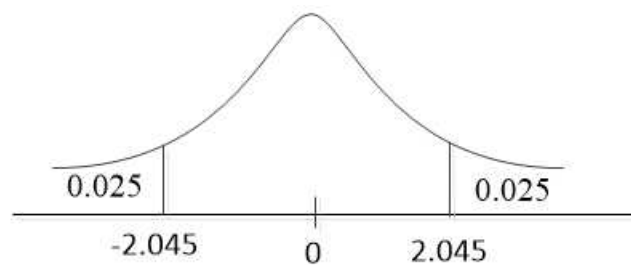
	P - values		
df	0.01	0.025	0.05
1	-31.821	-12.706	-6.314
2	-6.965	-4.303	-2.920
3	-4.541	-3.182	-2.353
4	-3.747	-2.776	-2.132
5	-3.365	-2.571	-2.015
6	-3.143	-2.447	-1.943
7	-2.998	-2.365	-1.895
8	-2.896	-2.306	-1.860
9	-2.821	-2.262	-1.833
10	-2.764	-2.228	-1.812
11	-2.718	-2.201	-1.796
12	-2.681	-2.179	-1.782

## Hypothesis test of single sample

13	-2.650	-2.160	-1.771
14	-2.624	-2.145	-1.761
15	-2.602	-2.131	-1.753
16	-2.583	-2.120	-1.746
17	-2.567	-2.110	-1.740
18	-2.552	-2.101	-1.734
19	-2.539	-2.093	-1.729
20	-2.528	-2.086	-1.725
21	-2.518	-2.080	-1.721
22	-2.508	-2.074	-1.717
23	-2.500	-2.069	-1.714
24	-2.492	-2.064	-1.711
25	-2.485	-2.060	-1.708
26	-2.479	-2.056	-1.706
27	-2.473	-2.052	-1.703
28	-2.467	-2.048	-1.701
29	-2.462	<b>-2.045</b>	-1.699
30	-2.46	-2.04	-1.70

The value of 2.045

We now have our rejection region as shown below



Now ask the question, does our observed value of 2.7386 fall in either left rejected region values or right rejected region?

## Hypothesis test of single sample

Look at the curve above. On the left tail we have all values  $\leq -2.045$  and on the right tail we have all values  $\geq 2.045$ . Both of these value ranges will be rejected as they constitute the 5% doubt of making type I error. And if our observed value (2.7386) falls in those rejected values, it means we have some evidence  $H_0$  could be rejected in favor of  $H_1$

The answer to the above question is yes. Our observed value from the sample falls in the right tail rejection region. Thus we reject the null hypothesis. We conclude that the mean females who underwent FGM is not exactly 15 as claimed by the government rather it is different than 15.

As a second example of left tailed test is as follows

Suppose the quality control agency claim a certain drug imported into the country contain Ibuprofen with mean quantity of 20mg per 100g. To test validity of their claim, a random sample of 50 drugs is tested in the laboratory with mean Ibuprofen content of 19mg and standard deviation 4.5mg

Test the null hypothesis that the mean Ibuprofen content is 20mg at 1% significance level

State the null hypothesis and alternative hypothesis first

$$H_0: \mu = 20$$

$$H_1: \mu < 20$$

List the observed sample values and calculate the sample observed value

$$\alpha = 0.01, n = 50, \bar{x} = 19, \sigma_s = 4.5$$

Using central limit theorem we assume the drug population is normally distributed

$$\frac{\bar{x} - \mu}{\sigma_s / \sqrt{n}} \sim Z(0, 1)$$

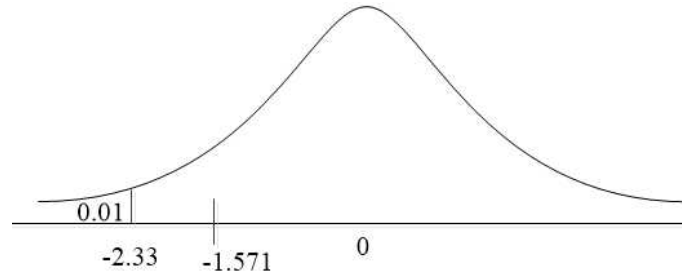
$$\frac{19 - 20}{4.5 / \sqrt{50}} = -1.5713$$

Now find the z-score corresponding to the left tail area of 0.01 to form the rejection region

You will get  $z = -2.33$ . Please try to locate this value using z – table given in the appendix. The rejection region is now illustrated below



## Hypothesis test of single sample



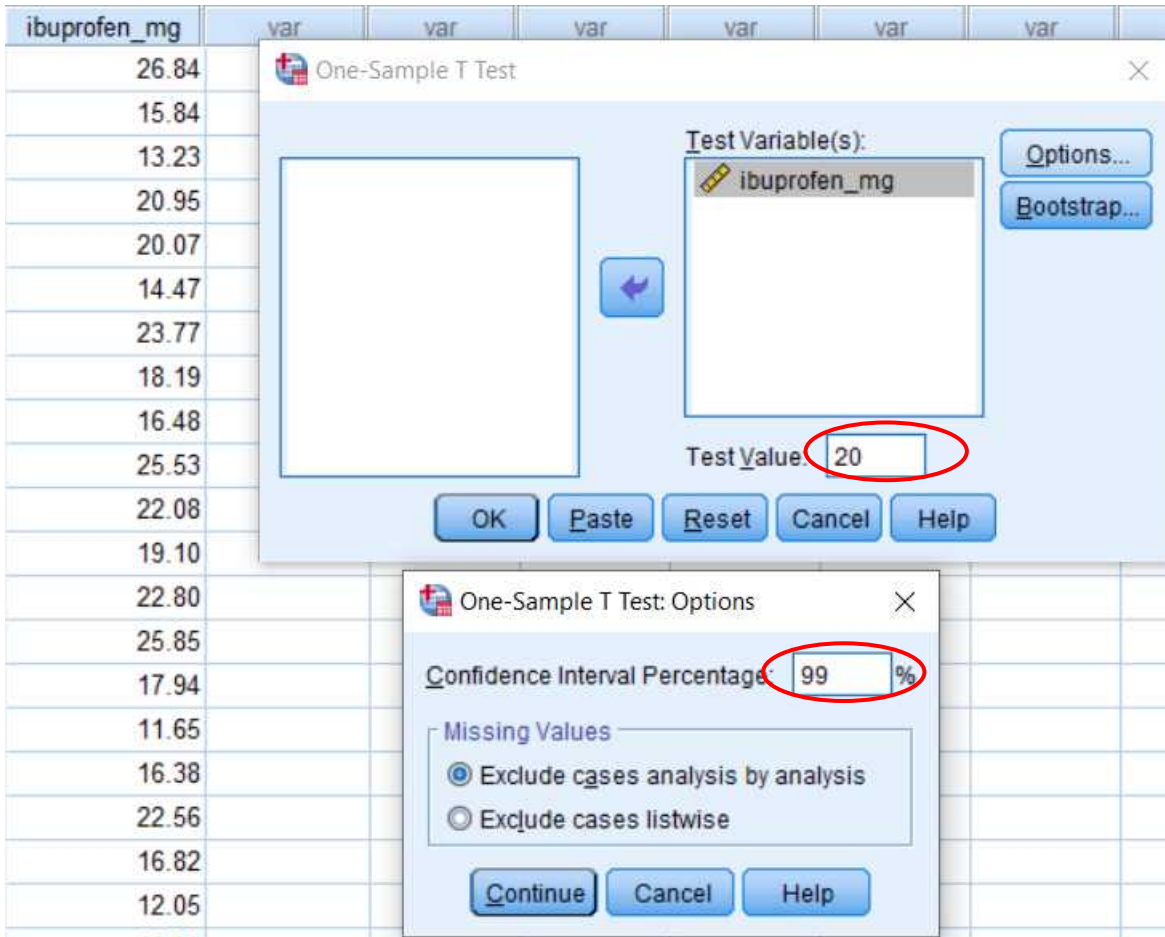
It can be seen that our observed sample value doesn't fall in the rejection region. Hence we accept the claim of the quality agency and conclude that there is sufficient evidence the mean Ibuprofen content is 20mg per 100g

To test this one sample t-test in SPSS, the sample data is as follows

26.84	22.08	16.88	20.07	20.02
15.84	19.1	19.98	19.55	19.18
13.23	22.8	13.35	22.92	25.84
20.95	25.85	16.68	11.97	18.38
20.07	17.94	20.92	18.27	16.89
14.47	11.65	18.01	20.29	12.26
23.77	16.38	20.22	20.83	20.51
18.19	22.56	21.4	25.29	5.22
16.48	16.82	7.23	13.8	20.84
25.53	12.05	17.37	21.52	21.19

After data entry, go to **analyze > compare means > one sample t-test**

## Hypothesis test of single sample



After click ok, the following output result window will be generated for interpretation

**One-Sample Statistics**

	N	Mean	Std. Deviation	Std. Error Mean
ibuprofen_mg	50	18.5887	4.58924	.64902

**One-Sample Test**

	Test Value = 20					
	t	df	Sig. (2-tailed)	Mean Difference	99% Confidence Interval of the Difference	
					Lower	Upper
ibuprofen_mg	-2.175	49	.035	-1.41135	-3.1507	.3280

## Hypothesis test of single sample

The results table shows that the p-value (0.035) is greater than test significance level ( $\alpha = 0.01$ ). This provides evidence to support the null hypothesis that the mean Ibuprofen content of the general population is 20mg

### Hypothesis test for population standard deviation

So far we have seen how to validate hypothesis about population mean using rejection region method. Now we will test a claim made about population standard deviation

Suppose that a bottled – water Production Company claims the sodium content is normally distributed with standard deviation of 7.2mg per liter of water. But we believe that their claim is not correct and sodium content is less. Thus we collected random sample of 20 water bottles and calculated standard deviation as 6.5mg. Test the null hypothesis that the standard deviation of the sodium content is 7.2mg at 5% significance level.

Our task is now to prove their claim and either accept it or reject it when sufficient evidence is available against their claim (the null hypothesis)

State the null hypothesis and alternative hypothesis for this test

$$H_0: \sigma = 7.2$$

$$H_1: \sigma < 7.2$$

We are given  $n = 20$   $\sigma = 7.2$   $\sigma_s = 6.5$   $df = 19$

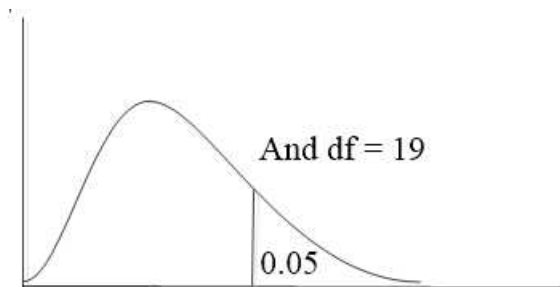
For standard deviation we will use chi – square test to evaluate the test statistic

$$\frac{(n - 1)\sigma_s^2}{\sigma^2} \sim \chi^2 \quad \text{with degrees of freedom } n - 1$$

$$\frac{(20 - 1)6.5^2}{7.2^2} = 15.485$$

Is this left tail or right tail test? Of course it is left tail because we want prove the standard deviation is less than the one claimed by the factory

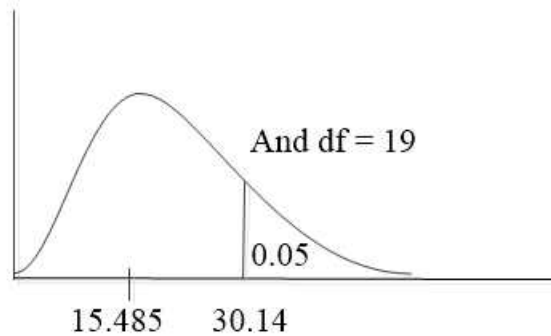
Now using Chi-square table, find critical value corresponding to  $df = 19$  and  $\alpha = 0.05$  (5%)



$$\chi^2 = ?$$

If you try using  $\chi^2$  – table in the appendix, you will find value of 30.14

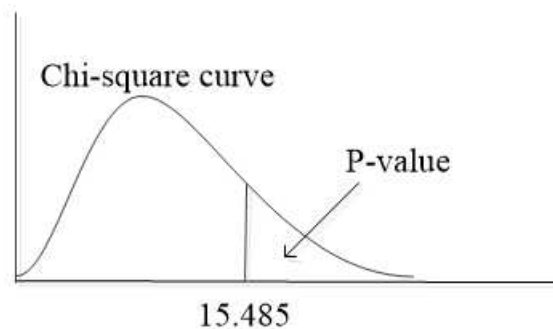
## Hypothesis test of single sample



The area to the right of the value 30.14 is the rejection region. Our observed value from the sample lies outside of the rejection region. Hence we accept the null hypothesis and conclude that the mean sodium content of the water is 7.2mg as claimed by the factory.

This further means our statistic value observed from the sample lies in the rejection region with a probability of 5%. If it happens what we observed lies in the rejection region, we can infer that null hypothesis to be rejected in favor of the alternative hypothesis.

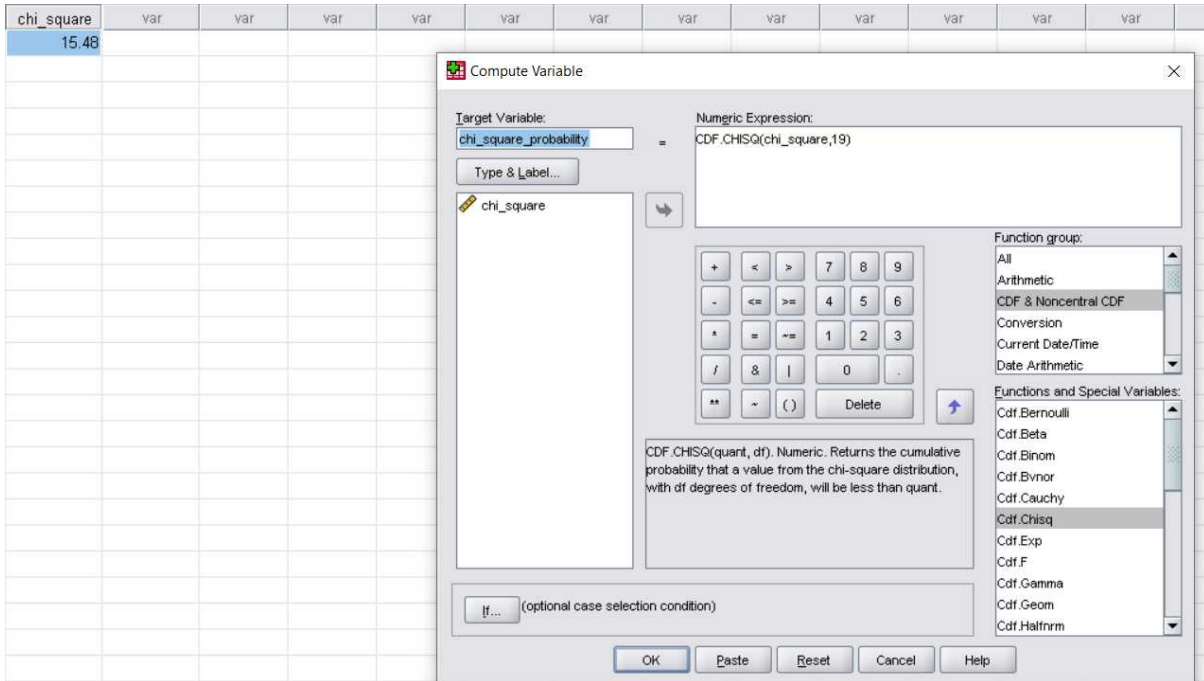
A second approach to the rejection region is what we call p – value. The p – value is the probability of getting value bigger than (right tail) our observed value. If the p – value is large than our significance level, then it will support the null hypothesis against the alternative hypothesis.



Using SPSS **cdf.chisq** in **transform > compute variable**, we will get p – value of 0.31 which is greater than  $\alpha$  ( $0.31 > 0.05$ ).

We therefore accept the null hypothesis and conclude that the mean sodium content of the water is 7.2mg as claimed.

## Hypothesis test of single sample



### Review questions

1. A weather agency in the country claims a prediction that average rainfall rate in next year will be 900mm. You made observation on daily rainfall rates throughout the year and your conclusion shows slightly higher average rainfall rate of 1000mm  
Test at 95% confidence interval the average rainfall rate will be 900mm
2. A certain telecommunication company claim 90% of the population use their network. A random sample of 20 people were interview and only 9 use that company's network
  - a. State the null hypothesis and alternative hypothesis
  - b. What is the test statistic you could use to infer population proportion?

## Chapter Eight

### **Goodness of fit test for normality**

---

After completing this section, you should be able to

- Understand the concept of goodness of fit
- Know that many parametric tests require data represented by the dependent sample to be normally distributed
- Appreciate how Chi-square statistic can be used to test sample data normality
- Appreciate Kolmogorov – Smirnov normality tests
- Run goodness of fit tests in SPSS for practice

## Goodness of fit test for normality

### Chi – square goodness of fit test for normality

In goodness of fit test, you are after the following question:

If I take random sample from certain population, is there a way I could test if the probability distribution of the sample is normally distributed?

Was the sample taken from normal, binomial or Poisson?

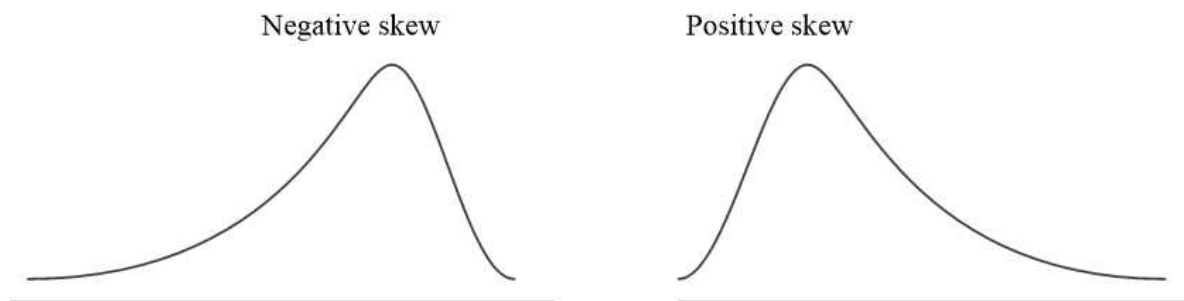
Such questions are answered by goodness of fit test. We will restrict this chapter to only goodness of fit for normality. Many parametric studies require normality of the independent data before executing the study analysis.

To explain goodness of fit in simple way, consider a categorical data of salary of 35 employees in certain production company. For simplicity, the salary is defined in continuous intervals

Salary range (in 1000)				
15	46	27	70	28
20	55	50	20	56
25	78	84	15	22
45	90	70	21	
60	67	10	28	
16	33	30	71	
49	29	35	67	
54	18	40	62	

As a first step in goodness of fit it is advised to run descriptive statistics and analyze Kurtosis and Skewness of the data distribution

Skewness indicates data is not symmetrical and tilted to the right or left as shown below



## Goodness of fit test for normality

Normal distribution has skewness close to zero (between -1 and +1)

On the other hand the normal distribution should have lower kurtosis as opposed to t – distribution which has heavy tails and thus higher kurtosis. Normal distribution has kurtosis of three

Now run **analyze > descriptive** with distribution boxes checked to obtain the following result



Skewness		Kurtosis	
Statistic	Std. Error	Statistic	Std. Error
.377	.398	-1.036	.778

From the result table above the data is not skewed as skewness lies between -1 and +1

Rule of thumb to use it

$$3 \times \text{Std. Error} > |\text{statistic}|$$

This is our case. Hence we conclude the data can be assumed to be normally distributed

With preliminary normality established under skewness and kurtosis, let us look at other methods at our disposal

Given this categorical data, the goodness of fit test will ask to provide evidence for the following null hypothesis

$$H_0: \text{salary of employee population follows normal distribution}$$

*“A certain department presents an expected values of data, we then collect sample and compute observed values. We then compare observed values with the expected values assuming the population follows normal distribution”*

See the frequency column of the table below. These are frequencies observed for each category. For example, 10 employees earn salary range 40 – 50 and 6 employees earn salary range of 30 – 40

Now what are the expected frequencies based on our sample? We have sample size n, our rule of thumb is at least 5 expectations in each interval, we then have  $n / 5$  range of values. Then 35 people with at least 5 expected people in each interval will yield interval sizes of  $35 / 5 = 7$ . Thus we will have expected frequency of at least 5 in each interval

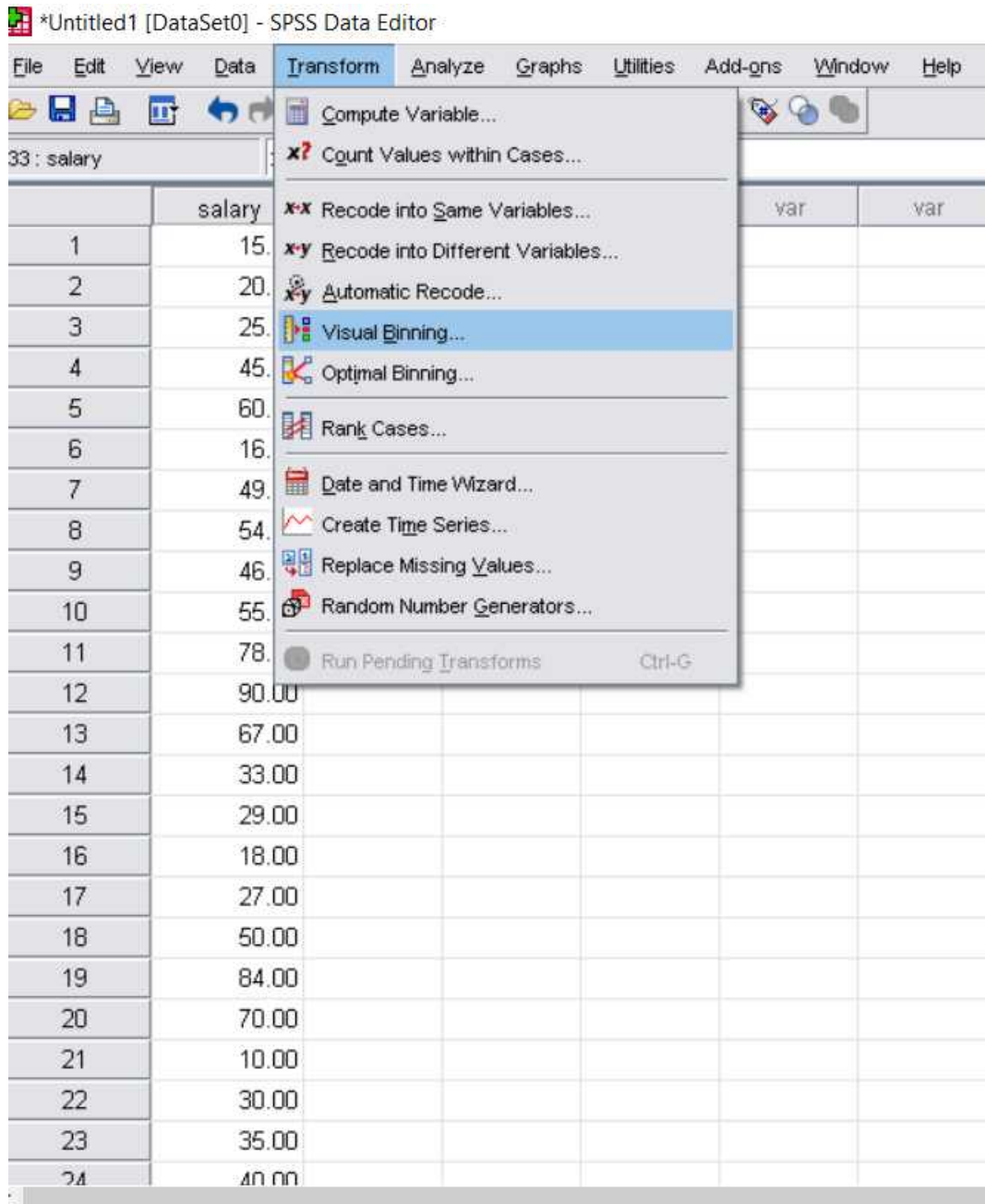


### Goodness of fit test for normality

Salary range (in 1000)	Observed frequency	Expected frequency
10 – 17	4	5
17 – 24	5	5
24 – 31	6	5
31 – 38	2	5
38 – 45	2	5
45 – 52	3	5
52 – 59	3	5
59 – 66	2	5
66 – 73	5	5
73 – 80	1	5
80 – 87	1	5
87 – 94	1	5

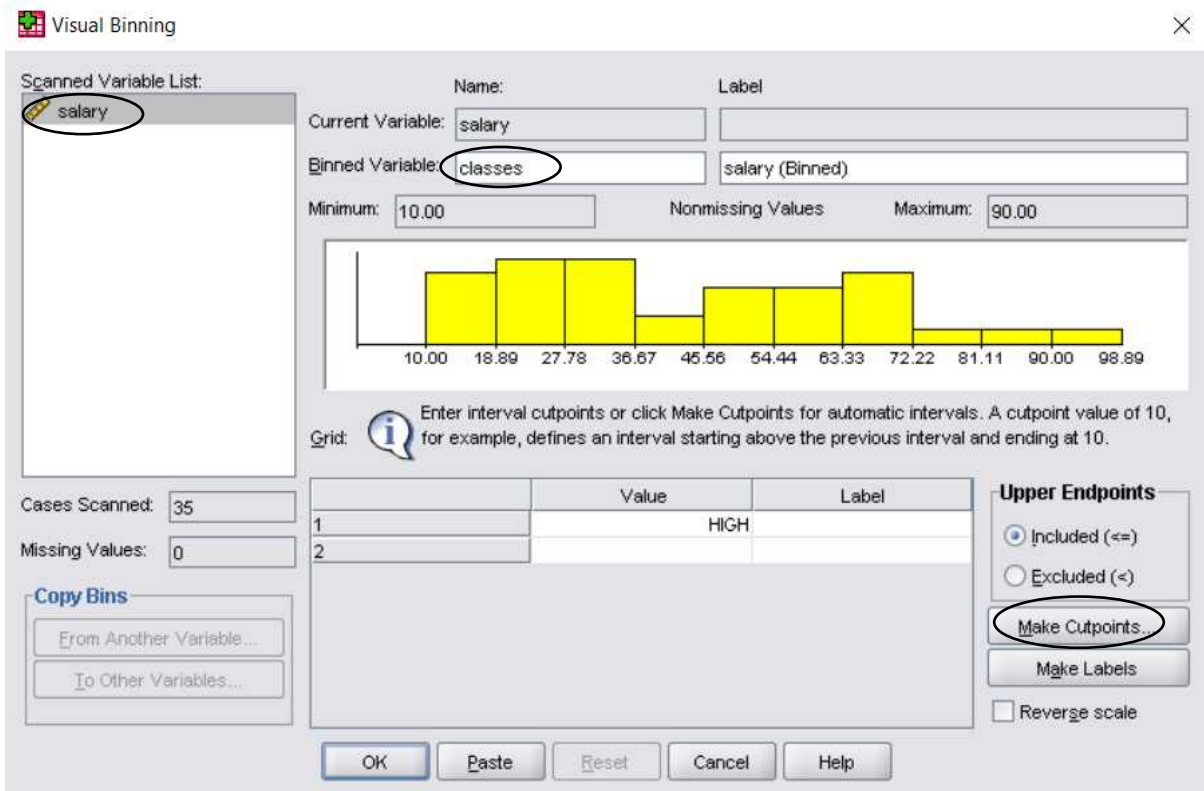
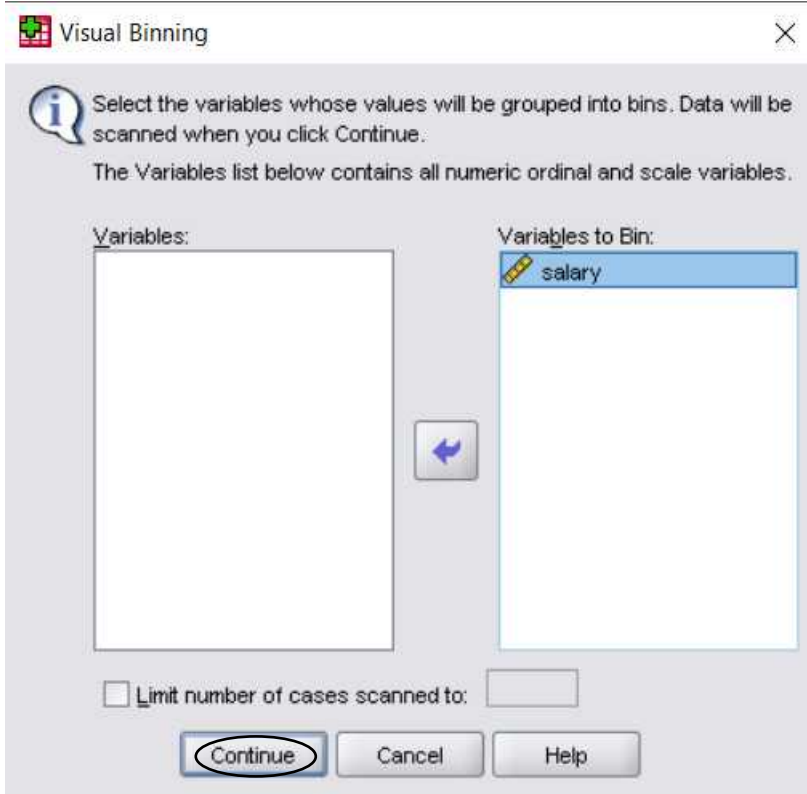
To verify and create this grouped table in SPSS, first create variable name salary in the **variable view**, and then enter that salary data into **data view**. Choose **transform** menu and then select **visual binning** as shown below

## Goodness of fit test for normality

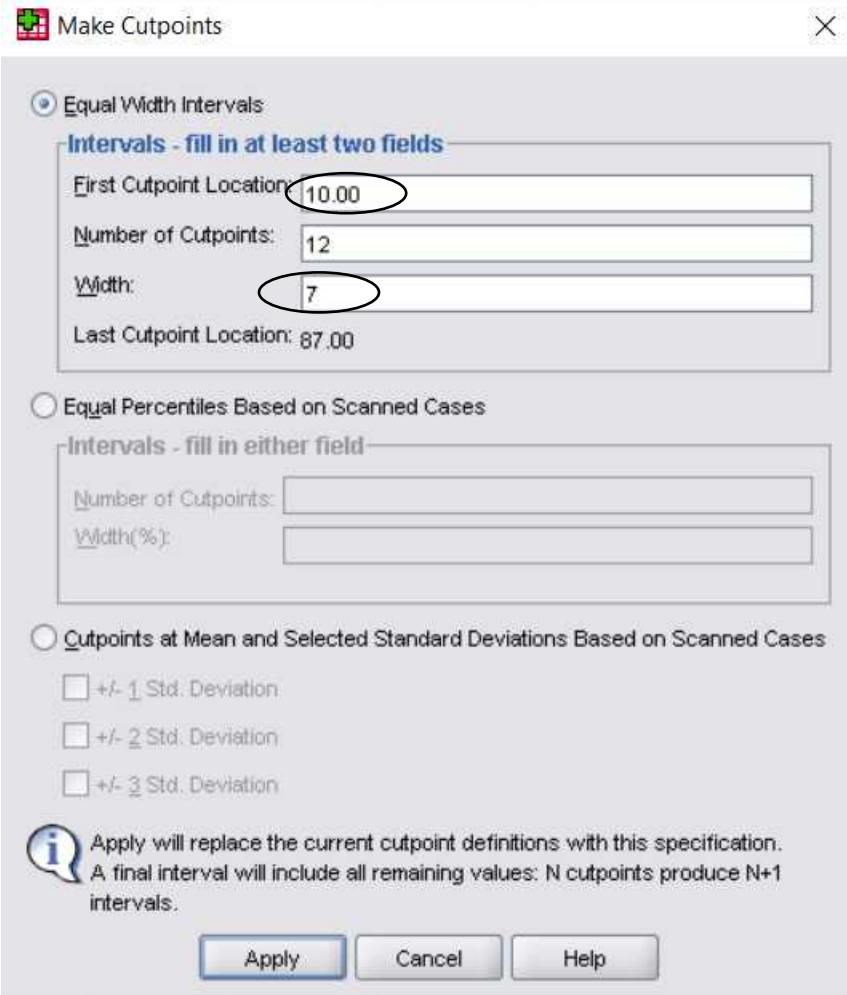


The visual binning dialog box will pop up. Forward salary in variable window to **variables to bin window** and then click continue

## Goodness of fit test for normality



## Goodness of fit test for normality

The image shows a software dialog box titled "Make Cutpoints" with a close button (X) in the top right corner. It contains three radio button options for defining intervals. The first option, "Equal Width Intervals", is selected. Below it, a sub-section titled "Intervals - fill in at least two fields" contains four input fields: "First Cutpoint Location" with the value "10.00", "Number of Cutpoints" with the value "12", "Width" with the value "7", and "Last Cutpoint Location" with the value "87.00". The second option, "Equal Percentiles Based on Scanned Cases", is unselected and has two empty input fields for "Number of Cutpoints" and "Width(%)". The third option, "Cutpoints at Mean and Selected Standard Deviations Based on Scanned Cases", is unselected and has three checkboxes for "+/- 1 Std. Deviation", "+/- 2 Std. Deviation", and "+/- 3 Std. Deviation". At the bottom, there is an information icon (i) followed by a paragraph of text: "Apply will replace the current cutpoint definitions with this specification. A final interval will include all remaining values: N cutpoints produce N+1 intervals." Below this text are three buttons: "Apply", "Cancel", and "Help".

