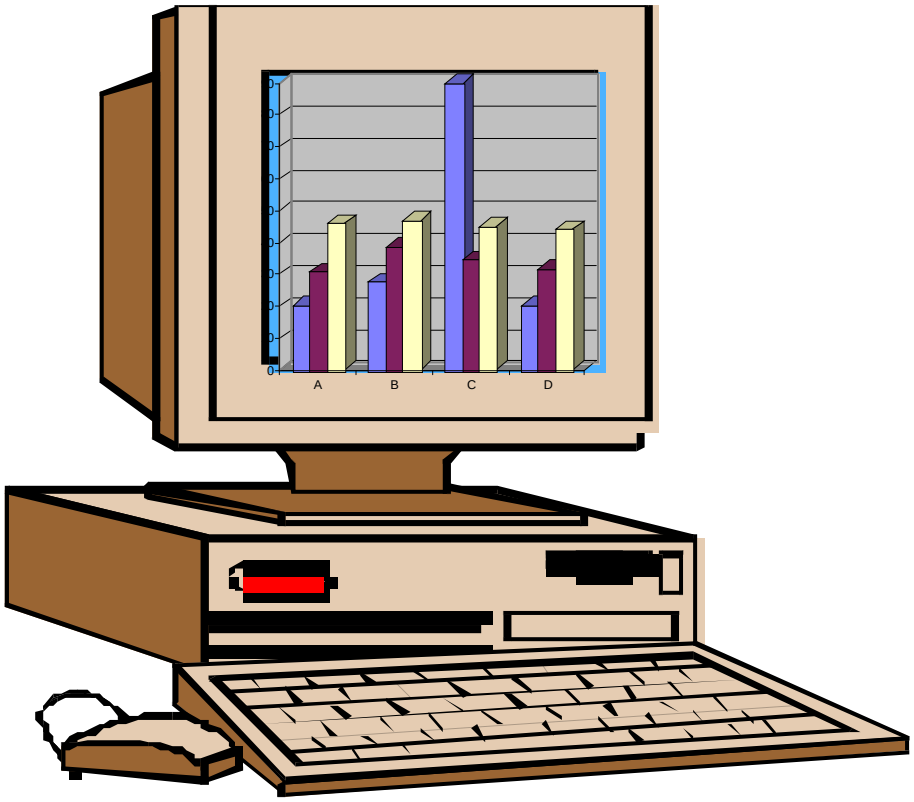
Department of Statistics

STA2601: Applied Statistics II



Workbook

Let wel: Hierdie studiemateriaal is slegs in Engels beskikbaar. As Afrikaans-ingeskrewe student staan dit u steeds vry om ’n werkopdrag of ’n eksamenvraestel in Afrikaans te beantwoord.

CONTENTS

ORIENTATION v

STUDY UNIT 1 (Revision of statistical distributions)

1.1 Introduction 1
1.2 General concepts of continuous and discrete distributions 1
1.3 Standard distributions 5

STUDY UNIT 2 (Concepts of estimation and inference)

2.1 Introduction 20
2.2 Defining a random sample and a statistic 20
2.3 Point estimation 24
2.4 Methods of finding estimators 32
2.5 Hypothesis testing 41
2.6 Confidence intervals 45
2.7 Simultaneous inference 54
2.8 Bayesian inference 54

STUDY UNIT 3 (Introduction to statistical software: JMP)

3.1 Introduction 55
3.2 Familiarising yourself with JMP 60
3.3 Generating random data 69

STUDY UNIT 4 (Testing for normality and goodness-of-fit tests in general)

4.1 Introduction 80
4.2 Graphical techniques 80
4.3 Goodness-of-fit test for normality 90
4.4 Goodness-of-fit tests in general 98
4.5 Using the method of moments to test for normality 102

STUDY UNIT 5 (Statistical independence)

5.1 The meaning of independence 116
5.2 Examples of dependence 118
5.3 Contingency table analysis 118
Exact test for a 2×2 table 122
5.4 Correlation 138

STUDY UNIT 6 (Inference on variances)

6.1 One-sample problem 161
6.2 Two independent samples 170
6.3 Paired observations 180
6.4 More than two independent samples 183

STUDY UNIT 7 (Inference on means)

7.1 One-sample problem 189
7.2 The power of the test and the noncentral t-distribution 197
7.3 Two-sample problem: independent samples 203
7.4 Paired observations 214
7.5 Independent samples with unequal variances 218
7.6 More than two independent samples 227

STUDY UNIT 8 (Regression)

8.1 Correlation and regression 235
8.2 The simple linear regression model 236
8.3 Estimation 239
8.4 Inference on the coefficients 243
8.5 Inference on the regression line 256
8.6 Relationship between tests for correlation and regression 260
8.7 Simple linear regression in matrix notation 261

STUDY UNIT 9 (Typical examination questions)

9.1 Worksheet 262
9.2 Worksheet: Solutions 290

ORIENTATION

Using the workbook

You will find that the workbook involves you in an *activity-feedback process.* As soon as you have to "do something" there will be a numbered *activity*. (In the study guide you have **examples** and **exercises** but in the workbook you have **activities**.) An empty space is always provided for your response. This will be followed by the *feedback icon* which indicates that a solution to your test will appear. Of course, if you peep at the solution before attempting to solve the problem on your own, you will gain very little from the *activity-feedback cycle.* We would like to suggest that you cover our feedback with a piece of paper to help you to be "strong"! (Perhaps you can even write the words **do it** – **I can** on this piece of paper and keep it as your bookmark.)

Do all the activities. They will help you understand and enjoy the work.

Both the study guide and workbook are divided into identical "sessions" or study units, covering specific topics, and they are also divided into identical *subsections.* However the workbook has an extra chapter of *typical examination questions.*

> This workbook should always be **used concurrently with the guide** as you proceed through the syllabus section by section . **You should first study the relevant subsection in the study guide** and then test your knowledge in the workbook.

The workbook provides you with an opportunity to apply yourself to the material covered in the guide. In other words, as you are gaining theoretical knowledge by working through the study guide, this *interactive* workbook will help you test and apply that knowledge. You will notice as we proceed through the study guide that whenever you have to perform a computer action, you will be guided step by step in the workbook on how to do it. Hence it seems logical that the workbook on some of the study units (see for example study unit 3) will be rather lengthy! Especially section 3.2 of workbook will guide you in detail through the first sessions on the computer and we do hope that you enjoy your introduction to statistical software!

The current **study guide** and **workbook** refer to *Sall, J, Creighton, L, and Lehman, A. (2007* **fourth edition) JMP™ Start Statistics** which is based on **JMP 7**. If Unisa issues you with the newer **JMP 9, the steps will not differ from JMP 8**.

Decimal comma or point?

We realise that in the South African schooling system the decimal comma is used and that you were even penalised at school when you used the decimal point. Now we sit between two fires: on the one hand your background and on the other hand computers! Because most computer packages use decimal points (one sometimes has the option to change it) and the authors *Sall, Creighton, and Lehman* also use the decimal point in our textbook (**JMP**^{TM} **Start Statistics)**, we will **use the decimal point in the workbook**.

Message to the student

Dear Student,

Studying by distance learning can be quite challenging and for some of you it is very expensive to come and meet your lecturer. This workbook is meant to bridge the gap between your studies and the lecturer. It shows step-by-step calculations. It was designed by Reina Nieuwoudt and the second revision is by Suwisa Muchengetwa.

Reina had been a statistics lecturer at Unisa for many years and always felt sorry for those students who were unable to pay her a visit at her office. She designed this workbook with the open distance learner in mind. Imagine that you have made an appointment and are sitting next to the lecturer whenever you work in your workbook. (If you do make a real appointment because you still experience some problems, make sure to bring your workbook along in order for us to see what you have written. This helps us to diagnose whether some important building blocks could be missing...) One of the learner characteristics identified is that most students request more detailed examples. You will find that the workbook not only provides additional examples, but we also try to help you through the "rough spots" or difficult sections in the study guide in an informal manner – almost as if you are sitting next to the lecturer!

There will be times when you feel you cannot go on or you might feel overwhelmed by the work. It is then that you have to remember that the big secret is to take it step by step. It is unpleasant to try to master statistics when you are under too much time pressure. It is also unpleasant to try to absorb too much in one go. Have you heard the saying: "If you fail to plan – you plan to fail"? Plan in advance when/where/how you are going to spend time with STA2601 and spend on average 10 hours per study unit. Budget an additional 30 hours for examination preparation and to complete your assignments.

Good luck with all your studies and **enjoy your statistics**!

STUDY UNIT 1

Revision of statistical distributions

1.1 Introduction

From your exposure to our first-level statistics courses, you must have gathered that statistical distributions form the backbone of any kind of application and are the basis for inference. You must feel very comfortable with the content of the first study unit and make sure that you absorb the terminology. We suppose that this new language sounded very unfamiliar at first, but that the secret to learning this new language was to proceed step by step: one step at a time. Make sure that you understand every new word and make it part of your vocabulary. If you feel overwhelmed by the condensed summary or if you are not sure that you understand a specific concept, revise that part in the relevant module of your first year.

1.2 General concepts of continuous and discrete distributions

Section 1.2 contains the important definitions which are necessary to build distribution theory. However, you must keep in mind that study unit 1 must be considered as our "toolbox" filled with whatever we need for the real job of applying our statistics knowledge. The emphasis is therefore not on deriving fancy distribution theorems (which is done in STA26032) but on **applying** this (supposedly known) theoretical work to observed data.

How do you apply distributions? The most important application is to obtain a desired probability.

In this section we not only aim to **revise your vocabulary** of statistical distributions but a few other concepts/terms/ideas are also given as a reference section.

Be sensitive for the "golden threa'd" or repetitive theme of hypothesis testing that will run through all the applications to come in future study units. There will be a "pattern of hypothesis testing". What is this pattern? For each *hypothesis to be tested* there must be a *test statistic* and **each test statistic must have a known distribution.** With each known distribution a *rare event*, in other words an event with a small probability, will be associated and this will give rise to the *critical values.* Critical values are all tabulated in *statistical tables*.

Activity 1.1

Indicate whether the following statements are correct or incorrect.

(a) The two main types of distribution functions are discrete and continuous distribution functions.....

.....

(b) A discrete distribution function could have jumps at a number of points called the discrete points of the distribution..

.....

(c) The probability function $f_X(x)$ of a discrete variable X has the property that $\sum_{x \in A} f_X(x) = 1$ where A is the set of discrete points of X.

.....

(d) For a discrete variable X, $f_X(x) = P(X = x) \geq 0$ only if A has finitely many elements.

.....

(e) The expected value of a discrete variable X is denoted by $\mu = E(X) = \sum_{x \in A} (x)\, f_X(x)$.

.....

Feedback  Feedback

(a) Correct.

(b) Correct.

(c) Correct.

(d) Incorrect. This is true for any value of x regardless of the size of A.

(e) Correct.

Activity 1.2

Suppose that the probability distribution for X is as is shown below.

x	$f_X(x)$
3	$\frac{1}{6}$
3.5	$\frac{2}{6}$
4	$\frac{2}{6}$
6	$\frac{1}{6}$

(a) Calculate the mean value of X.

(b) Calculate the variance of X.

(c) Would you say that the distribution is symmetrical?

Feedback → Feedback

(a) The *mean* value of X is calculated as:

$$\begin{aligned}\mu &= E(x) = \sum_x x\ f_X(x) \\ &= 3\left(\frac{1}{6}\right) + 3.5\left(\frac{1}{3}\right) + 4\left(\frac{1}{3}\right) + 6\left(\frac{1}{6}\right) \\ &= 4.0.\end{aligned}$$

(b) The *variance* of X is calculated as:

$$\begin{aligned}\sigma^2 &= \sum_{x\in A} (x-\mu)^2\ f_X(x) \\ &= (3-4)^2\left(\frac{1}{6}\right) + (3.5-4)^2\left(\frac{1}{3}\right) + (4-4)^2\left(\frac{1}{3}\right) + (6-4)^2\left(\frac{1}{6}\right) \\ &= 0.91667.\end{aligned}$$

(c) The *third central moment* of X is calculated as:

$$\begin{aligned}\mu_3 &= \sum_{x\in A} (x-\mu)^3\ f_X(x) \\ &= (3-4)^3\left(\frac{1}{6}\right) + (3.5-4)^3\left(\frac{1}{3}\right) + (4-4)^3\left(\frac{1}{3}\right) + (6-4)^3\left(\frac{1}{6}\right) \\ &= 1.125.\end{aligned}$$

The coefficient of skewness of X is calculated as $\beta_1 = \dfrac{\mu_3}{\sigma^3} = \dfrac{1.125}{(\sqrt{0.91667})^3} = 1.281\,8.$

If we draw a graph of the distribution, we can see that it is not symmetrical but seems to be positively skewed.

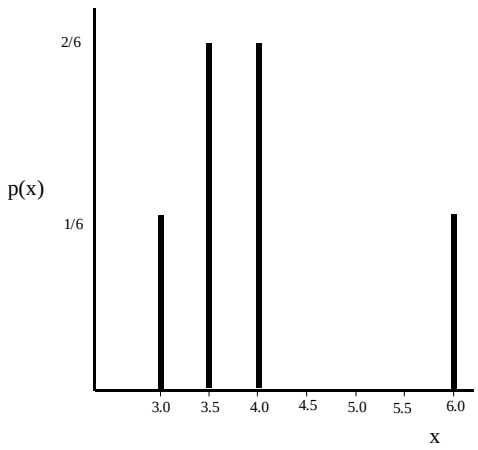


Figure 1.1

1.3 Standard distributions

Section 1.3 is given to revise the definition and important properties of some of the *most frequently used distributions* in statistics. Make sure that you are able to look up a value that is associated with a given fixed probability (for a specified distribution). In other words, this is where we start using our prescribed book of tables. It is very important that you know how to do this for the following distributions:

(i) normal distribution

(ii) t-distribution

(iii) χ^2-distribution

(iv) F-distribution

In the rest of the guide you will learn when and why a specific distribution can be applied, and as soon as the distribution is specified, you must be able to visualise the form of the density function and make a connection between a specific value and a probability.

The normal distribution plays such a major part in STA2601 that we need to make a few extra comments.

Extra comments on the normal distribution

Think about the given formal definition (see definition 1.16). What does it mean to you?

At first glance the very complicated mathematical formula for the probability density function (pdf.) might put you off. (Does it?) Well then, take time to make friends with it! Convince yourself that you could (with the help of a calculator) draw a graph of this function if you really had to! Represent x on the horizontal axis and $f_x(x)$ on the vertical axis.

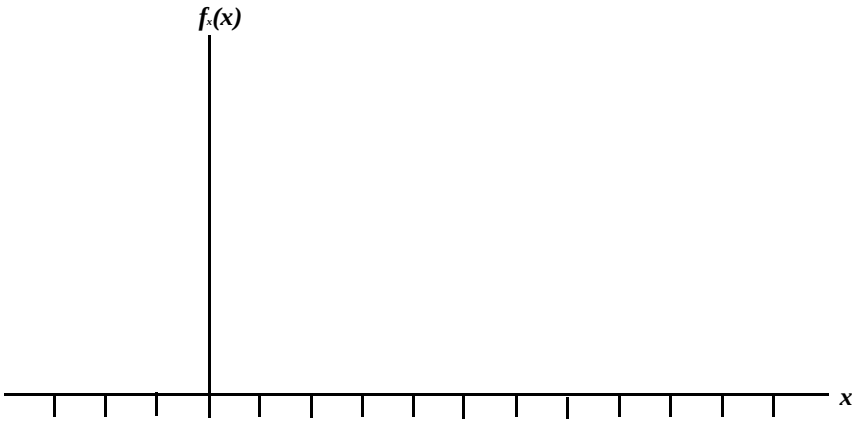


Figure 1.2

So, for any specific value of x (say $x = 12$) we could compute one single tiny little dot of the pdf. In other words, **provided** that we **know** what the numerical values of μ and σ^2 are, we can compute $f_x(x = 12)$ by replacing x by 12 in the formula.

Activity 1.3

Suppose that $\mu = 10$ and $\sigma = 10$. Show how you would compute $f_x(x = 12)$. (No need to simplify!)

.....

.....

.....

.....

.....

Feedback  Feedback

$$f_x(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2/\sigma^2}$$

Replace μ by 10 and σ by 10:

$$\therefore f_x(x) = \frac{1}{10\sqrt{2\pi}} e^{-\frac{1}{2}(x-10)^2/100}$$

Now replace x by 12:

$$f_x(x = 12) = \frac{1}{10\sqrt{2\pi}} e^{-\frac{1}{2}(12-10)^2/100}$$

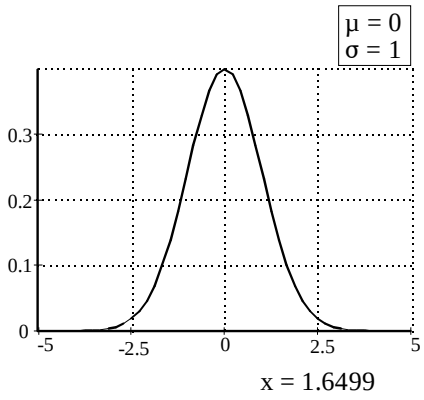
which can be computed on a calculator as the value 0.0391.

In other words, without knowing what the values of μ and σ are we cannot draw a graph of the pdf. That is why we say that the normal distribution has two parameters.

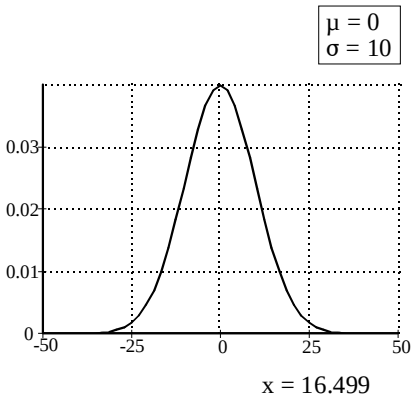
The graphs that follow are just a few examples of what $f_x(x)$ looks like for different values of μ and σ^2. I have given the values of μ and σ^2 in the upper right-hand corner. I have also calculated the value of X for which $P(X \geq x) = 0.05$.