**Make Cutpoints**

**Equal Width Intervals**

**Intervals - fill in at least two fields**

First Cutpoint Location: 10.00

Number of Cutpoints: 12

Width: 7

Last Cutpoint Location: 87.00

**Equal Percentiles Based on Scanned Cases**

**Intervals - fill in either field**

Number of Cutpoints:


Width(%):

**Cutpoints at Mean and Selected Standard Deviations Based on Scanned Cases**

+/- 1 Std. Deviation

+/- 2 Std. Deviation

+/- 3 Std. Deviation

 Apply will replace the current cutpoint definitions with this specification.  
A final interval will include all remaining values: N cutpoints produce N+1 intervals.

Click apply to go back to the visual binning dialog box and click **make labels** bottom

## Goodness of fit test for normality

Visual Binning

Scanned Variable List: salary

Name: Current Variable: salary Binned Variable: classes Label: salary (Binned)

Minimum: 10.00 Nonmissing Values Maximum: 90.00

Enter interval cutpoints or click Make Cutpoints for automatic intervals. A cutpoint value of 10, for example, defines an interval starting above the previous interval and ending at 10.

Grid:

	Value	Label
1	10.00	
2	17.00	
3	24.00	
4	31.00	
5	38.00	
6	45.00	
7	52.00	
8	59.00	

Cases Scanned: 35  
Missing Values: 0

Copy Bins:  
From Another Variable...  
To Other Variables...

Upper Endpoints:  
 Included ( $\leq$ )  
 Excluded ( $<$ )  
Make Cutpoints...  
Make Labels  
 Reverse scale

OK Paste Reset Cancel Help

Visual Binning

Scanned Variable List: salary

Name: Current Variable: salary Binned Variable: classes Label: salary (Binned)

Minimum: 10.00 Nonmissing Values Maximum: 90.00

Enter interval cutpoints or click Make Cutpoints for automatic intervals. A cutpoint value of 10, for example, defines an interval starting above the previous interval and ending at 10.

Grid:

	Value	Label
1	10.00	$\leq 10.00$
2	17.00	11.00 - 17.00
3	24.00	18.00 - 24.00
4	31.00	25.00 - 31.00
5	38.00	32.00 - 38.00
6	45.00	39.00 - 45.00
7	52.00	46.00 - 52.00
8	59.00	53.00 - 59.00

Cases Scanned: 35  
Missing Values: 0

Copy Bins:  
From Another Variable...  
To Other Variables...

Upper Endpoints:  
 Included ( $\leq$ )  
 Excluded ( $<$ )  
Make Cutpoints...  
Make Labels  
 Reverse scale

OK Paste Reset Cancel Help

Click ok

### Goodness of fit test for normality

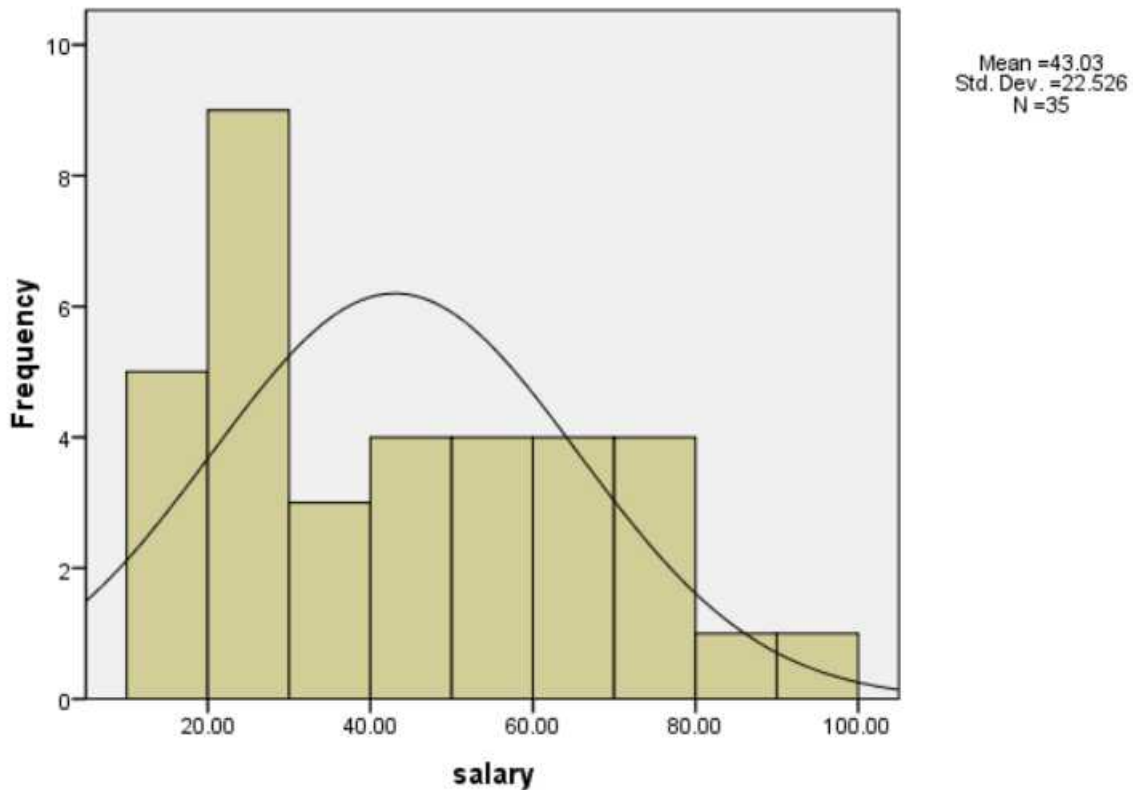
salary (Binned)

N	Valid	35
	Missing	0
Mean		6.20
Std. Deviation		3.279

**salary (Binned)**

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	<= 10.00	1	2.9	2.9	2.9
	11.00 - 17.00	3	8.6	8.6	11.4
	18.00 - 24.00	5	14.3	14.3	25.7
	25.00 - 31.00	6	17.1	17.1	42.9
	32.00 - 38.00	2	5.7	5.7	48.6
	39.00 - 45.00	2	5.7	5.7	54.3
	46.00 - 52.00	3	8.6	8.6	62.9
	53.00 - 59.00	3	8.6	8.6	71.4
	60.00 - 66.00	2	5.7	5.7	77.1
	67.00 - 73.00	5	14.3	14.3	91.4
	74.00 - 80.00	1	2.9	2.9	94.3
	81.00 - 87.00	1	2.9	2.9	97.1
	88.00+	1	2.9	2.9	100.0
	Total	35	100.0	100.0	

## Goodness of fit test for normality



The mean shown as 6.2 from the SPSS output is normalized mean when data is binned into classes. We multiply it by the class width of 7 which will then give mean of approximately 43. Similarly standard deviation is 22.5 (3.279 x 7)

Now we have data divided into classes to make it easier for Chi – square categorical analysis and we stated null hypothesis statement. Next is to find the Chi – square test statistic which is defined by the following formula

$$\chi^2 = \frac{(f_o - f_E)^2}{f_E} \quad \text{summed all over the classes}$$

Let the recall our grouped frequency table and calculate  $\chi^2$

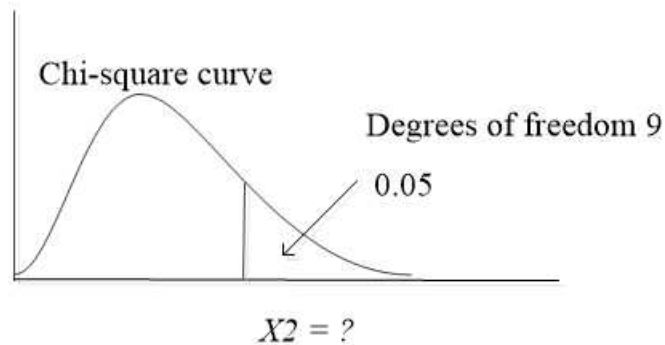
Salary range (in 1000)	Observed frequency ( $f_o$ )	Expected frequency ( $f_E$ )	$f_o - f_E$	$(f_o - f_E)^2$	$\frac{(f_o - f_E)^2}{f_E}$
10 – 17	4	5	-1	1	0.2
17 – 24	5	5	0	0	0
24 – 31	6	5	1	1	0.2
31 – 38	2	5	-3	9	1.8

### Goodness of fit test for normality

38 – 45	2	5	-3	9	1.8
45 – 52	3	5	-2	4	0.8
52 – 59	3	5	-2	4	0.8
59 – 66	2	5	-3	9	1.8
66 – 73	5	5	0	0	0
73 – 80	1	5	-4	16	3.2
80 – 87	1	5	-4	16	3.2
87 – 94	1	5	-4	16	3.2
					$\chi^2 = 17$

Now with null hypothesis stated and test statistic computed, test the salary distribution of the employees follow normal distribution at 5% significance level.

We need to find the rejection region with  $\chi^2 = 17$  and  $\alpha = 0.05$   $df = \text{number classes} - 3 = 12 - 3 = 9$  as shown in the diagram below

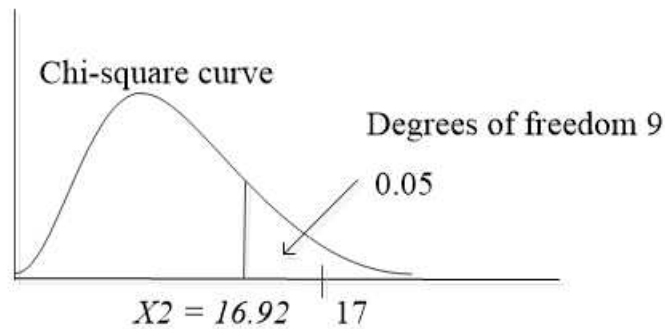


	P – values						
Degrees of freedom	0.01	0.025	0.05	0.10	0.90	0.95	0.99
1	6.63	5.02	3.84	2.71	0.02	0.00	0.00
2	9.21	7.38	5.99	4.61	0.21	0.10	0.02
3	11.34	9.35	7.81	6.25	0.58	0.35	0.11
4	13.28	11.14	9.49	7.78	1.06	0.71	0.30



### Goodness of fit test for normality

5	15.09	12.83	11.07	9.24	1.61	1.15	0.55
6	16.81	14.45	12.59	10.64	2.20	1.64	0.87
7	18.48	16.01	14.07	12.02	2.83	2.17	1.24
8	20.09	17.53	15.51	13.36	3.49	2.73	1.65
9	21.67	19.02	<b>16.92</b>	14.68	4.17	3.33	2.09
10	23.21	20.48	18.31	15.99	4.87	3.94	2.56
11	24.72	21.92	19.68	17.28	5.58	4.57	3.05
12	26.22	23.34	21.03	18.55	6.30	5.23	3.57



As can be seen our observed test statistic 17 is inside critical region and approximately equal to the critical value at 5% significance level. Hence we conclude that there is little evidence that the data assumes normal probability distribution as claimed in the null hypothesis statement. In other words, the data has not likely come from normal distribution population,

Let us take another example, but be very quick this time to cut the long explanation

Suppose you downloaded population of 40 cities from census board website. They published their result as below (per 1000)

70	90	105	66	35
120	50	110	75	40
20	42	67	70	25
79	65	80	60	30
84	89	55	85	36

## Goodness of fit test for normality

47	51	109	99	95
81	77	68	28	44
33	82	24	39	23

What do we expect from this census data?

$40 / 5 = 8$  gives class width as 8 and should be the same for each class interval

An SPSS output is shown below with the classes made

\* Visual Binning.

\*population\_census\_1000.

```
RECODE population_census_1000 (MISSING=COPY) (LO THRU 20=1)
      (LO THRU 28=2) (LO THRU 36=3) (LO THRU 44=4) (LO THRU 52=5)
      (LO THRU 60
```

```
      =6) (LO THRU 68=7) (LO THRU 76=8) (LO THRU 84=9) (LO THRU
      92=10) (LO THRU 100=11) (LO THRU 108=12) (LO THRU 116=13) (
      LO
```

```
      THRU HI=14) (ELSE=SYSMIS) INTO classes.
```

```
VARIABLE LABELS classes 'population_census_1000 (Binned)'.  
  
FORMAT classes (F5.0).
```

```
VALUE LABELS classes 1 '<= 20.00' 2 '21.00 - 28.00' 3 '29.00 - 36.00' 4 '37.00 - 44.00' 5 '45.00 - 52.00' 6 '53.00 - 60.00' 7 '61.0
```

```
0 - 68.00' 8 '69.00 - 76.00' 9 '77.00 - 84.00' 10 '85.00 - 92.00' 11 '93.00 - 100.00' 12 '101.00 - 108.00' 13 '109.00 - '+
```

```
'116.00' 14 '117.00+'.
```

```
MISSING VALUES classes ( ).
```

```
VARIABLE LEVEL classes (ORDINAL).
```

```
EXECUTE.
```

### Goodness of fit test for normality

Population (x 1000)	Observed frequency ( $f_o$ )	Expected frequency ( $f_E$ )	$f_o - f_E$	$(f_o - f_E)^2$	$\frac{(f_o - f_E)^2}{f_E}$
20.00 - 27.00	5	5	0	0	0
28.00 - 35.00	4	5	-1	1	0.2
36.00 - 43.00	4	5	-1	1	0.2
44.00 - 51.00	3	5	-2	4	0.8
52.00 - 59.00	2	5	-3	9	1.8
60.00 - 67.00	4	5	-1	1	0.2
68.00 - 75.00	3	5	-2	4	0.8
76.00 - 83.00	6	5	1	1	0.2
84.00 - 91.00	3	5	-2	4	0.8
92.00 - 99.00	2	5	-3	9	1.8
100.00 - 107.00	1	5	-4	16	3.2
108.00 - 116.00	3	5	-2	4	0.8
Total	40				$\chi^2 = 10.8$

Let us choose significance level of 10% for this test

Our test statistic is therefore 10.8 from the table computation above.

So find the critical value in the Chi – square table the value corresponding to  $\alpha = 0.1$  and  $df = 11$  and you will find 17.28

Since the test statistic < critical value (test statistic outside of critical region), we can accept the null hypothesis and conclude that the population census pertains to normal distribution statistics.

Following this examples let us recap our understanding of goodness of fit test. Given a sample of data collected from certain population, we can claim normality of the data under null hypothesis, and then calculate the expected value given the null hypothesis (the population was indeed following normal distribution) is true. To do this hypothesis test we used Chi-squared test.

## Goodness of fit test for normality

However when testing hypothesis for population normality we seen that it is long procedure for continuous distribution. There is easier approach which is called Kolmogorov – Smirnov (KS) test which can be run is SPSS

### Kolmogorov – Smirnov (KS) goodness of fit test

It is well suited for continuous distributions such the normal distribution. It is useful non-parametric test in the sense that it is free from “assume the population follows a certain distribution” presumption. It provides an efficient way of whether the data of your observation sample was from a normal population (it tests normality of your empirical data). The Chi – square test is cumbersome in that it needs data to be binned first (grouped into classes)

Given the N data is ordered (either ascending or descending), the KS test statistic is given by the maximum of the two values in the closed brackets

$$D = \max_{1 \leq i \leq N} \left\{ F(Y_i) - \frac{i-1}{N}, \frac{i}{N} - F(Y_i) \right\}$$

Because we are comparing our sample data to normal distribution, we covert all our scores to normal probabilities as shown in the column labeled with  $F(Y_i)_n$  in the table below  
In KS test we state the null hypothesis as follows

$$H_0: \text{the sample data follows normal distribution}$$

If the test statistic calculated from the observation is less than the critical value calculated at the selected significance level, accept the null hypothesis

$$D < D_c$$

If, however, the test statistic calculated from the observation is greater than the critical value calculated at the selected significance level, reject the null hypothesis

$$D > D_c$$

Let us revisit our last example of population census data of 40 cities (x1000) arranged in ascending order. Test the null hypothesis at significance level of 10%

$i$	$F(Y_i)$	$F(Y_i)_n$	$\frac{i}{N}$	$\frac{i-1}{N}$	$F(Y_i)_n - \frac{i-1}{N}$	$\frac{i}{N} - F(Y_i)_n$
1	20	0.055675159	0.025	0	0.0557	-0.03068
2	23	0.069056878	0.05	0.025	0.0441	-0.01906
3	24	0.074029967	0.075	0.05	0.0240	0.00097
4	25	0.079272128	0.1	0.075	0.0043	0.020728
5	28	0.096683426	0.125	0.1	-0.0033	0.028317

**Goodness of fit test for normality**

6	30	0.10975803	0.15	0.125	-0.0152	0.040242
7	33	0.131673585	0.175	0.15	-0.0183	0.043326
8	35	0.147861921	0.2	0.175	-0.0271	0.052138
9	36	0.156436074	0.225	0.2	-0.0436	0.068564
10	39	0.184083383	0.25	0.225	-0.0409	0.065917
11	40	0.193937111	0.275	0.25	-0.0561	0.081063
12	42	0.214585376	0.3	0.275	-0.0604	0.085415
13	44	0.236457557	0.325	0.3	-0.0635	0.088542
14	47	0.271446873	0.35	0.325	-0.0536	0.078553
15	50	0.308839745	0.375	0.35	-0.0412	0.06616
16	51	0.321786807	0.4	0.375	-0.0532	0.078213
17	55	0.375631945	0.425	0.4	-0.0244	0.049368
18	60	0.446383137	0.45	0.425	0.0214	0.003617
19	65	0.518888368	0.475	0.45	0.0689	-0.04389
20	66	0.533391294	0.5	0.475	0.0584	-0.03339
21	67	0.547850013	0.525	0.5	0.0479	-0.02285
22	68	0.562245542	0.55	0.525	0.0372	-0.01225
23	70	0.590772415	0.575	0.55	0.0408	-0.01577
24	70	0.590772415	0.6	0.575	0.0158	0.009228
25	75	0.659720669	0.625	0.6	0.0597	-0.03472
26	77	0.686008705	0.65	0.625	0.0610	-0.03601
27	79	0.711385123	0.675	0.65	0.0614	-0.03639
28	80	0.723700282	0.7	0.675	0.0487	-0.0237
29	81	0.73575187	0.725	0.7	0.0358	-0.01075
30	82	0.747529886	0.75	0.725	0.0225	0.00247
31	84	0.770229928	0.775	0.75	0.0202	0.00477
32	85	0.781136722	0.8	0.775	0.0061	0.018863

### Goodness of fit test for normality

33	89	0.821676456	0.825	0.8	0.0217	0.003324
34	90	0.831020687	0.85	0.825	0.0060	0.018979
35	95	0.872934963	0.875	0.85	0.0229	0.002065
36	99	0.900797646	0.9	0.875	0.0258	-0.0008
37	105	0.933800951	0.925	0.9	0.0338	-0.0088
38	109	0.950574226	0.95	0.925	0.0256	-0.00057
39	110	0.954186853	0.975	0.95	0.0042	0.020813
40	120	0.979877128	1	0.975	0.0049	0.020123
					Max = 0.0689	Max = 0.0885

What is the maximum value of the last two columns that represent the statistic?

$$\max_{1 \leq i \leq N} \left\{ F(Y_i) - \frac{i-1}{N}, \frac{i}{N} - F(Y_i) \right\}$$

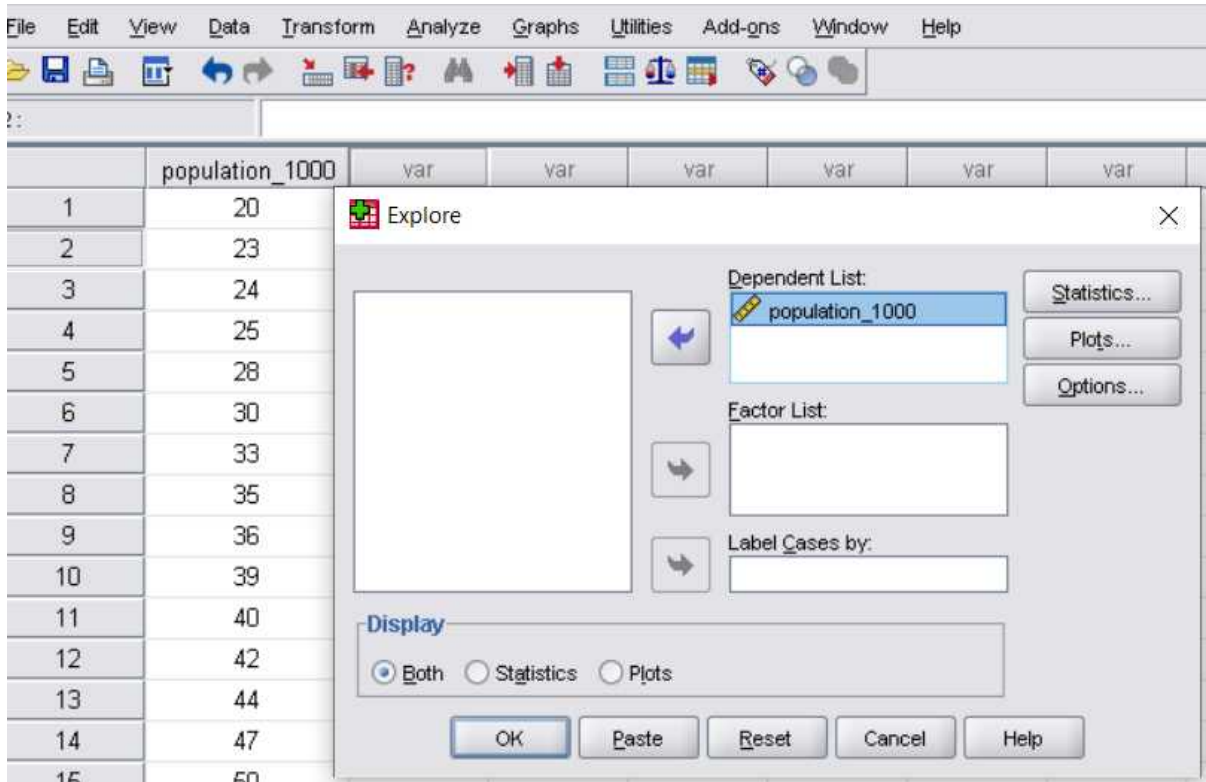
It is  $D = 0.089$

If you use one sample KS table you will get  $D_c = 0.189$  at  $\alpha = 0.1$ . You could search that table on the net and thus not provided in the appendix D

As can be seen  $D < D_c$  ( $0.089 < 0.189$ ) and therefore we accept the null hypothesis and conclude the sample exhibits characteristics of normal distribution.

To perform normality test in SPSS, first do data entry into the data window. Then from **analyze** menu select **descriptive > explore**

## Goodness of fit test for normality



The SPSS output of normality test will look like below

**Tests of Normality**

	Kolmogorov-Smirnov <sup>a</sup>			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
population_census_1000	.089	40	.200	.964	40	.225

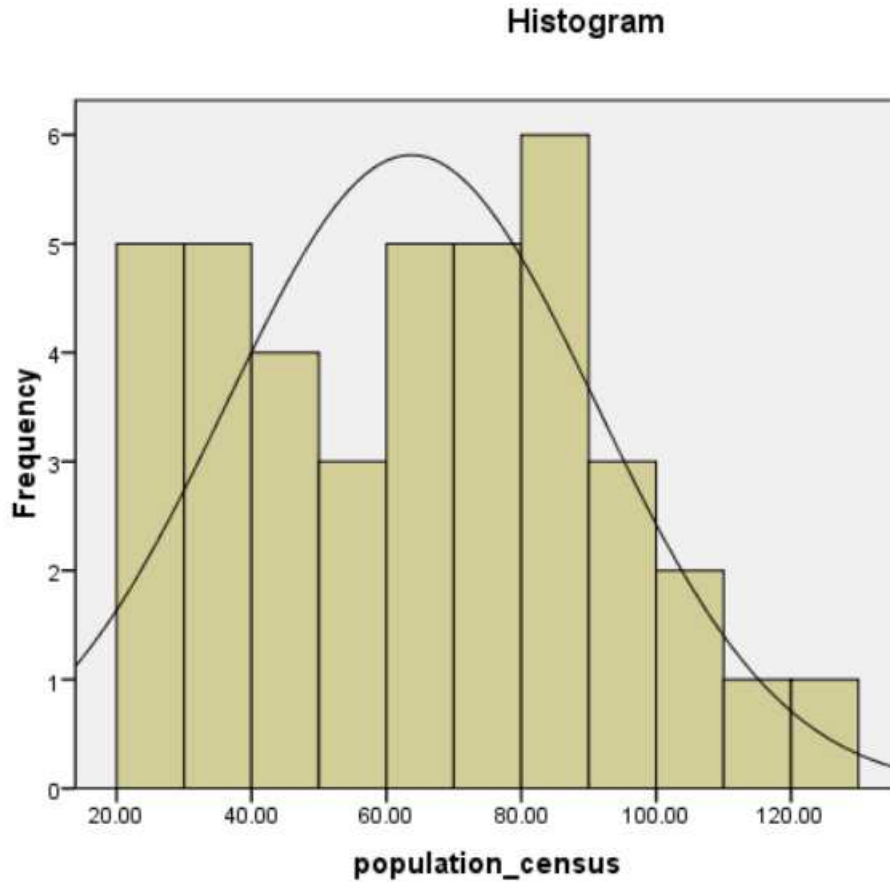
a. Lilliefors Significance Correction

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

As can be seen from the result table above critical value  $D_c$  (0.200) is greater than the statistic (0.089) or significance level which matches with the hand calculation. The conclusion is again to accept the null hypothesis that the sample data can be assumed to be normally distributed

The normality test will also plot histogram for visualization in the SPSS output window as below

## Goodness of fit test for normality



More or less the shape of the histogram looks a little bit like normal distribution

From the same output window one can also inspect the Q-Q plot below which if the data values align with the  $45^{\circ}$  line indicates normality. This is our case as shown below



## Goodness of fit test for normality



### Review questions

1. Write down all the steps you will take to perform normality test for student test score of your university. Try to simulate the mindset of a researcher
2. Suppose that a drug manufacturing company published Ibuprofen content per 100mg of drug of 51 sample

12	19	20	15	20	12
16	15	15	11	16	
17	14	14	13	15	
16	10	15	18	15	
11	10	16	10	10	
18	13	18	15	12	
13	10	20	15	12	
18	18	10	14	14	
17	11	20	13	11	
11	15	18	14	13	

Test the null hypothesis that the data can fit normal distribution at 5% significance level. (Remember to prefer using the SPSS or any other computer program as the simulation gives accurate results than the hand calculation, and in real life research you will be using simulation)

## Independent and dependent samples hypothesis test

### Chapter 9

## Independent and dependent samples hypothesis test

---

After completing this section, you should be able to

- Familiarize dependent samples as paired samples that draw conclusion “after-before” cases
- Conduct t – test hypothesis on paired (dependent) samples to test whether the difference between their means is statistically significant
- Conduct Chi – square hypothesis test on independent samples to test association between categorical data
- Practice examples in SPSS

## Independent and dependent samples hypothesis test

Consider the following statistical research situations

1. Suppose that temperature of new machine was tested before and after operation. 12 such machines were sampled for test.

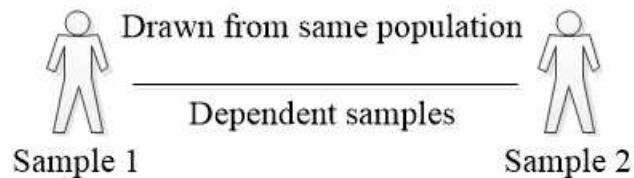
Before operation	20	15	10	12	19	21	16	11	9	20	23	13
After operation	20	6	19	14	14	9	8	11	7	19	8	9

Test at 5% level of significance, the claim that there is no significant difference between before operation and after operation mean temperatures

What are the characteristics of this type of test?

- We have two samples drawn from the same population (same machines)
- We are interested if “after sample” is dependent on or paired with the previous “before sample” or alternatively the mean difference before and after is zero

When the score of one sample is paired with another score of the same sample, we call the samples dependent samples. Another example of dependent samples are those in measurement experiments, where same variables are measured before and after.