Figure 1.3

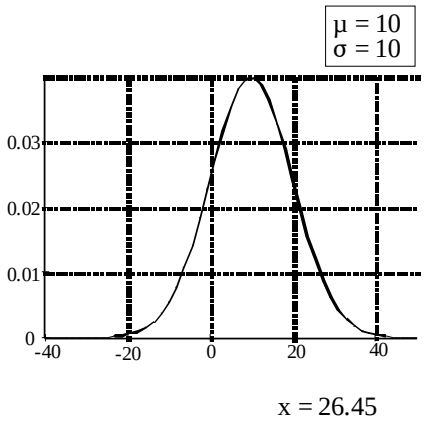
(a)



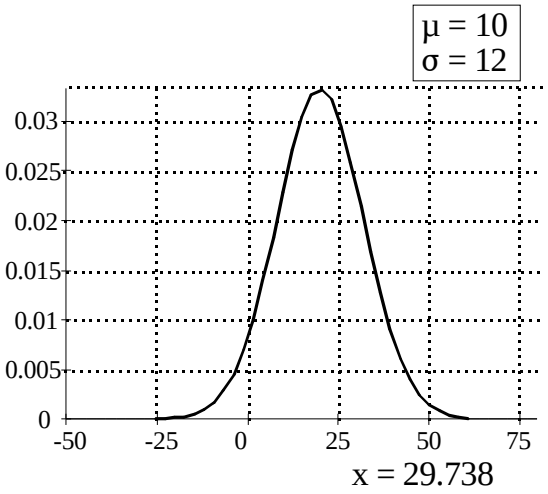
(b)



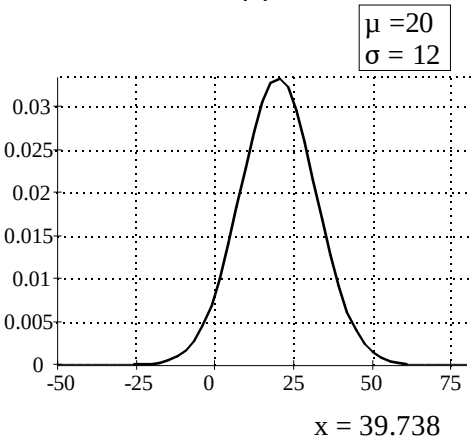
(c)



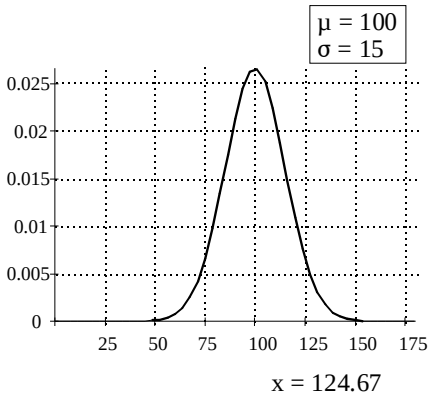
(d)



(e)



(f)



Why did we do this?

(1) So that you can get used to visualising any bell-shaped normal pdf. (Always try to make a little sketch!)

(2) So that you can appreciate the $Z = \dfrac{X - \mu}{\sigma}$ transformation that will help you to compute the x-values that we have given in the graphs (each associated with $\alpha = 0.05)$. Are you satisfied with the computations? Can you find the x-values?

(3) To try and convince you that the less you make use of recipes and the more you understand what the Stoker tables are all about, the more rewarding this section and the rest of the guide will be.

Activity 1.4

(a) Assume that $X \sim n(0;1)$. [See figure 1.3 (a).] Show that $P[X \geqslant 1.645] = 0.05$.

.....

.....

.....

.....

.....

(b) Assume that $X \sim n$(0;100). [See figure 1.3 (b).] Show that $P[X \geqslant 16.45] = 0.05$.

.....

.....

.....

.....

.....

(c) Assume that $X \sim n(10;100)$. [See figure 1.3 (c).] Show that $P[X \geqslant 26.45] = 0.05$.

.....

.....

.....

.....

.....

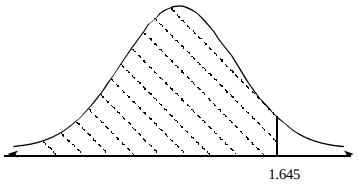
Feedback Feedback

(a) If $X \sim n(0;1)$ then $X \equiv Z$.

$\therefore P[Z \geqslant 1.645] = 1 - \Phi(1.645)$.

$\Phi(Z)$ is our special symbol for the distribution function of Z and is defined as

$\Phi(z) \equiv P[Z \leq z] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{1}{2}x^2} dx$.



The value $z = 1.645$ cannot be read off directly from TABLE I. We have to interpolate between $z = 1.64$ and $z = 1.65$.

$\therefore \frac{0.9495 + 0.9505}{2} = \frac{1.9000}{2} = 0.950 \Longrightarrow \Phi(1.645) = 0.95$.

So, $P[Z \geqslant 1.645] = 1 - 0.95 = 0.05$.

(b) Now $X \sim n(0;100)$, that is, $\mu = 0$ and $\sigma = \sqrt{100} = 10$.

Hence, $Z = \frac{X - \mu}{\sigma} = \frac{X - 0}{10}$.

$$\therefore P[X \geqslant 16.45] \equiv P\left[\frac{X - 0}{10} \geqslant \frac{16.45 - 0}{10}\right]$$

$$\equiv P[Z \geqslant 1.645]$$

(from this point onwards it is similar to question (a))

$$= 0.05.$$

(c) Now $X \sim n(10;100)$, that is, $\mu = 10$ and $\sigma = \sqrt{100} = 10$.

Hence, $Z = \frac{X - \mu}{\sigma} = \frac{X - 10}{10}$

$$\therefore P[X \geqslant 26.45] \equiv P\left[\frac{X - 10}{10} \geqslant \frac{26.45 - 10}{10}\right]$$

$$\equiv P[Z \geqslant 1.645]$$

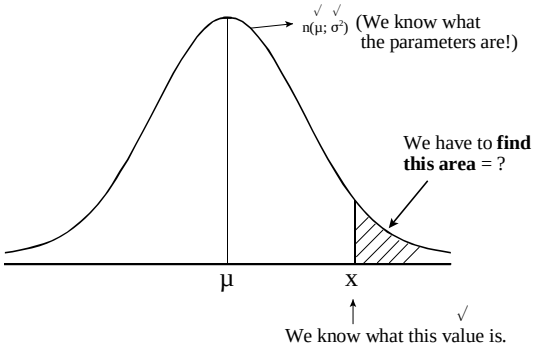
(from this point onwards it is similar to question (a))

$$= 0.05.$$

Whenever you have to apply the normal distribution you will have to do one of the following:

(1) **Find the appropriate probability** (probability $=$ area) that is associated with a given x-value on the horizontal axis.

Schematically:

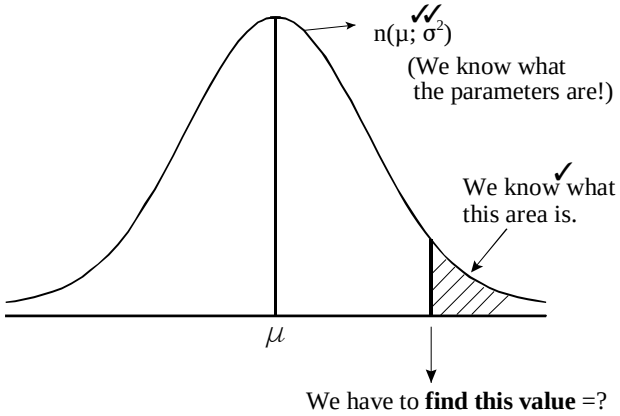


How? USE TABLE I.

OR

(2) **Find the appropriate x-value** (or z-value) on the horizontal axis that is associated with a given probability.

Schematically:



How? USE TABLE II.

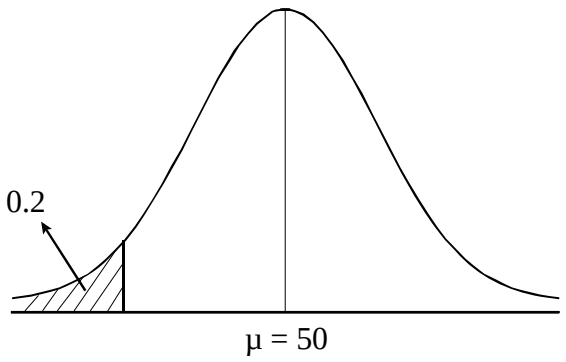
Activity 1.5

Suppose that $X \sim n(50;\ 169)$.

Find the x-value such that $P[X \leqslant x] = 0.2$.

Feedback Feedback

This is an example of the second scenario (which is the "more difficult" computation)!



$P[X \leqslant x] = 0.2$

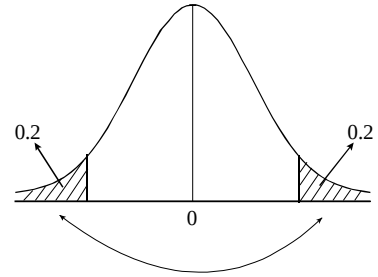
$$\therefore P\left[\frac{X-50}{\sqrt{169}} \leqslant \frac{x-50}{\sqrt{169}}\right] = 0.2$$

$$\therefore P\left[Z \leqslant \frac{x-50}{13}\right] = 0.2$$

Strictly speaking, we ought to be able to use TABLE II directly because we are given that $\Phi(Z) = 0.2$ and we want to find $\Phi^{-1}(z)$.

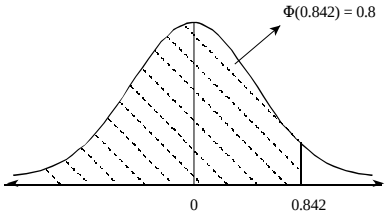
Now, unfortunately, because of the symmetry of the normal distribution, table II only starts at $z = 0.00$ and clearly the "wanted value" is less than 0.00. What now?

Work with the mirror image of the problem to convert it to a positive Z-value.

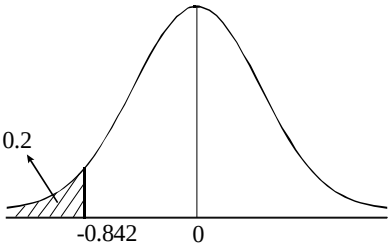


This will result in numerically the same value **except for the sign**.

From table II, the value associated with an area of 0.8, that is $P[Z \leqslant y] = 0.8$ is $y = 0.842$



and thus the corresponding negative value will yield $P[Z \leqslant -0.842] = 0.2$.



But keep in mind that the original question involves an unknown value that is associated with X. So, the last step is to transform back from a Z-value to an x-value.

Let $\dfrac{X - 50}{13} = -0.842$

$$\begin{aligned} \implies \quad X &= 13(-0.842) + 50 \\ &= 39.054. \end{aligned}$$

Finally we may write that

$$P[X \leqslant 39.054] = 0.20.$$

For both activities 1.4 and 1.5 the key to the solution was to make use of the z-transformation.

KEY TO THE SOLUTION WHEN YOU WORK WITH NORMAL VARIABLES:

Transform from $n(\mu;\ \sigma^2)$ to $n(0;\ 1)$ (or vice versa) by employing

$$Z = \frac{X - \mu}{\sigma}.$$

Please do not skip any of the examples of the normal distribution. If you do not work through them and do not understand them properly you will struggle with many other sections to follow, for example section 2.3 where all the important concepts of hypothesis testing are explained, assuming that you know the “workings” of the normal distribution. You will also struggle with the manual computations of performing a goodness-of-fit test for the normal distribution. And then of course there are all the other specific applications of the normal distribution where “something has an **approximate** normal distribution”. See for example study unit 5, where a function of the sample correlation coefficient has an **approximate** normal distribution and a function of the difference of two sample correlation coefficients has an **approximate** normal distribution.

The moral of the story? It is dangerous to skip anything in this section. We know you are not explicitly examined on study unit 1 but there are parts that you have to know. Apart from the normal distribution, there are a few other basic distributional properties that you must also know.

- For the χ^2-distribution it is important that you **know** result 1.1.
- For the F-distribution it is important that you **know** definition 1.21 and result 1.4.

If you ***summarise*** all the important results that you really have to know by heart on *one* piece of paper, you would not need a very large piece of paper! So please, consider them as important and really try to absorb them.

Have you heard about "mind mapping"? Tony Buzan developed it in 1970 and states that mind mapping engages the often inactive right hemisphere of the brain, emphasising spatial and visual language. Other authors like Astrid Brinkman, who teaches mathematics at the University of Duisburg, Germany, considers it a valuable tool in mathematics education. So, try to do more than merely a summary and draw yourself a mind map that shows the links between concepts.

Activity 1.6

Let $X_1,\ X_2,\ ...,\ X_7$ be independent $n\left(8;\ 9\right)$ variables. Are the following statements *correct or incorrect*?

(a) $\sum_{i=1}^{7} \frac{\left(X_i - 8\right)^2}{9}$ has a χ_7^2 distribution.

(b) $\sum_{i=1}^{7} \frac{\left(X_i - \overline{X}\right)^2}{9}$ has a χ_6^2 distribution.

(c) $E\left[\sum_{i=1}^{7} \frac{\left(X_i - \overline{X}\right)^2}{9}\right] = 6$

(d) $Var\left[\sum_{i=1}^{7} \frac{\left(X_i - 8\right)^2}{9}\right] = 12$

(e) $\frac{\left(X_1 - 8\right)^2}{9}$ has a X_1^2 distribution.

Feedback → Feedback

(a) **Correct**. From **result 1.2** it follows that $\sum_{i=1}^{n}\left[\frac{X_i-\mu}{\sigma}\right]^2 \sim \chi^2_n$

and thus $\sum_{i=1}^{7}\left[\frac{X_i-8}{3}\right]^2 = \sum_{i=1}^{7}\frac{(X_i-8)^2}{9} \sim \chi^2_7$.

(b) **Correct**. From **result 1.3** it follows that $\sum_{i=1}^{n}\left[\frac{X_i-\overline{X}}{\sigma}\right]^2 \sim \chi^2_{n-1}$

and thus $\sum_{i=1}^{7}\left[\frac{X_i-\overline{X}}{3}\right]^2 = \sum_{i=1}^{7}\frac{\left(X_i-\overline{X}\right)^2}{9} \sim \chi^2_{7-1}$.

(c) **Correct**. From **property (i) of result 1.1** it follows that

$E\left(\chi^2_n\right) = n.$

From question (b) above it holds that $E\left(\chi^2_6\right) = 6$.

(d) **Incorrect**. From **property (i) of result 1.1** it follows that

$Var\left(\chi^2_n\right) = 2n.$

From question (a) above it holds that $Var\left(\chi^2_7\right) = 2\,(7) = 14$.

(e) **Correct.** From **property (iv) of result 1.1** it follows that if $n = 1$ then

$$\left(\frac{X-\mu}{\sigma}\right)^2 \sim \chi^2_1, \text{ and thus } \frac{(X_1-8)^2}{9}$$

does indeed have a χ^2_1 distribution.

The idea behind the previous activity, was to make the distributional properties of the χ^2-distribution more concrete. Similarly, the idea behind the next activity, is to make the distributional properties of the F-distribution more concrete and to familiarise yourself with the tables of the χ^2-distribution and the F-distribution.

Activity 1.7

Let U_1 and U_2 be independent variables with $U_1 \sim \chi^2_{10}$ and $U_2 \sim \chi^2_{12}$. Are the following statements *correct or incorrect?*

[:-) Do not guess! Work them out!]

(a) $P\left(3.9403 \leq U_1 \leq 18.307\right) = 0.90$

(b) $\dfrac{U_1/10}{U_2/12} \sim F_{10;\,12}$

(c) $P\left[U_1/U_2 \geq 2.8083\right] = 0.025$

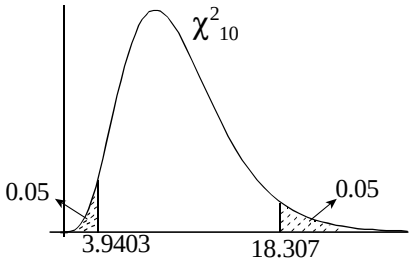
(d) Let $Y = \dfrac{U_1/10}{U_2/12}$ then $P\left[0.3436 \leq Y \leq 2.75\right] = 0.90$.

(e) Let $Y = \dfrac{U_1/10}{U_2/12}$ then $P\left[\dfrac{1}{Y} \geq 2.75\right] = 0.05$.

Feedback Feedback

It is given that $U_1 \sim \chi^2_{10}$ independent from $U_2 \sim \chi^2_{12}$.

(a) From TABLE IV – Stoker we can look up the following values:



$$P\left(\chi^2_{10} \geq 18.307\right) = 0.05$$

$$\Longrightarrow P\left(\chi^2_{10} \leq 18.307\right) = 0.95$$

and

$$P\left(\chi^2_{10} \geq 3.9403\right) = 0.95$$

$$\Longrightarrow P\left(\chi^2_{10} \leq 3.9403\right) = 1 - 0.95 \equiv 0.05$$

Thus $P\left(3.94 \leq U_1 \leq 18.307\right) = P\left(\chi^2_{10} \leq 18.307\right) - P\left(\chi^2_{10} \leq 3.94\right)$

$= 0.95 - 0.05$

$= 0.90.$

The statement is correct.

(b) In accordance with the definition in 1.21

$$\frac{U_1 / d_1}{U_2 / d_2} \sim F_{d_1;d_2}$$

provided that U_1 and U_2 are independent with $U_1 \sim \chi^2_{d_1}$ and $U_2 \sim \chi^2_{d_2}$.

Thus $\dfrac{U_1/10}{U_2/12} \sim F_{10;12}$ and the statement is correct.

(c) From (b) it follows that

$$P\left[\frac{U_1/10}{U_2/12} \geq F_{0.025;10;12}\right] = 0.025.$$

But $F_{0.025;10;12} = 3.37$ (from TABLE VI – Stoker).

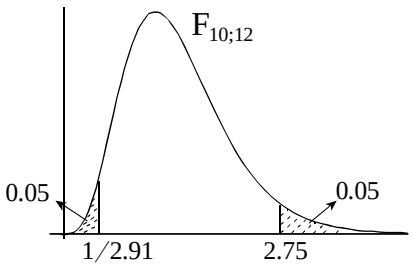
$$\therefore\ P\left[\frac{12\,U_1}{10\,U_2} \geq 3.37\right] = 0.025$$

$$\therefore\ P\left[U_1/U_2 \geq 2.8083\right] = 0.025$$

The statement is correct.

(d) $P\left[\dfrac{U_1/10}{U_2/12} \leq F_{0.95;10;12}\right] = 0.05$

But $F_{0.95;10;12} = \dfrac{1}{F_{0.05;12;10}} = \dfrac{1}{2.91}$ (from TABLE V).



$$Y = \frac{U_1/10}{U_2/12} \sim F_{10;12}$$

$$P\left[0.3436 \leq Y \leq 2.75\right] = 0.90$$

The statement is correct.

(e) If $Y = \frac{U_1/10}{U_2/12}$ then Y has an $F_{10;12}$ distribution whereas $\frac{1}{Y}$ has an $F_{12;10}$ distribution.

$$P\left[\frac{1}{Y} \geq F_{0,05;12;10}\right] = 0.05$$

But $F_{0,05;12;10} = 2.91$ (from TABLE V – Stoker) and not 2.75.