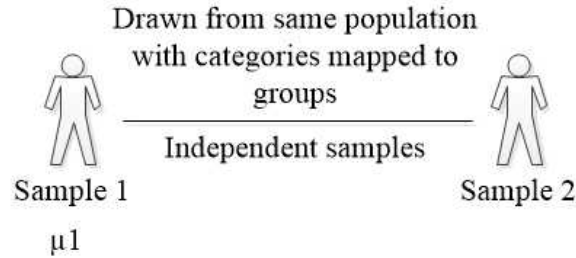


2. A TV broadcasting organization wants to study if different age groups of the population have different preferences of their programs. They sampled one group from adult category and another different group is young people under 25 years. They also sampled two TV programs A and B. In this case the two samples are drawn from two independent groups (age category group vs TV program group) of the same population

In independent samples test, we are interested if there is an association between different TV programs and age categories of the population

$H_0$ : *program ratings are independent of age categories (there is NO association)*

## Independent and dependent samples hypothesis test



Dependent (paired) samples	Independent nominal samples	Independent (unpaired) samples
$H_o$ : mean difference is zero	$H_o$ : the samples are not associated with each other	$H_o$ : mean difference is zero
Tested using student's t – test	Tested using Chi-square test	Tested using student's t – test
Samples drawn from the same sample subjects of a single population	Samples drawn from different groups or categories of single population	Samples drawn from two populations that are independent

### Dependent (paired) samples test

Let us take the first situation above as an example and add other columns

What we want is mean and standard deviation of the “difference” sample

Let the “difference” mean be  $\mu_d$

Let the “difference” standard deviation be  $\sigma_d$

Before	After	Difference (d) = after - before	$d - \bar{\mu}_d$	$(d - \bar{\mu}_d)^2$	$\sigma_d = \sqrt{\frac{\sum(d - \bar{\mu}_d)^2}{n - 1}}$
20	20	0	3.75	14.0625	1.278409
15	6	-9	-9	81	7.363636
10	19	9	9	81	7.363636
12	14	2	2	4	0.363636
19	14	-5	-5	25	2.272727

## Independent and dependent samples hypothesis test

21	9	-12	-12	144	13.09091
16	8	-8	-8	64	5.818182
11	11	0	0	0	0
9	7	-2	-2	4	0.363636
20	19	-1	-1	1	0.090909
23	8	-15	-15	225	20.45455
13	9	-4	-4	16	1.454545
		$\mu_d = \sum \frac{d}{n}$ = -3.75		Total = 659.0625	$\sigma_d = 7.74$

$$\mu_d = -3.75 \quad \sigma_d = 7.74$$

Before we proceed to the analysis, let us practice again normality test (from last chapter) to test whether we can assume the “difference” sample data is normally distributed since sample size is less than 30 and central limit theorem cannot be applied

Let us use histogram and KS normality tests to test goodness of fit for normal distribution at 95% significance level in SPSS

$H_0$ : the "difference" sample fits normal distribution

The SPSS KS normality test will look like this

Tests of Normality						
	Kolmogorov-Smirnov <sup>a</sup>			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
temperature_diff	.182	12	.200	.919	12	.275

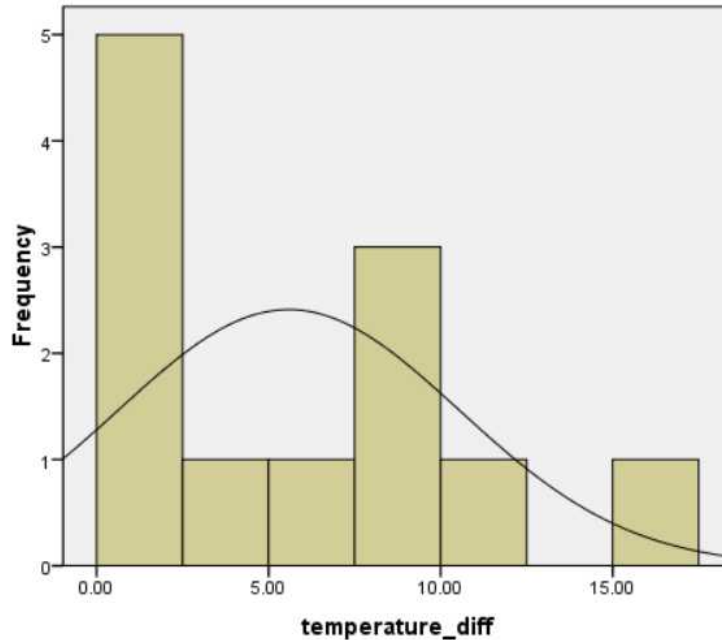
a. Lilliefors Significance Correction

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

The p-value (0.200) is greater than the significance (0.05) which indicates data normality

You could also view the histogram diagram which also indicates more or less normality of the sample

## Independent and dependent samples hypothesis test



Now let us go back to our problem of the dependency test (if samples are paired). At significance level of 5%, can we infer that mean temperature is not different before and after machine operation?

For this example, since our sample is relatively very small (only 12) let us use student's t – distribution with degrees of freedom of  $n - 1 = 12 - 1 = 11$

Given

$$\mu_d = -3.75 \quad \sigma_d = 7.74 \quad \alpha = 0.05 \quad df = 12 - 1 = 11$$

Use t-table to find the critical value corresponding to probability of 0.05

State hypothesis statements

$$H_0: \mu = 0 \text{ (no difference between paired data means)}$$

$$H_1: \mu \neq 0 \text{ (there is difference between paired data means)}$$

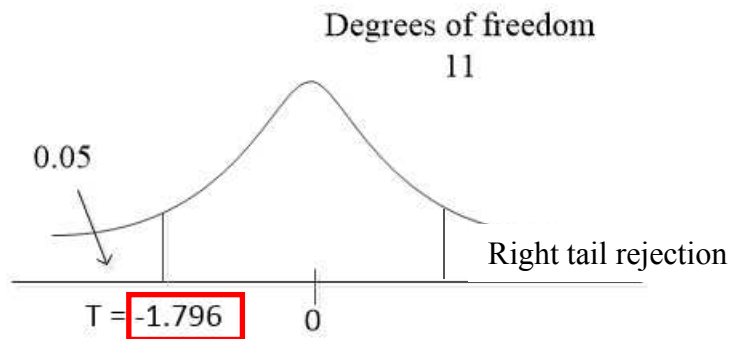
Remember in hypothesis test, you assume the null hypothesis is true, and we therefore claim that  $\mu$  is equal to 0 (i.e. the mean before taking the sample is the same as the mean after taking the sample.)

Using rejection region we can get critical value from the t – distribution table

	P - values		
df	0.01	0.025	0.05
1	-31.821	-12.706	-6.314

## Independent and dependent samples hypothesis test

2	-6.965	-4.303	-2.920
3	-4.541	-3.182	-2.353
4	-3.747	-2.776	-2.132
5	-3.365	-2.571	-2.015
6	-3.143	-2.447	-1.943
7	-2.998	-2.365	-1.895
8	-2.896	-2.306	-1.860
9	-2.821	-2.262	-1.833
10	-2.764	-2.228	-1.812
11	-2.718	-2.201	<b>-1.796</b>
12	-2.681	-2.179	-1.782
13	-2.650	-2.160	-1.771
14	-2.624	-2.145	-1.761
15	-2.602	-2.131	-1.753



Calculate test statistic t – value

$$t(11) = \frac{\mu_d - \mu}{\sigma_d / \sqrt{n}} = \frac{-3.75 - 0}{7.74 / \sqrt{10}} = -1.532$$

Is this calculated value outside of the rejection region? The answer is YES.

This gives strong evidence to the null hypothesis. It can be inferred that there is evidence to support machine temperature is not significantly different before and after operation.

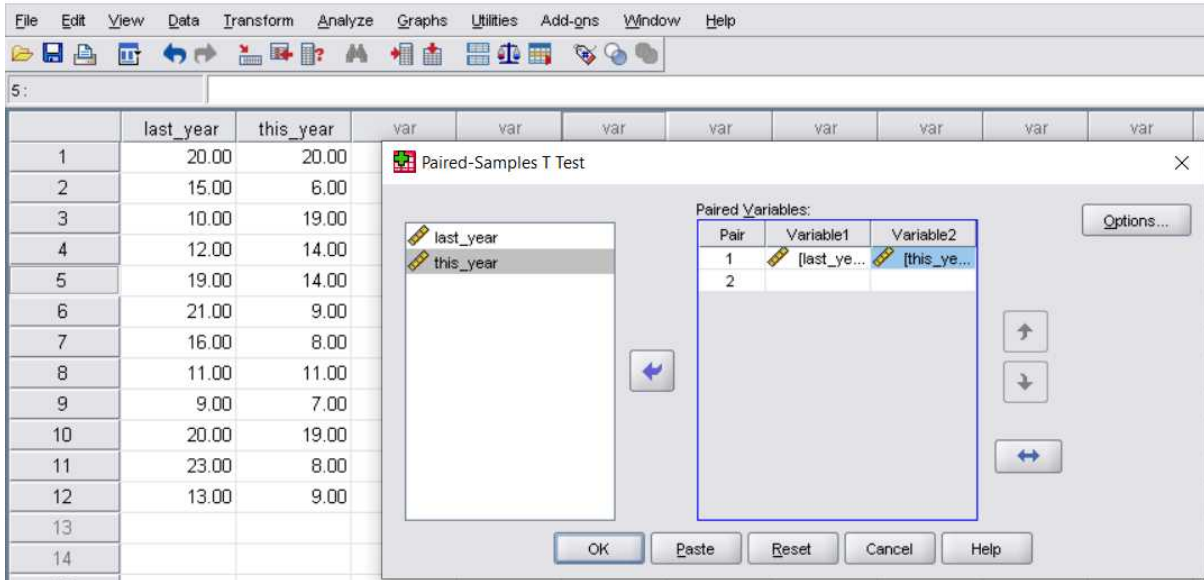
## Independent and dependent samples hypothesis test

To do dependent (paired) samples t – test in SPSS follow these steps

Enter before and after samples into data view

**Go to analyze > compare means > paired-samples t-test**

Put the first item into **variable 1** and the second data into **variable 2** box



After completing the analysis, the SPSS will generate this result

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	before_FGM_apply - after_FGM_apply	3.7500E0	6.579928	1.899462	- .430687	7.930687	1.974	11	.074

Look at the last three values of the table

$$t - \text{value} = 1.974 \quad df = 11 \quad \text{give } p - \text{value} = 0.74$$

Since the p – value is greater than ( $\alpha = 0.05$ ) we accept the null hypothesis and conclude that there is no average temperature difference between before and after machine operation.

### Independent samples t – test to compare means

In some cases we may want to study if mean difference between two independent populations is statistically significant. We draw two independent samples from the two



## Independent and dependent samples hypothesis test

groups of single population and using descriptive statistics to compute the test statistic using t – distribution

Give the following data

Sample 1 mean	$\bar{x}_1$	Population 1 mean	$\mu_1$	Population 1 $\sigma_1^2$ unknown
Sample 2 mean	$\bar{x}_2$	Population 2 mean	$\mu_2$	Population 2 $\sigma_2^2$ unknown
Sample 1 standard deviation	$\sigma_{s1}$			$\sigma_1^2 \neq \sigma_2^2$
Sample 2 standard deviation	$\sigma_{s2}$			
Sample 1 size	$N_1$			
Sample 2 size	$N_2$			

Let mean difference be  $\mu_1 - \mu_2 = \mu_d$

$H_0: \mu_1 - \mu_2 = 0$  there is no mean difference between populations

$H_1: \mu_1 - \mu_2 \neq 0$  there is mean difference between populations

The test statistic using t – distribution is given by

$$\frac{(\bar{x}_1 - \bar{x}_2) - \mu_d}{\sqrt{\frac{\sigma_{s1}^2}{N_1} + \frac{\sigma_{s2}^2}{N_2}}} \sim t(df)$$

Consider a study we want to know if there is difference between mean exam score of males and mean exam score of females. We take sample from each population of size 10 and tabulate it as below

Male	86	72	89	85	83	75	81	78	78	82
Female	64	69	52	80	56	66	57	50	53	63

Test if the mean difference between the two exam scores is statistically significant at 5%

The dependent variable (exam score) is in continuous scale and assumes normal distribution. In this study male students are population 1 while female students are population 2

Sample 1 mean	$\bar{x}_1 = 80.9$	Population 1 mean	$\mu_1$	Population 1 $\sigma_1^2$ unknown
Sample 2 mean	$\bar{x}_2 = 61$	Population 2 mean	$\mu_2$	Population 2 $\sigma_2^2$ unknown

## Independent and dependent samples hypothesis test

Sample 1 standard deviation	$\sigma_{s1} = 5.216$			$\sigma_1^2 \neq \sigma_2^2$
Sample 2 standard deviation	$\sigma_{s2} = 9.25$			

You could run the data in SPSS descriptive statistics and obtain the following result

	N	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
male_scores	10	80.9000	5.21643	-.211	.687	-.533	1.334
female_scores	10	61.0000	9.24962	.837	.687	.471	1.334
Valid N (listwise)	10						

We can now compute the t – test statistic using the above formula

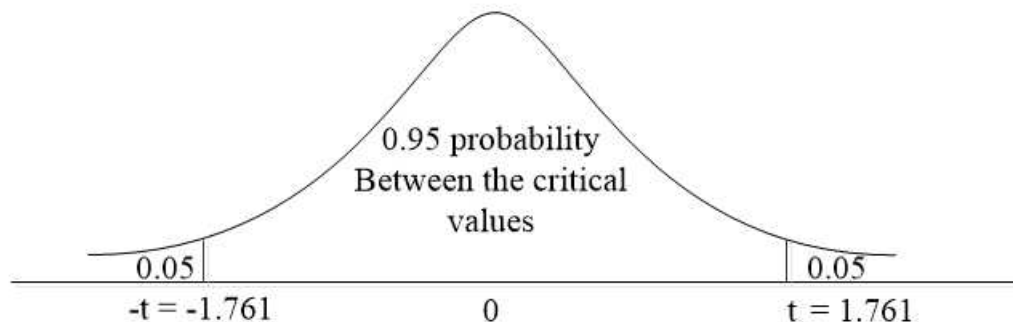
$$t = \frac{(80.9 - 61) - 0}{\sqrt{\frac{(5.216)^2}{10} + \frac{(9.25)^2}{10}}} = 5.926$$

The degrees of freedom of the two independent samples t – distribution is given by

$$df = \frac{\left[ \frac{\sigma_{s1}^2}{N_1} + \frac{\sigma_{s2}^2}{N_2} \right]^2}{\frac{\left( \frac{\sigma_{s1}^2}{N_1} \right)^2}{N_1 - 1} + \frac{\left( \frac{\sigma_{s2}^2}{N_2} \right)^2}{N_2 - 1}} = \frac{\left( \frac{(5.216)^2}{10} + \frac{(9.25)^2}{10} \right)^2}{\frac{\left( \frac{(5.216)^2}{10} \right)^2}{10 - 1} + \frac{\left( \frac{(9.25)^2}{10} \right)^2}{10 - 1}} = 14$$

To find the critical value corresponding to  $\alpha = 0.05$  and degrees of freedom of 14, using the t – distribution in appendix D

You should value critical value = -1.761



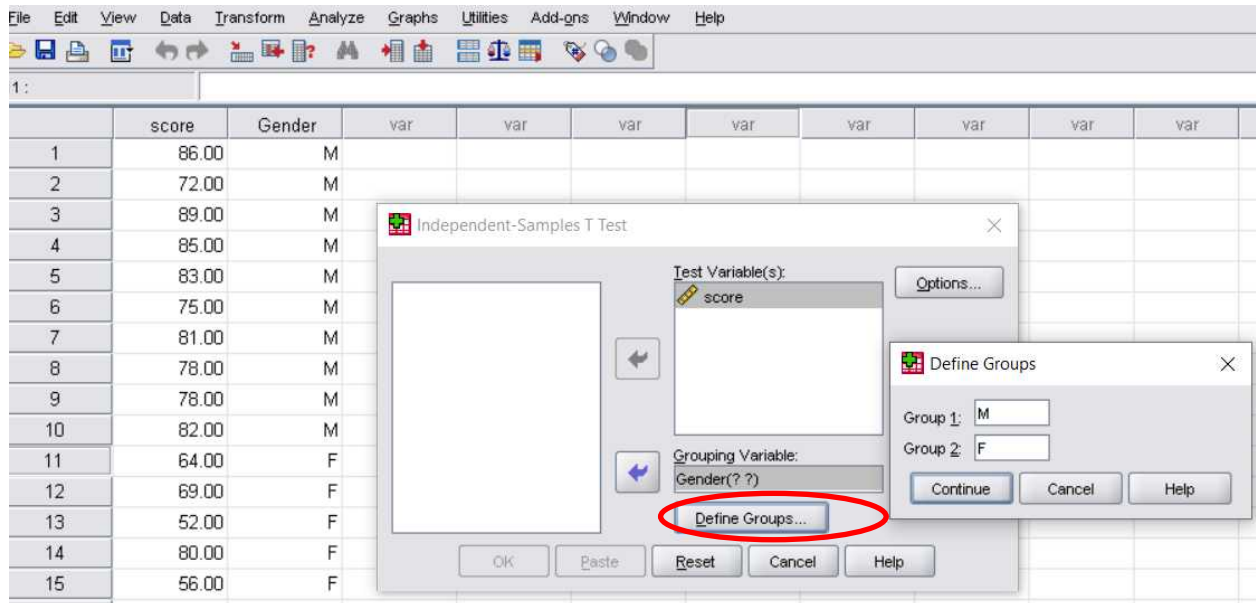
## Independent and dependent samples hypothesis test

Is our test statistic (5.926) inside the critical regions in the tails? The answer is YES

Hence we reject the null hypothesis, and conclude that there is evidence to support mean difference between the groups is not the same

Now let us analyze the same problem in SPSS. Remember first step is always variable creation and data entry

Run **analyze > compare means > independent samples t – test**



When you set everything as shown and click ok, you will see the following output result window

	Gender	N	Mean	Std. Deviation	Std. Error Mean
score	M	10	80.9000	5.21643	1.64958
	F	10	61.0000	9.24962	2.92499

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
score	Equal variances assumed	3.254	.088	5.926	18	.000	19.90000	3.35807	12.84495	26.95505
	Equal variances not assumed			5.926	14.199	.000	19.90000	3.35807	12.70711	27.09289

Looking at the last row of the table with equal populations variances not assumed, the p – value is much less than our significance level  $\alpha = 0.05$ . Hence we will reject the null

## Independent and dependent samples hypothesis test

hypothesis and again conclude that there is mean variation between the two independent populations.

### Chi-square test for nominal variables independence

We use Chi – square to understand how to test if categorical distribution of the population mapped into different groups is in association with one another. Let us take as an example, the second situation in the opening page of this chapter about TV program ratings

Suppose a particular TV operator airs two programs to audience. They want to conduct a research to know how program A and program B is favored among different ages of the viewers. Their observed result ( $f_o$ ) is published below after sampling 200 people

	Program A	Program B	
15 – 20 age	50	35	Total = 85
25 – 30 age	27	25	Total = 52
Adults	20	43	Total = 63
	Total = 97	Total = 103	

Test at 5% significance level there is no association between TV programs and viewer’s age

Look at the data table above. It is called contingency or cross-tabulation data where categorical (nominal) data is arranged into rows and columns

Here are a few things you need to know

- Use of Chi-square statistic needs expected values to be calculated assuming the claim is true (null hypothesis that no association exists between the groups)
- Degree of freedom =  $(\#rows - 1)(\#columns - 1) = (3 - 1)(2 - 1) = 2$
- Expected value of each cell =  $\frac{(\text{row total})(\text{column total})}{\text{sample size}}$
- Example in cell (program A, adults) you get  $\frac{(97)(63)}{200} \approx 30.6$

Now we can form contingency table of expected ( $f_E$ ) values given the null hypothesis is true

	Program A	Program B
15 – 20 age	41.225	43.775

## Independent and dependent samples hypothesis test

25 – 30 age	25.22	26.78
Adults	30.555	32.445

Do you remember the Chi – square test statistic? It is defined as

$$\chi^2 = \frac{(f_o - f_E)^2}{f_E}$$

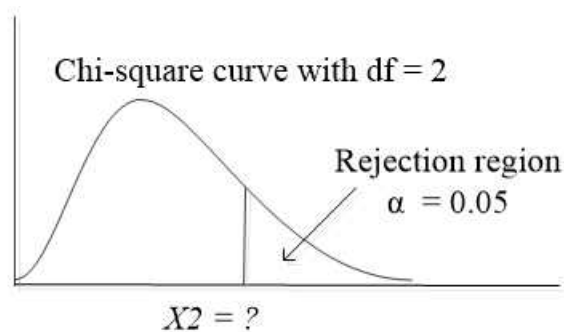
Now write hypothesis statements

$H_0$ : TV program is independent of age groups (no association)

$H_1$ : TV program is not independent of age groups (there is association)

Observed frequency ( $f_o$ )	Expected frequency ( $f_E$ )	$f_o - f_E$	$(f_o - f_E)^2$	$\frac{(f_o - f_E)^2}{f_E}$
50	41.225	8.775	77.00063	1.867814
27	25.22	1.78	3.1684	0.12563
35	43.775	-8.775	77.00063	1.759009
25	26.78	-1.78	3.1684	0.118312
20	30.555	-10.555	111.408	3.646147
43	32.445	10.555	111.408	3.43375
				$\chi^2 = 10.95$

Let us now form the rejection region



$\chi^2 = 5.59$  from chi – square table

## Independent and dependent samples hypothesis test

As can be seen our calculated  $\chi^2$  value from the table is inside of the rejection region which gives us strong evidence against the null hypothesis and hence *the alternative hypothesis is accepted that there is association between age and TV programs preference*

If you want to use p – value instead of rejection region, first get the test statistic value as 10.95 in our example. The p-value is the probability of observing values greater than the test statistic (10.95)

Using Chi – square table with degrees of freedom of 2, the closest value to our test statistic is 9.21 with associated p – value = 0.01 or 1%

*accept  $H_0$  if  $p - value > \alpha$*

*reject  $H_0$  if  $p - value \leq \alpha$*

In our case  $p - value \leq \alpha$  because of  $0.01 < 0.05$ , hence we reject the null hypothesis

Let us verify our hand calculation result with SPSS independence test analysis

Here is our real sampled value collection after conducting our research

age_ category	TV progra m	age category	TV progra m	age category	TV progra m	age category	TV progra m
15-20	A	15-20	A	25-30	A	adult	A
15-20	A	15-20	B	25-30	A	adult	A
15-20	A	15-20	B	25-30	A	adult	A
15-20	A	15-20	B	25-30	A	adult	A
15-20	A	15-20	B	25-30	A	adult	A
15-20	A	15-20	B	25-30	A	adult	A
15-20	A	15-20	B	25-30	A	adult	A
15-20	A	15-20	B	25-30	A	adult	A
15-20	A	15-20	B	25-30	A	adult	B
15-20	A	15-20	B	25-30	A	adult	B
15-20	A	15-20	B	25-30	A	adult	B
15-20	A	15-20	B	25-30	A	adult	B
15-20	A	15-20	B	25-30	A	adult	B
15-20	A	15-20	B	25-30	B	adult	B

### Independent and dependent samples hypothesis test

15-20	A	15-20	B	25-30	B	adult	B
15-20	A	15-20	B	25-30	B	adult	B
15-20	A	15-20	B	25-30	B	adult	B
15-20	A	15-20	B	25-30	B	adult	B
15-20	A	15-20	B	25-30	B	adult	B
15-20	A	15-20	B	25-30	B	adult	B
15-20	A	15-20	B	25-30	B	adult	B
15-20	A	15-20	B	25-30	B	adult	B
15-20	A	15-20	B	25-30	B	adult	B
15-20	A	15-20	B	25-30	B	adult	B
15-20	A	15-20	B	25-30	B	adult	B
15-20	A	15-20	B	25-30	B	adult	B
15-20	A	15-20	B	25-30	B	adult	B
15-20	A	15-20	B	25-30	B	adult	B
15-20	A	15-20	B	25-30	B	adult	B
15-20	A	15-20	B	25-30	B	adult	B
15-20	A	15-20	B	25-30	B	adult	B
15-20	A	15-20	B	25-30	B	adult	B
15-20	A	15-20	B	25-30	B	adult	B
15-20	A	15-20	B	25-30	B	adult	B
15-20	A	15-20	B	25-30	B	adult	B
15-20	A	15-20	B	25-30	B	adult	B
15-20	A	15-20	B	25-30	B	adult	B
15-20	A	15-20	B	25-30	B	adult	B
15-20	A	25-30	A	25-30	B	adult	B
15-20	A	25-30	A	25-30	B	adult	B
15-20	A	25-30	A	adult	A	adult	B
15-20	A	25-30	A	adult	A	adult	B
15-20	A	25-30	A	adult	A	adult	B

## Independent and dependent samples hypothesis test

15-20	A	25-30	A	adult	A	adult	B
15-20	A	25-30	A	adult	A	adult	B
15-20	A	25-30	A	adult	A	adult	B
15-20	A	25-30	A	adult	A	adult	B
15-20	A	25-30	A	adult	A	adult	B
15-20	A	25-30	A	adult	A	adult	B
15-20	A	25-30	A	adult	A	adult	B
15-20	A	25-30	A	adult	A	adult	B
		25-30	A	adult	A	adult	B
						adult	B

First our real sample data is entered into data window of the SPSS, then choose **analyze > descriptive statistics > crosstabs**

The screenshot shows the SPSS data editor window with a dataset named 'age\_category'. The data is entered as follows:

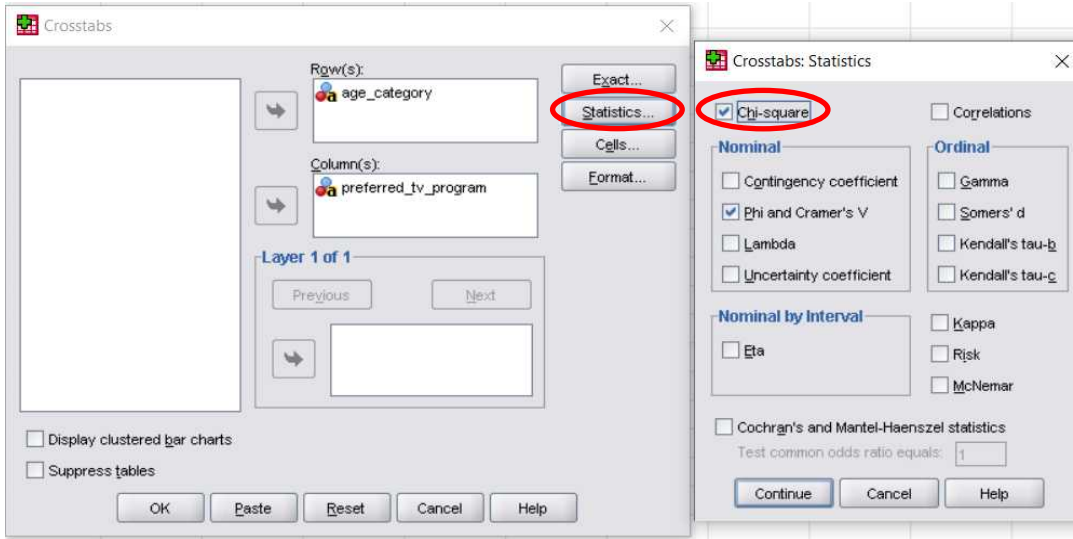
	age_category	preferred_tv_program
1	15-20	A
2	15-20	A
3	15-20	A
4	15-20	A
5	15-20	A
6	15-20	A
7	15-20	A
8	15-20	A
9	15-20	A
10	15-20	A
11	15-20	A
12	15-20	A
13	15-20	A
14	15-20	A
15	15-20	A
16	15-20	A
17	15-20	A
18	15-20	A
19	15-20	A

Overlaid on the data window is the 'Crosstabs' dialog box. The 'Row(s):' field contains 'age\_category' and the 'Column(s):' field contains 'preferred\_tv\_program'. The 'Layer 1 of 1' section has 'Previous' and 'Next' buttons. At the bottom, there are checkboxes for 'Display clustered bar charts' and 'Suppress tables', both of which are currently unchecked. The 'OK' button is highlighted.

Click the **statistics button** and make sure you check the Chi-square box



## Independent and dependent samples hypothesis test



Click **continue** and finally ok to run the analysis. The SPSS will output the result of the contingency table containing observed and expected values as well the Chi-square test result for interpretation

**age\_category \* preferred\_tv\_program Crosstabulation**

			preferred_tv_program		Total
			A	B	
age_category	15-20	Count	50	35	85
		Expected Count	41.2	43.8	85.0
	25-30	Count	27	25	52
		Expected Count	25.2	26.8	52.0
	adult	Count	20	43	63
		Expected Count	30.6	32.4	63.0
Total		Count	97	103	200
		Expected Count	97.0	103.0	200.0

**Chi-Square Tests**

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	10.951 <sup>a</sup>	2	.004
Likelihood Ratio	11.152	2	.004
N of Valid Cases	200		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 25.22.

The first row of the chi-square table needs attention. The value record shows the test statistic value same as our hand calculation (10.95)

The df record is the degree of freedom we calculated before as  $(\#rows - 1)(\#columns - 1) = (3 - 1)(2 - 1) = 2$

## Independent and dependent samples hypothesis test

The sigma value of 0.004 is the p – value much less than the assigned  $\alpha$  of 0.05

The conclusion is therefore to reject the null hypothesis and interpret that there is significant association between different programs and age categories of the population. The TV operator may now come up with decision to satisfy different age groups preferring different programs and ways to satisfy each group

### Review questions

1. A secondary school head has seen that morning statistics class performance is better than afternoon class. Morning class is taught by teacher A while afternoon class is taught by teacher B.  
The head asked his research department to sample 100 students, and test whether there is an association between class session and teacher preference (whether there is session-based preference for a particular teacher)

	Teacher A	Teacher B	
Morning class	23	27	
Afternoon class	30	20	

Test at 5% significance level there is no association between teacher preference and class session

2. A secondary school head is concerned with students getting low grades in particular subject. He believes that if new teacher is hired for this subject, student grade will improve. He asked his research department to conduct a test on a sample of 10 students before and after hiring a new teacher. The research department obtained the following result

Before grade	45	60	55	70	49	51	61	64	58
After grade	65	69	54	66	50	63	72	69	68

Test at 5% significance level the mean difference between the paired samples is zero such that there is no significance variation between exam grades as a result of teacher change

### Chapter Ten

#### Linear regression and correlation

---

After completing this chapter, you should be able to

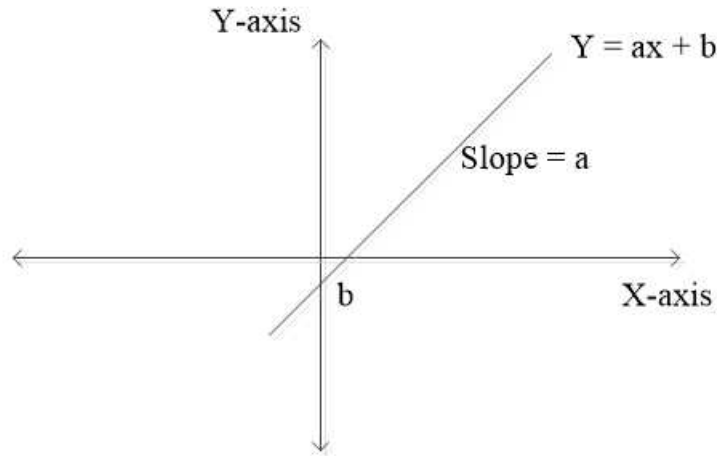
- Understand the difference between regression and correlation
- Appreciate studying relationship between two variables
- Understand regression least square method
- Obtain regression equation to establish prediction model between two variables
- Understand multiple linear regression
- Understand how to determine correlation coefficient

## Linear regression and correlation

From high school math, you studied linear equations as

$$y = ax + b$$

Where  $a$  is the slope of the straight line and  $b$  is the  $y$ -intercept on the  $x$ - $y$  graph as shown below



If we collect values of  $x$ , and substitute into the equation above, we will get corresponding values of  $y$  assuming  $a$ , and  $b$  are known. Hence  $x$  is called **independent variable** while  $y$  is called **dependent variable**.