So

$$\left.\begin{array}{rcl} P\left[Y \geq 2.75\right] & = & 0.05 \\ \text{OR} & & \\ P\left[\frac{1}{Y} \geq 2.91\right] & = & 0.05 \end{array}\right] \text{ Both correct}$$

But $P\left[\frac{1}{Y} \geq 2.75\right] = 0.05$ **is an incorrect statement**.

STUDY UNIT 2

Concepts of estimation and inference

2.1 Introduction

Always remember that **you should first study the relevant subsection in the study guide** and then test your knowledge in the workbook. If you find that the definitions and/or methods become too theoretical, we will try and explain them in easy English in the workbook and also try to elaborate by including numerical examples. We have to remind you that you will gain very little from the *activity-feedback cycle* if you peep at the solution before attempting to solve the problem on your own.

2.2 Defining a random sample and a statistic

Activity 2.1

What do you make of definition 2.1 of the study guide?

……

……

……

……

……

……

Feedback → Feedback

It is very important that you really understand what the following sentence (which is frequently used throughout the study guide) means:

"Let $X_1,\ ...,\ X_n$ be a random sample from a distribution with pdf. $f_X\,(x)$."

First of all, you must realise that the notation $X_1,\ ...,\ X_n$ is an "abstraction" of a real-life situation.

How? Well, in real life we will know how big the random sample is. In other words, we will know that $n = 100$ or that $n = 350$ or whatever! So, in concrete terms, $X_1,\ ...,\ X_n$ will be something like $20;\ 31;\ 42;\ ...;\ 27$ for a **particular sample**.

Secondly, from the definition of a **random sample** we have the right to say that the probability density function of X_1 will be exactly the same as the probability density function of X_2 which will be exactly the same as that of X_3, et cetera, up to X_n. In other words, if we are interested in a particular measurement X on the individuals in a population, the value of X for an individual selected at random from the population is a random variable and we refer to this distribution as the ***population distribution***. A random sample of size n is a sequence: $X_1,\ X_2,\ ...,\ X_n$ where X_i is the X for the i-th individual selected. The variables X_i are exchangeable and they have a common distribution – the population distribution.

Thirdly, for any given problem to be specifically solved, the "pdf $f_X\,(x)$" must be a known function.

To enhance your understanding of definition 2.1, please do the following activity.

Activity 2.2

Let $X_1,\ X_2,\ ...,\ X_{16}$ constitute a **random sample** from a $n\,(10;\ 9)$ distribution.

(a) Write down the pdf. of X_1.

.....

.....

.....

(b) Write down the pdf. of X_{12}.

.....

.....

.....

(c) What is $E(X_4)$?

.....

.....

.....

(d) What is $E(X_{13})$?

.....

.....

.....

(e) What is $var(X_2)$?

.....

.....

.....

(f) What is $var(X_{16})$?

.....

.....

.....

(g) What is the joint pdf. of $X_1,\ X_2,\ ...,\ X_{16}$?

.....

.....

.....

Feedback → Feedback

(a) $X_1 \sim n\,(10;\ 9)$

$\Rightarrow f_{X_1}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2/\sigma^2}$ (see definition 1.16 in the study guide)

Thus, $f_{X_1}(x) = \frac{1}{3\sqrt{2\pi}} e^{-\frac{1}{2}(x-10)^2/9}$ (since it is given that $\mu = 10$ and $\sigma^2 = 9$).

(b) $X_{12} \sim n\,(10;\ 9)$

X_1 and X_{12} have the same distribution! So, $f_{X_{12}}(x) = \frac{1}{3\sqrt{2\pi}} e^{-\frac{1}{2}(x-10)^2/9}$.

(c) $E(X_4) = \mu$ (for the population)
$= 10$

(d) $E(X_{13}) = \mu$
$= 10$

$X_1,\ X_2,\ ...X_{16}$ all have the same expected value (that is the expected value of the population).

(e) $var(X_2) = \sigma^2$
$= 9$

(f) $var(X_{16}) = \sigma^2$
$= 9$

They all have the same variance (that is the variance of the population).

(g) $f_{X_1, X_2, ..., X_{16}}(x_1,\ x_2,\ ...,\ x_{16}) = f_X(x_1) \cdot f_X(x_2)\ ...\ f_X(x_{16})$

2.3 Point estimation

:-) To enhance your understanding of definition 2.3 (unbiasedness) and to exercise your *"mathematical muscles"*, please do the following two activities.

The aim of the first activity is to prove **the hint that is given in exercise 2.1, question 1 of the study guide**. It also "refreshes" your concept of "**mathematical expectation**" in case it has become a little rusty!

Do you recall that **the expectation of a sum is equal to the sum of the expectations**?

Do you also recall that $E[aX+b] = aE[X] + b$?

Activity 2.3

Assume that

$$\left.\begin{array}{l} E(X_i) = \mu \\ var(X_i) = E\left[(X_i - \mu)^2\right] = \theta \end{array}\right\} \text{ for } i = 1;\ 2;\ ...;\ n$$

and that $E[X_i,\ X_j] = E[X_i]\,E[X_j]$ for $i \neq j$ (that is, they are mutually stochastically independent).

Show that

(a) $E\left(X_i^2\right) = \theta + \mu^2$

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

and

(b) $E\left(\overline{X}^2\right) = \frac{\theta}{n} + \mu^2$.

Feedback → Feedback

(a) $\theta = E\left[(X_i - \mu)^2\right]$ [*This is the definition of a variance*.]

$= E\left[X_i^2 - 2X_i\mu + \mu^2\right]$ (*multiply out*)

$= E\left(X_i^2\right) - E\left(2X_i\mu\right) + E\left(\mu^2\right)$

$= E\left(X_i^2\right) - 2\mu E\left(X_i\right) + \mu^2$

$= E\left(X_i^2\right) - 2\mu \cdot \mu + \mu^2$

$= E\left(X_i^2\right) - \mu^2$

Therefore $\boxed{E\left(X_i^2\right) = \theta + \mu^2.}$

Please note:

Before we do the (b) part, (that is to prove that $E\left(\overline{X}^2\right) = \frac{\theta}{n} + \mu^2$), we want you to **consider the following squares**:

$$\left(X_1 + X_2 + \underset{\substack{\uparrow \\ \text{note}}}{X_3} \right)^2 = X_1^2 + X_2^2 + X_3^2 + \text{[....\textbf{all the cross terms}...]}.$$

How many terms? ${}_{(3)}P_2 = 3 \times 2 = 6$

The cross terms are:

$$\underbrace{X_1X_2 + X_2X_1 + X_1X_3 + X_3X_1 + X_2X_3 + X_3X_2}_{\textit{6 cross terms}}$$

Similarly $\left(X_1 + X_2 + X_3 + \underset{\substack{\uparrow \\ \text{note}}}{X_4} \right)^2 = X_1^2 + X_2^2 + X_3^2 + X_4^2 + \text{[....\textbf{all the cross terms}...]}.$

How many terms? ${}_{(4)}P_2 = 4 \times 3$

The cross terms are:

$$\underbrace{X_1X_2 + X_2X_1 + X_1X_3 + X_3X_1 + X_1X_4 + X_4X_1 + ... + X_4X_3}_{\textit{12 cross terms}}$$

Do you agree that

$$\left(X_1 + X_2 + ... + X_{(n)}\right)^2 = X_1^2 + X_2^2 + ... + X_n^2 + \text{[....\textbf{all the cross terms}...]}$$

and the **number** of cross terms will be ${}_n\mathbf{P}_2 = \mathbf{n}\,(n-1)$?

So, $\left(\sum_{i=1}^{n} X_i \right)^2$ can be written in the following neat way:

$$\underbrace{(X_1 + X_2 + ... + X_n)^2}_{\left(\sum_{i=1}^{n} X_i \right)^2} = \underbrace{X_1^2 + X_2^2 + ... + X_n^2}_{\sum_{i=1}^{n} X_i^2} + \underbrace{\text{all the cross terms}}_{\sum_{\substack{i=1 \\ i \neq j}}^{n} \sum_{j=1}^{n} X_i X_j}$$

[**:-)** We hope this convinces you forever and ever that $\left(\sum_{i=1}^{n} X_i \right)^2 \neq \sum_{i=1}^{n} X_i^2$.]

Now we return to the (b) part of the question.

(b) $E\left(\overline{X}^2\right) = E\left[\left(\frac{\Sigma X_i}{n}\right)^2\right]$

$$= E\left[\frac{1}{n^2}\left(\Sigma X_i\right)^2\right]$$

$$= \frac{1}{n^2}E\left[\sum_{i=1}^{n} X_i^2 + \sum_{\substack{i=1 \\ i\neq j}}^{n}\sum_{j=1}^{n} X_i X_j\right]$$

$$= \frac{E\left[\sum_{i=1}^{n} X_i^2\right]}{n^2} + \frac{E\left[\sum_{i=1\neq}^{n}\sum_{j=1}^{n} X_i X_j\right]}{n^2}$$

$$= \frac{\sum_{i=1}^{n} E\left[X_i^2\right]}{n^2} + \frac{\sum_{i=1\neq}^{n}\sum_{j=1}^{n} E\left[X_i X_j\right]}{n^2}$$

But we know that $E\left[X_i X_j\right] = E\left[X_i\right]E\left[X_j\right]$ for $i \neq j$

$$= \mu \cdot \mu = \mu^2.$$

Thus $E\left[\overline{X}^2\right] = \frac{\sum_{i=1}^{n}\left[\theta + \mu^2\right]}{n^2} + \frac{n\left(n-1\right)\mu^2}{n^2}$ (from (a) and $\underset{i\neq j}{\Sigma\Sigma} = n\left(n-1\right)$)

$$= \frac{n\left[\theta + \mu^2\right]}{n^2} + \frac{n\left(n-1\right)\mu^2}{n^2}$$

$$= \frac{\theta}{n} + \frac{\mu^2}{n} + \mu^2 - \frac{\mu^2}{n}$$

$$= \frac{\theta}{n} + \mu^2.$$

Activity 2.4

Do question 1 in exercise 2.1 of the study guide.

Feedback  Feedback

Before we really get to the solution, please note the following:

- The exercise does not specify the pdf. of the population (that is, it does not say it is normal, or chi-square or anything). All we know is that $E(X_1) = \mu$ and $E(X_2) = \mu$ and $E(X_3) = \mu$, et cetera, up to $E(X_n) = \mu$.

How do we know this? From the words ... "**with expected value** μ" ... If you feel unsure about this concept, go back to section 1.2 or even to some of your first-year study guides.

- We also know that $E(X_1 - \mu)^2 = \theta$ and that $E(X_2 - \mu)^2 = \theta$ ... and that $E(X_n - \mu)^2 = \theta$.

How do we know this? From the words: ..."with expected value μ **and variance** θ". Variance is defined as $E[X - E(X)]^2$. Do you recall this from section 1.2?

:-) Did you realise that the first sentence of exercise 2.1 question 1 really gives us this much information?

We can summarise this information in mathematical notation:

$$\left.\begin{array}{l} E(X_i) = \mu \\ E(X_i - \mu)^2 = \theta \end{array}\right\} \text{ for } i = 1;\ 2;\ ...;\ n \qquad \ldots\ldots\ldots\ldots\ldots\ldots (A)$$

The X_is are mutually independent. ..... (B)

The second sentence instructs us what to do:

"**Prove** that $\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$ is an **unbiased** estimator for θ."

Solution

Let $T=\frac{1}{(n-1)}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$.

Now we have to prove that $E(T)=\theta$ (that is, we have to apply definition 2.3).

So,

$$E\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2\right] = E\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i^2-2X_i\overline{X}+\overline{X}^2\right)\right]$$

(multiply out and keep the E- symbol outside the brackets)

$$= E\left[\frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2-2\overline{X}\sum_{i=1}^{n}X_i+\sum_{i=1}^{n}\overline{X}^2\right)\right]$$

$$= E\left[\frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2-2\overline{X}\left(n\overline{X}\right)+n\overline{X}^2\right)\right]$$

$$= E\left[\frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2-2n\overline{X}^2+n\overline{X}^2\right)\right]$$

$$= \frac{1}{n-1}\left[E\left(\sum_{i=1}^{n}X_i^2\right)-E\left(n\overline{X}^2\right)\right]$$

expectation of a sum = sum of the expectations

$$= \frac{\sum_{i=1}^{n}E\left(X_i^2\right)}{n-1}-\frac{nE\left(\overline{X}^2\right)}{n-1}$$

$$= \frac{\sum_{i=1}^{n}\left[\theta+\mu^2\right]}{n-1}-\frac{n\left[\frac{\theta}{n}+\mu^2\right]}{n-1}$$

(follows from the hint)

$$= \frac{n\left[\theta+\mu^2\right]}{n-1}-\frac{n\left[\frac{\theta}{n}+\mu^2\right]}{n-1}$$

$$= \frac{n}{n-1}\theta+\frac{n}{n-1}\mu^2-\frac{\theta}{n-1}-\frac{n}{n-1}\mu^2$$

$$= \frac{n}{n-1}\theta-\frac{\theta}{n-1}$$

$$= \frac{\theta(n-1)}{n-1}$$

$$= \theta$$

:-) To enhance your understanding of definition 2.4 (efficiency) and to exercise your *"mathematical muscles"*, please do the following activity:

Activity 2.5

Do exercise 2.1 question 2. in the study guide.

[The first part of the exercise is fine – now try to do the second part and show how you apply the hint.]

Prove that T_1 has a smaller variance than T_2.

.....

.....

.....

.....

.....

.....

.....

.....

.....

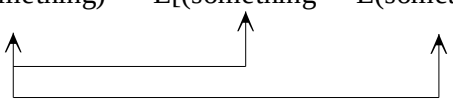
.....

Before I give the feedback, please note the following:

There are basically two ways of finding the variance of a variate which we will simply indicate as $var\,(\text{something})$.

Method I: Apply the definition of a variance:

$$\text{var(something)} = \text{E[(something - E(something))}^2\text{]}$$



If we apply this to T_1 it would mean:

$$\begin{aligned} var\,(T_1) &= E\left[\{T_1 - E\,[T_1]\}^2\right] \\ &= E\left[\left\{\frac{1}{n}\Sigma\,(X_i-\mu)^2 - E\left[\frac{1}{n}\Sigma\,(X_i-\mu)^2\right]\right\}^2\right] \end{aligned}$$

(and even if we substitute $E(T_1) = \theta$ we still have a complicated expression ...)

$$= \underbrace{E\left[\left\{\frac{1}{n}\Sigma(X_i-\mu)^2\right\}^2\right.}_{\text{Terrible to simplify!}} \left. -2\theta\left(\frac{1}{n}\Sigma(X_i-\mu)^2\right)+\theta^2\right]$$

So, what do we do now? We try the second method.

Method II: Try to express the "something" of which you have to find the variance in terms of "something else" but then (and this is the catch) you must **know** what the variance of "something else" is.

Let us apply this to T_1.

From our "toolbox" (see result 1.1) we know that if $Y \sim \chi^2_d$ then $var(Y) = 2d$. (Aha! So this is the variance that is **known** and which somebody else has computed!)

But, we also know (see result 1.2) that if

$$Y = \sum_{i=1}^{n} (X_i-\mu)^2/\sigma^2 \text{ then } Y \sim \chi^2_n.$$

In other words $\boxed{var(Y) = 2n \text{ and } E(Y) = n.}$

How do we transform T_1 to get to Y?

If $T_1 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2$ then $\frac{\mathbf{n}T_1}{\boldsymbol{\sigma}^2} = \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2} = Y$, in other words $T_1 = \left(\frac{\sigma^2}{n}\right)\cdot Y$.

So, $var(T_1) = var\left[\left(\frac{\sigma^2}{n}\right)Y\right]$ and $E(T_1) = \frac{\sigma^2}{n}\cdot E(Y) = \frac{\sigma^2}{n}\cdot n = \sigma^2$

$$= \left(\frac{\sigma^2}{n}\right)^2 var(Y) \qquad \text{(Do you recall that } var[aX] = a^2 var(X)\text{?)}$$

$$= \frac{\sigma^4}{n^2}\cdot 2n$$

$$= \frac{2\sigma^4}{n}.$$

Now you must try to do the same for $var(T_2)$.

2.4 Methods of finding estimators

A. Least squares estimation

Most of the more complicated applications of the method of least squares involve the solution of two or more unknowns from two or more equations. This is done more elegantly and efficiently if we employ matrix methods. I will include the optional matrix solution method for students who are at ease with inverses of matrices but I do not expect it from you.

Activity 2.6

Let $X_1, \ ..., \ X_n$ be independent random variables such that

$$\begin{aligned} E(X_i) &= \theta_1 c_i + \theta_2; & i = 1, 2, ..., n \\ Var(X_i) &= \sigma^2; & i = 1, 2, ..., n \end{aligned}$$

with σ^2, θ_1 and θ_2 unknown parameters and $c_{1,}\, c_{2,\, ...} c_n$ known constants. Find the least squares estimators of θ_1 and θ_2.