Now how about if we don't know values of  $a$ , and  $b$  and we want to predict value of  $y$  from a given a value of  $x$ ? – We will then use regression analysis to find  $a$  and  $b$  of the line of best fit

In regression analysis if we can predict the dependent variable from the independent variable, we say that we have done regression analysis. The linear equation that fits well into this prediction is called line of best fit (regression equation)

If on the other hand, we want to test if the variables are associated or correlated with each other and we can measure the strength of that association, we say that we have done correlation analysis. The measure that tells the strength of association between continuous variables is called correlation coefficient

Here is a summary about what we said so far to differentiate regression and correlation

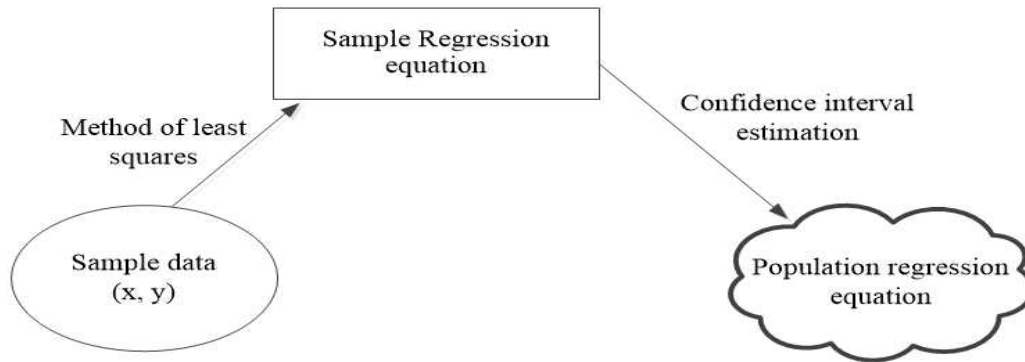
Regression	Correlation
Given one variable $x$ , find a linear equation to predict the value of variable $y$	Measure if an association between variables exist or does not exist
Regression equation	Correlation coefficient
Example is “can we predict relationship between age and income, such that if we	Example is “what is the degree of correlation between height and age”

## Linear regression and correlation

know the age of person, we can calculate the corresponding income””	
---	--

### Linear regression equation

In simple term we have data from sample which we can develop regression model. Then we can use the regression model or equation to make dependent variable predictions. We then use confidence interval estimation to find true population regression model.



Given dependent variable  $y$  and independent variable  $x$ , the linear regression equation for data sample is given by

$$y = a_1x + a_0 \quad \text{line of best fit form}$$

Late we will see how to use values of  $a_0$  and  $a_1$  as point estimator for building confidence interval for true values in the general population.

As an example suppose that we want establish relationship between yearly income and age of the population. To conduct this research we collect sample data of 20 people as shown by this table

Yearly income (x1000)	Age
38	49
43	59
48	65
46	62
41	33
43	66
29	23
41	61

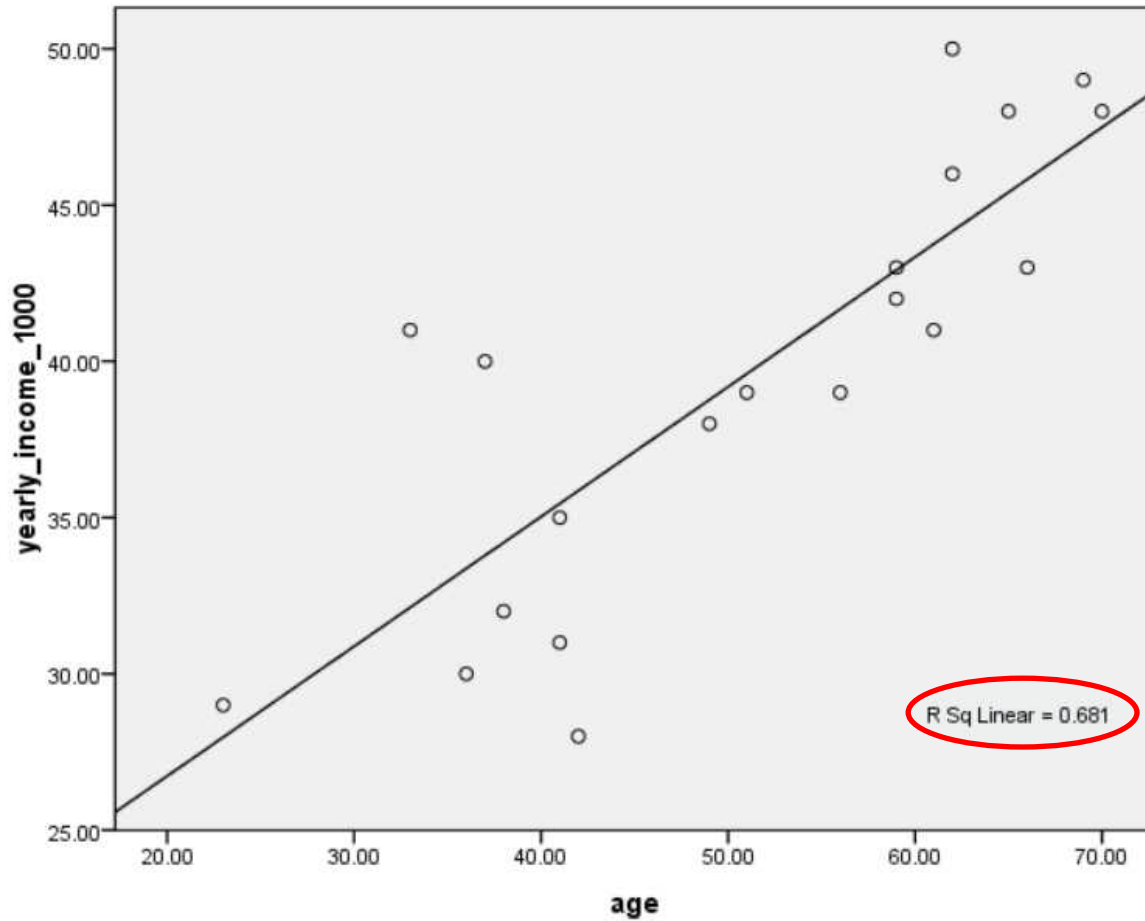
## Linear regression and correlation

35	41
39	56
31	41
50	62
48	70
49	69
28	42
39	51
42	59
32	38
30	36
40	37

The first question is which variable is x and which one is y. suppose we want to predict income of certain person of particular age. In that case x is age and y is the income we want to predict based on the value of x. Therefore we substitute x with age and y with income

Before going to develop the linear regression equation or line of best fit, let us plot scatter plot that would give preliminary hint about the relationship of the variables. Use the **graphs > chart builder** explained in chapter one and select scatter/dot plots from the gallery

## Linear regression and correlation



Visual inspection of this scatter shows that the strength of the line of fit is around 68% dependent variable predicted by the independent variable.

Now with scatter plot visualization done, let us establish regression equation which is the line of best fit for the data in the scatter plot

The least square method for finding sample regression equation is given by

$$a_1 = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2}$$

$$a_0 = \bar{y} - a_1\bar{x}$$

Let us populate our data table with all parts of the equations

y	x	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})(y - \bar{y})$	$(x - \bar{x})^2$
38	49	-2	-1.6	3.2	4
43	59	8	3.4	27.2	64
48	65	14	8.4	117.6	196

### Linear regression and correlation

46	62	11	6.4	70.4	121
41	33	-18	1.4	-25.2	324
43	66	15	3.4	51	225
29	23	-28	-10.6	296.8	784
41	61	10	1.4	14	100
35	41	-10	-4.6	46	100
39	56	5	-0.6	-3	25
31	41	-10	-8.6	86	100
50	62	11	10.4	114.4	121
48	70	19	8.4	159.6	361
49	69	18	9.4	169.2	324
28	42	-9	-11.6	104.4	81
39	51	0	-0.6	0	0
42	59	8	2.4	19.2	64
32	38	-13	-7.6	98.8	169
30	36	-15	-9.6	144	225
40	37	-14	0.4	-5.6	196
$\bar{y} = \frac{\sum y}{n}$ = 39.6	$\bar{x} = \frac{\sum x}{n}$ = 51			Sum = 1488	Sum = 3584

$$a_1 = \frac{1488}{3584} = 0.415$$

$$a_0 = 39.6 - 0.415(51) = 18.42$$

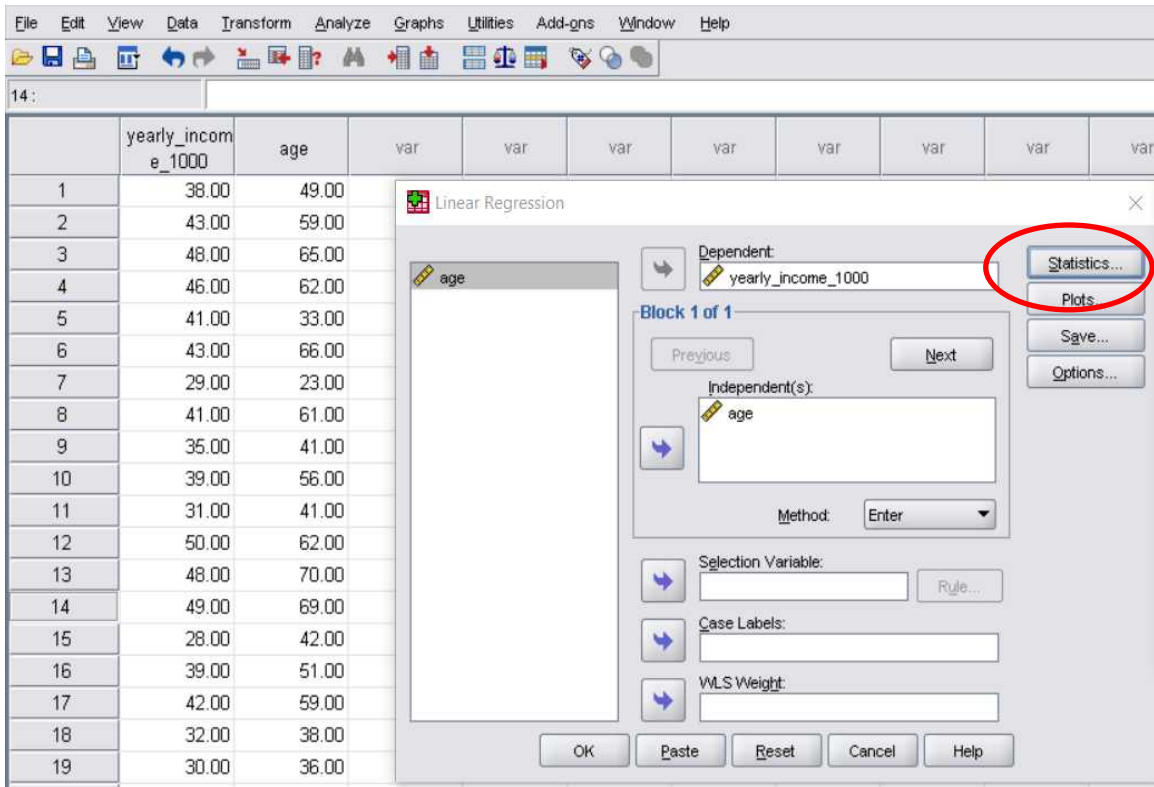
Our regression equation is thus  $y = 0.415x + 18.42$  *y is predicted from x*

Taking glance at this equation shows that the slope (0.415) is positive which indicates positive model between the variables. This means if we increase the value the independent variable (x), the predicted dependent variable (y) correspondingly increases

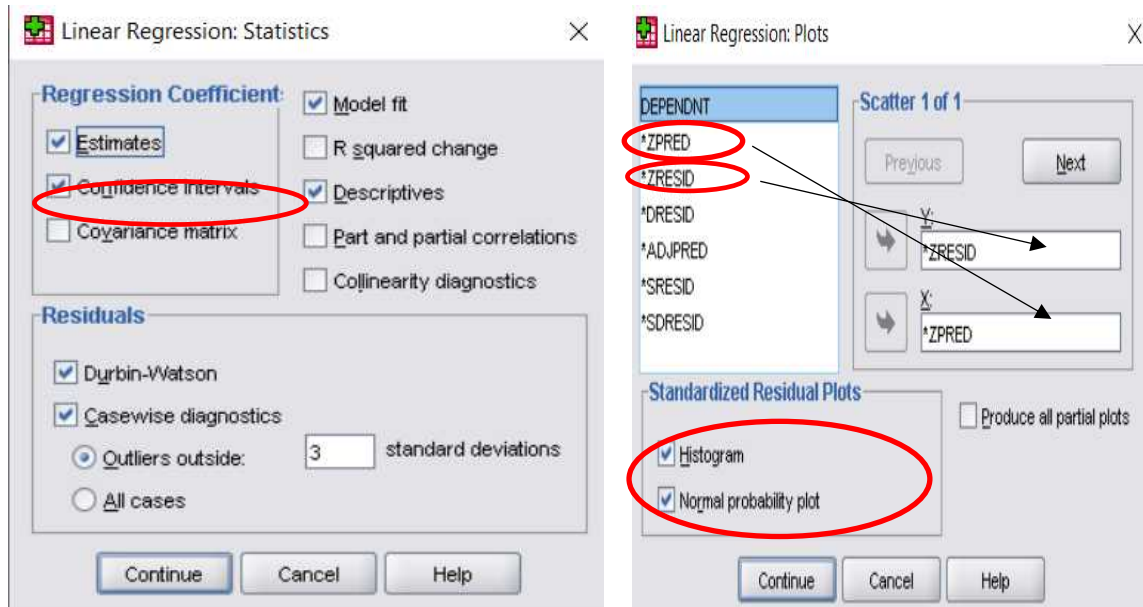


## Linear regression and correlation

To do regression analysis in SPSS go to **analyze > regression > linear** to pop up linear regression dialog box



Click **statistics** and **plots** buttons, and set them as shown below



Once everything is set as shown click Ok to bring up results dialog for interpretation

## Linear regression and correlation

Let us first get Residuals statistic dialog to check for any outlier that will make the regression analysis less effective

**Residuals Statistics<sup>a</sup>**

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	27.9750	47.4884	39.6000	5.70220	20
Residual	-7.86339	8.87321	.00000	3.90016	20
Std. Predicted Value	-2.039	1.383	.000	1.000	20
Std. Residual	-1.962	2.214	.000	.973	20

a. Dependent Variable: yearly\_income\_1000

The last row of **standard residuals** should not be less than -3 (minimum) or greater than 3 (maximum) to consider everything is normal. Residual is the distance of data points from the line of best fit. If the standard residual is large it means the data point is far from the line of best fit and can be considered an outlier. Also you could check any data entry error to find the cause of the outlier

Finally the coefficient table for values of the constants  $a_0$  and  $a_1$ . See how it matches with hand calculation

**Coefficients<sup>a</sup>**

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
1 (Constant)	18.426	3.529		5.221	.000	11.011	25.840
age	.415	.067	.825	6.203	.000	.275	.556

a. Dependent Variable: yearly\_income\_1000

But this relationship is for sample. The next question is whether this relationship is significant at the general population level

First if the estimated sample regression equation is

$$y = a_1x + a_0$$

Then the population regression equation is

$$y = \beta_1x + \beta_0$$

The test of regression model significance can be stated as

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

Reject  $H_0$  if p-value  $\leq \alpha$

To test this hypothesis, the test statistic using t-distribution is given by

$$t - \text{statistic} = \frac{\text{slope}}{\text{standard error}} = \frac{0.415}{0.067} = 6.20$$

## Linear regression and correlation

The p-value corresponding to the extreme of the test statistic is given by using the t-table with degrees of freedom 19. You will get much small value like 0.00001

Thus the p-value < 0.05 and we therefore reject the null hypothesis and conclude that the regression model is significant

### Goodness of fit for the regression equation

If you go back to our scatter plot above you will see the value circled which say “R sq Linear” is 0.681. What does this value represent and what does its interpretation suggest?

This is a ratio called coefficient of determination of regression and it measures how well the regression equation can predict the dependent variable (income in our case). The value produced by SPSS scatter plot shows only 68% of the dependent variable can be predicted using the regression equation.

Now let us define new quantities and populate our table

- **Residual = observation – predicted**

For example by substituting the first x row value of our data into the regression equation we get

$$y = 0.415(49) + 18.42 = 38.755 \quad \text{predicted value}$$

But our observed value for y was 38

Hence residual = 38 – 38.755 = -0.755

- **Observation (y) – mean observation (39.6) = (income value – mean income) and then square it**

For example the first row will result (38 – 39.6) = -1.6 which if squared will give 2.56

- **Predicted – mean observation (39.6) and then square it**

For example the first row will result (38.755 – 39.6) = -0.845 which if squared will give 0.714025

Y (observation)	X (predictor)	Predicted	Residual	Residual squared	(observation – 39.6) <sup>2</sup>	(predicted – 39.6) <sup>2</sup>
38	49	38.755	-0.755	0.570025	2.56	0.714025
43	59	42.905	0.095	0.009025	11.56	10.92303
48	65	45.395	2.605	6.786025	70.56	33.58202
46	62	44.15	1.85	3.4225	40.96	20.7025
41	33	32.115	8.885	78.94323	1.96	56.02523

### Linear regression and correlation

43	66	45.81	-2.81	7.8961	11.56	38.5641
29	23	27.965	1.035	1.071225	112.36	135.3732
41	61	43.735	-2.735	7.480225	1.96	17.09823
35	41	35.435	-0.435	0.189225	21.16	17.34723
39	56	41.66	-2.66	7.0756	0.36	4.2436
31	41	35.435	-4.435	19.66923	73.96	17.34723
50	62	44.15	5.85	34.2225	108.16	20.7025
48	70	47.47	0.53	0.2809	70.56	61.9369
49	69	47.055	1.945	3.783025	88.36	55.57703
28	42	35.85	-7.85	61.6225	134.56	14.0625
39	51	39.585	-0.585	0.342225	0.36	0.000225
42	59	42.905	-0.905	0.819025	5.76	10.92303
32	38	34.19	-2.19	4.7961	57.76	29.2681
30	36	33.36	-3.36	11.2896	92.16	38.9376
40	37	33.775	6.225	38.75063	0.16	33.93063
$\bar{y} = \frac{\sum y}{n}$ = 39.6	$\bar{x} = \frac{\sum x}{n}$ = 51			Sum of residuals squared (SSE) = 289.018 9	Total sum of deviation squares (SST) = 906.8	Sum of regression squares (SSR) = 617.2589

In summary we have defined three new quantities as

$$SSE = \sum (\text{observation} - \text{predicted})^2$$

$$SST = \sum (\text{observation} - \text{mean observation})^2$$

$$SSR = \sum (\text{predicted} - \text{mean observation})^2$$

Now our goodness of fit for regression equation is given by the following equation

$$R^2 = \frac{SSR}{SST} = \frac{617.2589}{906.8} = 0.681$$

## Linear regression and correlation

This value matches our SPSS scatter plot value. It simply says using our obtained regression equation, we can only predict 68% of the dependent variable correctly

Instead of the scatter plot, you can find the value of  $R^2$  from model summary table in the regression analysis output window as shown below

**Model Summary<sup>b</sup>**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.825 <sup>a</sup>	.681	.664	4.00704	2.132

a. Predictors: (Constant), age  
b. Dependent Variable: yearly\_income\_1000

This indicates our regression model is not much significantly good fit.  $R^2 > 90\%$  for your model to be more significant

### Confidence interval for regression equation slope

Recall from chapter 6 we said that

$$\text{confidence interval for population} = \text{sample statistic} \pm \text{margin of error}$$

The regression equation obtained above was based on sample data. We now take test statistic from the sample data as slope of the regression line

$$\text{sample statistic: } a_1 = 0.415$$

The standard error of the regression line equation is

$$s_{a1} = \frac{s_y}{\sqrt{\sum(\text{predictor} - \text{mean predictor})^2}} = \frac{s_y}{\sqrt{\sum(x - \bar{x})^2}}$$

Where  $s_y$  is standard deviation of the dependent variable y

$$s_y = \sqrt{\frac{\sum(\text{observation} - \text{predicted})^2}{n - 2}} = \sqrt{\frac{SSE}{n - 2}} = \sqrt{\frac{289.0189}{20 - 2}} = 4$$

Hence

$$s_{a1} = \frac{4}{\sqrt{3584}} = 0.0668$$

SPSS regression analysis coefficient table can verify our standard error as shown below

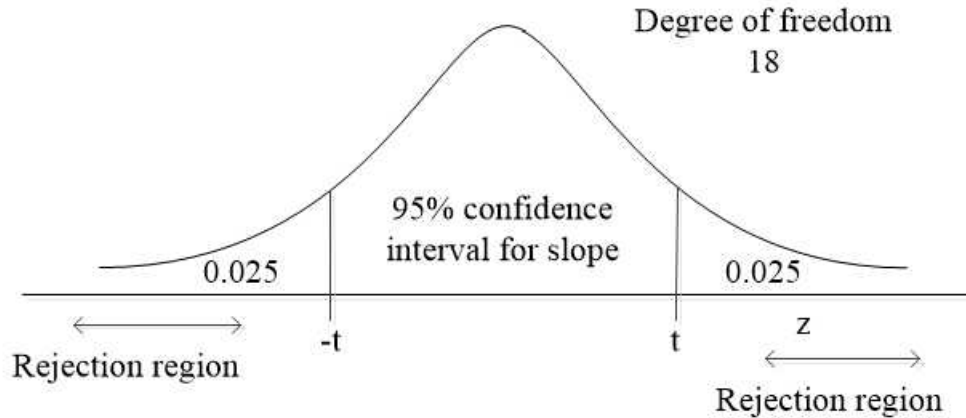
## Linear regression and correlation

**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	18.426	3.529		5.221	.000	11.011	25.840
	age	.415	.067	.825	6.203	.000	.275	.556

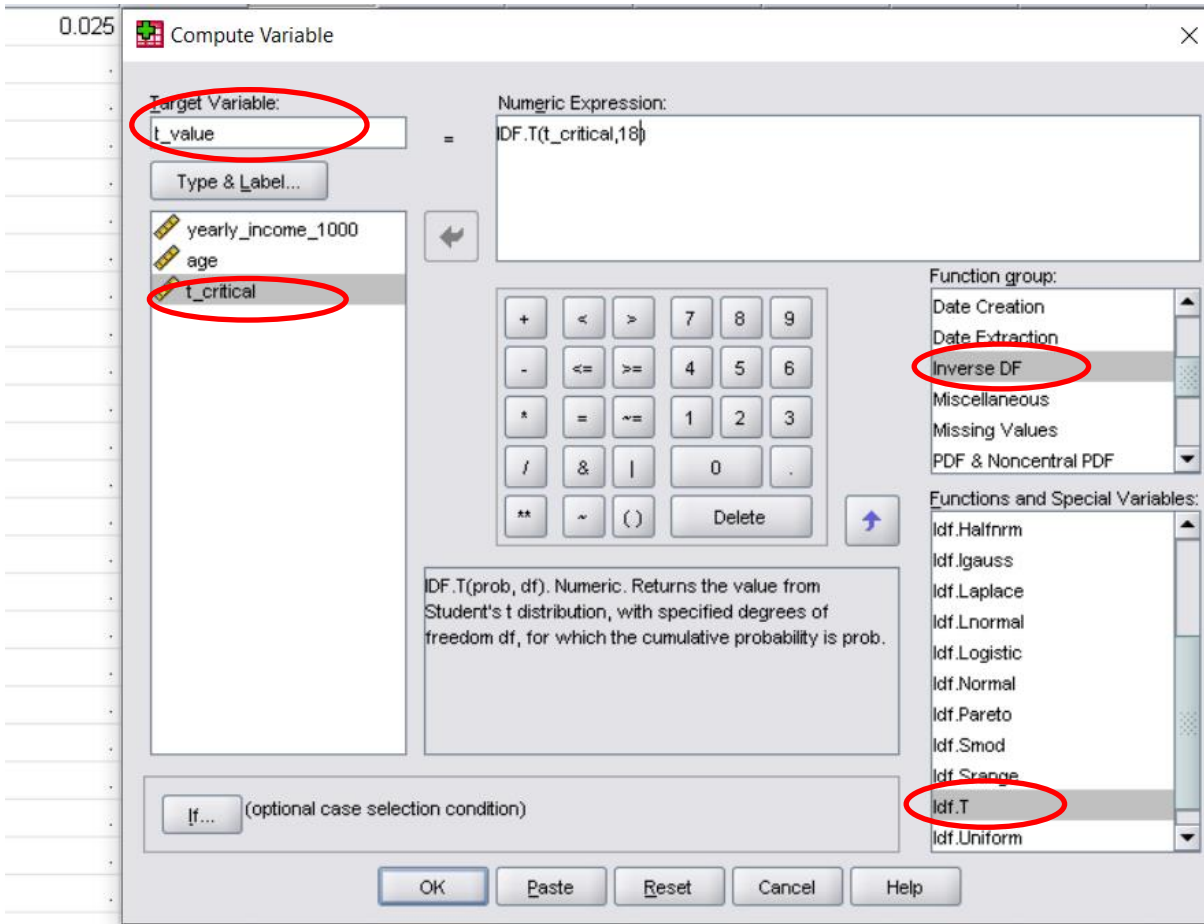
a. Dependent Variable: yearly\_income\_1000

Now using  $t$  – distribution, find the confidence interval of the regression equation slope at 95% confidence level



From the diagram critical values are 2.101 and -2.101. We can confirm using SPSS as shown below which will give you same result

## Linear regression and correlation



*confidence interval for regression slope =  $0.415 \pm (2.101)(0.0668)$*

This means the true slope of the regression line will lie in the interval [0.275, 0.555]

SPSS regression analysis coefficient table can verify our confidence interval for regression analysis slope as shown below

Coefficients <sup>a</sup>								
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	18.426	3.529		5.221	.000	11.011	25.840
	age	.415	.067	.825	6.203	.000	.275	.556

a. Dependent Variable: yearly\_income\_1000

From this table just above you can also extract confidence interval for the constant term of the regression equation ( $a_0$ ) which shows that it lies in the interval [11.011, 25.840]

### Multiple linear regression

Our discussion on regression equation from last section was considering on one independent variable. We saw that income was linearly related with age and the relationship was positive. We considered only age as independent variable. However, there

## Linear regression and correlation

are many cases in which it is helpful to consider the effect of multiple independent variable. For example we can study how income is related to age and education level. In this case we have two independent variable namely age and education level. But education level is not quantitative but ordinal data. To use ordinal data in regression analysis, we use dummy variables in SPSS.

If simple linear regression equation for one independent variable is

$$y = a_1x + a_0$$

Then multiple regression equation for k independent variables is

$$y = a_1x_1 + a_2x_2 + \cdots a_kx_k + a_0$$

We revisit our example above but now add one more column of education level coded as High school = 1, Bachelor = 2, Master = 3, and PhD = 4

Yearly income (x1000)	Age	Education level
38	49	3
43	59	1
48	65	3
46	62	1
41	33	3
43	66	4
29	23	2
41	61	1
35	41	4
39	56	4
31	41	1
50	62	2
48	70	4
49	69	1
28	42	4
39	51	3
42	59	1



## Linear regression and correlation

32	38	1
30	36	3
40	37	2

To do multiple regression in SPSS go to **analyze > regression > linear** and populate the variables boxes as shown below

The screenshot shows the SPSS Linear Regression dialog box. The dependent variable is 'income of respondent [incom...]' and the independent variables are 'age of respondent [age]' and 'education level [ed]'. The method is set to 'Enter'. The dialog box also includes buttons for 'Statistics...', 'Plots...', 'Save...', 'Options...', 'Style...', and 'Bootstrap...'. The 'Previous' and 'Next' buttons are also visible.

Once click ok, the following output result will be generated

First the model summary table shows high R square value (77%) which indicates this model is good fit of the predicted data

**Model Summary<sup>b</sup>**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.878 <sup>a</sup>	.771	.744	3.496

a. Predictors: (Constant), education level, age of respondent

b. Dependent Variable: income of respondent

## Linear regression and correlation

The ANOVA table shows the model is significant overall as shown below. The p-value is much less than the test significance (0.05)

**ANOVA<sup>a</sup>**

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	699.050	2	349.525	28.601	.000 <sup>b</sup>
	Residual	207.750	17	12.221		
	Total	906.800	19			

a. Dependent Variable: income of respondent

b. Predictors: (Constant), education level, age of respondent

To test the significance of each independent variable, we can check the p-value in the coefficient table below. As can be seen both age and education independent variable show p-value less than the test significance (0.05). This suggest the regression model is statistically significant.

**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	15.269	3.313		4.608	.000	8.279	22.260
	age of respondent	.297	.074	.590	3.989	.001	.140	.453
	education level	3.348	1.298	.381	2.579	.020	.609	6.087

a. Dependent Variable: income of respondent

From the coefficients table above we can also build the estimated regression equation as below

$$y = 0.297x_1 + 3.348x_2 + 15.269$$

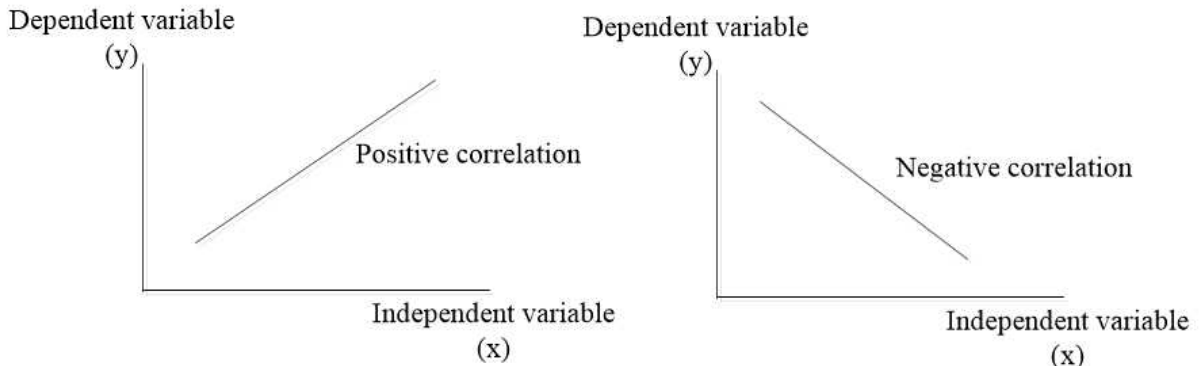
Where  $x_1$  is the age variable and  $x_2$  is the education variable

## Linear regression and correlation

### Correlation coefficient $r$

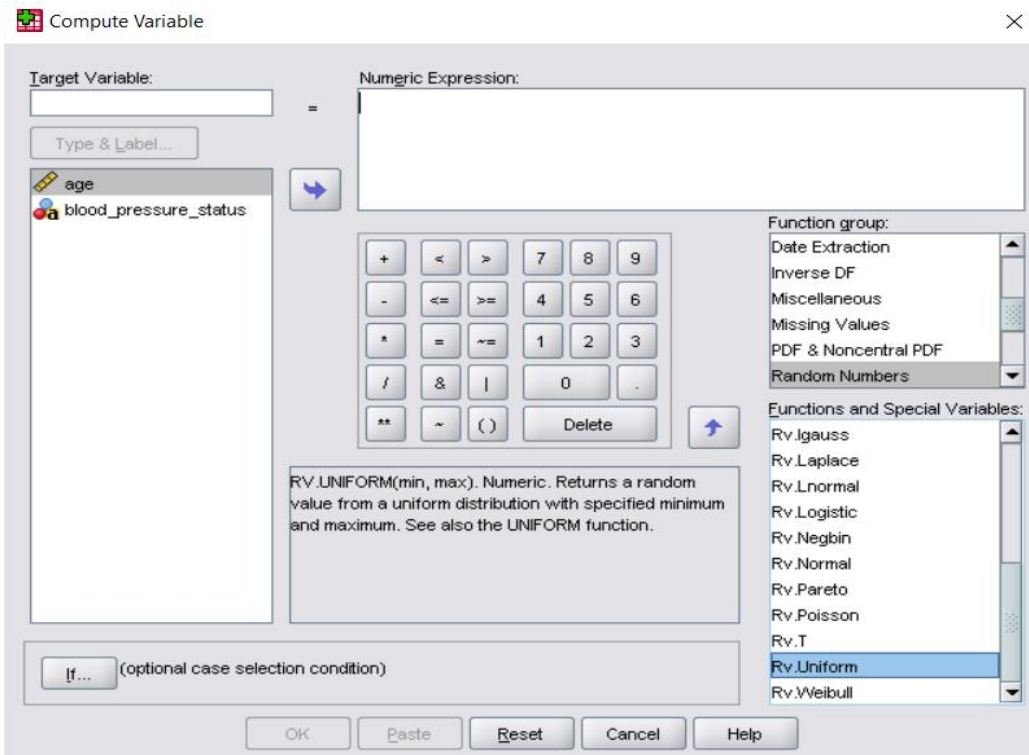
Enough we said about linear regression in the last few pages. Now we turn our attention to finding if two continuous variables of particular interest are correlated and how strong is the correlation

If the dependent variable linearly increases with the independent variable, the correlation is positive and negative otherwise



Suppose that a clinical researcher might be interested in finding if there is correlation between height and age of a person. He takes a sample of 20 people then prepares the sample data as shown in the table below

Let us generate random number to use as our data for age and height using **SPSS transform > computer variable** and random numbers in the function group



## Linear regression and correlation

Age	Height (cm)
44	139
24	163
22	138
40	165
32	163
37	136
35	149
53	165
25	139
31	153
43	148
42	163
48	157
27	145
28	136
39	146
29	169
57	168
20	158
36	136

The coefficient of correlation has a range of values between [-1, 1]

The Pearson correlation coefficient is defined as

$$r = \frac{s_{xy}}{s_x s_y}$$

Where

## Linear regression and correlation

$$s_{xy} = \frac{\sum(x - \bar{x})(y - \bar{y})}{n - 1}$$

$$s_x = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

$$s_y = \sqrt{\frac{\sum(y - \bar{y})^2}{n - 1}}$$

Let us population these equation into our table above

x	y	$(x - \bar{x})^2$	$(y - \bar{y})^2$	$(x - \bar{x})(y - \bar{y})$
44	139	64	169	-104
24	163	144	121	-132
22	138	196	196	196
40	165	16	169	52
32	163	16	121	-44
37	136	1	256	-16
35	149	1	9	3
53	165	289	169	221
25	139	121	169	143
31	153	25	1	-5
43	148	49	16	-28
42	163	36	121	66
48	157	144	25	60
27	145	81	49	63
28	136	64	256	128
39	146	9	36	-18
29	169	49	289	-119
57	168	441	256	336
20	158	256	36	-96

## Linear regression and correlation

36	136	0	256	0
$\bar{x}$ $= \frac{\sum x}{n}$ $= 36$	$\bar{y} = \frac{\sum y}{n}$ $= 152$	Sum = 2002	Sum = 2720	Sum = 706

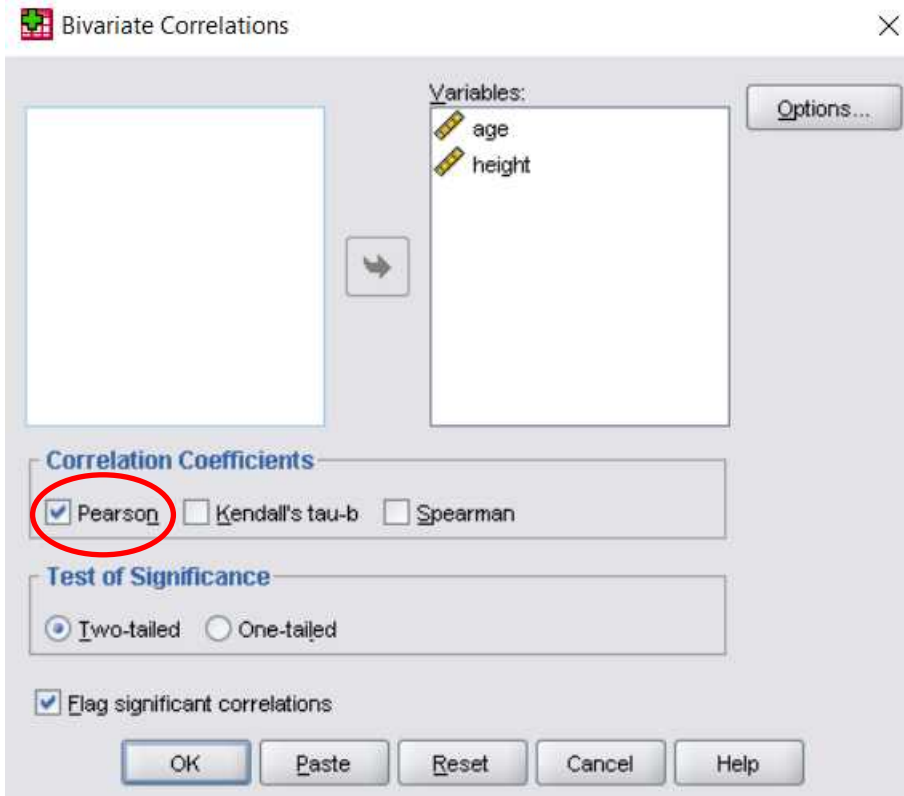
$$s_x = \sqrt{\frac{2002}{19}} = 10.265$$

$$s_y = \sqrt{\frac{2720}{19}} = 11.965$$

$$s_{xy} = \frac{706}{19} = 37.158$$

$$r = \frac{37.158}{(10.265)(11.965)} = 0.302$$

To get the same value is SPSS run **analyzer > correlate > bivariate** to bring up bivariate correlation dialog box, and make sure Pearson is checked under correlation coefficients



## Linear regression and correlation

Click Ok to generate the analysis in the SPSS output windows below

		age	height
age	Pearson Correlation	1	.302
	Sig. (2-tailed)		.195
	N	20	20
height	Pearson Correlation	.302	1
	Sig. (2-tailed)	.195	
	N	20	20

The correlation coefficient  $r = 0.302$  in the SPSS output matches with the value we calculated using our hand

Now with the value of the correlation coefficient being computed, what can we conclude about the variables age and height?

Remember we said  $r$  is between  $-1$  and  $+1$  and as it gets close to  $-1$  or  $+1$ , it indicate strong correlation between the variables under study. In our case we got very small value of  $r = 0.302$  and we therefore conclude there is weak correlation (or there is weaker correlation) between the variables

One more thing is that this result is for the sample data. How can infer if this correlation is significant at the population level. If the  $p$ -value in the correlation table is less than the test significance, we conclude the correlation is statistically significant. This is not our case as  $p$ -value ( $0.195$ ) is greater than  $0.05$ . we therefore conclude the correlation is not statistically significance

### Review questions

1. The aviation agency is concerned with correlation between number flights per day and passenger number. On one week observation they collected number of flights and number of passenger flying each day as shown below

Number of flights	14	9	21	12	17
Number of passengers	192	178	99	102	105

Using Pearson correlation coefficient, is there correlation between number of flights and number of passengers?

## **One – way Analysis of variance (one – way ANOVA)**

### Chapter Eleven

## **One – way Analysis of variance (one – way ANOVA)**

---

After completing this chapter, you should be able to

- Understand the concept of random experimental design
- Use one-way ANOVA to test equality of population means
- Demonstrate clear understanding of how to use F-distribution to compare means of more than two independent samples
- Understand case of repeated measures ANOVA for paired k samples
- Practice ANOVA analysis in SPSS



## One – way Analysis of variance (one – way ANOVA)

### Randomized experimental design

In the last few chapters on inferential statistics we observed sample data and then used it to draw conclusion about the general population. There are, however, cases in statistics where you may want to do an experiment (rather than observational research) to study how one factor can control result of the experiment. Since you cannot conduct the experiment on all the subjects of the population (experimental units), you need a method that can help you analyze random data collected from controlled experiment

Suppose that the ministry of agriculture is planning to increase agricultural production of the country because of food shortage. They want to decide if particular fertilizer can have same effect on three crop types.

The ministry of agriculture sets up small agricultural land to test if average growth height of three different crop populations varies with same fertilizer. They decided to plant three different types of crops in which each crop type ten samples were planted. They then applied the fertilizer equally for the three months of the spring

As can be seen from this experiment we have more than two samples and we cannot use methods used in chapter 9. To analyze more than two samples case, we run one-way ANOVA using statistic called F-distribution. SPSS provides a much robust technique called Welch’s test. The underlying assumptions in this tests are normality, homogeneity (claim that variances of the groups are equal) and independence of the population groups. So a researcher will always test these first before proceeding to ANOVA for comparing equality of means. Ways of testing normality are explored in chapter 8. When running ANOVA in SPSS, homogeneity test done by Levene’s is displayed as part of the analysis. A large value of sigma value in Levene’s test indicates homogeneity of variances

The following table presents randomized experiment terms and how they relate to our study

Term	Definition	Example to our case
Factor	the independent variable (the controlled variable) manipulated by the experimenter	Crop type
Treatment	The measured dependent variable	Crop height
Single factor experiment	Experiment is associated with one factor	Only crop height measured
Experimental units	Measurement variables	30 crops
Randomized design	Subjects are selected randomly	

## One – way Analysis of variance (one – way ANOVA)

Replication	Number of times of experimental units us	One crop but replicated 10 times for each sample
-------------	--	--

### One-way ANOVA test for three samples means equality

Shown below are heights (cm) of three crop types with each of the crop type sampled 10 times. The three crop types are treated equally by the same fertilizer

Crop A	Crop B	Crop C
72	84	72
69	58	62
70	54	97
73	79	92
74	91	63
76	59	89
96	84	98
99	100	91
90	73	80
83	76	77
Mean = 80.2	Mean = 75.8	Mean = 82.1
Variance =123.5	Variance = 227	Variance =177.9

In this experiment we can have two types of variations

- Variation of crop height within the same crop type (within ten samples of the crop A for example)
- Variation of crop height among three types of crops (among crop A, crop B, crop C)

Look at how we designed this random experiment in which we are studying three different populations (three different crop populations in this case) each one of them normally distributed. We then experiment replicated samples from each population

In general we can say that to run one-way experimental ANOVA for population means equality

- Select k populations with means  $\mu_1, \mu_2, \mu_3 \dots \mu_k$

## One – way Analysis of variance (one – way ANOVA)

- Assume k populations are independent (chapter 9) and normally distributed (chapter 8)
- Select samples from each population with each sample replicated n times

To begin analyzing our given problem we make the assumption (null hypothesis) that there is not mean differences among heights of different crop populations which translates to that their average heights will be equal

$$H_0: \mu_1 = \mu_2 = \mu_3$$

The alternative hypothesis is that there is at least one mean difference among different crop heights when same particular fertilizer is applied

$$H_1: \textit{at least two of the means not equal}$$

Looking at our data table above, we found mean of crop A sample is 80.2, mean of crop B sample is 75.8, mean of crop C sample is 82.1. This means are called mean within treatment

The mean of the 30 experimental units (total samples) is  $(80.2 + 75.8 + 82.1) / 3 = 79.37$ . This mean is called mean between treatments. When the sample sizes of the treatments are different, add all the experimental units and divide by total sample size

Let us now define new quantity called mean square due to treatment (MSTR) defined as

$$MSTR = \frac{\sum n(\textit{mean within treatment} - \textit{mean between treatments})^2}{k - 1} \textit{ for all } k$$

Where n is number of replications per treatment (10 in our case)

Calculation of MSTR using the above formula is showed in the following table using values from our example

	<b>(mean within treatment A – mean between treatments)</b>	<b>(mean within treatment B – mean between treatments)</b>	<b>(mean within treatment C – mean between treatments)</b>	<b>Total</b>
	0.83	-3.57	2.73	
Squared	0.6889	12.745	7.453	
(n = 10)	(10)(0.6889)	(10)(12.745)	(10)(7.453)	
(k = 3)	(10)(0.6889) / 2 =		(10)(7.453) / 2 =	<b>MSTR</b>
(k – 1) = 2		(10)(12.745) / 2 =		<b>= 104.4</b>

Let us now define another new quantity called mean square due to error (MSE)

## One – way Analysis of variance (one – way ANOVA)

$$MSE = \frac{\sum(n - 1)sample\ variance}{n - k} \quad \text{for all } k$$

Where the n in the numerator is replication per treatment (10 in our case) and n in the denominator is total experimental units (30 crops total)

	Variance =123.5	Variance = 227	Variance =177.9	Total
(n = 10 samples) (n - 1) = 9	(9)(123.5)	(9)(227)	(9)(177.9)	
(k = 3)				
/(n - k) = 30 – 3 = 27	(9)(123.5) / 27	(9)(227) / 27	(9)(177.9) / 27	<b>MSE = 176</b>

Now we come to defining the test statistic we will use to test equality of treatments means using the F- distribution

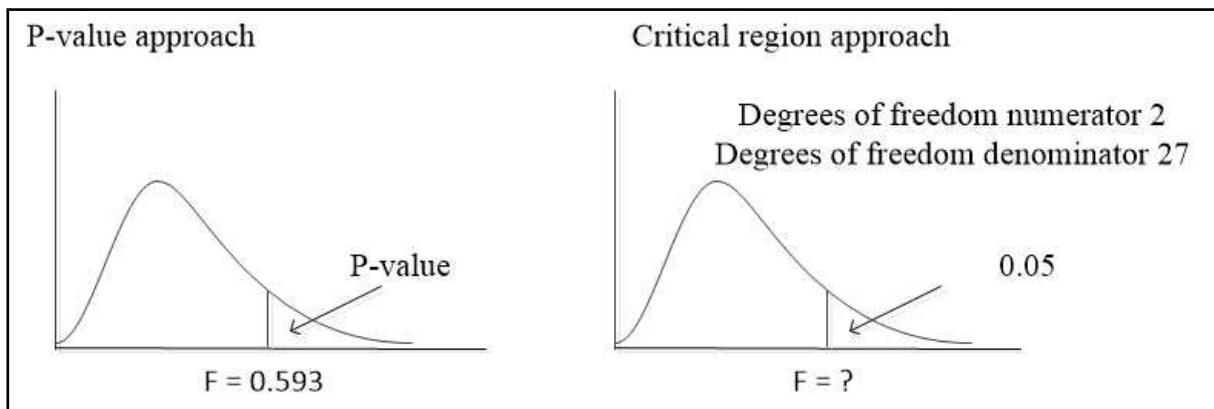
$$F = \frac{MSTR}{MSE} = \frac{104.4}{176} = 0.593 \quad \text{test statistic}$$

If we have the test statistic, can you guess what final piece we are missing to reach final conclusion? Yes we need to get the critical value

Remember we are testing the null hypothesis at  $\alpha = 5\%$

The F – distribution graph is skewed graph that has only positive critical values. We can either use critical region approach to find if the test statistic (0.593) falls in the rejection region or p – value  $> \alpha$  approach

Let us use both as shown below



To use the rejection region approach let use F – distribution table to find F – value corresponding to  $\alpha = 0.05$

## One – way Analysis of variance (one – way ANOVA)

The F – distribution table has degrees of freedom in the numerator and degrees of freedom in the denominator

Degrees of freedom in numerator =  $k - 1 = 3 - 1 = 2$

Degrees of freedom in denominator =  $n - k = 30 - 3 = 27$

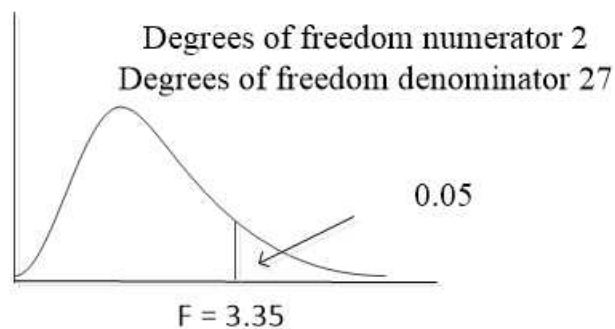
Where  $n$  = number treatment replication and  $k$  = number of populations

	Numerator degrees of freedom										
Right tail probability	Denominator df	1	2	3	4	5	6	7	8	9	10
0.05	1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88
0.05	2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
0.05	3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
0.05	4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
0.05	5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
0.05	6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
0.05	7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
0.05	8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
0.05	9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
0.05	10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
0.05	11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
0.05	12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
0.05	13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
0.05	14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
0.05	15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
0.05	16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
0.05	17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
0.05	18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
0.05	19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38

## One – way Analysis of variance (one – way ANOVA)

0.05	20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
0.05	21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32
0.05	22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
0.05	23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27
0.05	24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25
0.05	25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24
0.05	26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
0.05	27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20
0.05	28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19

Thus we get critical value of **3.35** when we use (2, 27) degrees of freedom intersection on the table and  $\alpha = 0.05$



Do the test statistic (0.593) fall in the rejection region or not?

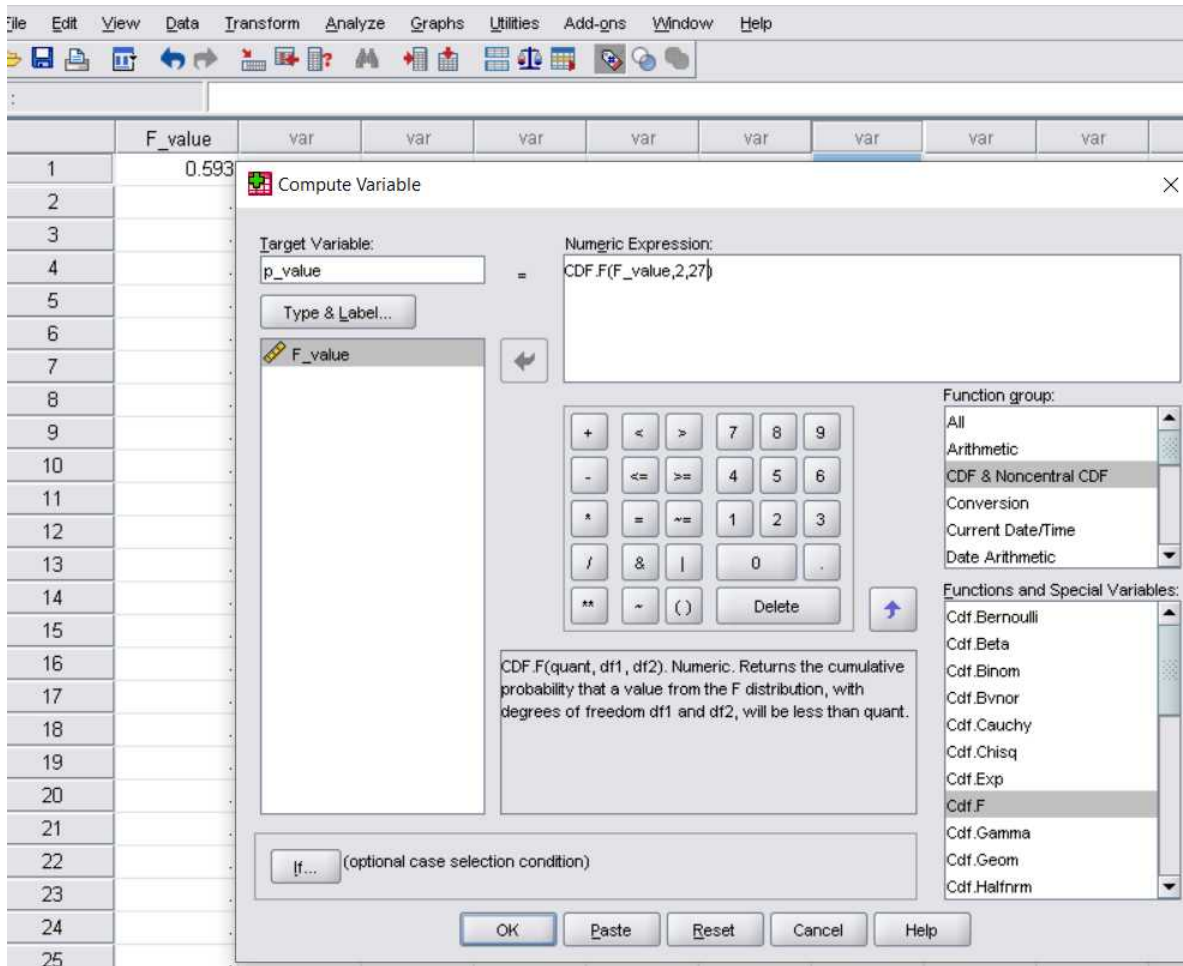
The answer is NO as  $0.593 < 3.35$

This gives evidence to accept the null hypothesis and conclude that all population means are equal. The three crops growing to equal mean heights is justified and the same fertilizer could be ordered on large scale

We can also use the p –value approach which is finding probability of getting greater than our test statistic (0.593)

Use SPSS the find the p – value as show below using **CDF.F** from function group

## One – way Analysis of variance (one – way ANOVA)



You will get  $p$  – value = 0.44

As you can see  $p$  – value  $> \alpha$  ( $0.44 > 0.05$ ), we therefore accept the null hypothesis and conclude that all population means will be equal for all crop types when that same fertilizer is applied.

Now time has come to run one – factor ANOVA in SPSS. First thing is data entry. Then select **analyze > compare means > one-way ANOVA**

## One – way Analysis of variance (one – way ANOVA)

The screenshot shows the SPSS One-Way ANOVA dialog box overlaid on a data table. The data table has columns for 'crop\_category' and 'height'. The dialog box shows 'height' in the 'Dependent List' and 'crop\_category' in the 'Factor' field.

	crop_category	height	var	var	var	var	var	var
1	A	72.00						
2	A	69.00						
3	A	70.00						
4	A	73.00						
5	A	74.00						
6	A	76.00						
7	A	96.00						
8	A	99.00						
9	A	90.00						
10	A	83.00						
11	B	84.00						
12	B	58.00						

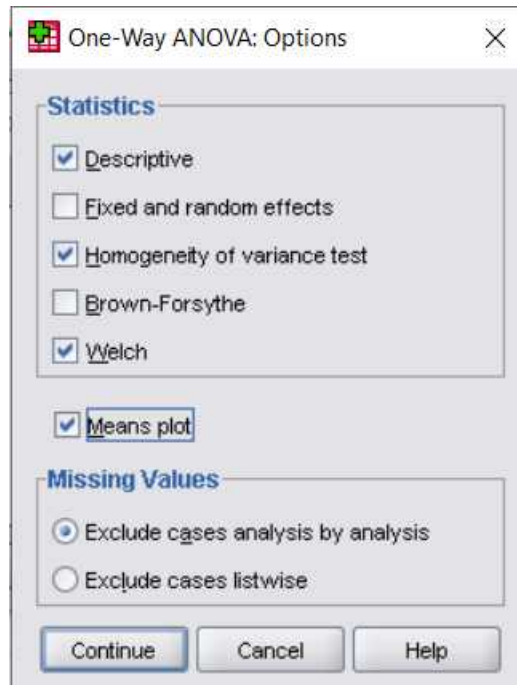
Since we have equal replication in each group, select **Tukey** for equal variances assumed under **post\_hoc** button

The screenshot shows the 'One-Way ANOVA: Post Hoc Multiple Comparisons' dialog box. Under 'Equal Variances Assumed', the 'Tukey' checkbox is selected and circled in red. Under 'Equal Variances Not Assumed', the 'Games-Howell' checkbox is selected and circled in red. The 'Significance level' is set to 0.05.

Click options in one-way ANOVA dialog and selection the following options



## One – way Analysis of variance (one – way ANOVA)



Click ok and SPSS output window will generate

Let us walk through all results one by one

First descriptive statistics table will summarize results for mean and standard deviation of the treatment and also 95% confidence for the true population means

**Descriptives**

height

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
A	10	80.2000	11.11356	3.51441	72.2498	88.1502	69.00	99.00
B	10	75.8000	15.06873	4.76515	65.0205	86.5795	54.00	100.00
C	10	82.1000	13.33708	4.21756	72.5592	91.6408	62.00	98.00
Total	30	79.3667	13.08456	2.38890	74.4808	84.2525	54.00	100.00

Next Levene's test of homogeneity (equal variances) is displayed as it is needed assumption of the on-way ANOVA test

The higher sigma value (0.7) indicate homogeneity of the different crop populations is respected

**Test of Homogeneity of Variances**

height

Levene Statistic	df1	df2	Sig.
.361	2	27	.700

## One – way Analysis of variance (one – way ANOVA)

The one-way ANOVA report table using F-statistic is shown below along with robust test of means equality using Welch test

**ANOVA**

height					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	208.867	2	104.433	.593	.560
Within Groups	4756.100	27	176.152		
Total	4964.967	29			

**Robust Tests of Equality of Means**

height				
	Statistic <sup>a</sup>	df1	df2	Sig.
Welch	.489	2	17.714	.622

a. Asymptotically F distributed.

The last column of the ANOVA table with label Sig is the p – value which is greater than  $\alpha$  ( $0.560 > 0.05$ ). Hence we accept the null hypothesis which matches with our hand calculation

### Fisher LSD multiple comparisons

The F statistic will indicate whether the null hypothesis is true and the population means are equal or otherwise reject it. It will not indicate where the differences lie if the null hypothesis is rejected. Fisher least significant difference (LSD) can be used to test the alternative hypothesis that at least two of the means are not equal.

From the example in this chapter we tested three populations. Let sample means of any two out of the three groups be  $\bar{x}_i$  and  $\bar{x}_j$  while that of their corresponding populations means is  $\mu_i$  and  $\mu_j$ . Sample I size is  $n_i$  and sample j size is  $n_j$ . We select significance level of 5%

The test statistic of fisher LSD is then the mean difference  $\bar{x}_i - \bar{x}_j$

For example we take populations 1 and 2 (crop A and crop B)

The null hypothesis is  $H_0: \mu_i = \mu_j$

The test rule is

$$\text{reject } H_0 \text{ if } p - \text{value} < \alpha$$

Where test statistic is

## One – way Analysis of variance (one – way ANOVA)

$$t = \frac{\bar{x}_i - \bar{x}_j}{\sqrt{MSE \left( \frac{1}{n_i} + \frac{1}{n_j} \right)}}$$

Using this equation we can compute test statistic for crop A and crop B, borrowing value of MSE from last section (176)

$$t = \frac{80.2 - 75.8}{\sqrt{176 \left( \frac{1}{10} + \frac{1}{10} \right)}} = 0.742$$

To compute p-value from this t-score, the degrees of freedom = total experimental units – number of populations = 30 – 3 = 27

Using SPSS you will get p-value of 0.7678 at t-score = 0.742 and df = 27

Since the p-value is greater than test significance, we accept the null hypothesis and conclude there is no significant mean difference.

To run multiple comparisons test in SPSS of the crops example, go to **analyze > compare means > one way ANOVA** and select LSD from post-hoc button

## One – way Analysis of variance (one – way ANOVA)

The image shows two dialog boxes from SPSS. The top dialog is 'One-Way ANOVA' with 'crop\_height' in the 'Dependent List' and 'crop\_type' in the 'Factor' list. The 'Post Hoc...' button is circled in red. The bottom dialog is 'One-Way ANOVA: Post Hoc Multiple Comparisons' with 'LSD' checked under 'Equal Variances Assumed'. Other options like Bonferroni, Tukey, etc., are unchecked. The 'Significance level' is set to 0.05.

The generated result is shown below

### Multiple Comparisons

Dependent Variable: crop height

LSD

(I) crop type	(J) crop type	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
A	B	4.40000	5.93552	.465	-7.7787	16.5787
	C	-1.90000	5.93552	.751	-14.0787	10.2787
B	A	-4.40000	5.93552	.465	-16.5787	7.7787
	C	-6.30000	5.93552	.298	-18.4787	5.8787
C	A	1.90000	5.93552	.751	-10.2787	14.0787
	B	6.30000	5.93552	.298	-5.8787	18.4787

## One – way Analysis of variance (one – way ANOVA)

The p-value greater than the test significance indicates not to reject the null hypothesis and the pairwise means is the same

### Repeated measures one-way ANOVA

From the last section, the independent variables were different subjects (different crops). Sometime we are interested to perform the experiment on the same subjects repeated over and over again. This is the concept of repeated measures ANOVA which kind like an extension of paired samples t-test seen in chapter 9

We have seen that one-way ANOVA requires three assumptions to be valid

- Independence of the populations
- Normality of the populations
- Unknown but equal variances among the groups

In repeated measures ANOVA, the groups are related. Hence the independence assumption is not required. Repeated measures ANOVA is also called within-subjects ANOVA

Let us revisit the crop example again, but this time we only experiment crop-A sample over three times and test the hypothesis that the measured height of the three groups are not different from each other.

Crop A	Crop A	Crop A
72	79	75
69	58	72
70	59	87
73	70	82
74	81	73
76	59	79
96	80	88
99	90	71
90	73	80
83	76	77
Mean = 80.2	Mean = 75.8	Mean = 82.1
Variance =123.5	Variance = 227	Variance =177.9

## One – way Analysis of variance (one – way ANOVA)

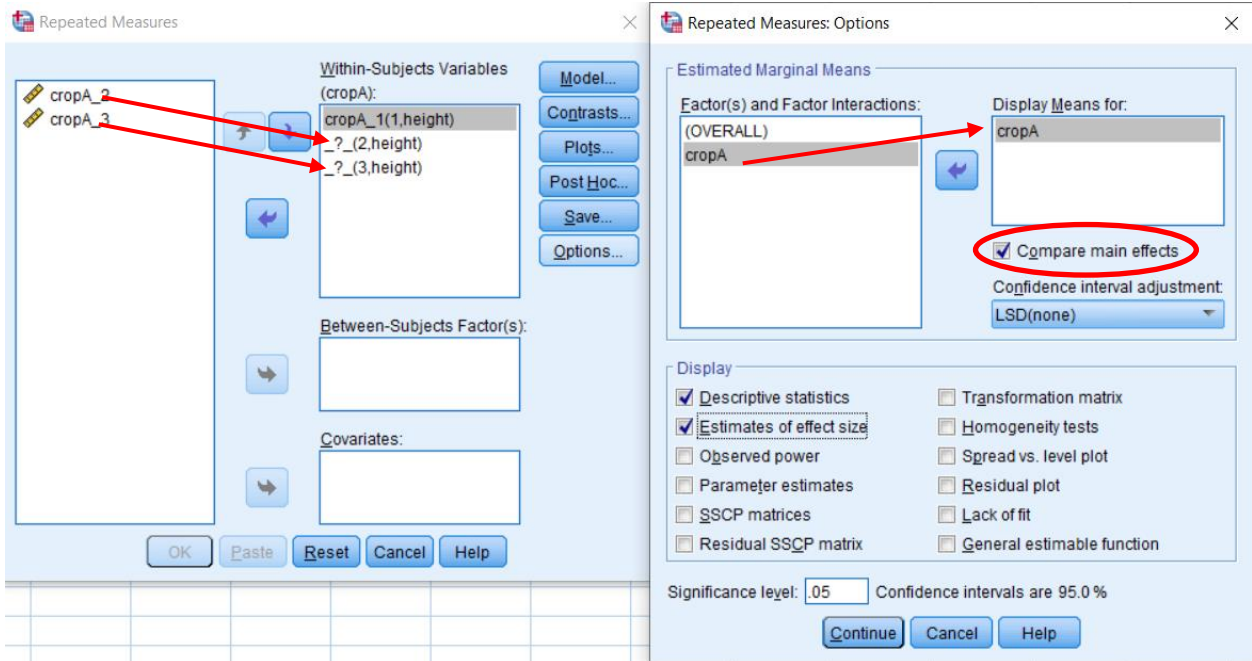
To perform repeated measures ANOVA in SPSS go to **analyze > General linear models > repeated measures**. Set the **within-subject factor** and **measure name** boxes as shown below

	cropA_1	cropA_2	cropA_3
1	72.00	79.00	75.00
2	69.00	58.00	72.00
3	70.00	59.00	87.00
4	73.00	70.00	82.00
5	74.00	81.00	73.00
6	76.00	59.00	79.00
7	96.00	80.00	88.00
8	99.00	90.00	71.00
9	90.00	73.00	80.00
10	83.00	76.00	77.00
11			
12			
13			
14			
15			
16			
17			
18			

The screenshot shows the 'Repeated Measures Define Factor(s)' dialog box in SPSS. The 'Within-Subject Factor Name' field is set to 'cropA', and the 'Number of Levels' is set to 3. The factor list contains 'cropA(3)'. The 'Measure Name' field is empty, and the measure list contains 'height'. Red circles highlight the 'cropA' field, the 'Number of Levels' field, and the 'height' field.