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

Feedback → Feedback

$E(X_i) = \theta_1 c_i + \theta_2; \qquad i = 1, 2, ..., n$

$$\begin{aligned} Q(\theta_1; \theta_2) &= \sum_{i=1}^{n} (X_i - E(X_i))^2 \\ &= \sum_{i=1}^{n} (X_i - \theta_1 c_i - \theta_2)^2 \end{aligned}$$

$$\begin{aligned} \frac{dQ}{d\theta_1} &= \sum_{i=1}^{n} 2(X_i - \theta_1 c_i - \theta_2)(-c_i) \\ &= -2 \sum_{i=1}^{n} (c_i X_i - c_i^2 \theta_1 - c_i \theta_2) \end{aligned}$$

$$\begin{aligned} \frac{dQ}{d\theta_2} &= \sum_{i=1}^{n} 2(X_i - \theta_1 c_i - \theta_2)(-1) \\ &= -2 \sum_{i=1}^{n} (X_i - \theta_1 c_i - \theta_2) \end{aligned}$$

Let $\frac{dQ}{d\theta_1} = 0$

$$\therefore \qquad \theta_1 \sum_{i=1}^{n} c_i^2 + \theta_2 \sum_{i=1}^{n} c_i = \sum_{i=1}^{n} c_i X_i \qquad (1)$$

Let $\frac{dQ}{d\theta_2} = 0$

$$\therefore \qquad \theta_1 \sum_{i=1}^{n} c_i + n\theta_2 = \sum_{i=1}^{n} X_i \qquad (2)$$

From equation (2) we find that

$$\begin{aligned} n\theta_2 &= \sum_{i=1}^{n} X_i - \theta_1 \sum_{i=1}^{n} c_i \\ \theta_2 &= \frac{\sum_{i=1}^{n} X_i - \theta_1 \sum_{i=1}^{n} c_i}{n}. \end{aligned} \tag{3}$$

Substitute (3) into (1):

$$\theta_1 \sum_{i=1}^{n} c_i^2 + \left[\frac{\sum_{i=1}^{n} X_i - \theta_1 \sum_{i=1}^{n} c_i}{n} \right] \sum_{i=1}^{n} c_i = \sum_{i=1}^{n} c_i X_i$$

$$\theta_1 \sum_{i=1}^{n} c_i^2 - \theta_1 \frac{\left(\sum_{i=1}^{n} c_i \right)^2}{n} = \sum_{i=1}^{n} c_i X_i - \frac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} c_i}{n}$$

$$\theta_1 \left[\sum_{i=1}^{n} c_i^2 - \frac{\left(\sum_{i=1}^{n} c_i \right)^2}{n} \right] = \frac{n \sum_{i=1}^{n} c_i X_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} c_i}{n}$$

$$\theta_1 \left[\frac{n \sum_{i=1}^{n} c_i^2 - \left(\sum_{i=1}^{n} c_i \right)^2}{n} \right] = \frac{n \sum_{i=1}^{n} c_i X_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} c_i}{n}$$

$$\theta_1 = \frac{n \sum_{i=1}^{n} c_i X_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} c_i}{n \sum_{i=1}^{n} c_i^2 - \left(\sum_{i=1}^{n} c_i \right)^2} \tag{4}$$

Substitute (4) back into (3):

$$\theta_2 = \frac{\sum_{i=1}^{n} X_i - \theta_1 \sum_{i=1}^{n} c_i}{n} \qquad from \tag{3}$$

$$= \frac{\sum_{i=1}^{n} X_i}{n} - \frac{\sum_{i=1}^{n} c_i}{n} \left[\frac{n \sum_{i=1}^{n} c_i X_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} c_i}{n \sum_{i=1}^{n} c_i^2 - \left(\sum_{i=1}^{n} c_i \right)^2} \right]$$

(no need to simplify further)

Matrix solution: from equations (1) and (2) we may write in matrix notation that

$$\begin{bmatrix} \sum_{i=1}^{n} c_i^2 & \sum_{i=1}^{n} c_i \\ \sum_{i=1}^{n} c_i & n \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} c_i X_i \\ \sum_{i=1}^{n} X_i \end{bmatrix}$$

$$\begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} c_i^2 & \sum_{i=1}^{n} c_i \\ \sum_{i=1}^{n} c_i & n \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^{n} c_i X_i \\ \sum_{i=1}^{n} X_i \end{bmatrix} \quad \text{where}$$

$$\begin{bmatrix} \sum_{i=1}^{n} c_i^2 & \sum_{i=1}^{n} c_i \\ \sum_{i=1}^{n} c_i & n \end{bmatrix}^{-1} = \frac{1}{n(\sum_{i=1}^{n} c_i^2) - (\sum_{i=1}^{n} c_i)^2} \begin{bmatrix} n & -\sum_{i=1}^{n} c_i \\ -\sum_{i=1}^{n} c_i & \sum_{i=1}^{n} c_i^2 \end{bmatrix} \quad \text{and thus}$$

$$\begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} = \frac{1}{n(\sum_{i=1}^{n} c_i^2) - (\sum_{i=1}^{n} c_i)^2} \begin{bmatrix} n & -\sum_{i=1}^{n} c_i \\ -\sum_{i=1}^{n} c_i & \sum_{i=1}^{n} c_i^2 \end{bmatrix} \begin{bmatrix} \sum_{i=1}^{n} c_i X_i \\ \sum_{i=1}^{n} X_i \end{bmatrix}.$$

$$\text{Hence, } \theta_1 = \frac{n\sum_{i=1}^{n} c_i X_i - (\sum_{i=1}^{n} c_i)(\sum_{i=1}^{n} X_i)}{n(\sum_{i=1}^{n} c_i^2) - (\sum_{i=1}^{n} c_i)^2}$$

$$\text{and } \theta_2 = \frac{-(\sum_{i=1}^{n} c_i)(\sum_{i=1}^{n} c_i X_i) + (\sum_{i=1}^{n} c_i^2)(\sum_{i=1}^{n} X_i)}{n(\sum_{i=1}^{n} c_i^2) - (\sum_{i=1}^{n} c_i)^2}.$$

B. Maximum likelihood estimation

Our formal definition of the likelihood function (see definition 2.5) states that

$$L(\theta) = f_X(X_1;\ \theta) \cdot f_X(X_2;\ \theta) \ldots \cdot \cdot f_X(X_n;\ \theta).$$

↖

What does this mean?

It means theoretically that we have to substitute the value of each X_i (considered as a **specific outcome for** X_i of the random sample) in the formula of the pdf. (which then leaves us with only **one unknown** which is the parameter θ that has to be estimated). So, although we still work with $X_i s$, they are not considered to be unknowns!

Additional comments, steps and explanation for example 2.8 of the study guide

It is given that $f(x) = cx^{-c-1}$ for $x > 1$.

Does it bother you that we use a "c" and not a "θ" for the unknown parameter? If it does, you can replace the "c" with a "θ". It would not change anything!

Now we know that $f_X(X_1;\ c) = cX_1^{-c-1}$ and $f_X(X_2;\ c) = cX_2^{-c-1}$, et cetera ... up to $f_X(X_n;\ c) = cX_n^{-c-1}$.

The product $f_X(X_1;\ c) \cdot f_X(X_2;\ c) \ldots f_X(X_n;\ c)$ then becomes

$$\left(cX_1^{-c-1}\right) \cdot \left(cX_2^{-c-1}\right) \ldots \left(cX_n^{-c-1}\right).$$

It takes time and practice to get used to the notation $\prod_{i=1}^{n} f(X_i;\ c)$. It is much safer to write out all the terms if you feel uneasy in the beginning! Then you can "see" how to simplify the expression. But you must also practise to use this shorthand!

So $L(c) = c^{\mathbf{n}} \left(X_1^{-c-1}\right)\left(X_2^{-c-1}\right) \ldots \left(X_n^{-c-1}\right)$.

Keep in mind that this is an abstraction of a real-life situation. We will know in real life what the value of n is, and $X_1;\ X_2;\ \ldots;\ X_n$ will be known numerical values so that $L(c)$ is truly only a function of c. We must, however, be able to find the maximum value of $L(c)$ **in algebraic terms**, and you cannot do this unless you know the rules of algebra!

Rewrite $L(c)$ as:

$L(c) = c^n [X_1 \cdot X_2 \ldots X_n]^{-c-1}$ (→ *we can even simplify this to* $-(c+1)$)

$\therefore \log L(c) = n \log c - (c+1) \log [X_1 \cdot X_2 \ldots X_n]$

apply the rule $\log(x^a) = a \log(x)$

$\therefore \log L(c) = n \log c - (c+1)[\log X_1 + \log X_2 + ... + \log X_n]$

apply the rule $\log(x \cdot y) = \log x + \log y$

So $\log L(c) = n \log c - (c+1) \sum_{i=1}^{n} \log X_i$.

(You can even write this as $n \log c - c \sum_{i=1}^{n} \log X_i - \sum_{i=1}^{n} \log X_i$ if it makes the differentiation with respect to c easier!) The rest follows in the study guide.

Activity 2.7

Let $X_1, X_2, ..., X_n$ be a random sample from a distribution with pdf

$$f_X(x) = \theta(1-\theta)^{x-1} \text{ for } x > 1.$$

Find the maximum likelihood estimator for θ.

$$L(\theta) = \prod_{i=1}^{n} f\left(X_i;\ \theta\right) = \prod_{i=1}^{n} \theta\left(1-\theta\right)^{X_i - 1}.$$

Or

$$\begin{aligned} L\left(\theta\right) &= \theta\left(1-\theta\right)^{X_1-1} \cdot \theta\left(1-\theta\right)^{X_2-1} \cdots \theta\left(1-\theta\right)^{X_n-1} \\ &= \theta^n (1-\theta)^{X_1+X_2+\ldots+X_n-n} \end{aligned}$$

$$\therefore \log L\left(\theta\right) = n\log\theta + \left[\sum_{i=1}^{n} X_i - n\right]\log(1-\theta)$$

$$\begin{aligned} \frac{d\log L\left(\theta\right)}{d\theta} &= \frac{n}{\theta} + \left[\sum_{i=1}^{n} X_i - n\right]\frac{-1}{(1-\theta)} \\ &= \frac{n(1-\theta) - \theta\left[\sum_{i=1}^{n} X_i - n\right]}{\theta(1-\theta)} \end{aligned}$$

Set $\dfrac{d\log L\left(\theta\right)}{d\theta} = 0$

$$\begin{aligned} \therefore \qquad n(1-\theta) - \theta\left[\sum_{i=1}^{n} X_i - n\right] &= 0 \\ n - n\theta - \theta\sum_{i=1}^{n} X_i + n\theta &= 0 \\ \theta\sum_{i=1}^{n} X_i &= n \\ \therefore \widehat{\theta} &= \frac{n}{\Sigma X_i} \\ \widehat{\theta} &= \frac{1}{\overline{X}} \end{aligned}$$

AN IMPORTANT NOTE FOR BCom and BA STUDENTS:

Before 1999 the module STA205-Q: "Statistical Distributions" was a compulsory parallel module when you enrolled for STA2601 and it covered all the necessary differential calculus needed for STA2601. Unfortunately this is no longer the case and now there seems to be a few gaps in what you are taught in first-year statistics and what you need to do study unit 2 properly, especially **finding the derivatives of certain functions**. The very cryptic summary below might help fill this gap.

The following standard results are the derivatives of some of the standard functions most often used in statistics:

[1] $\frac{d}{dx}(ax+b) = a$

[2] $\frac{d}{dx}(x^n) = nx^{n-1}$ where $n \neq 0$

[3] $\frac{d}{dx}(e^x) = e^x$

[4] $\frac{d}{dx}(\log x) = \frac{1}{x}$ where $x > 0$

[5] $\frac{d}{dx}(b^x) = b^x \log x$ where $b > 0$

[6] $\frac{d}{dx}[f(x) + g(x)] = \frac{d}{dx}f(x) + \frac{d}{dx}g(x)$ (addition rule)

Sometimes we write $\frac{d}{dx}[f(x)]$ as $f'(x)$.

[7] $\frac{d}{dx}[f(x)g(x)] = \left[\frac{d}{dx}f(x)\right]g(x) + f(x)\left[\frac{d}{dx}g(x)\right]$ (product rule)

[8] $\frac{d}{dx}[f(g(x))] = \{f'(g(x))\}\left\{\frac{d}{dx}g(x)\right\}$ (chain rule)

For example:

$$\begin{aligned}\frac{d}{dx}[(1+x^3)^5] &= \{5(1+x^3)^4\}\left\{\frac{d}{dx}(1+x^3)\right\} \\ &= \{5(1+x^3)^4\}\{3x^2\} \\ &= 15x^2(1+x^3)^4 \text{ (after simplification)}\end{aligned}$$

Another example:

$$\begin{aligned}\frac{d}{dx}[\log(3x^2+5x-6)] &= \left\{\frac{1}{(3x^2+5x-6)}\right\}\left\{\frac{d}{dx}(3x^2+5x-6)\right\} \\ &= \left\{\frac{1}{(3x^2+5x-6)}\right\}\{6x+5\} \\ &= \frac{6x+5}{3x^2+5x-6}\end{aligned}$$

Applying the chain rule leads to the following generalisations of results [1] to [5]:

[9] $\frac{d}{dx}(ag(x)+b) = ag'(x)$

[10] $\frac{d}{dx}g(x)^n = ng(x)^{n-1}.g'(x)$ where $n \neq 0$

[11] $\frac{d}{dx}(e^{g(x)}) = e^{g(x)}g'(x)$

[12] $\frac{d}{dx}(\log g(x)) = \frac{1}{g(x)}g'(x)$ where $x > 0$

[13] $\frac{d}{dx}(b^{g(x)}) = b^{g(x)}(\log bg'(x))$ where $b > 0$

2.5 Hypothesis testing

You must not get confused between the *overall seven steps* of hypothesis testing in general, which have been mentioned in this section, which are

Step 1. The brainwave
Step 2. Choice of a model
Step 3. Specification of the hypothesis and significance level
Step 4. Planning of the experiment
Step 5. Choice of a test
Step 6. The experiment
Step 7. Analysis of the data

and the "three-step-approach" which we gave as a recipe in first-year modules, that is

the first step is to **specify the hypotheses** to be tested;
the second step is to decide on a **test statistic;**
the third step is to find a **decision rule** that would enable you to make the correct decision, and there are two routes:

Route A is only possible when you have a computer program to compute the *p-value*.
Route B is compulsory when you perform a hypothesis test manually, and you have "**to obtain a rejection region**".

The "*overall seven steps"* summarises the idealistic theoretical frame whilst the *"three-step approach"* of our first year summarises the practicality of hypothesis testing.

The essence of a statistical test of H_0 is a comparison of actual sample data with what might be expected when H_0 is true. Such a comparison is based on the outcome of a relevant sample statistic, which in a hypothesis testing setup, is called the **test statistic**. The sampling distribution of this test statistic depends on what is assumed as the population distribution (in other words the statistical model that fits the hypothesis). The sampling distribution of a test statistic under H_0 is called its null distribution.

Thus the starting point is to have a *known sampling distribution* of a *test statistic* and to use this information in order **to specify values for the test statistic that are extreme under** H_0 (of course this will be in the direction of H_1, suggesting that H_1 provides a better explanation of the data than H_0).

Of course "to specify values for the test statistic that are extreme under H_0" means "**to obtain a rejection region**".

Our piglet example (see example 2.9) is nothing new to you because we dealt with it in the first year as **"testing the population mean when the population standard deviation is known".** This

means the assumed population distribution is a $n\left(\mu;\ \sigma^2\right)$ distribution. The relevant test statistic that would be able to discriminate between H_0 and H_1 will be $\overline{X}$. Thus the sampling distribution of $\overline{X}$ depends on the assumption that the population distribution is $n\left(\mu;\ \sigma^2\right)$, that is we can derive (see theorem 1.2) the following:

If $X \sim n\left(\mu;\ \sigma^2\right)$ and if we assume that σ^2 **is known**, then $\overline{X} \sim n\left(\mu;\ \sigma^2/n\right)$.

Or

If $X \sim n\left(\mu;\ \sigma^2\right)$ and if we assume that σ^2 is unknown, then $\dfrac{\overline{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$.

Usually the most laborious step in hypothesis testing will be to specify values of the test statistic that are extreme under H_0 (and then of course in the direction of H_1, suggesting that H_1 provides a better explanation of the data than H_0). This so-called critical region with an associated probability cannot be specified unless the null distribution is known.

How we construct such test statistics and find their null distributions is what the rest of the study guide is all about! Do not feel discouraged if at first you struggle with the concepts of hypothesis testing. The more test statistics we derive and the more hypotheses we test the easier all these concepts will become!

Another concept which is difficult to grasp (usually because it is too difficult to compute manually) is that of *the power of the test*. You will see how clear this concept becomes once we start working with a computer program to do all the difficult calculations!

Activity 2.8

Refer to example 2.9 of the study guide (the piglet example).

Show that if $X \sim n\left(40;\ 4\right)$ then $P\left[X \geq 43.290\right] = 0.05$.

.....

.....

.....

.....

.....

.....

.....

.....

.....

Feedback → Feedback

$$
\begin{aligned}
P[X \geq 43.290] &\equiv P\left[\frac{X-\mu}{\sigma} \geq \frac{43.290-40}{\sqrt{4}}\right] \\
&= P\left[Z \geq \frac{3.290}{2}\right] \\
&= P[Z \geq 1.645] \\
&= 0.05
\end{aligned}
$$

Activity 2.9

Refer to example 2.9(d) of the study guide (the continuation of the piglet example).

(a) Show that if $X \sim n(40.5;\ 4)$ then $P[X \geq 43.29] = 0.0816$.

.....

.....

.....

.....

(b) Show that if $X \sim n(43.29;\ 4)$ then $P[X \geq 43.29] = 0.5$.

.....

.....

.....

.....

(c) Show that if $X \sim n(45.362;\ 4)$ then $P[X \geq 43.29] = 0.85$.

.....

.....

.....

.....

(a) $P[X \geq 43.29] = P\left[\frac{X-\mu}{\sigma} \geq \frac{43.29-40.5}{2}\right]$

$= P[Z \geq 1.395]$

$= 1 - \Phi(1.395)$

$= 1 - \left[\frac{0.9177+0.9192}{2}\right]$ (Table I)

$= 1 - 0.9184$

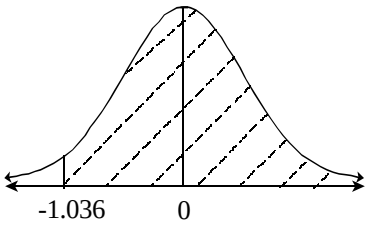
$= 0.0816$

(b) $P[X \geq 43.29] = P\left[\frac{X-\mu}{\sigma} \geq \frac{43.29-43.29}{2}\right]$

$= P[Z \geq 0]$

$= 0.5$

(c) $P[X \geq 43.29] = P\left[\frac{X-\mu}{\sigma} \geq \frac{43.29-45.362}{2}\right]$



$= P[Z \geq -1.036]$

$= P[Z \leq 1.036]$ (Symmetry of normal distribution)

$= \Phi(1.036)$

$= 0.8500$ (Table I)

INTERPOLATION:

$\Phi(1.036) = \Phi(1.03) + \frac{6}{10}[\Phi(1.04) - \Phi(1.03)]$

$= 0.8485 + 0.6[0.8508 - 0.8485]$

$= 0.8485 + 0.6[0.0023]$

$= 0.8485 + 0.00138$

$= 0.84988$

$= 0.8500$ (Rounded to three decimals)

2.6 Confidence intervals

In section 2.3 we derived point estimates for unknown population parameters. One problem with point estimation is that the single computed value does not include any indication of how high the probability is that the value of the estimate is close to the unknown true value. In interval estimation we derive two points, defining an interval, containing the unknown population parameter and most importantly, we attach a probability or a degree of certainty to our result. Hypothesis testing and the computation of a confidence interval for an unknown population parameter usually go hand in hand. Most textbooks prefer to organise study contents in such a manner that the derivation of a confidence interval is followed directly by the derivation of the hypothesis test *for the same, specific sampling distribution* or vice versa. Whichever way, the starting point is to have a *known sampling distribution* of a *sample statistic* and to **manipulate a probability expression** whereby we create a lower and an upper bound for the interval.

The initial discussion in this section is rather abstract and theoretical but example 2.10 and exercise 2.8 are practical illustrations. These two illustrations are borrowed from the work we covered in the first year and you will encounter them again in study unit 6.
The following two activities are aimed at reviewing some of the concepts and at highlighting the difference between the distribution of a single variable and the sample mean.

Activity 2.10

Psychologists make the claim that IQ scores (for a specific population) follow a normal distribution with $\mu = 100$ and $\sigma = 15$.

How do you (statistically) interpret this information?

.....

.....

.....

.....

.....

.....

.....

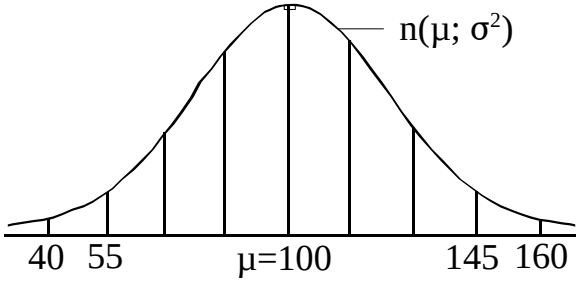
.....

.....

.....

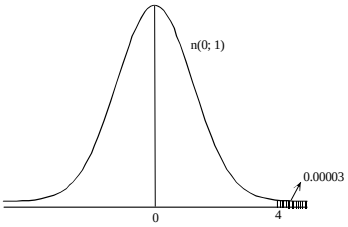
Feedback Feedback

With the wording “normal distribution” we immediately picture a bell-shaped symmetrical distribution. If $X =$ IQ score for **an individual** from this population then $X \sim n\left(100;\ (15)^2\right)$.

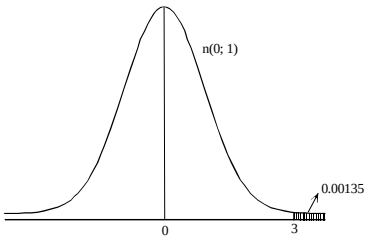


From table 1 we find that $\Phi(4.000) = 0.99997$.

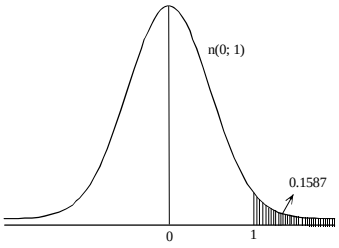
⇒ We are 99.994% sure that all the IQ scores will fall between 40 and 160 (in other words four standard deviations above and below the mean).



⇒ We are 99.73% sure that all the IQ scores will fall between 55 and 145 (in other words three standard deviations above and below the mean).



⇒ 68.26% of the population will have IQs between 85 and 115.



We translate the words “with a probability of 0.9973” to the words “we are 99.73% sure”

Activity 2.11

Refer to activity 2.10.

Suppose we draw a random sample of size $n = 100$ from this population.

What do you expect your sample to look like?

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

Feedback → Feedback

(a) I would expect more or less 50 values to be above 100 and 50 values to be below 100.

(b) I would expect more or less 68.26% of the sample to be between 85 and 115.

(c) I would expect the mean, $\overline{X} = \frac{\sum_{i=1}^{100} X_i}{100}$, to be close to 100.

Activity 2.12

Refer to activities 2.10 and 2.11.

Now focus your attention only on $\overline{X} = \frac{\sum_{i=1}^{100} X_i}{100}$ (and not on the individual values of the sample).

Indicate whether the following statements are true or false:

(a) $\overline{X}$ will be exactly 100.

.....

.....

.....

(b) $\overline{X}$ will be close to 100.

.....

.....

.....

(c) If the sample size was $n = 1\,000$ then $\overline{X}$ will be exactly 100.

.....

.....

.....

(d) We have the same probability of finding a sample mean of 145 as we have of drawing an individual value of 145.

.....

.....

.....

(e) The spread of $\overline{X}$ is a function of σ and of n.

.....

.....

.....

Feedback → Feedback

(a) False! $\overline{X}$ could work out to be exactly 100, or it could be $\overline{X} = 101.24$ or $\overline{X} = 99.98$ or ... any value close to 100. Theoretically $E\left(\overline{X}\right) = \mu$ which implies that $\overline{X}$ varies in the vicinity of μ.

(b) True.

(c) False! Even if $n = 1\,000$ the value of $\overline{X}$ need not be exactly equal to 100. But now we feel that $\overline{X}$ will vary even closer to μ.

(d) False! We intuitively feel that the *very large values* of individual Xs will "cancel" the *very small values* of individual X-observations to create the mean. It is almost impossible (if the X-observations came from a $n\left(100;\ (15)^2\right)$ distribution) that $\overline{X} = 145,$ but it is not impossible that an individual $X = 145$.

(e) True! The larger the value of n, the closer $\overline{X}$ will vary about μ.

The activities above were designed with the aim of clarifying **result 1.3** and the application of **theorem 1.4** in the study guide (which will be given combined as **theorem 7.1** in study unit 7**).** They must make complete sense! Carefully note what is given, what is different and what is the result.

Let $X_1,\ X_2,\ ...,\ X_n$ be independent $n\left(\mu;\ \sigma^2\right)$ variates (in other words a random sample from a $n\left(\mu;\ \sigma^2\right)$ population) and let

$\overline{X} = \frac{1}{n}\Sigma X_i;$ and $S^2 = \frac{1}{(n-1)}\Sigma\left(X_i - \overline{X}\right)^2.$ Then

$\overline{X}$ is an $n\left(\mu;\ \sigma^2/n\right)$ variate

$\Rightarrow \frac{\left(\overline{X} - \mu\right)}{\sqrt{\sigma^2/n}} \sim n\left(0;\ 1\right)$(which is only usable if σ^2 is known).....**Result I**

$T = \frac{\overline{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$ (which is always usable since s^2 is known).....**Result II**

Activity 2.13

Rewrite in your own words what is the meaning of results I and II.

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

Feedback  Feedback

These results specify the distribution of $\overline{X}$.
In other words, if we know (or assume) that we have a random sample from a $n\left(\mu;\ \sigma^2\right)$ population and we compute $\overline{X}$, then $\overline{X}$ will again be a random variable that is also normally distributed and for which

$$\begin{aligned} E\left(\overline{X}\right) &= \mu \\ Var\left(\overline{X}\right) &= \sigma^2/n. \end{aligned}$$

So, if $\overline{X} \sim n\left(\mu;\ \sigma^2/n\right)$ then we can standardise $\overline{X}$ as

$$Z_{\overline{X}} = \frac{\overline{X} - E\left(\overline{X}\right)}{\sqrt{var\left(\overline{X}\right)}} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}\left(\overline{X} - \mu\right)}{\sigma} \sim n\left(0;\ 1\right).$$

If we want to use $Z_{\overline{X}}$ to perform a hypothesis test or to derive a confidence interval, it implies that **the value of σ must be known**.

If not? If σ is replaced by the unbiased estimate $\hat{\sigma} = S = \sqrt{\frac{1}{n-1}\Sigma\left(X_i - \overline{X}\right)^2}$
then we cannot claim normality anymore and the standardised variate has a t_{n-1}-distribution.

So, in example 2.10 we use result II (that is the t-distribution where σ is replaced by s) and in exercise 2.8 we use result I.

Activity 2.14

Refer to exercise 2.8 in the study guide.

Show that $\left(\overline{X} - 1.96\sigma/\sqrt{n};\ \overline{X} + 1.96\sigma/\sqrt{n}\right)$ is a 95% confidence interval for θ if $X_1,\ X_2,\ ...,\ X_n$ is a random sample from a $n\left(\theta;\ \sigma^2\right)$ distribution.

.....

.....

.....

.....

.....

.....

.....

.....

Feedback Feedback

What has changed in our **notation** from activity 2.13 to activity 2.14?

We have replaced μ with θ.

Why? This is to emphasise that μ is the unknown parameter for which we have to derive a confidence interval!

What is "$U\left(\theta\right)$" for this example?

$U\left(\theta\right) = \dfrac{\overline{X} - \theta}{\sigma/\sqrt{n}} \sim n\left(0;\ 1\right)$ [In this case we assume that we know what the value of σ is!]

Please note:

For each and every confidence interval you always have to start off with

$$P\left[a \leq U\left(\theta\right) \leq b\right] = 1 - \alpha = 0.95 \text{ (if } \alpha = 0.05\text{)}.$$

From this point onwards it would be impossible to solve the problem, in other words to derive a confidence interval, if we do not have a known statistical distribution to work with! But we do! We know that $U(\theta) = Z_{\overline{X}}$ and that the $n(0;\ 1)$-distribution is well tabulated.

So, $P[-1.96 \leq Z \leq 1.96] = 0.95$

$$\therefore P\left[-1.96 \leq \frac{\overline{X} - \theta}{\sigma/\sqrt{n}} \leq 1.96\right] = 0.95$$

$$\therefore P\left[-1.96\sigma/\sqrt{n} \leq \overline{X} - \theta \leq 1.96\sigma/\sqrt{n}\right] = 0.95$$

$$\therefore P\left[-\overline{X} - 1.96\sigma/\sqrt{n} \leq -\theta \leq 1.96\sigma/\sqrt{n} - \overline{X}\right] = 0.95$$

$$\therefore P\left[\overline{X} + 1.96\sigma/\sqrt{n} \geq \theta \geq \overline{X} - 1.96\sigma/\sqrt{n}\right] = 0.95.$$

Therefore the interval $\left(\overline{X} - 1.96\sigma/\sqrt{n};\ \ \overline{X} + 1.96\sigma/\sqrt{n}\right)$ is a $100(1-\alpha)\,\% = 95\%$ confidence interval for θ.

Activity 2.15

How would you use result I (which we defined as such in the *feedback on activity 2.12*) (also see theorem 7.1(a)) to **perform a hypothesis test** (instead of deriving a confidence interval as in activity 2.14)?

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

Feedback Feedback

The test statistic is the same old $Z_{\overline{X}} = \dfrac{\overline{X} - \theta}{\sigma/\sqrt{n}} \sim n\,(0;\ 1)\,.$

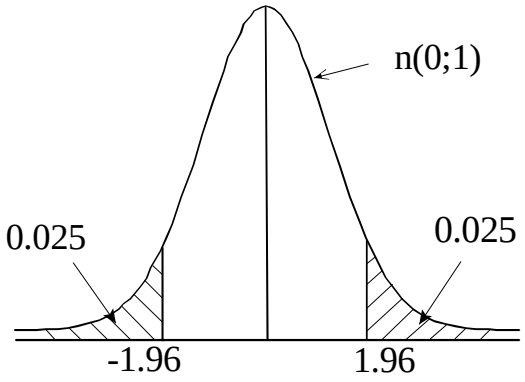
Please note that if we assume that the value of σ is known, the only other unknown value in the expression is θ. [For a specific real-life example we will know the sample size n and be able to compute the numerical value of $\overline{X}$.]

But ... As soon as we specify $H_0 : \theta = \theta_0$ (which is now our fancy mathematical way of saying θ is given a particular value) nothing is unknown in $Z_{\overline{X}}$ and for a given sample $Z_{\overline{X}}$ will simplify to a fixed numerical value (say $Z_{\overline{X}} = 0.893$).

The numerical value of $Z_{\overline{X}}$ will now be compared to the critical values of the $n\,(0;1)$ distribution that are associated with the rare event probability of α.

Suppose that $\alpha = 0.05$ and that we have two-sided testing (which implies that the alternative hypothesis is $H_1 : \theta \neq \theta_0$).

We will reject H_0 if $Z_{\overline{X}} \leq -1.96$ or if $Z_{\overline{X}} \geq 1.96$.



Do you grasp the connection between hypothesis testing and confidence intervals?
Don't feel disheartened at this stage! You will become more and more familiar with these difficult concepts as we proceed through the guide.

2.7 Simultaneous inference

In the last section of study unit 7 we will do a concrete example of simultaneous inference. You will also see that there are other different practical choices of how to perform simultaneous testing when we use statistical software.

2.8 Bayesian inference

Although we do not pursue this topic further in our module STA2601, you should take notice of this very important branch in statistics.

STUDY UNIT 3

Introduction to statistical software: JMP

3.1 Introduction

This module forms part of the new statistics curriculum and it could be the first time that you are introduced to a statistical package. It could even be the first time that you are introduced to a computer! For the highly sophisticated, computer-literate students, many of the activities in this section might be very boring. It could even be frustrating working through some of the activities but we urge you to be patient with the students who might need to catch up. You can pace yourself and *skim through the elementary computer stuff* and skip the rest of this section (including activity 3.1).

If you are using your computer for the first time

It can be very bewildering to use a computer for the first time. If you have never touched a computer before, don't panic. Focus on the end result of having mastered something new. It is most rewarding to know that you have mastered something daunting and unknown! A possible (and very reassuring option) is to ask an old hand at computers to help you. When you are learning to drive a car, there is usually somebody to help you and to explain a few basic do's and don'ts. There are also a few basic standard operations and computer terminology you should understand and know before you try to do the hands-on examples of JMP. (Keep in mind that one does not have to understand how an engine works to be able to drive a car – the same is true for a computer!)

Working through activity 3.1 could clarify the specifications (stipulated in the orientation to the study guide) regarding "access to a suitable computer".

Activity 3.1

Revision of basic computer terminology

Explain each of the following computer terms. If appropriate demonstrate how to use/do it.

hardware

software

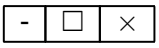
operating system

computer languages

starting your PC and navigating the desktop

click; double-click and drag

task bar

title bar and its icons 

move and resize windows

scroll bar

menu bar and toolbars

manipulate/manage files

navigate through folders

edit functions: select, copy and paste



Hardware is basically everything *you can see* and touch, provided it is attached to a computer system. For example, the **screen** in front of you (which is usually called a *monitor*), the **central processing unit** (CPU – which we could describe as the main computer box), the *keyboard* and the *mouse* are known as hardware. (This mouse does not eat cheese but is the small curved box that you put your hand on and use to point electronically at something.) Also what you *can see* inside the PC should you open it with a screwdriver, is hardware – the motherboard (a flat board connecting most other components), the disk drives, small electrical cables et cetera. If you look at the back of your CPU you will see many different connections or ports for possible connections with other hardware. There are the power supply (usually close to the fan), the connections for the mouse, the keyboard and the monitor. If you have your own printer it will be connected to one of the serial or parallel ports.

Software is a program or an application that users run on their systems to perform various tasks. For example MS Word, Internet Explorer or JMP6, that is present on the system, are categorised as software. Without software a computer system is hardly more than a box filled with a flat board containing microscopically small electrical devices and switches, implying that your hardware is useless without software! We require software to perform the various tasks we intend the computer to do – such as document writing, preparing worksheets or producing graphical images. When you turn your system on, the different software programs take over the control to make your computer perform.

An **operating system** is a special kind of software. In fact, all operating systems are software. They make it easier for you to interact with your computer. They bring it to the ready mode and from there on help you run any software you want. My computer has the WINDOWS_XP operating system (some new computers have WINDOWS 7). We mentioned in the orientation that JMP will not run on Windows 98 or ME.

Computer languages: To write a program for a computer we need to use some "language" to be able to communicate with the computer. However, since the computer does not understand simple English commands, there are special computer languages, which are required to develop software. Computer languages are categorised as high-level and low-level languages. Low-level languages work on the binary system of 1s and 0s and hence are easier for the computer to interpret. It can be learnt only if the programmer has a detailed knowledge of the hardware specifics. High-level languages are easier for the user to understand in the sense that the codes are usually very close to human language. They do not require the programmer to have detailed knowledge of the internal operations of a computer. C++, Java and Cobol are some examples of high-level languages.

Starting your PC and navigating the desktop

If you press the *power button* on the CPU your monitor will display technical detail (called startup information), looking like white text on a black background scrolling before your eyes. This will be followed by the Windows XP (or Windows 7) logo as Windows loads until it eventually displays the **desktop** with the different **icons**. Moving your mouse will make the mouse pointer (looking like an arrow point) move around on the screen. If you do this to purposely indicate and select something specific it is called navigation.

There are various operations you could perform with your mouse:

You can **click** an object to select it. This is usually one click on the left button of the mouse with your index finger. To open an application you have to point the mouse pointer over the icon and then **double-click** the object (click twice in quick succession on the left button of the mouse with your index finger). To display detail of an icon, for example additional information, or a shortcut menu or the context, you can **right-click** the object (one click on the right-hand button of the mouse with your index/middle finger). *Click and drag* is when you move something you have selected. **:-)** The best mouse training exercises are to play computer games like FreeCell, Solitaire, et cetera.

The **task bar** is normally *at the bottom* of your computer's desktop. It has the start button in the left-hand corner and will show a button for each other program that is currently running or opened. In the right-hand corner of the task bar you will find the "notification area" displaying for example the system clock. When you *open a program*, it appears as a window on the desktop. You can open more than one application or more than one document at the same time and thus have many open windows. However, **only one window can be active at any given time**. Active means it will respond to what you do with the mouse pointer or what you type on the keyboard. You will notice that the active window usually has a darker colour on the task bar than the other windows that are "dormant" on the task bar.

The **title bar** is usually the *blue area* at the top of your document where the name is displayed. It tells you what software program is in operation as well as the name of the document. The *three icons* in the right-hand corner of the title bar [_ | ☐ | ×] are from left to right the minimising button, the maximising button and the close button.
The **minimizing button** [_] minimizes your window/page – clicking on this button will not close your program but it will make the window disappear and show it as an active button on the task bar. If you click on it in the task bar it will open your window to fill the desktop again.

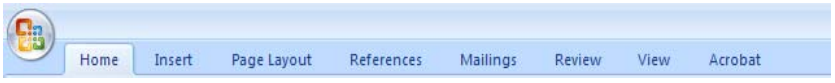
The **maximizing/restore** button [☐] maximizes your opened window. If the window is maximised, you can reduce its size by clicking on this icon.

The **close** button [×] will close your program/window – clicking on this button is the *shortcut to close your program* (and you will have to reactivate it for further use). To close a window you can also click on  and then either on **Close** or **Exit**.

Scroll bar

On the right of an open window you will find a vertical bar, called the scroll bar. In this scroll bar is a *scroll button* which you can *click and drag* to quickly move in large increments through the open window. However, at the top and bottom of this scroll bar are the *scroll arrows*: one pointing up and one pointing down to move through the document line by line (or in small increments).