Click define button and set as shown below

## One – way Analysis of variance (one – way ANOVA)



Click continue and then ok to pop up output results windows

**Mauchly's Test of Sphericity<sup>a</sup>**

Measure: height

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon <sup>b</sup>		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
cropA	.784	1.952	2	.377	.822	.982	.500

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. Design: Intercept  
Within Subjects Design: cropA

b. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

The above table shows that sphericity assumption is valid

**Tests of Within-Subjects Effects**

Measure: height

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
cropA	Sphericity Assumed	324.467	2	162.233	2.305	.128	.204
	Greenhouse-Geisser	324.467	1.644	197.351	2.305	.141	.204
	Huynh-Feldt	324.467	1.963	165.275	2.305	.130	.204
	Lower-bound	324.467	1.000	324.467	2.305	.163	.204
Error(cropA)	Sphericity Assumed	1266.867	18	70.381			
	Greenhouse-Geisser	1266.867	14.797	85.617			
	Huynh-Feldt	1266.867	17.669	71.701			
	Lower-bound	1266.867	9.000	140.763			

## One – way Analysis of variance (one – way ANOVA)

The above table of tests within-subjects effects shows there is no significant mean difference within the groups as p-value (0.128) is greater than the test significance

If instead in the table above you found out the p-value was less the test significance and there was evidence for overall mean difference, you could look at pairwise comparisons below to further investigate where the differences lie

### Pairwise Comparisons

Measure: height

(I) cropA	(J) cropA	Mean Difference (I-J)	Std. Error	Sig. <sup>b</sup>	95% Confidence Interval for Difference <sup>b</sup>	
					Lower Bound	Upper Bound
1	2	7.700*	2.829	.024	1.301	14.099
	3	1.800	3.872	.653	-6.960	10.560
2	1	-7.700*	2.829	.024	-14.099	-1.301
	3	-5.900	4.385	.211	-15.821	4.021
3	1	-1.800	3.872	.653	-10.560	6.960
	2	5.900	4.385	.211	-4.021	15.821

Based on estimated marginal means

\*. The mean difference is significant at the .05 level.

b. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

### Review questions

1. A particular recruiting manager is much concerned with whether mean project completion time (in hours) differs among three candidates (A, B, C) so as to award one OF them full time contract. Each candidate was tested for 5 consecutive days and task completion time was recorded for each day

Candidate A	Candidate B	Candidate C
1	3	2
2	3	4
2	1	1
2	5	2
1	1	1

Perform one-way ANOVA test in SPSS at 5% if there is mean completion time difference between the candidates is statistically significant

2. A new training method was proposed by the ministry of health to train country medical staff. To evaluate the effectiveness of the method, 10 medical staff were



### One – way Analysis of variance (one – way ANOVA)

selected. The results of scores for pretest, posttest1, and posttest2 were recorded as follows

Pretest	posttest1	posttest2
67	65	83
87	50	71
64	72	83
59	77	59
66	72	70
70	53	57
70	53	72
76	83	73
75	87	90
63	52	54

Using SPSS, Perform repeated measures ANOVA at 5% level? What could you conclude?

## Non-parametric tests

### Chapter 12

#### Non-parametric tests

---

After completing this section you should be able to

- Differentiate between parametric and non-parametric test
- Appreciate the usefulness of not parametric test when the underlying assumption of sample distribution cannot be established
- Understand how to use Wilcoxon signed rank test as an alternative to paired samples t- test
- Understand how to use Mann-Whitney test as an alternative to independent samples t-test
- Understand how to use Kruskal-Wallis test an alternative to one-way ANOVA F test
- Understand how to use Spearman rank correlation coefficient as an alternative to Pearson correlation coefficient with added capability of ordinal data sets
- Understand how to use Freidman test as an extension to parametric paired samples t-test when analyzing repeated measures ANOVA test

## Non-parametric tests

Many statistical tests require variables to be normally distributed so that test statistic could be calculated using standardized distributions such normal and student's t – distribution. This type of tests are called parametric tests as the hypothesis is based on parameters of the hypothesized distribution

For example in chapter 11 we studied if the difference between means of three population is statistically significant. We assumed that three populations to be normally distributed and used F-distribution to calculate the test statistic. We also studied in chapter 9 cases of two samples related or unrelated and used the assumption of normality to use t-distribution

As a statistician times will exist when you will find out that variables under test don't follow normal distribution in which case you cannot rely on parametric tests for accurate result. In such case, you will use distribution – free tests called non-parametric test that do not require assumption of normality to compute the test statistic

In the table below we summarize non-parametric tests we will discuss in this chapter along with their parametric counterparts

Non – parametric test	Application	Equivalent Parametric test
Wilcoxon signed rank test	When the difference sample is not normally distributed and not skewed	Paired samples t-test
Mann-Whitney U test	When the assumption of normality is violated	Unpaired independent samples t-test
Spearman's rank correlation coefficient	When the relationship is monotonic and not linear	Pearson correlation coefficient
Kruskal-Wallis H test	When assumption of normality is violated	One-way ANOVA

### Kruskal – Wallis test

This is the equivalent non-parametric test of ANOVA using F-distribution as discussed in chapter 11. In Kruskal – Wallis test you rank sample elements of the combined group from lowest to highest and sum rank of each group separately. You will then use H statistic to calculate the p-value

The test statistic of Kruskal – Wallis test is given by the following equation

$$H = \frac{12}{n(n+1)} \sum \frac{R_g}{n_g} - 3(n+1)$$

Where n = total sample elements of combined groups

And  $R_g$  = sum of each group rank totaled separately

And  $n_g$  = sample elements of each group

## Non-parametric tests

To illustrate non-parametric application of H test, revisit crop example in chapter 11 about ANOVA. In that chapter we solved the problem using F – statistic assuming groups were normally distributed. We concluded that not to reject the null hypothesis and the three crop populations were growing to equal heights after being treated by same fertilizer. Now using H statistic we should show the same result of accepting the null hypothesis.

The table of crop heights is again shown below

Crop A	Crop B	Crop C
72	84	72
69	58	62
70	54	97
73	79	92
74	91	63
76	59	89
96	84	98
99	100	91
90	73	80
83	76	77
Mean = 80.2	Mean = 75.8	Mean = 82.1
Variance =123.5	Variance = 227	Variance =177.9

In chapter 11 ANOVA using F statistic we based hypothesis statements on population means. In the non-parametric test using H statistic we will base hypothesis on population medians rather than means provided independent populations have identical shape

We choose significance level for this test as  $\alpha = 0.05$  and state null hypothesis as below

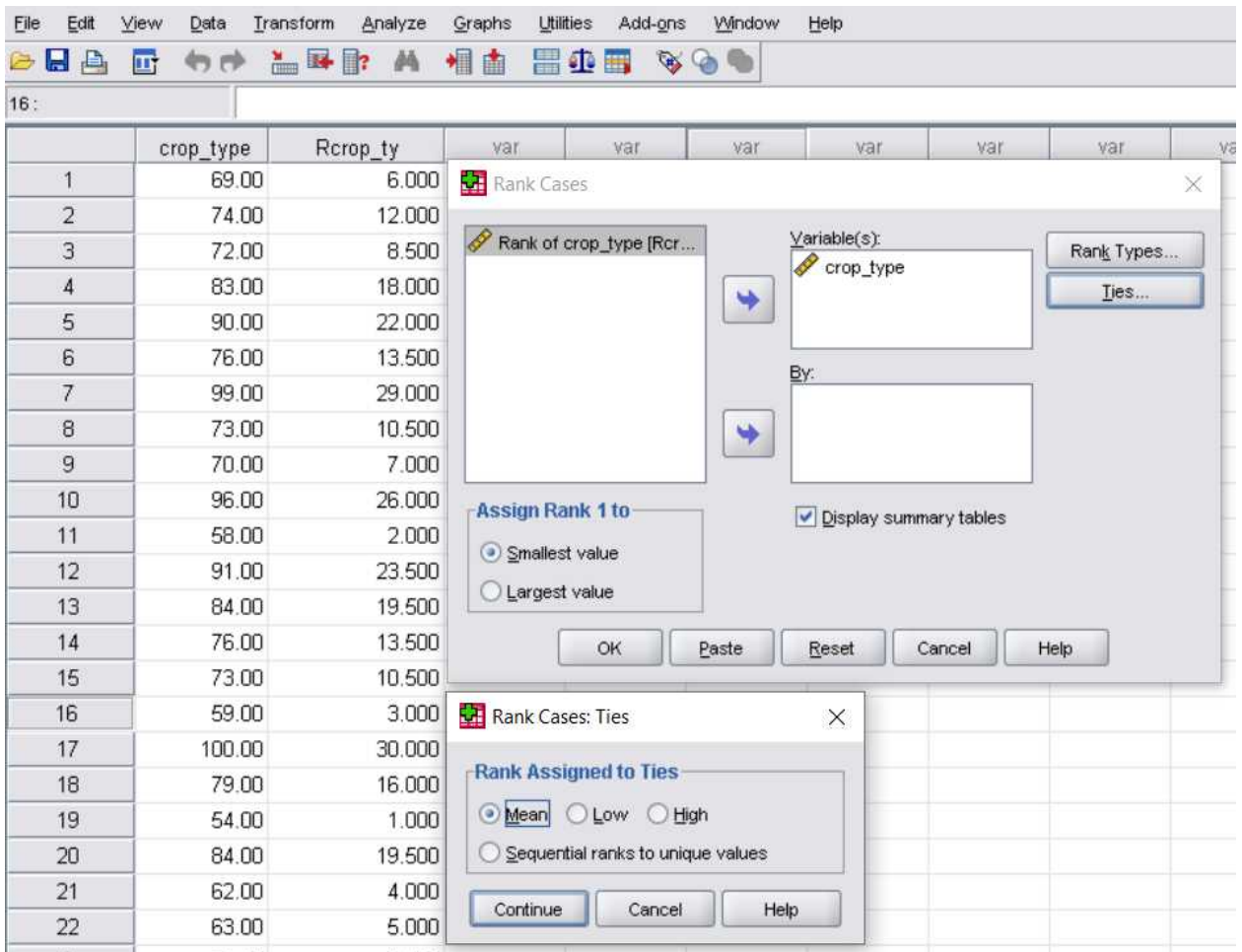
$$H_0: \text{the three crop populations are not significantly different}$$

We will reject the null hypothesis if the **test statistic** > **the critical value**

A good practice is to order the groups and then give rankings. In this case we have 30 sample elements. Hence we rank elements from 1 to 30

To do ranking in SPSS go to **transform** > **rank cases** then set everything as shown below. Click on the ties button and choose mean to break the ties

## Non-parametric tests



Shown below is our data table added with new columns of ranking

crop_A	Rcrop_A	crop_B	Rcrop_B	crop_C	Rcrop_C
69	6	58	2	62	4
74	12	91	23.5	63	5
72	8.5	84	19.5	72	8.5
83	18	76	13.5	77	15
90	22	73	10.5	80	17
76	13.5	59	3	89	21
99	29	100	30	91	23.5
73	10.5	79	16	92	25
70	7	54	1	97	27

## Non-parametric tests

96	26	84	19.5	98	28
	Total R = 152.5		Total R = 138.5		Total R = 174

Now let us calculate H using the above formula

$$H = \frac{12}{30(30 + 1)} \left[ \frac{(152.5)^2}{10} + \frac{(138.5)^2}{10} + \frac{(174)^2}{10} \right] - 3(30 + 1)$$

$$H = 0.8017 \quad \text{test statistic}$$

When the sample size of group is greater than 5 we can use Chi-square distribution to calculate the critical value. But chi-square needs two parameters which are degrees of freedom and significance level

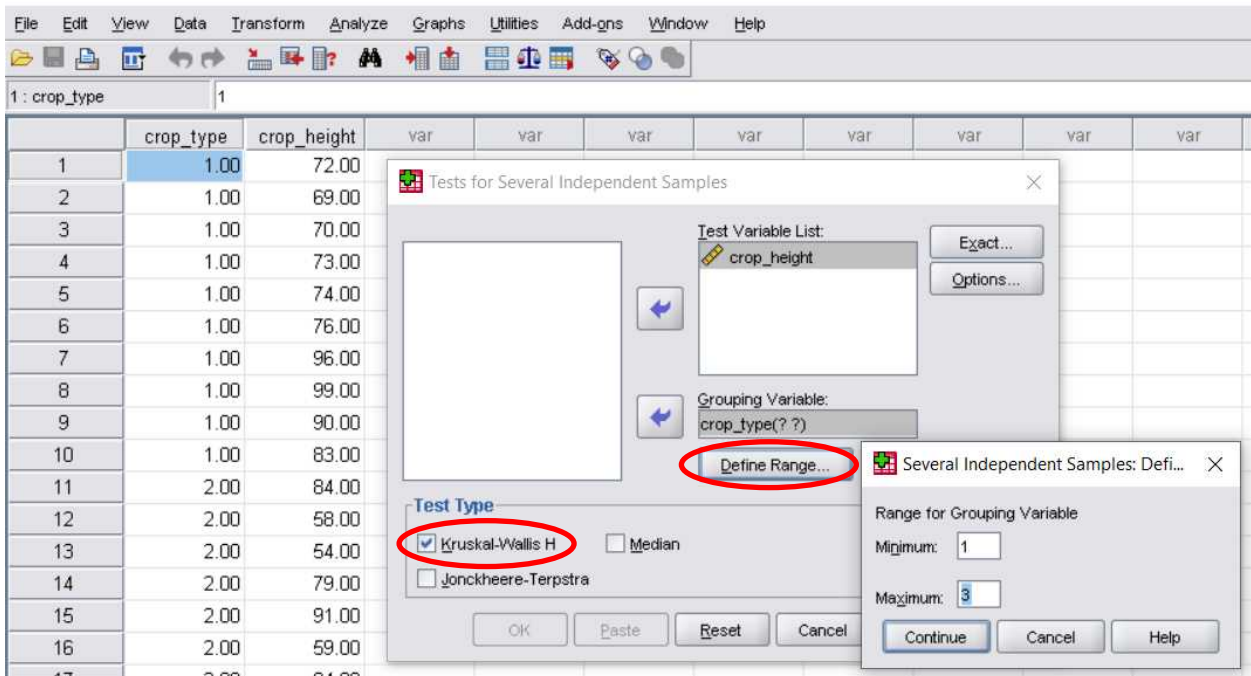
Degrees of freedom =  $k - 1$  when  $k$  is the number of groups (3 groups in our case)

Now find the critical value corresponding to degrees of freedom of 2 and  $\alpha = 0.05$  using chi-square table in appendix D. you should find value of 5.99

Since test statistic < critical value ( $0.8017 < 5.99$ ) we accept the null hypothesis and conclude that there is no significant mean difference among the three crop populations. This result matches with its corresponding parametric study using F-statistic ANOVA in chapter 11

Let us using SPSS do Kruskal-Wallis H test and interpret the result

Go to **Analyze > non parametric test > K independent samples**. When you click define range button as below set maximum to the maximum number of groups to be compared.



## Non-parametric tests

Once set as shown above click ok to generate the analysis output as below

### Kruskal-Wallis

	cro...	N	Mean Rank
crop_height	A	10	15.25
	B	10	13.85
	C	10	17.40
	Total	30	

	crop_height
Chi-Square	.826
df	2
Asymp. Sig.	.662

a. Kruskal Wallis Test

b. Grouping Variable: crop\_type

From the ranks table we see that there is not significant mean rank difference among the groups as they are nearly equal to each other.

From the test statistic table we see  $p\text{-value} > \alpha$  ( $0.662 > 0.05$ ), we therefore conclude not to reject the null hypothesis and that the three crop populations are identical

### Mann-Whitney U test

The non-parametric test is useful to test if two independent populations are identical or different. It does not require the assumptions of populations normality as seen in chapter 9

In this test we rank the two independent samples and find total rank of each sample. The test statistic U is then given by

$$U = \min(U_1, U_2)$$

Where

$$U_1 = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - \sum R_1$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2 + 1)}{2} - \sum R_2$$

The null hypothesis is stated as below using significance level of  $\alpha = 0.05$

$H_0$ : *there is not test score difference between male and female populations*

We will reject the null hypothesis if the **test statistic** < **the critical value**

## Non-parametric tests

Let us revisit independent unpaired samples test covered in chapter 9 which studied the null hypothesis that grade of female students were identical of grade of male students

Male	86	72	89	85	83	75	81	78	78	82
Female	64	69	52	80	56	66	57	50	53	63

Again we rank the two groups from lowest to highest and total rank R of each group as shown below

Male	R1	Female	R2
86	19	64	7
72	10	69	9
89	20	52	2
85	18	80	14
83	17	56	4
75	11	66	8
81	15	57	5
78	12.5	50	1
78	12.5	53	3
82	16	63	6
	$\sum R_1 = 151$		$\sum R_2 = 59$

$$U_1 = (10)(10) + \frac{10(10 + 1)}{2} - 151 = 4$$

$$U_2 = (10)(10) + \frac{10(10 + 1)}{2} - 59 = 96$$

We take the test statistic U as the smaller between the two values above

Hence  $U = 4$

Next step is to find the critical value and compare it with the test statistic U.

Using Mann-Whitney table you should get critical value of 23 for  $\alpha = 0.05$  and  $n = 10$  for both groups

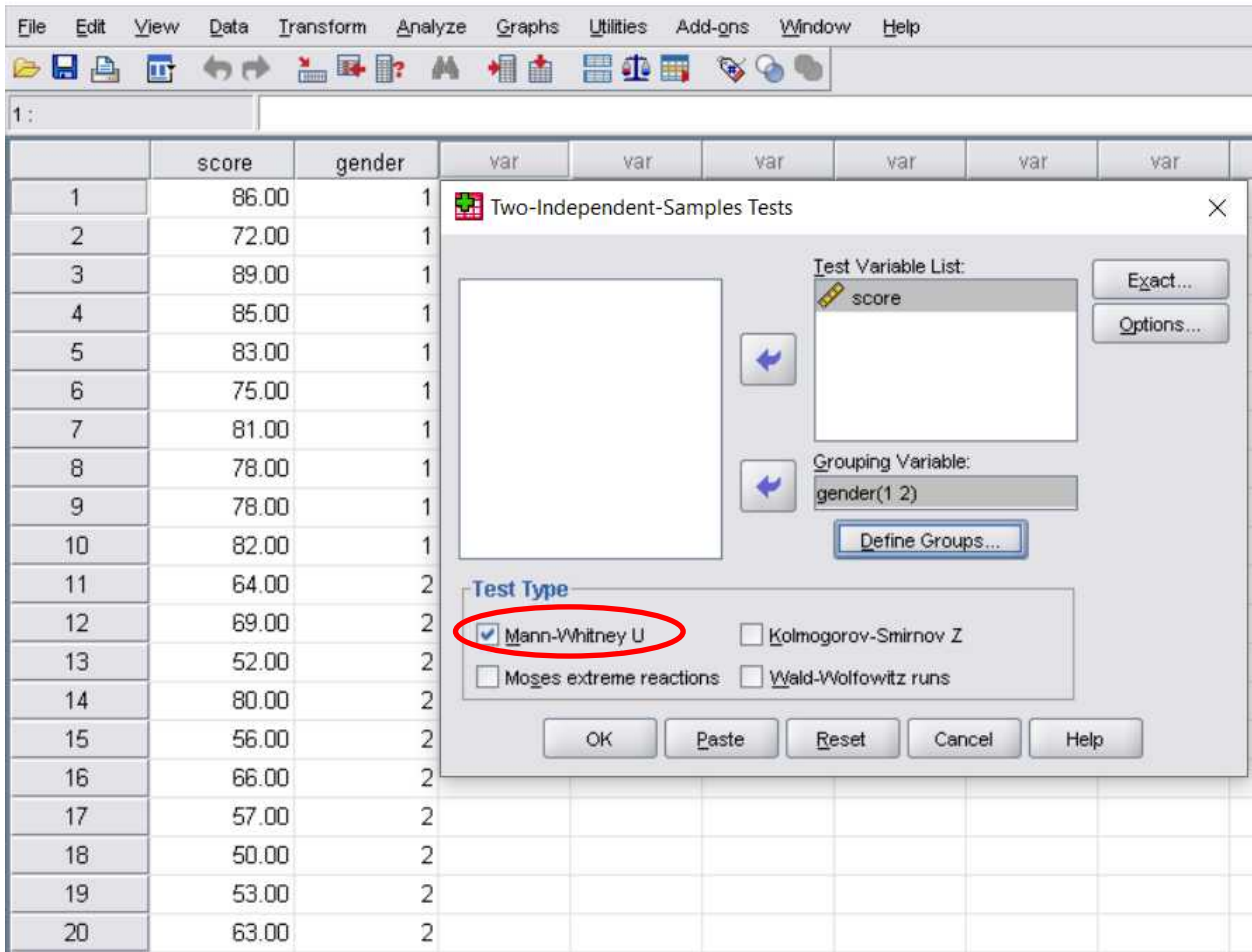


## Non-parametric tests

Now we see that test statistic < critical value, which gives evidence to reject the null hypothesis. We conclude that there is difference between test scores of male and female independent populations

Let us now run Mann-Whitney U test in SPSS. First step is always data entry with not error.

Then go to **analyze > non parametric test > 2-independent samples**



The screenshot shows the SPSS 'Two-Independent-Samples Tests' dialog box overlaid on a data editor window. The data editor contains a table with columns 'score' and 'gender'. The dialog box has the following settings:

- Test Variable List:** score
- Grouping Variable:** gender(1 2)
- Test Type:**  Mann-Whitney U,  Kolmogorov-Smirnov Z,  Moses extreme reactions,  Wald-Wolfowitz runs

The 'Mann-Whitney U' checkbox is circled in red. The dialog box also includes buttons for 'Exact...', 'Options...', 'Define Groups...', 'OK', 'Paste', 'Reset', 'Cancel', and 'Help'.

	score	gender
1	86.00	1
2	72.00	1
3	89.00	1
4	85.00	1
5	83.00	1
6	75.00	1
7	81.00	1
8	78.00	1
9	78.00	1
10	82.00	1
11	64.00	2
12	69.00	2
13	52.00	2
14	80.00	2
15	56.00	2
16	66.00	2
17	57.00	2
18	50.00	2
19	53.00	2
20	63.00	2

After you run the test, results will be generated in separate window as shown below

## Non-parametric tests

### Mann-Whitney

score	ge...	N	Mean Rank	Sum of Ranks
M		10	15.10	151.00
F		10	5.90	59.00
Total		20		

	score
Mann-Whitney U	4.000
Wilcoxon W	59.000
Z	-3.479
Asymp. Sig. (2-tailed)	.001
Exact Sig. [2*(1-tailed Sig.)]	.000 <sup>a</sup>

a. Not corrected for ties.

b. Grouping Variable: gender

From the top table named with ranks, you can infer significant mean rank difference between the two independent samples. From the test statistic table below you can see the test static  $U = 4$  matches with hand calculation. You can also see p-value much less than the significance level ( $0.001 < 0.05$ ) which gives strong evidence against the null hypothesis. We therefore conclude that there is significant difference between the two independent populations.

### Wilcoxon signed rank test

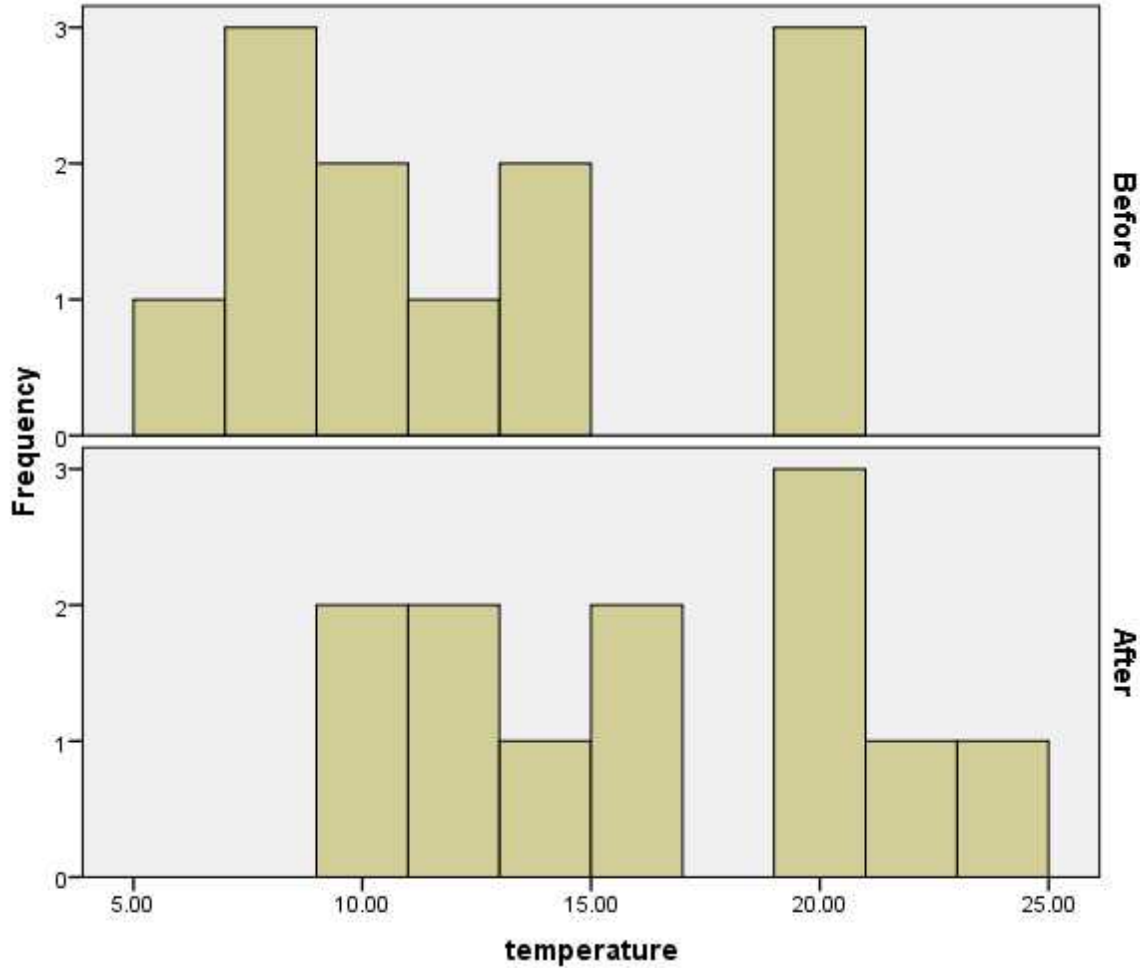
This test is parametric equivalent of paired dependent samples t-test. When using the t-test in chapter 9, the difference sample was assumed to be normally distributed. This assumption is not required in the non-parametric case. The Wilcoxon signed rank test requires the shape of the paired samples distributions to be identical in shape but not necessarily in location. The other difference is that we used mean difference in the parametric t-test to test the hypothesis in chapter 9. In the non-parametric  $W$  test we use median difference between the paired data to test the hypothesis

In Wilcoxon signed rank test, we first compute the absolute difference sample while ignoring zero difference cases. We then rank the difference sample and give sign to ranks based the difference sample sign

Let us recall the example in chapter of testing if there were mean difference between machine temperature before and after operation.

Let us plot the before and after machine temperature to explore how their histograms look like such that their shapes look identical but can be in different location

## Non-parametric tests



The shapes look nearly identical but different in their locations. So we can go ahead with Wilcoxon signed rank test procedure.

We state the null hypothesis as below using significance level of  $\alpha = 0.05$

$H_0$ : *there is no median difference between the paired samples*

We will reject the null hypothesis if the **p-value < significance level**

Before	After	Difference (d) = after - before	Absolute difference with zero omitted	Rank	Signed rank
20	20	0			
15	6	-9	9	7.5	-7.5
10	19	9	9	7.5	7.5
12	14	2	2	2.5	2.5
19	14	-5	5	5	-5

## Non-parametric tests

21	9	-12	12	9	-9
16	8	-8	8	6	-6
11	11	0			
9	7	-2	2	2.5	-2.5
20	19	-1	1	1	-1
23	8	-15	15	10	-15
13	9	-4	4	4	-4
					Sum of positive ranks = 10

When there is tie, the rank value is given by mean of their individual ranks

The test statistic is taken as the sum of positive rank value  $W = 10$

To find the critical value we will assume the sampling distribution of the signed ranks follow normal distribution as the difference sample size is greater than 10 which is our case for now.