The **menu bar** is usually the second line, just below the *blue area* of the title bar, at the top of your document. It does not have a standard form because the main menu depends on the specific software. For example, when you work with **Microsoft Word**, the menu bar will show the buttons:



but, when you work with another word processor for example ScientificWorkPlace, the menu bar will show the buttons:

File Edit Insert View Go Tag Typeset Tools Compute Window Help

The **toolbars** are also specifically designed for different programs, and typically appear just below the menu bar. A complex program like ScientificWorkPlace has many different toolbars which can even be "customised" in the sense that you can drag them to whatever edge of the window you prefer. (To see what possible toolbars there are for an application you can click on the **View** menu and then click on **Toolbars**.)

Each of the menus on the menu bar opens a *submenu* when you select a specific item on the main menu. For example when you work with Microsoft Word and you click on the  button, you get a *submenu* enabling you to choose from the following:

New = open a new document.

Open = open an existing document.

Save = save the opened document without changing its location or name; it saves the changes made to the original document.

Save as = save the opened document with a new name or at a different location.

Print = preview and print the currently opened document.

Prepare = prepare the document for distribution.

Send = send a copy of the document to other people.

Publish = distribute the document to other people.

Close = close recent documents.

Manipulating/managing files

For the smooth functioning of your own work on a PC, every document you create needs to be stored on your hard drive/floppy drive as a **file**. How will you organise your files in order to access them again? In older versions of Windows, we talked about directories which could be considered as virtual containers to keep your files. We now call them **folders**. (You can even talk about a subfolder when you store one folder inside another folder.)

Your computer also needs to keep its programs, functions and data organised but it is even more important to keep them protected. For the functioning of your PC there will be many system folders which you cannot open and the programs are stored in the program folder. Never try to fiddle with these files or folders! Windows XP will provide folders for the user where you can organise and save your files. It is good practice to save your files (which are not pictures, music or videos) under "My Documents" (in Windows 7, you save files under libraries, then documents).

To **navigate through folders** means that you scroll through the filing system to get to a specific document or file. Double click a folder to display its contents or click on the + sign to expand a folder that looks collapsed. Don’t panic if you cannot find a file because there is always a "search" you could perform to find the one you need!

All the edit functions that are associated with word processing programs are applicable to and helpful for statistical software programs as well. The following three functions could help in the process of creating data tables or analysis reports:

Select

You can select all text at once by clicking **Edit**, and then clicking **Select All**.

Copy

To copy text so that you can paste it in another location, select the text, and then on the **Edit** menu, click **Copy**.

Paste

To paste text you have cut or copied, click in the document where you want to paste the text, and then on the **Edit** menu, click **Paste**.

3.2 Familiarise yourself with JMP

Always keep in mind that our primary aim with STA2601 is to teach you Statistics. You must think of the statistical software as your super-duper calculator. What do you do with your pocket calculator? You use it for statistical calculations on data! You will do the same with your computer software program JMP. It is simply more sophisticated, more powerful, more structured and more organised than a pocket calculator to help you "do your statistics".

Now, to *analyse* data implies you have to *have* data. By "having data" we are not referring to methods of sampling, surveying or experimenting to obtain data but to *the capturing of values* so that you (and the computer) have them "in front of you" in the sense of "seeing them".

How can you get the data on your screen? There are basically three ways (or routes) you need to understand and they will become obvious and familiar as we move along this study unit.

The easiest way to have data is to open an existing data file. But, of course, before you can do this you have to activate JMP.

Installing JMP on your computer

Insert the CD you received from Unisa at registration. Begin installation by double-clicking the installer program, then follow the instructions given on each screen. After successful installation your desktop will display a **JMP 8 application icon** which you double-click every time you want to start the program. The installation instructions are normally given in Tutorial letter 102.

Opening an existing data file

- Double-click the JMP 8 **application icon** to start the program. Close the **Tip of the Day** window and have the **JMP Starter** window open.
- Now is the time to open your **prescribed textbook:** *Sall, J, Creighton, L and Lehman, A. (2007 4^{th} edition)* **JMP^{TM} Start Statistics.**

WORK THROUGH

Sall, Creighton and Lehman, **Chapter 2 JMP Right In**

Start reading on page 8 "*First Session*" and read up to "*Open a JMP Data Table*".

Follow the three "mouse symbol" steps under the heading *Open a JMP Data Table* but **do not continue beyond the third step** before you have done activity 3.2.

Please note:

Many beginners get confused by the textbook showing *three different possible software*

applications. $\nearrow$ Microsoft Windows
$\longrightarrow$ Macintosh
$\searrow$ Linux

(See for example figures $\nearrow$ $2.2 =$*Microsoft Windows*
$\longrightarrow$ $2.3 =$*Macintosh*)
$\searrow$ $2.4 =$*Linux*

The computer you are working on will only have one of these operating systems and Unisa's computers are mostly Microsoft Windows XP (or Windows 7) driven which means that all the "snapshots" I will be using in the rest of the guide and workbook will be for Microsoft Windows.

Please note:

If the **Look in** window does not show the folder **Sample Data**, you will have to navigate to the location where you installed your files for **JMP8**, for example C:**Program Files****SAS****JMP****8****Support Files English****Sample Data.**

Activity 3.2

With the data table open before you, (that is the *JMP file Big class.jmp)* try to answer the following questions by merely **scrolling** up and down or left to right. Do not yet click on any further menus or choose any other commands.

(a) How big is the data set? (Using statistical jargon we could ask: "How many experimental units are there?")

(b) How many variables are there?

(c) Are they all the same type of variables?

(d) Do you think it is important to think about the variable type?

Feedback  Feedback

(a) JMP refers to a data set as a *data table***,** because gathered data are now organised as rows and columns**.** It is important to notice that each horizontal line represents an independent experimental unit. If we scroll down to the end of the file we see that there are 40 cases, ending with **LAWRENCE**.

(b) There are *four different variables*. The variables are the different *vertical columns*. We cannot consider the first column as a variable since it merely serves as an identifier. The only proper variables are age, sex, height and weight appearing in columns 2 to 5.

(c) No, they are not all the same type of variables. We may classify them as age=numerical, sex=categorical, height=numerical and weight=numerical. (Revise your first-year textbooks on the classification of variables.)

Quick revision of variable types

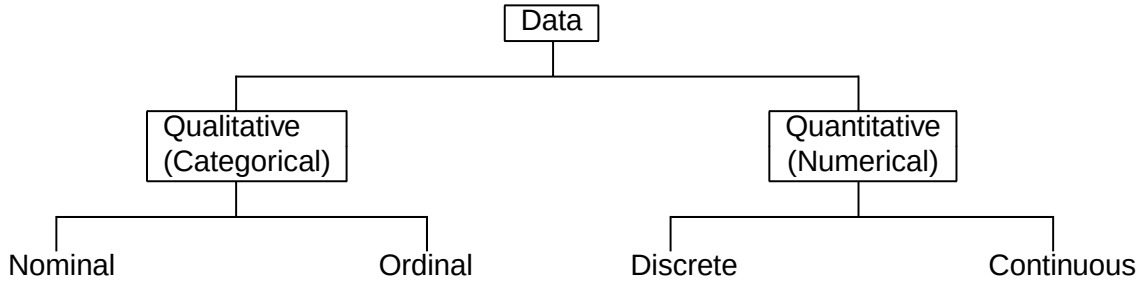
A measurement scale is something whereby some physical reality is represented by a number. Sometimes the numbers merely denote categories and sometimes the numbers are obtained by measuring or counting or calculating something.

A **qualitative variable** is the result of an experiment where the absence or presence of some or other quality characteristic is observed. Usually the outcome of such an "experiment" falls into one of k **categories**, for example hair colour; political party to which a person belongs or the gender of the individual. When we capture such data on a spreadsheet we could use symbols such as a; b; c; et cetera for the different categories or we could use the values 1; 2; 3; et cetera which are arbitrarily

assigned to each of the different categories. It is important to note that these categories have **no logical order**. This type of categorical data is also called **nominal data.**
A higher level of nominal data is where the categories have a logical or natural **order**. One category is "higher" or "better" than the next one. "Excellent" is better than "Good" and "Good" is better than "Average" – but we still don't know the magnitude of the differences between the groups. We call this type of categorical data **ordinal data** but it is still considered as *qualitative data.*

On the other hand, the term *quantitative data* implies that the characteristic is measurable on a numerical scale. Some textbooks (and computer packages) also refer to numerical variables as scale variables and they include **discrete** and **continuous** variables.

Schematically, the different data types can be represented by the following diagram:



The tips of the roots display from left to right the strength of the measurement scale.

(end of quick revision)

(d) Yes, you will see as we progress with the analyses to follow, that the type of variable dictates what can be done or not be done!

Creating your own data file

So, opening an existing JMP file was the easiest and most direct way to have a data table in front of you. Somebody else did all the laborious work to type in the data – value by value! Also falling under this method of acquiring data is to *import data from another application*, for example, importing an Excel data file, or an ASCII-file or whatever. It is difficult to teach you this method because I do not know what other software you have available. If need be you can experiment on your own to do this. However, in real life, when you are a practising statistician you will mostly have to create your own files. In the next activity you will learn to *create a new data table* and *save it* as a **.jmp-file**.

Activity 3.3

(a) **Capture** the following data in a JMP data table.

(b) **Save** the file for future use.

name	gender	age	residence	weight	height
Ashraf	2	21	2	57.2	1.68
Basadien	1	19	4	65.1	1.94
Conroy	1	24	3	45.1	1.56
De Wet	2	20	1	46.2	1.61
Ebrahim	1	18	3	63.2	1.87
Fourie	1	21	4	49.3	1.72
Gonane	2	22	5	42.1	1.6
Holomisa	1	19	2	54.8	1.83
Isaacs	2	20	5	45.9	1.55
Johnson	1	18	1	48.2	1.74
Kani	1	32	3	51.3	1.78
Levi	2	21	4	40.6	1.45

For the variable **gender**: 1 = Male and 2 = Female.

For the variable **residence**: 1 = House; 2 = Room; 3 = Apartment; 4 = Hotel and 5 = Other.

Feedback  Feedback

Step 1: Open a new data grid.

The snapshot on page 8 of the textbook (*Sall, Creighton and Lehman* **Chapter 2 JMP Right In)** shows the **JMP Starter menu**. In the previous activity we used the second button, but now we click the first button – **New Data Table** which will open a blank new grid in a new window. Notice that it shows the name **Untitled** in the title bar.

Step 2: Name the file and save it.

Chicken-and-egg situation. Do we first capture and then save or do we first save and then capture and save every now and then?

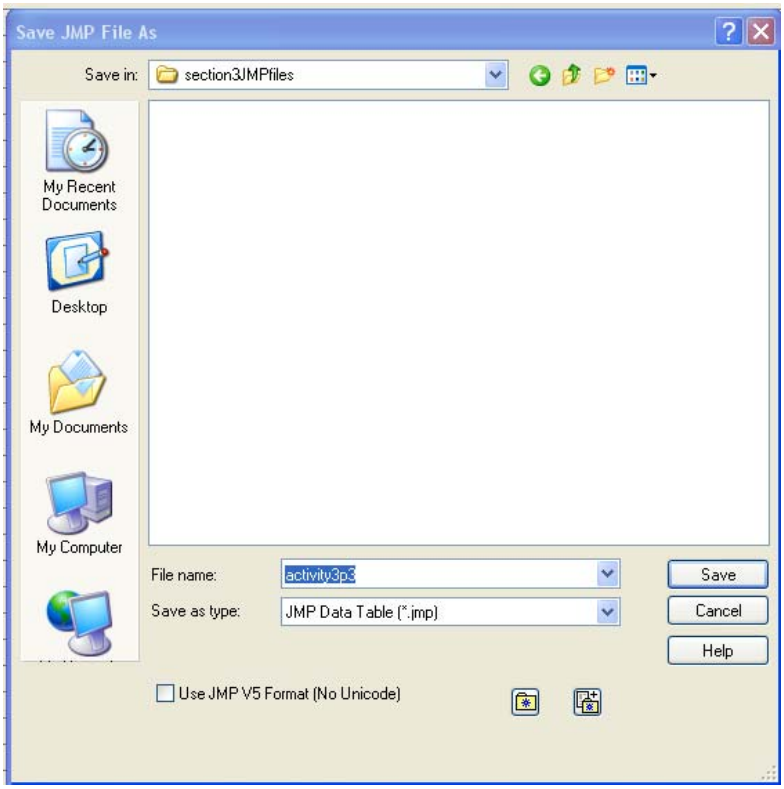
:-) I realise that each person develops his/her own style when working with or organising files. Maybe it is the frequent power failures that caused me to be obsessed with saving the work I am busy doing – hence my first step is usually to give a new blank file a name of my choice (for this example I named it *activity3p3*) and I will save it in an appropriate folder even before I start typing the data. (Either you save it on your own hard disk (**C: disk**) or directly on a flash disk if you work in one of Unisa's computer laboratories.)

[*For beginners*: *How to create a subdirectory/folder on your local C: Disk:*
Let the mouse point to **Start** and right-click to get **Explore.** With the **Start Menu** open, click on **File** ⟶ **New Folder.** (I have created a folder called *section3JMPfiles* but you are free to name it whatever you like.) Now go back to JMP.]

[*For beginners*: *How to save a file:*
When you have your blank untitled JMP grid open before you, select **File** on the main menu. Under the possible file operations you select **Save** A**s**... which will open the **Save JMP File As** dialog box.
:-) Please note: To be less long-winded going through the steps, our agreement could be that I type such a string of instructions as:

JMP [Untitled] → **File** → **Save** A**s**... → complete dialog box.
If I type "complete dialog box" it means you have to supply the computer with the necessary information in the open boxes where you see the drop-down arrows. How do you do this?

You could either directly type the information or navigate (with the drop-down arrow) to the appropriate directory. The snapshot below shows that I have navigated to the folder *section3JMPfiles* for the **Save in**: box, and I have typed in the name *activity3p3* in the **File name** box. The default for the **Save as type** box is that it will be saved as a JMP file and only where you want to export it to another application will you choose something different.



If you click on the **Save** button, your data grid page will automatically change from [**Untitled**] to [**activity3p3**]. This data table is now kept in memory and as you proceed to type your data you can simply click on the **Save** button to save your work.]

Step 3: My next step is usually to **give the columns their appropriate variable names** before I start typing the data.

You will notice that the table has only one column with the default name **Column 1**. What do we do now?

Firstly, we have to add more columns to accommodate our data which need six columns.

[*For beginners*: *How to add more columns*:

The fifth button on the main menu is **Cols**. Click this button and then select **Add Multiple Columns.**

This will activate the **JMP: Add Multiple Columns** dialog box which you have to complete. We want six columns in total and you might think to tell the computer to add five more columns, but the **How many columns to add** must be the total number you want and you have to type in **6**. Ignore the prompts of **Add Where** and **Data Type** and stick to the defaults, in other words the computer will automatically take it as:

Add Where? After last column

Data Type? Numeric.

If you click **OK** your data grid will have the default names of **Column 1, Column 2, Column 3, Column 4, Column 5** and **Column 6**. We will change the names in the next step.]

[*For beginners*: *How to change column names*:

Suppose we want to change the name of Column 1 to our first variable which is *name*. Highlight column 1 (that is, put your arrow there and click it – the colour when selected changes to blue and is not grey anymore). Now simply type **name** on your keyboard. As soon as you start typing, something interesting happens! The colour now changes from blue to white and the name starts to change! If you click on any other spot in the table your first name change is complete and your column 1 now has the heading *name*.]

Do the same for columns 2 to 6, to have the names of your columns appear as

name	gender	age	residence	weight	height

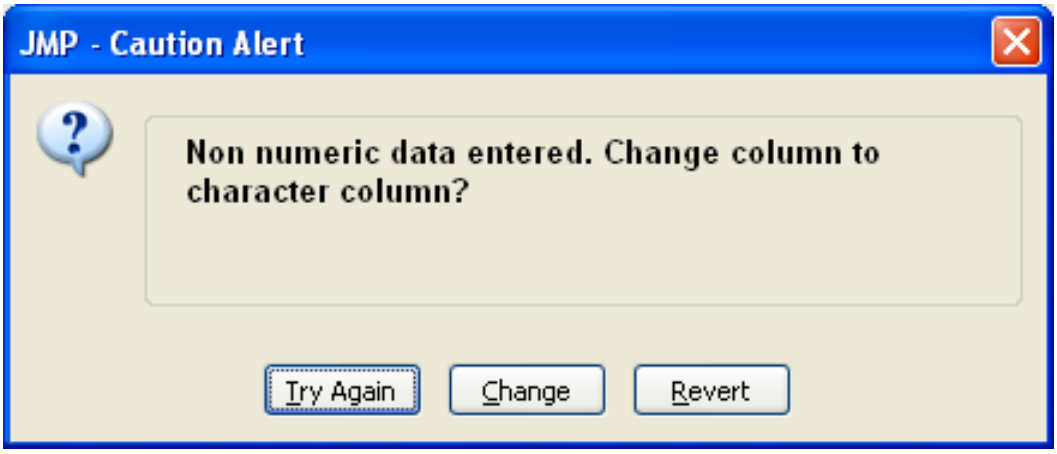
Step 4: The next step is to key in the data from the given table.

[*For beginners*: How to enter data:

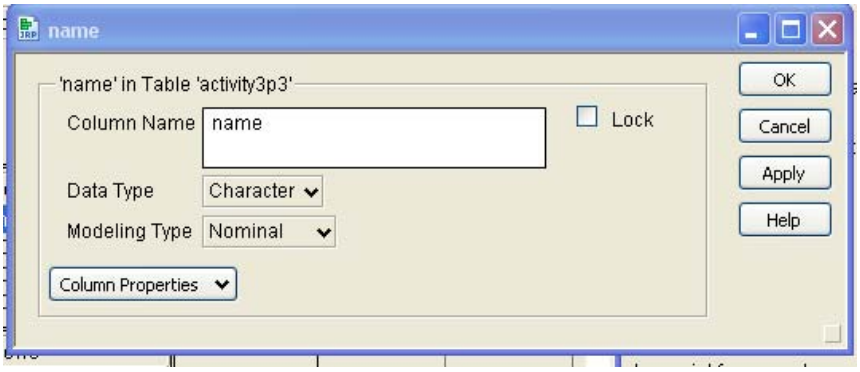
We have to type the names, values, et cetera in the appropriate **table cells.**

Move your cursor into the **first cell** of the variable **name** (in other words the first column). Double-click to get edit mode. Type the name **Ashraf**.

Do you notice what happens? The computer is not happy and you get a warning! We could have prevented this by specifying the correct data type when we named the columns in step 3 but because we ignored the prompt **Data Type?** it automatically assumed **Numeric**. Don't panic! It does not break anything if you get a warning...]



If you click the **Change** button, JMP will automatically change the data type to **Character**. (You can check this kind of variable information by double-clicking on any column. Ignore the automatic **Modeling Type** which is assumed to be **Nominal** because the first column is really only an identifier and will never be used in any kind of analysis. Click **OK** or close the window.