To approximate to normal distribution we need to find the mean and standard deviation and then calculate the z-score

$$\mu = \frac{n(n+1)}{4} = \frac{10(10+1)}{4} = 27.5$$

$$\sigma = \sqrt{\frac{n(n+1)(2n+1)}{24}} = \sqrt{\frac{10(10+1)(20+1)}{24}} = 9.81$$

Using standard normal formula the p-value which is probability of getting greater than the test statistic is obtained by first getting the z-score

$$z = \frac{\bar{x} - \mu}{\sigma} = \frac{10 - 27.5}{9.81} = -1.78$$

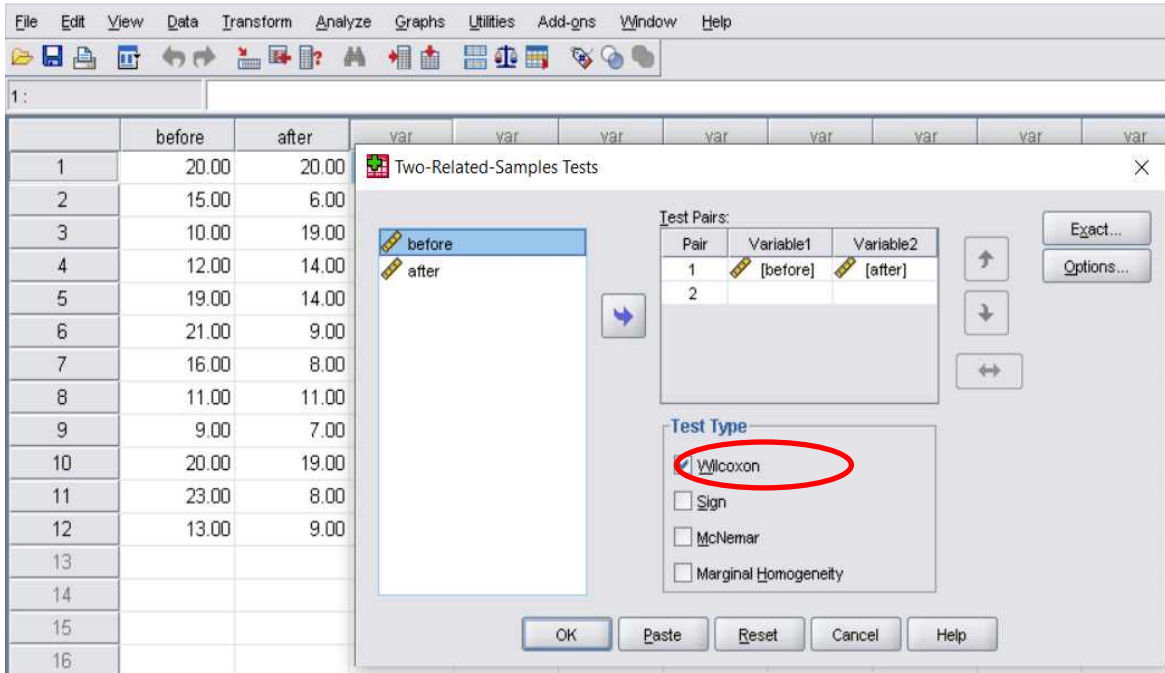
Now the p-value is  $P(z < -1.78) = 0.0375$

Since the test is two tailed, the p-value =  $2 \times 0.0375 = 0.075$

Since p-value is greater than the significance level ( $0.075 > 0.05$ ) we accept the null hypothesis and conclude that there is no statistically significant difference between before and after machine operation

.To run the analysis in SPSS go to **analyze > non parametric tests > k related samples**

## Non-parametric tests



After run the analysis, the following output result will be generated for interpretation

### Wilcoxon Signed Ranks

		Ranks		
		N	Mean Rank	Sum of Ranks
after - before	Negative Ranks	8 <sup>a</sup>	5.62	45.00
	Positive Ranks	2 <sup>b</sup>	5.00	10.00
	Ties	2 <sup>c</sup>		
	Total	12		

a. after < before

b. after > before

c. after = before

#### Test Statistics<sup>b</sup>

	after - before
Z	-1.786 <sup>a</sup>
Asymp. Sig. (2-tailed)	.074

a. Based on positive ranks.

b. Wilcoxon Signed Ranks Test

As can be seen from the test statistic table, the p-value (0.074) is greater than the significance level (0.05), hence we conclude not to reject the null hypothesis

## Non-parametric tests

### Spearman rank correlation coefficient

This is non parametric equivalent of the Pearson correlation coefficient covered in chapter 10 for quantitative data. In Spearman rank correlation we can test both quantitative and ordinal variables.

The spearman rank correlation coefficient is defined as

$$r_s = 1 - \frac{6 \sum d_i^2}{n^3 - n}$$

Where  $d_i$  is the difference between rank pairs of each row of the data. Those difference rank pairs are then squared and totaled. The value of n is the size of the sample and the data is ranked from highest to lowest (highest value given the first rank of 1)

The coefficient  $r_s$  has value between -1 and +1. When it approaches to +1 it indicates positive strong correlation, and when it approaches to -1 it indicates negative strong correlation. A value of 0 will indicates no correlation at all.

As an example suppose that we want to study correlation between reputation of 8 hospitals and number of patients checked in.

Reputation	4	7	2	9	5	1	3	8
# patients checked in	101	126	111	170	145	152	153	164

In this example we have two variables, reputation and number of patient checked-in. The variable reputation is ordinal. Let us denote reputation by y and patients checked in by x

The exercise is to rank the patient row from highest to lowest as the hospital reputation is already ranked. Then determine the correlation coefficient and significance of the correlation.

X	Rank of x	Y	Rank of y	$d_i$	$d_i^2$
4	4	101	8	-4	16
7	7	126	6	1	1
2	2	111	7	-5	25
8	8	170	1	7	49
5	5	145	5	0	0
1	1	152	4	-3	9
3	3	153	3	0	0
6	6	164	2	4	16
					$\sum d_i^2$ = 116

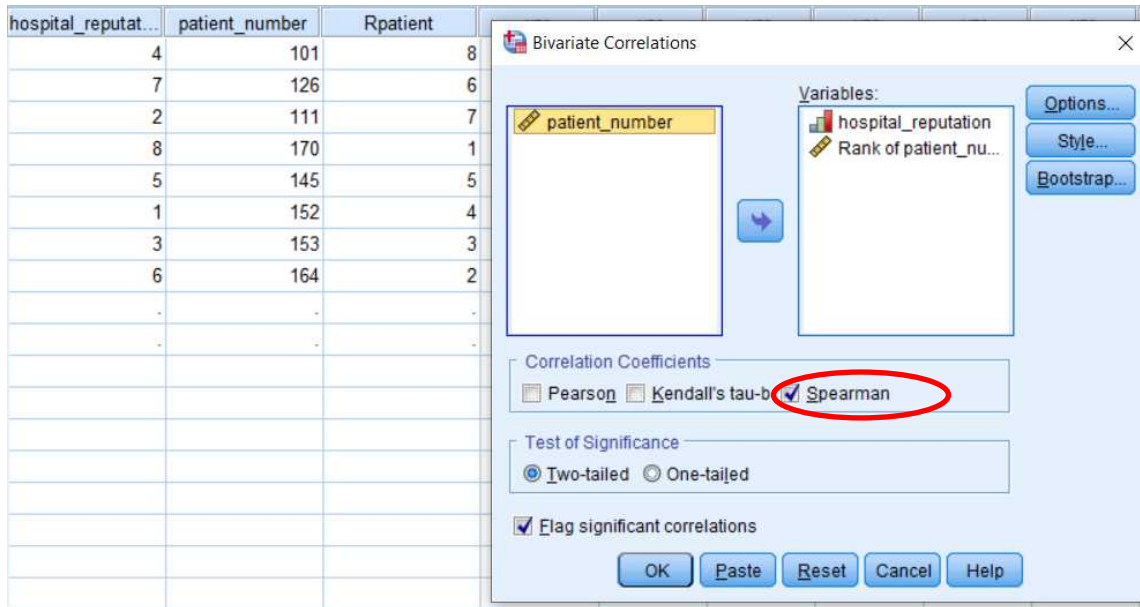
$$r_s = 1 - \frac{6(116)}{8^3 - 8} = -0.381$$

## Non-parametric tests

The spearman rank correlation coefficient shows that there weak negative relationship between hospital reputation and number of patients checked-in each hospital. This could be due to high reputation hospitals tending to be expensive.

But before we conclude let us test Spearman rank correlation in SPSS and also test if the relationship is significant at population level.

Go to **analyze > bivariate > correlate**



After you click ok, the following output result will be generated

Correlations				
			hospital_reputation	Rank of patient_number
Spearman's rho	hospital_reputation	Correlation Coefficient	1.000	<b>-.381</b>
		Sig. (2-tailed)	.	.352
		N	8	8
	Rank of patient_number	Correlation Coefficient	-.381	1.000
		Sig. (2-tailed)	.352	.
		N	8	8

To test the significance of this correlation, we test the null hypothesis that

$$H_0: \text{population coefficient is zero}$$

At  $\alpha = 0.05$ , we reject this null hypothesis if  $p\text{-value} < \alpha$

From the correlations table above p-value is 0.352 which is greater than the test significance (0.05). We therefore accept the null hypothesis and conclude that although there is correlation at the sample level, there is evidence that there is no significant correlation at the population level.

## Non-parametric tests

### Friedman test

This is the non-parametric alternative of repeated measures one-way ANOVA for  $k$  paired samples.

The Friedman tests the following hypothesis

$$H_0: k \text{ paired populations are identical}$$

The samples drawn from the  $k$  populations does not need to be normally distributed as was for repeated measures ANOVA. Friedman can test both continuous and ordinal data sets.

Let us revisit again the crop example treated over three times to see how Friedman result can match with that we saw in chapter 11 repeated measures ANOVA

To run Friedman test in SPSS, go to **analyze > non-parametric test > k related samples**

cropA_1	cropA_2	cropA_3
72.00	79.00	75.00
69.00	58.00	72.00
70.00	59.00	87.00
73.00	70.00	82.00
74.00	81.00	73.00
76.00	59.00	79.00
96.00	80.00	88.00
99.00	90.00	71.00
90.00	73.00	80.00
83.00	76.00	77.00

After you click ok, the following results window will appear for interpretation

N	10
Chi-Square	3.800
df	2
Asymp. Sig.	.150

a. Friedman Test

We can conclude from Friedman analysis, the  $p$ -value (0.150) is greater than the test significance level. Hence the null hypothesis is not rejected which indicates statistical evidence that the groups are identical.



## Non-parametric tests

### Review questions

1. After approval of new education curriculum, the ministry of education directed its research department to test the effectiveness of this new curriculum. The research department sampled 30 students and divided them into three independent groups each of 10 students. They recorded their pretest scores as well as posttest of the next two years.

Pretest score	Posttest score1	Posttest score2
72	84	92
69	68	82
54	74	97
73	79	92
74	91	93
76	79	89
76	84	98
59	86	96
90	73	80
83	76	87

Test at 5% if there difference between the groups is statistically significant

2. A telecom company is interested to study correlation between different internet connections and the number of subscribers that use each connection

Internet connection	Internet rate (Mbps)	Subscriber size (x1000)
EDGE	0.256	34
WIFI	1	65
DSL	2	23
3G	21	45
4G	100	80
5G	1Gbps	42

What is the suitable test for this study? Is there correlation at 5% level? What can you conclude about the significance of the correlation?

## Appendix A: questionnaire design for collecting sample data

### Appendix A: questionnaire design for collecting sample data

In observational research, you are taking proper samples from a population in order to arrive conclusion about that will be true for all the population. In other words you will be designing questionnaire to collect sample data

You will basically be writing “tick-box” sheet and distribute it to sample subjects in which they fill. You will need to pay attention to missing values in case a subject doesn’t provide complete answer. Another efficient way is to embed the sheet in your website and collect responses in web database

For example the research example in chapter 9 looked at TV programs using inferential statistics to research how the programs they air are preferred among different age categories of the population. They need to collect data about each viewer sampled such as name, age, gender, preferred program etc.

#### Gender:

Male

Female

#### Age:

Under 20

20 – 30

Above 30

#### Address:

Telephone \_\_\_\_\_

Email \_\_\_\_\_

#### Preferred program:

A

B

## Appendix B: SPSS data entry

### Appendix B: SPSS data entry

Once you collect sample data the next step is enter the data into SPSS. SPSS has two window views. Variable view and data view tabs. You first create variable names in variable view. Variable names for TV programs association with age categories would be **tv\_program** and **age\_group** with both data set to nominal in this case. If you are studying correlation between two quantitative variable you will set both variables to measure to scale as they both are continuous quantitative variables.

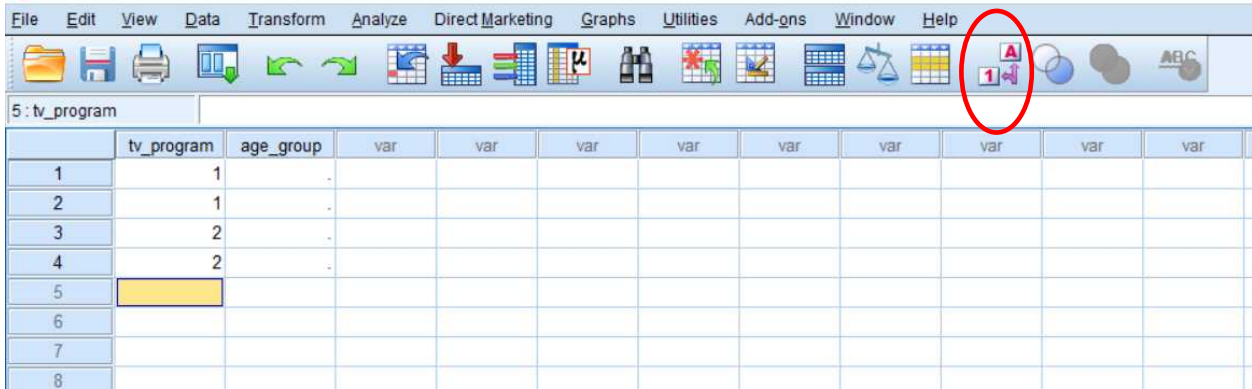
Since both variables in TV program preference in chapter 9 are nominal we use numbering labeling to avoid entering string

The screenshot shows the SPSS Variable View window. The variable list includes 'tv\_program' and 'age\_group', both set to 'Numeric' type with a width of 8 and 0 decimals. The 'Values' column for both is set to 'None'. A 'Value Labels' dialog box is open for 'tv\_program', showing 'Value: 2' and 'Label: B' entered, with '1 = "A"' listed in the list below. The 'Data View' tab is selected at the bottom.

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	tv_program	Numeric	8	0		None	None	8	Right	Nominal	Input
2	age_group	Numeric	8	0		None	None	8	Right	Nominal	Input
3											
4											
5											
6											
7											
8											
9											
10											
11											
12											
13											
14											
15											
16											
17											
18											
19											
20											
21											
22											
23											
24											
25											
26											
27											
28											
29											

In the data view window you will enter the coded label of the variable instead of the actual values. In this case we enter 1 for A and 2 for B

## Appendix B: SPSS data entry



Use the toggle button to view actual values A and B

### Multiple response questions

These type of question sets are called dichotomies as the respondents ticks all the options possible. As an example suppose that a bachelor graduating students is studying on ways of improving education quality in the country. He sets up his data collection questionnaire as follows

**How do you think education quality could be improved in the country? Please tick all boxes that apply**

Teaching training programs development

Salary rise for teachers

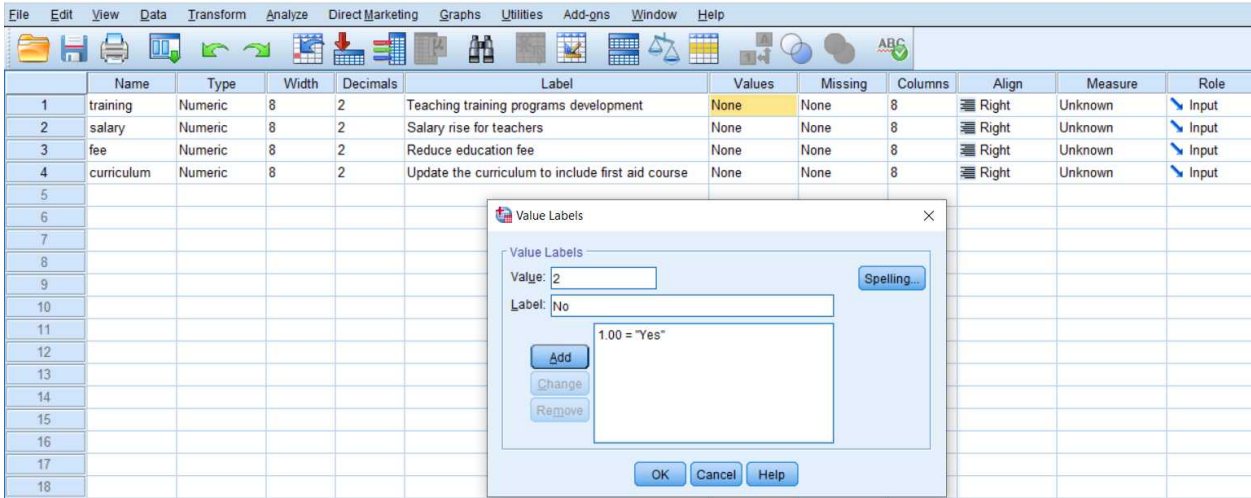
Reduce education fee

Update the curriculum to include first aid course

How would you do data entry for multiple response questions in SPSS?

The answer is you create variable for each option. In our case above, we will create four variables in data entry window. Then for each variable code 1 = Yes for respondent ticking the box and 2 = No for respondent not ticking it.

## Appendix B: SPSS data entry

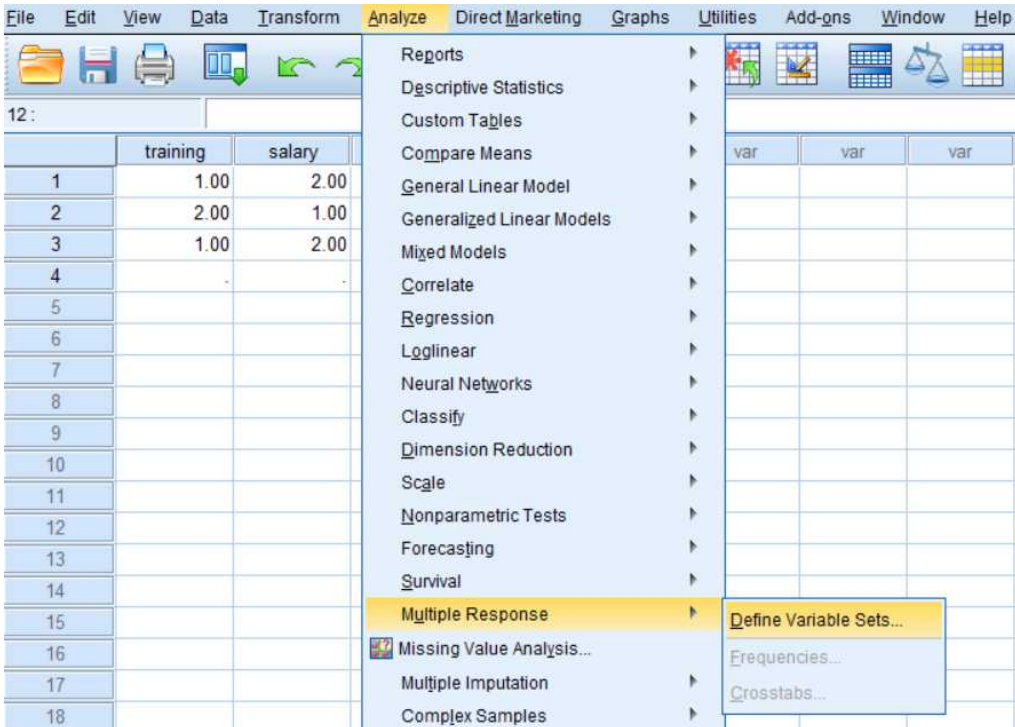


After coding as shown above, the data entry for the first three respondents is shown below

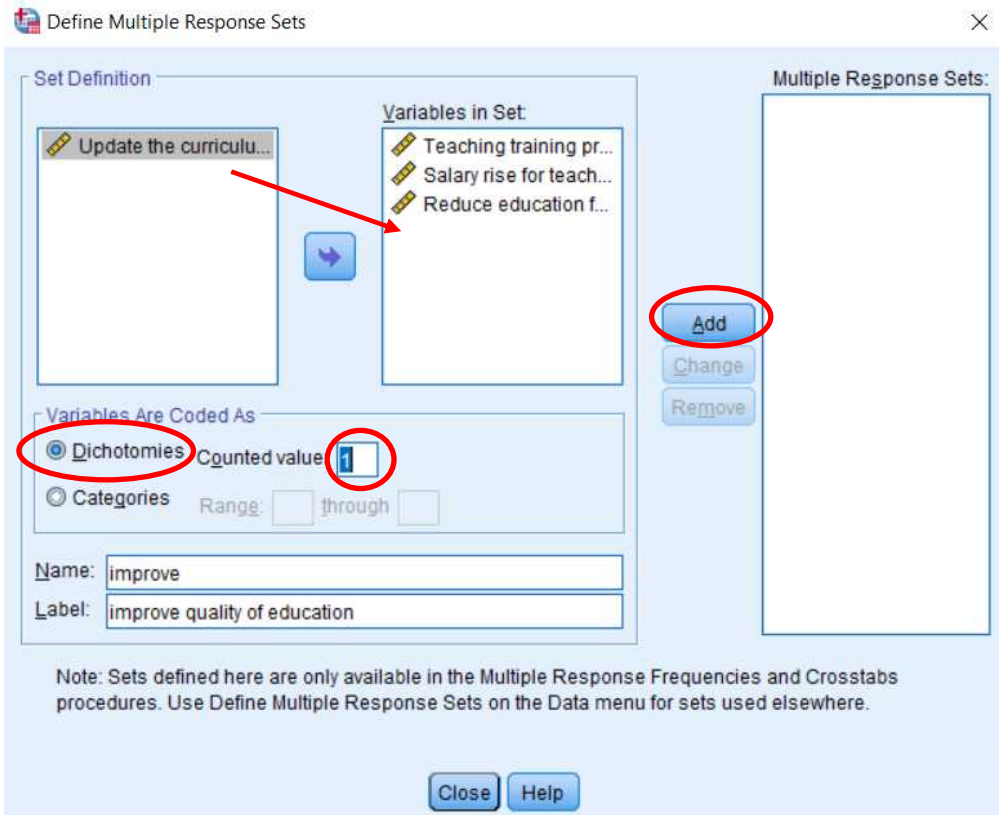
The screenshot shows the SPSS Data View dialog box. The 'training' variable has values 1.00, 2.00, and 1.00 for respondents 1, 2, and 3 respectively. The 'salary' variable has values 2.00, 1.00, and 2.00. The 'fee' variable has values 1.00, 1.00, and 1.00. The 'curriculum' variable has values 1.00, 2.00, and 2.00. There are also several empty 'var' columns.

	training	salary	fee	curriculum	var	var	var	var	var	var	var
1	1.00	2.00	1.00	1.00							
2	2.00	1.00	1.00	2.00							
3	1.00	2.00	1.00	2.00							

To analyze multiple response question in SPSS go to **analyze > multiple response > variable sets** as shown below



## Appendix B: SPSS data entry



Now go to **analyze > multiple response > frequencies** to analyze the multiple response statistics

**\$improve Frequencies**

		Responses		Percent of Cases
		N	Percent	
improve quality of education <sup>a</sup>	Teaching training programs development	2	28.6%	66.7%
	Salary rise for teachers	1	14.3%	33.3%
	Reduce education fee	3	42.9%	100.0%
	Update the curriculum to include first aid course	1	14.3%	33.3%
Total		7	100.0%	233.3%

a. Dichotomy group tabulated at value 1.

You can now interpret the results. For example the first row shows two people checked that teaching training programs will improve quality of education, while all respondents favored reduction of education fee.

## Appendix B: SPSS data entry

### Likert scale

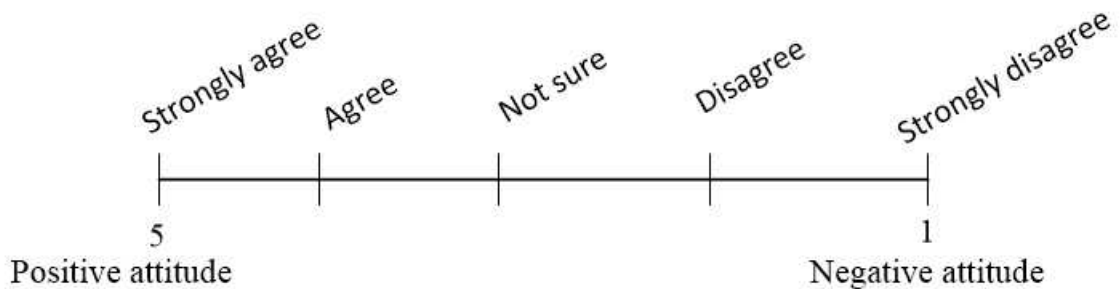
This type of questionnaire is useful in studying attitudes of respondents. For example suppose in certain data collection, people were asked whether they agree on new government taxation system or not. Using Likert attitude chart we can design a questionnaire to answer the following questions

**Do you agree on the government new taxation system?**

- |                   |                          |
|-------------------|--------------------------|
| Strongly agree    | <input type="checkbox"/> |
| agree             | <input type="checkbox"/> |
| Not sure          | <input type="checkbox"/> |
| Disagree          | <input type="checkbox"/> |
| Strongly disagree | <input type="checkbox"/> |

Strongly agree has the highest attitude toward the new taxation system while strongly disagree has the lowest attitude. We code 5 = strongly agree and 1 = strongly disagree in SPSS.

The Likert scale is represented as follows



## Appendix C: Guide to selecting appropriate statistical test

Name	Type	Width	Decimals	Label	Values
Private_vs_Public	Numeric	8	2		None

## Appendix C: Guide to selecting appropriate statistical test

Statistical study	Statistical tool	Chapter
Confidence interval for population mean	Standard normal distribution	6
	Student's t – distribution	6
Confidence interval for population standard deviation	Chi – square distribution	6
Confidence interval for population proportion	Standard normal distribution	6
Normality test	Kurtosis / skewness	8
	Q – Q plot	8
	Shapiro – Wilk for small sample	8
	Kolmogorov – Smirnov for large sample	8
One sample hypothesis test about population parameter	Student's t - distribution	7
Two samples paired / unpaired test	Student's t – distribution	9



### Appendix C: Guide to selecting appropriate statistical test

Two samples association test	Chi – square test	9
Linear regression goodness of fit test	Coefficient of determination $R^2$	10
Confidence interval for regression line	Student’s t – distribution	10
Correlation between two quantitative variables	Pearson coefficient r	10
Comparing means of more than two independent normally distributed samples	One – way ANOVA F – distribution	11
	Levene’s test for homogeneity	11
	Welch’s test for robustness	11
Multiple comparison of the groups differences	Fisher LSD test	11
Comparing means of more than two paired samples	Repeated measures one-way ANOVA (Sphericity and normality assumed)	11
Non parametric test two independent samples	Mann-Whitney U test	12
Non parametric test of two paired samples	Wilcoxon signed rank test	12
Non parametric test of more than two independent samples	Kurskal-Wallis H test	12
Non parametric test of more than two paired samples	Freidman test	12
Non parametric correlation test for ordinal data	Spearman rho and Kendall tau test	12

## Appendix D: Statistical tables

### Appendix D: Statistical tables

The following partial tables contain probability for distributions that you will use in your statistical analysis. Please note that the table's values are not complete but partial to help you answer review exercises in this guide.

Standard normal left tail probabilities  $P(Z \leq -z)$

	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.3	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985

### Appendix D: Statistical tables

-1.1	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

Right tailed P ( $z < z$ )

	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
3.3	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.2	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.1	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
2.9	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
2.8	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.7	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.6	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.5	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.4	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.3	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.2	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.1	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
1.9	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
1.8	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.7	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.6	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.5	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.4	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319

### Appendix D: Statistical tables

1.3	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.2	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.1	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
0.9	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
0.8	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.7	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.6	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.5	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.4	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.3	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.2	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.1	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753

Chi – square  $\chi^2$  distribution table right tailed

	P – values						
Degrees of freedom	0.01	0.025	0.05	0.10	0.90	0.95	0.99
1	6.63	5.02	3.84	2.71	0.02	0.00	0.00
2	9.21	7.38	5.99	4.61	0.21	0.10	0.02
3	11.34	9.35	7.81	6.25	0.58	0.35	0.11
4	13.28	11.14	9.49	7.78	1.06	0.71	0.30
5	15.09	12.83	11.07	9.24	1.61	1.15	0.55
6	16.81	14.45	12.59	10.64	2.20	1.64	0.87
7	18.48	16.01	14.07	12.02	2.83	2.17	1.24
8	20.09	17.53	15.51	13.36	3.49	2.73	1.65
9	21.67	19.02	16.92	14.68	4.17	3.33	2.09
10	23.21	20.48	18.31	15.99	4.87	3.94	2.56
11	24.72	21.92	19.68	17.28	5.58	4.57	3.05
12	26.22	23.34	21.03	18.55	6.30	5.23	3.57
13	27.69	24.74	22.36	19.81	7.04	5.89	4.11

### Appendix D: Statistical tables

14	29.14	26.12	23.68	21.06	7.79	6.57	4.66
15	30.58	27.49	25.00	22.31	8.55	7.26	5.23
16	32.00	28.85	26.30	23.54	9.31	7.96	5.81
17	33.41	30.19	27.59	24.77	10.09	8.67	6.41
18	34.81	31.53	28.87	25.99	10.86	9.39	7.01
19	36.19	32.85	30.14	27.20	11.65	10.12	7.63
20	37.57	34.17	31.41	28.41	12.44	10.85	8.26
21	38.93	35.48	32.67	29.62	13.24	11.59	8.90
22	40.29	36.78	33.92	30.81	14.04	12.34	9.54
23	41.64	38.08	35.17	32.01	14.85	13.09	10.20
24	42.98	39.36	36.42	33.20	15.66	13.85	10.86
25	44.31	40.65	37.65	34.38	16.47	14.61	11.52
26	45.64	41.92	38.89	35.56	17.29	15.38	12.20
27	46.96	43.19	40.11	36.74	18.11	16.15	12.88
28	48.28	44.46	41.34	37.92	18.94	16.93	13.56
29	49.59	45.72	42.56	39.09	19.77	17.71	14.26
30	50.89	46.98	43.77	40.26	20.60	18.49	14.95
31	52.19	48.23	44.99	41.42	21.43	19.28	15.66
32	53.49	49.48	46.19	42.58	22.27	20.07	16.36
33	54.78	50.73	47.40	43.75	23.11	20.87	17.07
34	56.06	51.97	48.60	44.90	23.95	21.66	17.79
35	57.34	53.20	49.80	46.06	24.80	22.47	18.51
36	58.62	54.44	51.00	47.21	25.64	23.27	19.23
37	59.89	55.67	52.19	48.36	26.49	24.07	19.96
38	61.16	56.90	53.38	49.51	27.34	24.88	20.69
39	62.43	58.12	54.57	50.66	28.20	25.70	21.43
40	63.69	59.34	55.76	51.81	29.05	26.51	22.16

### Appendix D: Statistical tables

41	64.95	60.56	56.94	52.95	29.91	27.33	22.91
42	66.21	61.78	58.12	54.09	30.77	28.14	23.65
43	67.46	62.99	59.30	55.23	31.63	28.96	24.40
44	68.71	64.20	60.48	56.37	32.49	29.79	25.15
45	69.96	65.41	61.66	57.51	33.35	30.61	25.90
46	71.20	66.62	62.83	58.64	34.22	31.44	26.66
47	72.44	67.82	64.00	59.77	35.08	32.27	27.42
48	73.68	69.02	65.17	60.91	35.95	33.10	28.18
49	74.92	70.22	66.34	62.04	36.82	33.93	28.94
50	76.15	71.42	67.50	63.17	37.69	34.76	29.71
51	77.39	72.62	68.67	64.30	38.56	35.60	30.48
52	78.62	73.81	69.83	65.42	39.43	36.44	31.25
53	79.84	75.00	70.99	66.55	40.31	37.28	32.02
54	81.07	76.19	72.15	67.67	41.18	38.12	32.79
55	82.29	77.38	73.31	68.80	42.06	38.96	33.57
56	83.51	78.57	74.47	69.92	42.94	39.80	34.35
57	84.73	79.75	75.62	71.04	43.82	40.65	35.13
58	85.95	80.94	76.78	72.16	44.70	41.49	35.91
59	87.17	82.12	77.93	73.28	45.58	42.34	36.70
60	88.38	83.30	79.08	74.40	46.46	43.19	37.48
61	89.59	84.48	80.23	75.51	47.34	44.04	38.27
62	90.80	85.65	81.38	76.63	48.23	44.89	39.06
63	92.01	86.83	82.53	77.75	49.11	45.74	39.86
64	93.22	88.00	83.68	78.86	50.00	46.59	40.65
65	94.42	89.18	84.82	79.97	50.88	47.45	41.44
66	95.63	90.35	85.96	81.09	51.77	48.31	42.24
67	96.83	91.52	87.11	82.20	52.66	49.16	43.04