Resume your typing by moving with your mouse pointer (or with your arrow keys) to the next cell,that is to cell 2 in column 1. Type the name **Basadien**. What else do you notice? As soon as you start typing the names in the first column, JMP will automatically place dots in the other columns and the rows will also be numbered as you progress down the column. Continue entering data in column 1 until you have typed **Levi** in the 12th row. Now you move your mouse pointer to the first cell of column 2 and type the number **2**; move down one cell and type the number **1**; move down and type **1**, et cetera until your second column is completed and until the whole table is completed.]

Your JMP file should look like this:

	name	gender	age	residence	weight	height
1	Ashraf	2	21	2	57.2	1.68
2	Basadien	1	19	4	65.1	1.94
3	Conroy	1	24	3	45.1	1.56
4	De Wet	2	20	1	46.2	1.61
5	Ebrahim	1	18	3	63.2	1.87
6	Fourie	1	21	4	49.3	1.72
7	Gonane	2	22	5	42.1	1.6
8	Holomisa	1	19	2	54.8	1.83
9	Isaacs	2	20	5	45.9	1.55
10	Johnson	1	18	1	48.2	1.74
11	Kani	1	32	3	51.3	1.78
12	Levi	2	21	4	40.6	1.45

Question to ponder: Are you comfortable with the 1s and 2s for gender and the numerical values 1 to 5 for the different choices of residence? We will come back to the issue of *changing the properties of a column* to give **labels** to the different numerical values. You also have to think about the **Data Type** and **Modeling Type** implications of having numerical values for these variables.

To enhance all these ideas and to add some extra detail on how to create a new JMP table, you can now work through the textbook's section on this procedure.

Activity 3.4

Capture the raw data of figure 3.2 in the textbook as your own JMP data file. Name and save this file as **BPStudy.jmp**.

WORK THROUGH

Sall, Creighton and Lehman, **Chapter 3 Data Tables, Reports and Scripts**

Start reading at "*Creating a New JMP Table*" and read up to

"*Figure 3.5 Finished Blood Pressure Study Table*".

Follow all the instructions and steps under this heading but *do not continue with* "The New Column Command".

Feedback → Feedback

Your output should be the following data table:

	Month	Control	Placebo	300 mg	450 mg
1	March	165	163	166	168
2	April	162	159	165	163
3	May	164	158	161	153
4	June	162	161	158	151
5	July	166	158	160	148
6	August	163	158	157	150

I suppose that you now feel that since we have broken the ice with the preliminaries you would like to start analysing the data? Not so fast – there is a third way *of obtaining a data file in front of you***.** The third way is to *generate data*, which implies that we have sampling and data capturing, performed by the computer, all in one step.

3.3 Generating random data

JMP has various random number functions of which the Uniform and the Normal will be used most often by us. But, before we get to the technical detail of the chapter on the "Formula Editor" we will do an exercise in chapter 7 of the textbook that shows you how to obtain your own random data. In other words we will *simply open a file* that already contains the formulas used to generate the following four statistical distributions: Uniform, Normal, Exponential and Double Exponential. (The theoretical detail of these distributions will be covered in your modules on distribution theory – at this stage of STA2601 we only want to focus on the Normal distribution.)

Activity 3.5

Reproduce a table similar to the output in figure 7.5 in the textbook.
Perform the following **two steps**:

Open Randdist.Jmp. (See ***Sall, Creighton and Lehman,*** chapter 7 page 126 at the bottom.)

Choose Rows⟶Add Rows. (See page 126 at the bottom.)

(Do not continue as yet with **Analyze⟶Distribution.)**

Feedback  Feedback

- Step 1: **Open Randdist.Jmp**

[*For beginners*: To open something in JMP, you have to choose the second option on the **File** menu of **JMP Starter**. This is the option **Open Data Table**. Choosing **Open Data Table** opens a dialog box called **Open Data File**. Whenever you get a dialog box it means the computer is talking to you and needs information back from you! You have to supply various facts to the computer before you can finally click the **Open** button. (Please see the note just above activity 3.2 if you experience problems.)

1] **Look in:** Direct the computer to where the file is saved. With the *drop-down arrow* you can navigate to the directory *where the file is to be found***,** if it is not part of the possible files that are shown. For this exercise you select **JMP IN Data** which will then automatically appear in **Look in:** (Please see the note just above activity 3.2 if you experience problems.) [**Note file in Program Files\SAS\JMP\8\Support Files English\Sample Data**.]

2] **File name:** You have to supply the specific file name. (You can either *type* this or *select* it from the given list. Since the directory **JMP IN Data** shows an alphabetical list, you must scroll to the right to find the file **Randdist.Jmp**.) Click on this name to select it as the file name and it will automatically appear in **File name:**

3] **Files of type:** Select the appropriate type if it is not correct. (The extension after the dot is the file type, in this case a **.JMP** file.)

The **Open Data File** dialog box is now completed and the **Open** button may be clicked.]

:-) Please note: From the activities in *Sall, Creighton and Lehman* you must have gathered that whenever the authors of *JMP Start Statistics* want you to do something active on the computer, there will be the little *icon looking like the computer's electronic mouse* (even with a curly bit of wire looking like a real mouse tail). They have copyright on this icon (and I anyhow cannot reproduce it) so whenever I want to direct you to one of their specific action steps, I will use the expression "mouse icon".

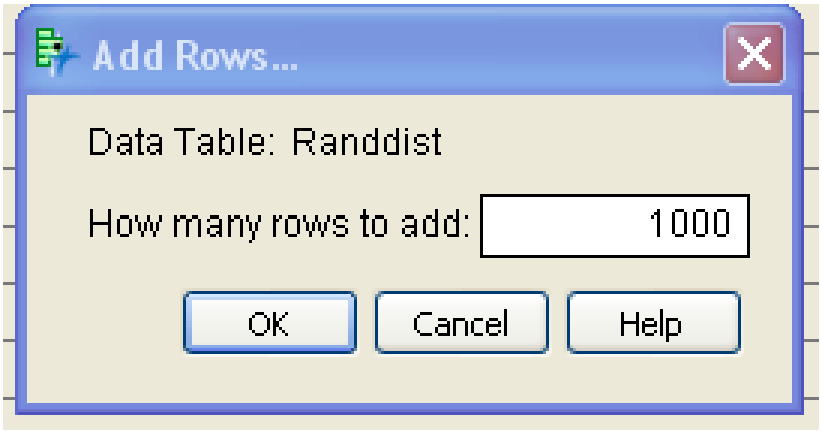
- Step 2: Now continue with the mouse icon instruction **Choose Rows⟶Add Rows.** (See page 126 at the bottom.)

[*For beginners*: On the main tool bar you will find from left to right the buttons **File**, **Edit**, **Tables**, **Rows**,**Cols**, **DOE** ... et cetera. Choose the *fourth* option on the main menu. This will open the **Rows submenu** where the last two options are

Add Rows...

Move Rows...

Click on **Add Rows** and enter the value **1000** when you get the dialog box.



I did not even bother to print my output because it is *similar* to the output in figure 7.5 in the textbook but not the same. In fact, at first glance it appears to have totally different values! I suppose the same happened to you? This is because each time the table is created it uses a different "random seed" created by the random number generator. It is as if the computer rolled a few dice to decide where to start the simulated sampling from the four specific distributions.

Generating a random sample from a normal distribution

Suppose we would like to have a random sample from a normal distribution. How could we go about obtaining such a sample?

Prompted by what we did in the previous activity, we could use the generated data set and *simply delete the three redundant columns* containing the Uniform, Exponential and Double Exponential.

Activity 3.6

Refer to activity 3.5.

Edit the data table to contain only the one column with the **Normal** as heading, in other words, *delete the three redundant columns* containing the Uniform, Exponential and Double Exponential.

This leaves you with a data table containing a random sample of $n = 1\,000$ values from a $n(\mu; \sigma^2)$ distribution.

(Do not continue as yet with **Analyze**$\longrightarrow$**Distribution.)**

Feedback  Feedback

[*For beginners*: Perform the following column operations on your data table:

Highlight the first column. The heading **Uniform** changes to the colour blue and the data values change to the colour white. (This means that the computer now considers this column as **active** and will wait for further instructions from you.)

On the main tool bar you will find from left to right the buttons **File**, **Edit**, **Tables**, **Rows**,**Cols**, **DOE** et cetera. Choose the *fifth* option on the main menu. This will open the **Columns submenu** where the last two options are

Delete Columns...

Recode...

Click on **Delete Columns** and your first column will disappear.

Do the same with the other two columns containing the headings **Exponential** and **Double Exponential**.]

Please note that it is not practical to give a snapshot of this data table since the file is too long.

Activity 3.7

Make use of the **Random Number Functions** to generate a random sample of size $n = 200$ from a normal distribution with $\mu = 100$ and $\sigma = 15$.

A similar exercise is dealt with on page 86 of the textbook.

Feedback  Feedback

[*For beginners*:
It is tricky to follow the cryptic instructions on page 86 (which in all fairness begins on page 84 with the heading "Random Number Functions"). We will do it step by step to create a sample from a normal $n(100; 225)$ distribution.

[**:-)** Please note: If X has a normal distribution with mean μ and variance σ^2 the convention in our study guides is to write it as $X \sim n(\mu; \sigma^2)$. If we know that $\mu = 100$ and $\sigma^2 = 225$ we write this as $X \sim n(100; 225)$. However, the *standard deviation* is $\sqrt{\sigma^2} = \sqrt{225} = 15$.]

- Choose **File**$\longrightarrow$**New** to create a new data table.
- Highlight the first column. (The heading **Column 1** changes to the colour blue which means that the computer now considers this column as active and will wait for further instructions from you.) Choose **Cols**$\longrightarrow$**Column Info** (the fifth option on the main menu will open the **Columns submenu** where you will find *fourth from the top* the option **Column Info**).
- Choose **Column Properties** $\longrightarrow$**Formula.** The **Column Info** dialog box has a field called **Column Properties** at the bottom of the **Column Info** box which opens a pop-up menu:

  Formula

  Notes

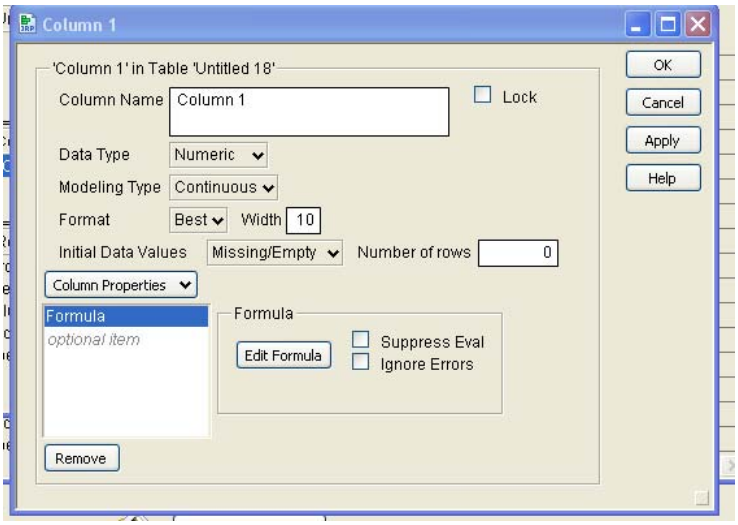
  Range Check

  $\downarrow$

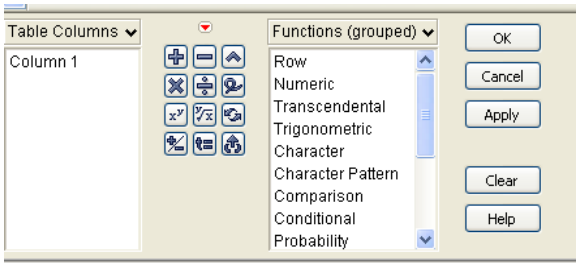
  et cetera.

DO NOT WORRY ABOUT ANY OF THE OTHER FIELDS AND DO NOT TRY TO FILL THEM!

- Selecting **Formula** enlarges the **Column Info** dialog box to show the following:



Click on the **Edit Formula** button which shows another dialog box with mathematical operators in the second column and (grouped) Functions in the third column. Use the scroll bar (or arrow) to find **Random**.



- Click on **Random** to find a pop-up menu with 18 different random functions**.** Select **Random Normal.** You will see that **Random Normal**[ ] appears in the big blank space where you have to create a formula**.**

- The **Random Normal** function will allow you to type two optional arguments. The first value you type is the **mean** of the distribution and the second value is the **standard deviation.** (If you do not supply the optional arguments the default is the $n(0;1)$ distribution.) Modify your **Random Normal**[ ] formula by double-clicking to get the edit mode. Change it to **Random Normal**$[100, 15]$. Make sure about the comma**.** Close the formula editor by also clicking OK.

- Don't panic if you see the empty first column of the data table! Although you cannot see it, you now have a table template for creating a sample from a $n(100;225)$ distribution. All you have to do is add as many rows as you want.

- You can now give this table a name and save it.

- With your blank table open, Choose **Rows** $\longrightarrow$**Add Rows** and add 200 rows.

What do you notice? Column 1 fills with 200 values which we may consider as being drawn from a normal population with $\mu = 100$ and $\sigma^2 = 225$. Every time you activate this sampling formula you will obtain a different sample.

We will in future use this sample to illustrate the **Analyze** menu of JMP.

The snapshot of the data table below is part of my output for this activity. Since it is difficult to read (and to capture the whole long column of 200 values) the 200 values are given in the condensed table.

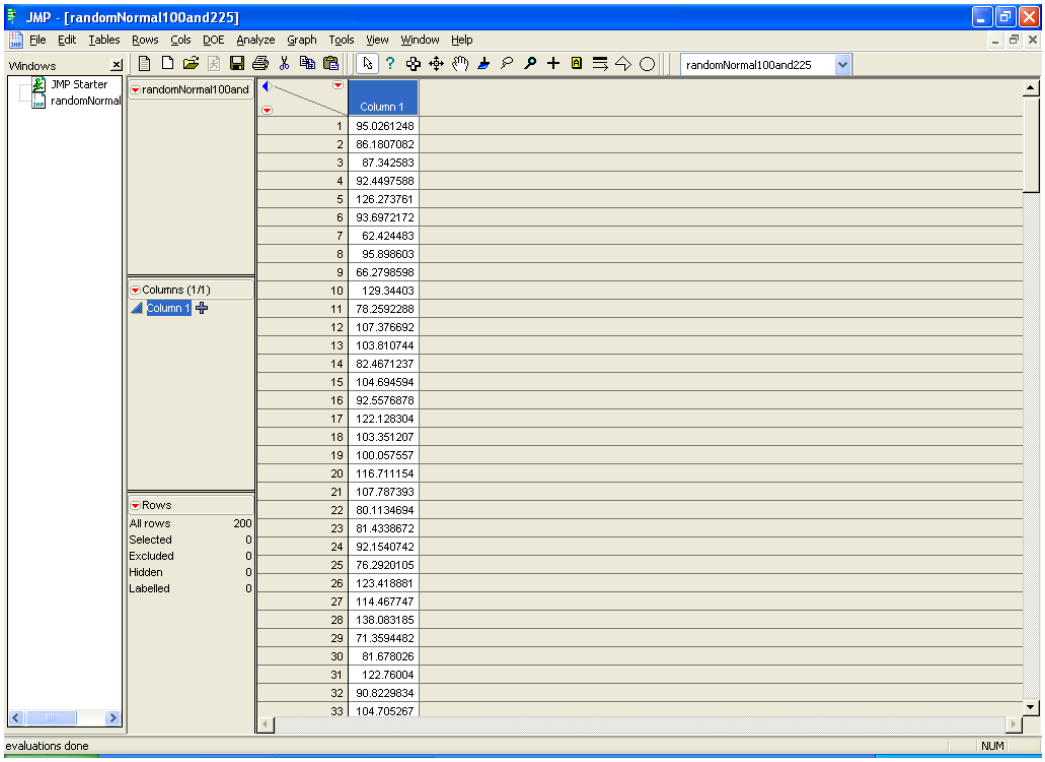


Table 3.1
Random sample of size $n = 200$ from a $n(100; 225)$ distribution

95.0261248	108.673703	105.207311	110.408931	98.9287494
86.1807082	96.2371431	106.816181	107.087295	82.90295529
87.342583	117.648554	101.143465	112.640705	105.2734719
92.4497588	72.3515123	104.007397	103.663943	103.7060315
126.273761	114.532135	115.480004	98.1667769	117.8774362
93.6972172	111.960113	104.757367	94.9082184	95.22628788
62.424483	103.720267	90.1301899	85.0839447	99.14806757
95.898603	141.104107	99.9642251	139.267817	119.396389
66.2798598	106.326517	99.9886544	99.3918981	102.4184193
129.34403	112.684672	110.858622	84.3003421	94.67080258
78.2592288	89.9198078	87.2774164	101.775315	108.8261539
107.376692	119.114798	101.275262	93.4032787	108.1526353
103.810744	105.537401	99.1559323	91.1256385	96.85622495
82.4671237	66.7534075	83.6001245	123.476955	89.6895426
104.694594	111.037693	87.4646249	90.6632368	80.52620825
92.5576878	102.564492	101.181145	87.1398378	83.05907006
122.128304	133.365777	84.5410086	72.9854585	92.24317298
103.351207	128.352053	80.2313952	74.2713204	99.72134987
100.057557	90.1970603	104.810991	74.1823075	130.6309563
116.711154	109.026082	90.1970885	89.200611	112.2280531
107.787393	106.157907	88.6430137	125.523816	89.41168103
80.1134694	110.778756	83.9120401	97.8748482	89.99950408
81.4338672	93.5407744	136.327297	77.8211061	114.2349559
92.1540742	84.0132972	104.421598	65.8111435	117.2293545
76.2920105	88.5202731	87.2445124	75.8592492	69.99247073
123.418881	108.651368	123.113381	90.6374495	78.14600043
114.467747	101.456106	95.646771	76.3839556	95.44716284
138.083185	94.5051387	110.344685	82.1228564	100.3929758
71.3594482	119.834766	77.9323831	99.9087132	92.66709382
81.678026	106.062303	73.1089372	106.766464	139.1394087
122.76004	102.337755	98.3795826	113.471127	82.39709488
90.8229834	101.591442	99.4115982	100.816091	108.5461405
104.705267	109.518999	105.527789	97.0364265	99.11854042
105.897602	77.0303965	95.9793677	82.4503798	79.25422393
85.5100171	125.286512	105.539517	106.99005	126.8929464
103.604254	83.9999384	93.2895136	106.061077	97.64914891
92.8974992	115.522558	91.3454019	95.3115182	93.40381505
98.1679818	113.289158	112.63634	121.9373	113.7370493
117.295325	99.6360811	99.2160423	114.7977	107.8258572
117.922741	123.820001	106.642021	99.7218702	87.18216867

Having a sample in front of us is the first step to all the exciting stuff that we can now finally do! For example, how can we *visualise* (in other words graphically display) the sample? How can we *summarise* the sample? What statistics can be computed for this sample?