### Appendix D: Statistical tables

68	98.03	92.69	88.25	83.31	53.55	50.02	43.84
69	99.23	93.86	89.39	84.42	54.44	50.88	44.64
70	100.43	95.02	90.53	85.53	55.33	51.74	45.44
71	101.62	96.19	91.67	86.64	56.22	52.60	46.25
72	102.82	97.35	92.81	87.74	57.11	53.46	47.05
73	104.01	98.52	93.95	88.85	58.01	54.33	47.86
74	105.20	99.68	95.08	89.96	58.90	55.19	48.67
75	106.39	100.84	96.22	91.06	59.79	56.05	49.48
76	107.58	102.00	97.35	92.17	60.69	56.92	50.29
77	108.77	103.16	98.48	93.27	61.59	57.79	51.10
78	109.96	104.32	99.62	94.37	62.48	58.65	51.91
79	111.14	105.47	100.75	95.48	63.38	59.52	52.72
80	112.33	106.63	101.88	96.58	64.28	60.39	53.54
81	113.51	107.78	103.01	97.68	65.18	61.26	54.36
82	114.69	108.94	104.14	98.78	66.08	62.13	55.17
83	115.88	110.09	105.27	99.88	66.98	63.00	55.99
84	117.06	111.24	106.39	100.98	67.88	63.88	56.81
85	118.24	112.39	107.52	102.08	68.78	64.75	57.63
86	119.41	113.54	108.65	103.18	69.68	65.62	58.46
87	120.59	114.69	109.77	104.28	70.58	66.50	59.28
88	121.77	115.84	110.90	105.37	71.48	67.37	60.10
89	122.94	116.99	112.02	106.47	72.39	68.25	60.93
90	124.12	118.14	113.15	107.57	73.29	69.13	61.75
91	125.29	119.28	114.27	108.66	74.20	70.00	62.58
92	126.46	120.43	115.39	109.76	75.10	70.88	63.41
93	127.63	121.57	116.51	110.85	76.01	71.76	64.24
94	128.80	122.72	117.63	111.94	76.91	72.64	65.07

### Appendix D: Statistical tables

<b>95</b>	129.97	123.86	118.75	113.04	77.82	73.52	65.90
<b>96</b>	131.14	125.00	119.87	114.13	78.73	74.40	66.73
<b>97</b>	132.31	126.14	120.99	115.22	79.63	75.28	67.56
<b>98</b>	133.48	127.28	122.11	116.32	80.54	76.16	68.40
<b>99</b>	134.64	128.42	123.23	117.41	81.45	77.05	69.23
<b>100</b>	135.80	129.56	124.34	118.50	82.36	77.92	70.06

Student's t – distribution left tailed

df	P - values		
	<b>0.01</b>	<b>0.025</b>	<b>0.05</b>
<b>1</b>	-31.821	-12.706	-6.314
<b>2</b>	-6.965	-4.303	-2.920
<b>3</b>	-4.541	-3.182	-2.353
<b>4</b>	-3.747	-2.776	-2.132
<b>5</b>	-3.365	-2.571	-2.015
<b>6</b>	-3.143	-2.447	-1.943
<b>7</b>	-2.998	-2.365	-1.895
<b>8</b>	-2.896	-2.306	-1.860
<b>9</b>	-2.821	-2.262	-1.833
<b>10</b>	-2.764	-2.228	-1.812
<b>11</b>	-2.718	-2.201	-1.796
<b>12</b>	-2.681	-2.179	-1.782
<b>13</b>	-2.650	-2.160	-1.771
<b>14</b>	-2.624	-2.145	-1.761
<b>15</b>	-2.602	-2.131	-1.753
<b>16</b>	-2.583	-2.120	-1.746
<b>17</b>	-2.567	-2.110	-1.740



### Appendix D: Statistical tables

18	-2.552	-2.101	-1.734
19	-2.539	-2.093	-1.729
20	-2.528	-2.086	-1.725
21	-2.518	-2.080	-1.721
22	-2.508	-2.074	-1.717
23	-2.500	-2.069	-1.714
24	-2.492	-2.064	-1.711
25	-2.485	-2.060	-1.708
26	-2.479	-2.056	-1.706
27	-2.473	-2.052	-1.703
28	-2.467	-2.048	-1.701
29	-2.462	-2.045	-1.699
30	-2.46	-2.04	-1.70

F-distribution probabilities right tailed

When using this table the F-score is given by **inverse F function** in SPSS and you in the form of F (probability, numerator degrees of freedom, and denominator degrees of freedom). For example  $INV.F(0.05, 3, 1) = 215.71$

		Numerator degrees of freedom									
Right tail probability	Denominator df	1	2	3	4	5	6	7	8	9	10
0.05	1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88
0.05	2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
0.05	3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
0.05	4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
0.05	5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74

### Appendix D: Statistical tables

0.05	6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
0.05	7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
0.05	8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
0.05	9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
0.05	10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
0.05	11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
0.05	12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
0.05	13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
0.05	14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
0.05	15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
0.05	16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
0.05	17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
0.05	18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
0.05	19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
0.05	20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
0.05	21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32
0.05	22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
0.05	23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27
0.05	24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25
0.05	25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24
0.05	26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
0.05	27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20
0.05	28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19
0.05	29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18
0.05	30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16
0.05	31	4.16	3.30	2.91	2.68	2.52	2.41	2.32	2.25	2.20	2.15
0.05	32	4.15	3.29	2.90	2.67	2.51	2.40	2.31	2.24	2.19	2.14
0.05	33	4.14	3.28	2.89	2.66	2.50	2.39	2.30	2.23	2.18	2.13

### Appendix D: Statistical tables

0.05	34	4.13	3.28	2.88	2.65	2.49	2.38	2.29	2.23	2.17	2.12
0.05	35	4.12	3.27	2.87	2.64	2.49	2.37	2.29	2.22	2.16	2.11
0.05	36	4.11	3.26	2.87	2.63	2.48	2.36	2.28	2.21	2.15	2.11
0.05	37	4.11	3.25	2.86	2.63	2.47	2.36	2.27	2.20	2.14	2.10
0.05	38	4.10	3.24	2.85	2.62	2.46	2.35	2.26	2.19	2.14	2.09
0.05	39	4.09	3.24	2.85	2.61	2.46	2.34	2.26	2.19	2.13	2.08
0.05	40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08
0.05	41	4.08	3.23	2.83	2.60	2.44	2.33	2.24	2.17	2.12	2.07
0.05	42	4.07	3.22	2.83	2.59	2.44	2.32	2.24	2.17	2.11	2.06
0.05	43	4.07	3.21	2.82	2.59	2.43	2.32	2.23	2.16	2.11	2.06
0.05	44	4.06	3.21	2.82	2.58	2.43	2.31	2.23	2.16	2.10	2.05
0.05	45	4.06	3.20	2.81	2.58	2.42	2.31	2.22	2.15	2.10	2.05
0.05	46	4.05	3.20	2.81	2.57	2.42	2.30	2.22	2.15	2.09	2.04
0.05	47	4.05	3.20	2.80	2.57	2.41	2.30	2.21	2.14	2.09	2.04
0.05	48	4.04	3.19	2.80	2.57	2.41	2.29	2.21	2.14	2.08	2.03
0.05	49	4.04	3.19	2.79	2.56	2.40	2.29	2.20	2.13	2.08	2.03
0.05	50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03
0.05	51	4.03	3.18	2.79	2.55	2.40	2.28	2.20	2.13	2.07	2.02
0.05	52	4.03	3.18	2.78	2.55	2.39	2.28	2.19	2.12	2.07	2.02
0.05	53	4.02	3.17	2.78	2.55	2.39	2.28	2.19	2.12	2.06	2.01
0.05	54	4.02	3.17	2.78	2.54	2.39	2.27	2.18	2.12	2.06	2.01
0.05	55	4.02	3.16	2.77	2.54	2.38	2.27	2.18	2.11	2.06	2.01
0.05	56	4.01	3.16	2.77	2.54	2.38	2.27	2.18	2.11	2.05	2.00
0.05	57	4.01	3.16	2.77	2.53	2.38	2.26	2.18	2.11	2.05	2.00
0.05	58	4.01	3.16	2.76	2.53	2.37	2.26	2.17	2.10	2.05	2.00
0.05	59	4.00	3.15	2.76	2.53	2.37	2.26	2.17	2.10	2.04	2.00
0.05	60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99
0.05	61	4.00	3.15	2.76	2.52	2.37	2.25	2.16	2.09	2.04	1.99

**Appendix D: Statistical tables**

<b>0.05</b>	62	4.00	3.15	2.75	2.52	2.36	2.25	2.16	2.09	2.03	1.99
<b>0.05</b>	63	3.99	3.14	2.75	2.52	2.36	2.25	2.16	2.09	2.03	1.98
<b>0.05</b>	64	3.99	3.14	2.75	2.52	2.36	2.24	2.16	2.09	2.03	1.98
<b>0.05</b>	65	3.99	3.14	2.75	2.51	2.36	2.24	2.15	2.08	2.03	1.98
<b>0.05</b>	66	3.99	3.14	2.74	2.51	2.35	2.24	2.15	2.08	2.03	1.98
<b>0.05</b>	67	3.98	3.13	2.74	2.51	2.35	2.24	2.15	2.08	2.02	1.98
<b>0.05</b>	68	3.98	3.13	2.74	2.51	2.35	2.24	2.15	2.08	2.02	1.97
<b>0.05</b>	69	3.98	3.13	2.74	2.50	2.35	2.23	2.15	2.08	2.02	1.97
<b>0.05</b>	70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.02	1.97
<b>0.05</b>	71	3.98	3.13	2.73	2.50	2.34	2.23	2.14	2.07	2.01	1.97
<b>0.05</b>	72	3.97	3.12	2.73	2.50	2.34	2.23	2.14	2.07	2.01	1.96
<b>0.05</b>	73	3.97	3.12	2.73	2.50	2.34	2.23	2.14	2.07	2.01	1.96
<b>0.05</b>	74	3.97	3.12	2.73	2.50	2.34	2.22	2.14	2.07	2.01	1.96
<b>0.05</b>	75	3.97	3.12	2.73	2.49	2.34	2.22	2.13	2.06	2.01	1.96
<b>0.05</b>	76	3.97	3.12	2.72	2.49	2.33	2.22	2.13	2.06	2.01	1.96
<b>0.05</b>	77	3.97	3.12	2.72	2.49	2.33	2.22	2.13	2.06	2.00	1.96
<b>0.05</b>	78	3.96	3.11	2.72	2.49	2.33	2.22	2.13	2.06	2.00	1.95
<b>0.05</b>	79	3.96	3.11	2.72	2.49	2.33	2.22	2.13	2.06	2.00	1.95
<b>0.05</b>	80	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	2.00	1.95
<b>0.05</b>	81	3.96	3.11	2.72	2.48	2.33	2.21	2.12	2.05	2.00	1.95
<b>0.05</b>	82	3.96	3.11	2.72	2.48	2.33	2.21	2.12	2.05	2.00	1.95
<b>0.05</b>	83	3.96	3.11	2.71	2.48	2.32	2.21	2.12	2.05	1.99	1.95
<b>0.05</b>	84	3.95	3.11	2.71	2.48	2.32	2.21	2.12	2.05	1.99	1.95
<b>0.05</b>	85	3.95	3.10	2.71	2.48	2.32	2.21	2.12	2.05	1.99	1.94
<b>0.05</b>	86	3.95	3.10	2.71	2.48	2.32	2.21	2.12	2.05	1.99	1.94
<b>0.05</b>	87	3.95	3.10	2.71	2.48	2.32	2.20	2.12	2.05	1.99	1.94
<b>0.05</b>	88	3.95	3.10	2.71	2.48	2.32	2.20	2.12	2.05	1.99	1.94
<b>0.05</b>	89	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94

### Appendix D: Statistical tables

<b>0.05</b>	90	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94
<b>0.05</b>	91	3.95	3.10	2.70	2.47	2.31	2.20	2.11	2.04	1.98	1.94
<b>0.05</b>	92	3.94	3.10	2.70	2.47	2.31	2.20	2.11	2.04	1.98	1.94
<b>0.05</b>	93	3.94	3.09	2.70	2.47	2.31	2.20	2.11	2.04	1.98	1.93
<b>0.05</b>	94	3.94	3.09	2.70	2.47	2.31	2.20	2.11	2.04	1.98	1.93
<b>0.05</b>	95	3.94	3.09	2.70	2.47	2.31	2.20	2.11	2.04	1.98	1.93
<b>0.05</b>	96	3.94	3.09	2.70	2.47	2.31	2.19	2.11	2.04	1.98	1.93
<b>0.05</b>	97	3.94	3.09	2.70	2.47	2.31	2.19	2.11	2.04	1.98	1.93
<b>0.05</b>	98	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.98	1.93
<b>0.05</b>	99	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.98	1.93
<b>0.05</b>	100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93
<b>0.01</b>	1	4052. 18	499 9.50	540 3.35	562 4.58	576 3.65	585 8.99	592 8.36	598 1.07	602 2.47	605 5.85
<b>0.01</b>	2	98.50	99.0 0	99.1 7	99.2 5	99.3 0	99.3 3	99.3 6	99.3 7	99.3 9	99.4 0
<b>0.01</b>	3	34.12	30.8 2	29.4 6	28.7 1	28.2 4	27.9 1	27.6 7	27.4 9	27.3 5	27.2 3
<b>0.01</b>	4	21.20	18.0 0	16.6 9	15.9 8	15.5 2	15.2 1	14.9 8	14.8 0	14.6 6	14.5 5
<b>0.01</b>	5	16.26	13.2 7	12.0 6	11.3 9	10.9 7	10.6 7	10.4 6	10.2 9	10.1 6	10.0 5
<b>0.01</b>	6	13.75	10.9 2	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87
<b>0.01</b>	7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62
<b>0.01</b>	8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81
<b>0.01</b>	9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26
<b>0.01</b>	10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85
<b>0.01</b>	11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54
<b>0.01</b>	12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30
<b>0.01</b>	13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10
<b>0.01</b>	14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94

### Appendix D: Statistical tables

0.01	15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80
0.01	16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69
0.01	17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59
0.01	18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51
0.01	19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43
0.01	20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37
0.01	21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31
0.01	22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26
0.01	23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21
0.01	24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17
0.01	25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13
0.01	26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09
0.01	27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06
0.01	28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03
0.01	29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00
0.01	30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98
0.01	31	7.53	5.36	4.48	3.99	3.67	3.45	3.28	3.15	3.04	2.96
0.01	32	7.50	5.34	4.46	3.97	3.65	3.43	3.26	3.13	3.02	2.93
0.01	33	7.47	5.31	4.44	3.95	3.63	3.41	3.24	3.11	3.00	2.91
0.01	34	7.44	5.29	4.42	3.93	3.61	3.39	3.22	3.09	2.98	2.89
0.01	35	7.42	5.27	4.40	3.91	3.59	3.37	3.20	3.07	2.96	2.88
0.01	36	7.40	5.25	4.38	3.89	3.57	3.35	3.18	3.05	2.95	2.86
0.01	37	7.37	5.23	4.36	3.87	3.56	3.33	3.17	3.04	2.93	2.84
0.01	38	7.35	5.21	4.34	3.86	3.54	3.32	3.15	3.02	2.92	2.83
0.01	39	7.33	5.19	4.33	3.84	3.53	3.30	3.14	3.01	2.90	2.81
0.01	40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80
0.01	41	7.30	5.16	4.30	3.81	3.50	3.28	3.11	2.98	2.87	2.79
0.01	42	7.28	5.15	4.29	3.80	3.49	3.27	3.10	2.97	2.86	2.78
0.01	43	7.26	5.14	4.27	3.79	3.48	3.25	3.09	2.96	2.85	2.76

### Appendix D: Statistical tables

0.01	44	7.25	5.12	4.26	3.78	3.47	3.24	3.08	2.95	2.84	2.75
0.01	45	7.23	5.11	4.25	3.77	3.45	3.23	3.07	2.94	2.83	2.74
0.01	46	7.22	5.10	4.24	3.76	3.44	3.22	3.06	2.93	2.82	2.73
0.01	47	7.21	5.09	4.23	3.75	3.43	3.21	3.05	2.92	2.81	2.72
0.01	48	7.19	5.08	4.22	3.74	3.43	3.20	3.04	2.91	2.80	2.71
0.01	49	7.18	5.07	4.21	3.73	3.42	3.19	3.03	2.90	2.79	2.71
0.01	50	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78	2.70
0.01	51	7.16	5.05	4.19	3.71	3.40	3.18	3.01	2.88	2.78	2.69
0.01	52	7.15	5.04	4.18	3.70	3.39	3.17	3.00	2.87	2.77	2.68
0.01	53	7.14	5.03	4.17	3.70	3.38	3.16	3.00	2.87	2.76	2.68
0.01	54	7.13	5.02	4.17	3.69	3.38	3.16	2.99	2.86	2.76	2.67
0.01	55	7.12	5.01	4.16	3.68	3.37	3.15	2.98	2.85	2.75	2.66
0.01	56	7.11	5.01	4.15	3.67	3.36	3.14	2.98	2.85	2.74	2.66
0.01	57	7.10	5.00	4.15	3.67	3.36	3.14	2.97	2.84	2.74	2.65
0.01	58	7.09	4.99	4.14	3.66	3.35	3.13	2.96	2.83	2.73	2.64
0.01	59	7.08	4.98	4.13	3.65	3.34	3.12	2.96	2.83	2.72	2.64
0.01	60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63
0.01	61	7.07	4.97	4.12	3.64	3.33	3.11	2.95	2.82	2.71	2.63
0.01	62	7.06	4.96	4.11	3.64	3.33	3.11	2.94	2.81	2.71	2.62
0.01	63	7.06	4.96	4.11	3.63	3.32	3.10	2.94	2.81	2.70	2.62
0.01	64	7.05	4.95	4.10	3.63	3.32	3.10	2.93	2.80	2.70	2.61
0.01	65	7.04	4.95	4.10	3.62	3.31	3.09	2.93	2.80	2.69	2.61
0.01	66	7.04	4.94	4.09	3.62	3.31	3.09	2.92	2.79	2.69	2.60
0.01	67	7.03	4.94	4.09	3.61	3.30	3.08	2.92	2.79	2.68	2.60
0.01	68	7.02	4.93	4.08	3.61	3.30	3.08	2.91	2.78	2.68	2.59
0.01	69	7.02	4.93	4.08	3.60	3.29	3.08	2.91	2.78	2.68	2.59
0.01	70	7.01	4.92	4.07	3.60	3.29	3.07	2.91	2.78	2.67	2.59
0.01	71	7.01	4.92	4.07	3.60	3.29	3.07	2.90	2.77	2.67	2.58
0.01	72	7.00	4.91	4.07	3.59	3.28	3.06	2.90	2.77	2.66	2.58

### Appendix D: Statistical tables

0.01	73	7.00	4.91	4.06	3.59	3.28	3.06	2.89	2.77	2.66	2.57
0.01	74	6.99	4.90	4.06	3.58	3.28	3.06	2.89	2.76	2.66	2.57
0.01	75	6.99	4.90	4.05	3.58	3.27	3.05	2.89	2.76	2.65	2.57
0.01	76	6.98	4.90	4.05	3.58	3.27	3.05	2.88	2.75	2.65	2.56
0.01	77	6.98	4.89	4.05	3.57	3.26	3.05	2.88	2.75	2.65	2.56
0.01	78	6.97	4.89	4.04	3.57	3.26	3.04	2.88	2.75	2.64	2.56
0.01	79	6.97	4.88	4.04	3.57	3.26	3.04	2.87	2.75	2.64	2.55
0.01	80	6.96	4.88	4.04	3.56	3.26	3.04	2.87	2.74	2.64	2.55
0.01	81	6.96	4.88	4.03	3.56	3.25	3.03	2.87	2.74	2.63	2.55
0.01	82	6.95	4.87	4.03	3.56	3.25	3.03	2.87	2.74	2.63	2.54
0.01	83	6.95	4.87	4.03	3.55	3.25	3.03	2.86	2.73	2.63	2.54
0.01	84	6.95	4.87	4.02	3.55	3.24	3.02	2.86	2.73	2.63	2.54
0.01	85	6.94	4.86	4.02	3.55	3.24	3.02	2.86	2.73	2.62	2.54
0.01	86	6.94	4.86	4.02	3.55	3.24	3.02	2.85	2.73	2.62	2.53
0.01	87	6.94	4.86	4.02	3.54	3.24	3.02	2.85	2.72	2.62	2.53
0.01	88	6.93	4.85	4.01	3.54	3.23	3.01	2.85	2.72	2.62	2.53
0.01	89	6.93	4.85	4.01	3.54	3.23	3.01	2.85	2.72	2.61	2.53
0.01	90	6.93	4.85	4.01	3.53	3.23	3.01	2.84	2.72	2.61	2.52
0.01	91	6.92	4.85	4.00	3.53	3.23	3.01	2.84	2.71	2.61	2.52
0.01	92	6.92	4.84	4.00	3.53	3.22	3.00	2.84	2.71	2.61	2.52
0.01	93	6.92	4.84	4.00	3.53	3.22	3.00	2.84	2.71	2.60	2.52
0.01	94	6.91	4.84	4.00	3.53	3.22	3.00	2.84	2.71	2.60	2.52
0.01	95	6.91	4.84	3.99	3.52	3.22	3.00	2.83	2.70	2.60	2.51
0.01	96	6.91	4.83	3.99	3.52	3.21	3.00	2.83	2.70	2.60	2.51
0.01	97	6.90	4.83	3.99	3.52	3.21	2.99	2.83	2.70	2.60	2.51
0.01	98	6.90	4.83	3.99	3.52	3.21	2.99	2.83	2.70	2.59	2.51
0.01	99	6.90	4.83	3.99	3.51	3.21	2.99	2.83	2.70	2.59	2.51
0.01	100	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59	2.50



### Appendix D: Statistical tables

<b>0.10</b>	1	39.86	49.5 0	53.5 9	55.8 3	57.2 4	58.2 0	58.9 1	59.4 4	59.8 6	60.1 9
<b>0.10</b>	2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39
<b>0.10</b>	3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23
<b>0.10</b>	4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92
<b>0.10</b>	5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30
<b>0.10</b>	6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94
<b>0.10</b>	7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70
<b>0.10</b>	8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54
<b>0.10</b>	9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42
<b>0.10</b>	10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32
<b>0.10</b>	11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25
<b>0.10</b>	12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19
<b>0.10</b>	13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14
<b>0.10</b>	14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10
<b>0.10</b>	15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06
<b>0.10</b>	16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03
<b>0.10</b>	17	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00
<b>0.10</b>	18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98
<b>0.10</b>	19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96
<b>0.10</b>	20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94
<b>0.10</b>	21	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92
<b>0.10</b>	22	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90
<b>0.10</b>	23	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89
<b>0.10</b>	24	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88
<b>0.10</b>	25	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87
<b>0.10</b>	26	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86
<b>0.10</b>	27	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85
<b>0.10</b>	28	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84

### Appendix D: Statistical tables

0.10	29	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86	1.83
0.10	30	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82
0.10	31	2.87	2.48	2.27	2.14	2.04	1.97	1.92	1.88	1.84	1.81
0.10	32	2.87	2.48	2.26	2.13	2.04	1.97	1.91	1.87	1.83	1.81
0.10	33	2.86	2.47	2.26	2.12	2.03	1.96	1.91	1.86	1.83	1.80
0.10	34	2.86	2.47	2.25	2.12	2.02	1.96	1.90	1.86	1.82	1.79
0.10	35	2.85	2.46	2.25	2.11	2.02	1.95	1.90	1.85	1.82	1.79
0.10	36	2.85	2.46	2.24	2.11	2.01	1.94	1.89	1.85	1.81	1.78
0.10	37	2.85	2.45	2.24	2.10	2.01	1.94	1.89	1.84	1.81	1.78
0.10	38	2.84	2.45	2.23	2.10	2.01	1.94	1.88	1.84	1.80	1.77
0.10	39	2.84	2.44	2.23	2.09	2.00	1.93	1.88	1.83	1.80	1.77
0.10	40	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76
0.10	41	2.83	2.44	2.22	2.09	1.99	1.92	1.87	1.82	1.79	1.76
0.10	42	2.83	2.43	2.22	2.08	1.99	1.92	1.86	1.82	1.78	1.75
0.10	43	2.83	2.43	2.22	2.08	1.99	1.92	1.86	1.82	1.78	1.75
0.10	44	2.82	2.43	2.21	2.08	1.98	1.91	1.86	1.81	1.78	1.75
0.10	45	2.82	2.42	2.21	2.07	1.98	1.91	1.85	1.81	1.77	1.74
0.10	46	2.82	2.42	2.21	2.07	1.98	1.91	1.85	1.81	1.77	1.74
0.10	47	2.82	2.42	2.20	2.07	1.97	1.90	1.85	1.80	1.77	1.74
0.10	48	2.81	2.42	2.20	2.07	1.97	1.90	1.85	1.80	1.77	1.73
0.10	49	2.81	2.41	2.20	2.06	1.97	1.90	1.84	1.80	1.76	1.73
0.10	50	2.81	2.41	2.20	2.06	1.97	1.90	1.84	1.80	1.76	1.73
0.10	51	2.81	2.41	2.19	2.06	1.96	1.89	1.84	1.79	1.76	1.73
0.10	52	2.80	2.41	2.19	2.06	1.96	1.89	1.84	1.79	1.75	1.72
0.10	53	2.80	2.41	2.19	2.05	1.96	1.89	1.83	1.79	1.75	1.72
0.10	54	2.80	2.40	2.19	2.05	1.96	1.89	1.83	1.79	1.75	1.72
0.10	55	2.80	2.40	2.19	2.05	1.95	1.88	1.83	1.78	1.75	1.72
0.10	56	2.80	2.40	2.18	2.05	1.95	1.88	1.83	1.78	1.75	1.71
0.10	57	2.80	2.40	2.18	2.05	1.95	1.88	1.82	1.78	1.74	1.71

### Appendix D: Statistical tables

0.10	58	2.79	2.40	2.18	2.04	1.95	1.88	1.82	1.78	1.74	1.71
0.10	59	2.79	2.39	2.18	2.04	1.95	1.88	1.82	1.78	1.74	1.71
0.10	60	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71
0.10	61	2.79	2.39	2.18	2.04	1.94	1.87	1.82	1.77	1.74	1.71
0.10	62	2.79	2.39	2.17	2.04	1.94	1.87	1.82	1.77	1.73	1.70
0.10	63	2.79	2.39	2.17	2.04	1.94	1.87	1.81	1.77	1.73	1.70
0.10	64	2.79	2.39	2.17	2.03	1.94	1.87	1.81	1.77	1.73	1.70
0.10	65	2.78	2.39	2.17	2.03	1.94	1.87	1.81	1.77	1.73	1.70
0.10	66	2.78	2.38	2.17	2.03	1.94	1.87	1.81	1.77	1.73	1.70
0.10	67	2.78	2.38	2.17	2.03	1.94	1.86	1.81	1.76	1.73	1.70
0.10	68	2.78	2.38	2.17	2.03	1.93	1.86	1.81	1.76	1.73	1.69
0.10	69	2.78	2.38	2.16	2.03	1.93	1.86	1.81	1.76	1.72	1.69
0.10	70	2.78	2.38	2.16	2.03	1.93	1.86	1.80	1.76	1.72	1.69
0.10	71	2.78	2.38	2.16	2.03	1.93	1.86	1.80	1.76	1.72	1.69
0.10	72	2.78	2.38	2.16	2.02	1.93	1.86	1.80	1.76	1.72	1.69
0.10	73	2.78	2.38	2.16	2.02	1.93	1.86	1.80	1.76	1.72	1.69
0.10	74	2.77	2.38	2.16	2.02	1.93	1.86	1.80	1.75	1.72	1.69
0.10	75	2.77	2.37	2.16	2.02	1.93	1.85	1.80	1.75	1.72	1.69
0.10	76	2.77	2.37	2.16	2.02	1.92	1.85	1.80	1.75	1.72	1.68
0.10	77	2.77	2.37	2.16	2.02	1.92	1.85	1.80	1.75	1.71	1.68
0.10	78	2.77	2.37	2.16	2.02	1.92	1.85	1.80	1.75	1.71	1.68
0.10	79	2.77	2.37	2.15	2.02	1.92	1.85	1.79	1.75	1.71	1.68
0.10	80	2.77	2.37	2.15	2.02	1.92	1.85	1.79	1.75	1.71	1.68
0.10	81	2.77	2.37	2.15	2.02	1.92	1.85	1.79	1.75	1.71	1.68
0.10	82	2.77	2.37	2.15	2.01	1.92	1.85	1.79	1.75	1.71	1.68
0.10	83	2.77	2.37	2.15	2.01	1.92	1.85	1.79	1.75	1.71	1.68
0.10	84	2.77	2.37	2.15	2.01	1.92	1.85	1.79	1.74	1.71	1.68
0.10	85	2.77	2.37	2.15	2.01	1.92	1.84	1.79	1.74	1.71	1.67
0.10	86	2.76	2.37	2.15	2.01	1.92	1.84	1.79	1.74	1.71	1.67

### Appendix D: Statistical tables

<b>0.10</b>	87	2.76	2.36	2.15	2.01	1.91	1.84	1.79	1.74	1.70	1.67
<b>0.10</b>	88	2.76	2.36	2.15	2.01	1.91	1.84	1.79	1.74	1.70	1.67
<b>0.10</b>	89	2.76	2.36	2.15	2.01	1.91	1.84	1.79	1.74	1.70	1.67
<b>0.10</b>	90	2.76	2.36	2.15	2.01	1.91	1.84	1.78	1.74	1.70	1.67
<b>0.10</b>	91	2.76	2.36	2.14	2.01	1.91	1.84	1.78	1.74	1.70	1.67
<b>0.10</b>	92	2.76	2.36	2.14	2.01	1.91	1.84	1.78	1.74	1.70	1.67
<b>0.10</b>	93	2.76	2.36	2.14	2.01	1.91	1.84	1.78	1.74	1.70	1.67
<b>0.10</b>	94	2.76	2.36	2.14	2.01	1.91	1.84	1.78	1.74	1.70	1.67
<b>0.10</b>	95	2.76	2.36	2.14	2.00	1.91	1.84	1.78	1.74	1.70	1.67
<b>0.10</b>	96	2.76	2.36	2.14	2.00	1.91	1.84	1.78	1.74	1.70	1.67
<b>0.10</b>	97	2.76	2.36	2.14	2.00	1.91	1.84	1.78	1.73	1.70	1.67
<b>0.10</b>	98	2.76	2.36	2.14	2.00	1.91	1.84	1.78	1.73	1.70	1.66
<b>0.10</b>	99	2.76	2.36	2.14	2.00	1.91	1.83	1.78	1.73	1.70	1.66
<b>0.10</b>	100	2.76	2.36	2.14	2.00	1.91	1.83	1.78	1.73	1.69	1.66