Activity 3.8

Use your own generated sample from a $n(100; 225)$ distribution and **draw a histogram**. (I will use the data in the table above.)

Perform the **steps given in *Sall, Creighton and Lehman,*** chapter 7 page 127.

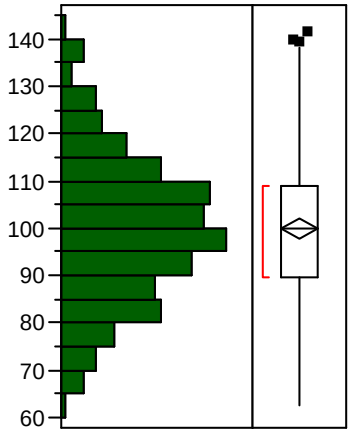
Feedback ⟶ Feedback

Aha! This is where you are finally allowed to click on **Analyze⟶Distribution.**

You will notice that the dialog box has the heading: **Report Distribution** and waits for you to specify the columns (that is the variables) for which the graphical display (histogram) of the distribution must

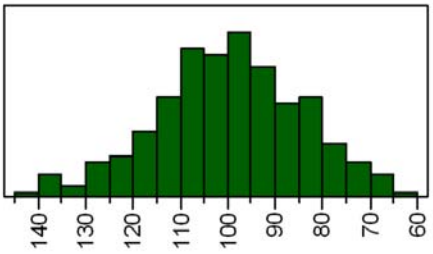
be drawn. Since we only have the one column with the name **normal**, we select this and then click on **Y, Columns** before we click on **OK**.

Did you obtain a histogram similar to the one below for your sample?



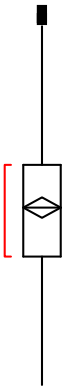
This is the *default output* you will get without having to give any additional fancy options. On the left you see the histogram as we know it from your first year, although it looks turned by 90 degrees!

[For finicky students: To look more like the horizontal graph we are used to, you can *rotate it by* 90^0 *to the left.* However, it is unnecessary, extra work and you must recondition your mind to get used to the first graph!]



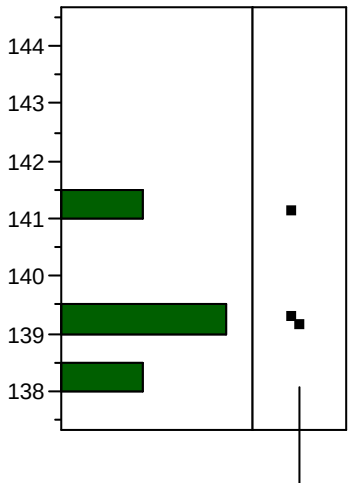
If this was a real-life exercise your interpretation of the graph is what counts and could be something along the following lines: This graph looks fairly symmetrical, it has only one peak and could pass as a bell-shaped normal distribution! (Of course we would not expect anything else because we know for sure that the sample was generated from a normal distribution!)

What about the display on the right of the *default output?*



This is an **Outlier and Quantile Box Plot** which is discussed on page 130 of the textbook.

Notice that the graph for my sample (on the previous page) has three points near the top. Single points are considered as "outliers". If this was a real-life exercise this output could be utilised in many ways. For example, it could be seen as a warning of possible errors when you captured your data set. If you move your mouse over the points, it will label the rows where these outliers occurred (it directed me to rows 48, 128 and 190) and you could then go back to your data table to make sure about the entering of the data. We know that for our specific example these points are not errors but only that they represent the individual x-values of a normal distribution associated with very small probabilities of occurring. You could further click on the **Magnifying glass** tool button to enlarge the area of the output (which automatically enlarges the corresponding histogram) to show the intervals where these values lie. (Read more about "Special Tools" on page 16 of the textbook.)



If you want to have the exact values of the "box" for the *Outlier and Quantile Box Plot*, you will find them under **Quantiles** in the output. For this sample we see that the middle 50% of the values lie between $Q_1 = 89.48$ and $Q_3 = 108.98$.

Quantiles		
100.0%	maximum	141.10
99.5%		141.09
97.5%		136.25
90.0%		121.73
75.0%	quartile	108.98
50.0%	median	99.82
25.0%	quartile	89.48
10.0%		79.34
2.5%		70.03
0.5%		62.44
0.0%	minimum	62.42

Activity 3.9

Use your own generated sample from a $n(100; 225)$ distribution (I will use the data in table 3.1 above) and compute the following sample statistics:

(a) The mean

(b) The standard deviation

Feedback ⟶ Feedback

I hope you noticed that we automatically obtained a summary of some of the important statistics when we performed **Analyze**⟶**Distribution**? Scroll down to the output labelled **Moments.**

Moments	
Mean	99.903554
Std Dev	15.714819
Std Err Mean	1.1112055
upper 95% Mean	102.0948
lower 95% Mean	97.712305
N	200

We will come back to discuss this output in detail in further study units. At this stage you may report back that

Mean $= 99.903554$

Std Dev $= 15.714819.$

I trust that your first sessions with JMP8 got you excited and keen to know more about statistics!

STUDY UNIT 4

Testing for normality and goodness-of-fit tests in general

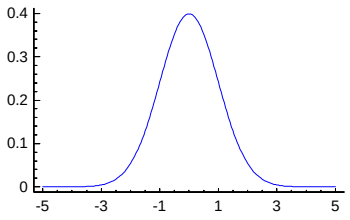
4.1 Introduction

In study unit 3 we discussed the behaviour of a sample from a normal distribution when we looked at the generated sample. Of course the generated sample from the $n(\mu;\sigma^2)$ population would behave as one would expect from such a sample because it came from a normal distribution! What tests could be applied to an unknown sample to decide whether it could pass for a "normal distribution sample"? In this study unit you will formally learn about graphical and analytical techniques to test for normality.

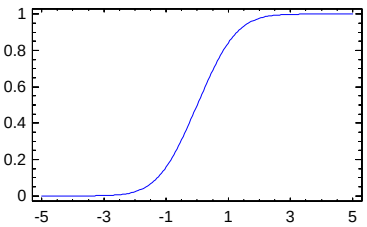
4.2 Graphical techniques

The words "graphical techniques" already indicate that we are going to conclude something from a picture. More specifically, we are hoping that the *picture of the sample* will resemble the *picture of the theoretical model.* What pictures do we have for the theoretical model of a normal distribution? In section 1.3 of study unit 1 of the study guide, we showed the two possible graphs that capture the properties of a standardised normal distribution (see figures 1.6 and 1.7).

$\phi(z)$ $\Phi(z)$



probability density function



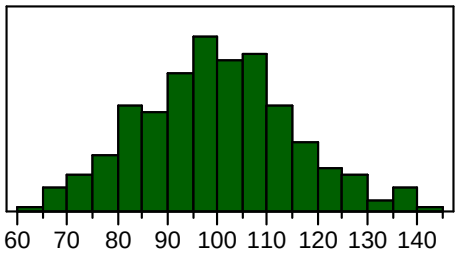
distribution function

What typical graphical displays do we have of sample data?

We saw in the previous study unit that JMP automatically displays the *histogram*.

Some people feel that if the form of the histogram looks more or less like the bell-shaped curve of the probability density function, we may conclude that the sample comes from a normal population. Is this comparing apples with apples? In other words is it fair to compare a histogram with the picture of the probability density function? We know that the total area under the probability density function equals one and if we divide the observed frequency of each interval of the histogram by the sample size, it would feel like a conversion of a histogram to a probability density function. So, it makes intuitive sense that the picture of the relative frequency histogram could be compared to the picture of the probability density function (provided that the sample is large enough for relatively narrow intervals (the bigger size taking care of non-empty intervals). Vice versa, we could convert the probability density function to a nonstandardised version fitting over the histogram. This is what JMP does when we perform the procedure of *superimposing a normal curve over the histogram.*

For example, in activity 3.8 we obtained the following histogram for the generated sample from the normal $\mu = 100; \sigma^2 = 225$ population:



By simply observing the *form of the histogram*, we have our first intuitive graphical test. Of course this histogram looks very symmetrical and bell-shaped because the sample came from a normal population! Now you will see how this conclusion can be enhanced by superimposing a normal curve over the histogram.

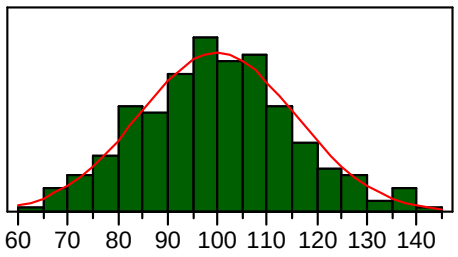


Figure 4.1

Activity 4.1

Use your own generated sample from a $n(100; 225)$ population (for which you already drew a histogram) and **superimpose a normal curve** over the histogram.
You have to repeat activity 3.8 to have the histogram displayed as output.
Perform the steps **given in *Sall, Creighton and Lehman,*** chapter 7 page 128.

Feedback  Feedback

[*For beginners*: How to superimpose a normal curve over the histogram:
Have your own generated sample of size $n = 200$ from a $n(100; 225)$ population open as the active data table.
Click on **Analyze**$\longrightarrow$**Distribution.**

In the dialog box **"Distribution"** select the column with the name normal and click on **Y, Columns** before you click on **OK**.

Click on the histogram bar (there is a red triangle before the variable name normal where you have to click. This opens a submenu where **Fit Distribution** opens the list of possible choices where you select **Normal**.]

We all agree that the normal curve fits the histogram rather well. When will you decide that the fit is not so good? There could be situations where one person will say the fit is OK but the other person might say the fit is not OK! The graphical method is rather subjective. To see what happens when a sample is not normally distributed, we have the following activity:

Activity 4.2

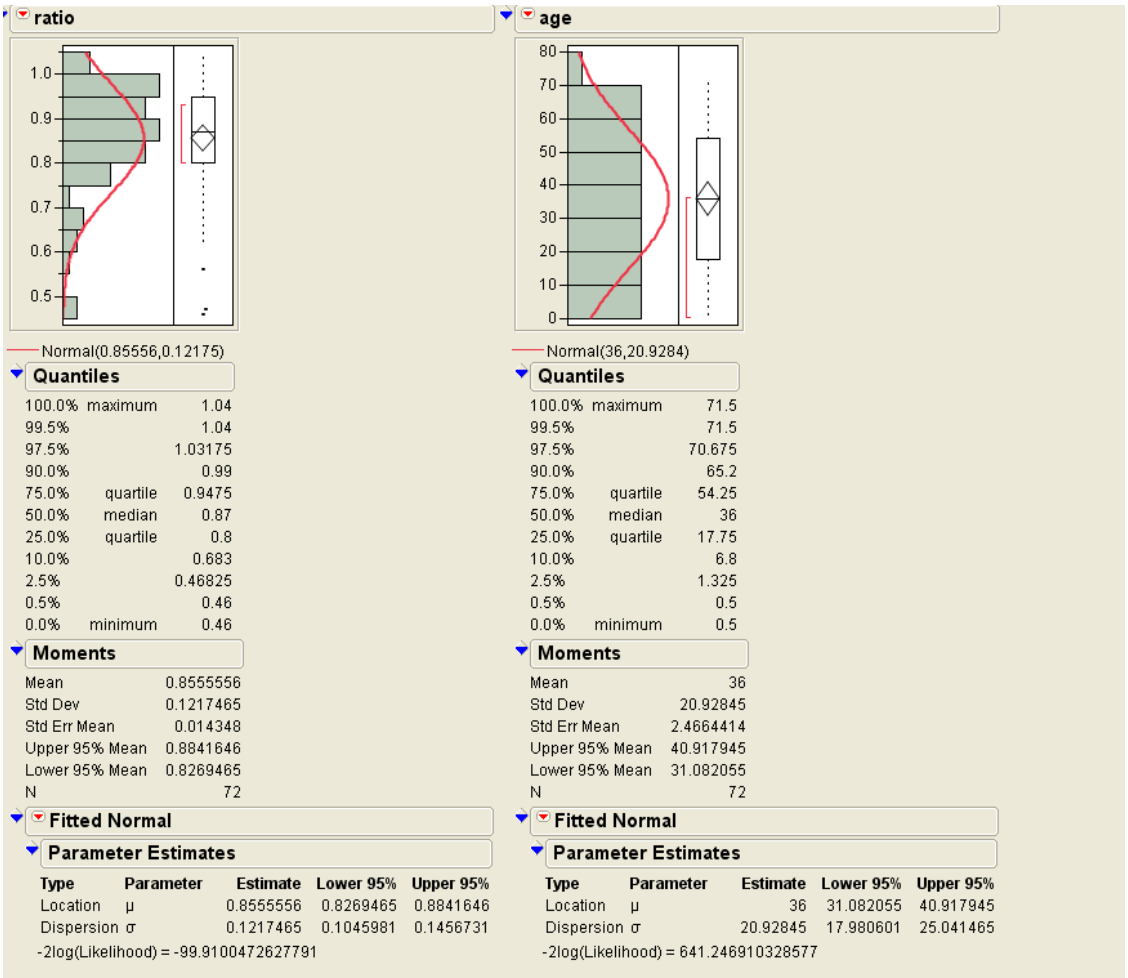
Open the data set *growth.jmp* from the **JMP Sample Data** and repeat activity 4.1 for this sample. **Superimpose the normal curve** over both the histograms for the variables "*ratio*" and "*age*".

Feedback ⟶ Feedback

[*For beginners*: How to reproduce the output given below.
Open the data set *growth.jmp* from the **JMP Sample Data.** If the **Look in** window does not show the folder **Sample Data**, you will have to navigate to the location where you installed your files for **JMP8**. For example **Program Files\SAS\JMP\8\Support Files English\Sample Data.**
Click on **Analyze⟶Distribution.**

With the dialog box **"Distribution"** open, select (highlight) both the variables "*ratio*" and "*age*" and click on **Y, Columns** before you click on **OK**. This will result in the simultaneous output for two variables showing two histograms alongside each other.

Click on the histogram bar of "*ratio*" (there is a red triangle before the variable name "*ratio*" where you have to click) and open the submenu **continuous fit** where you select **Normal**. Do the same for the variable "*age*".]



Do you agree that neither of the normal curves fit the histograms? (Please note that this was purely an illustration of variables that do not look normal and we are not going to discuss how the sample was drawn or what the experiment entailed.)

What did you learn from the activities above?

Making a decision about the fit of a normal distribution based on the histogram as graphical output, is a rather subjective process because there could be situations where one person will say the fit is good enough but the other person might say the fit is not good enough!

So far we have basically only focused on the *probability density function* of the normal distribution for comparison purposes. Do you think that the second possible graphical display (ie the *distribution function*) of a normal distribution could lead to a more objective conclusion?
It is a bit more complicated to comprehend what kind of pictorial comparison we could make if we want to employ the **distribution function.** You might argue that we could simply draw the cumulative relative frequency distribution of the sample (called the **ogive**) and then also superimpose the distribution function of the normal distribution over this graph. However, correct as this might feel, it is even more difficult to make a conclusion of "a good fit".

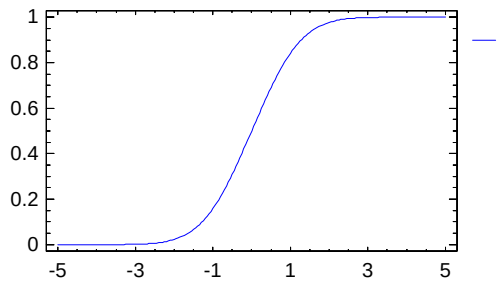
What now? We could try to change the *curve* to a *straight line*, as is the trick with the special probability paper, because it is easier to detect deviations from a straight line or we could follow a "different angle" of employing the distribution function which leads to the "**Normal Quantile Plot**".

The following activities are designed to enhance your understanding of this technique.

Activity 4.3

[:-) This is an **optional activity** – for the inquisitive, mathematically minded students!]
The figure shown below is simply a reprint of figure 1.7 of study unit 1 of the study guide, and shows the standardised normal distribution function $\Phi(z)$. On the vertical axis we see $\Phi(z)$ and on the horizontal axis we see z.

$\Phi(z)$



How is it possible that this curved graph could be changed into a straight line as depicted in figure 4.1 of study unit 4 of the study guide?

.....

.....

.....

.....

.....

.....

.....

.....

.....

Feedback → Feedback

Look carefully at the scales of the two graphs. What do you notice? In figure 1.7 of the study guide we have a *linear scale* on both the X-axis and the Y-axis, but in figure 4.1 of study unit 4 the scale of the Y-axis has been **distorted.**

To really compare the two graphs (ie apples with apples!) you need to keep in mind that figure 4.1 of study unit 4 shows $100\Phi(z)$ and not just $\Phi(z)$ as in figure 1.7 of the study guide. Another difference is that $100\Phi(z)$ of figure 4.1 starts at 0.01, which would imply that $\Phi(z)$ should start at $\frac{1}{100}(0.01) = 0.0001$. So to compare the two we either need to divide the scale of figure 4.1 by 100 or we need to multiply the scale of figure 1.7 by 100.

In figure 1.7 we have a *linear scale*. This means that the distance on the scale from 0.0 to 0.2 is the same as the distance on the scale from 0.2 to 0.4, et cetera. (Should we multiply by 100 to compare with figure 4.1 it means that the distance on the scale from 0 to 20 is the same as the distance on the scale from 20 to 40, is the same as the distance on the scale from 40 to 60, et cetera.

In figure 4.1 we have a *nonlinear scale*. Do you agree that the distance on the scale from 0.01 to 20 is not the same as the distance on the scale from 20 to 40 which is also not the same as the distance on the scale from 40 to 60, et cetera? This clever distortion caused the curved graph of the normal distribution function to look like a straight line as depicted in figure 4.1. All this trouble because it is easier to detect deviations from a straight line when you plot $100\Phi(z) = P(Z \leq z)$. This graphical technique would imply that we take the theoretical straight line of $100\Phi(z) = 100P(Z \leq z)$ as the "gauge" to measure how well our sample compares to a normal distribution.

Basically the only difference between the technique of using the special graph paper and drawing a "**Normal Quantile Plot**" is that the first technique uses a nonlinear scale (figure 4.1) and the **Normal Quantile Plot** converts the same values to be plotted on a linear scale!

To enhance your understanding of the concept "**percentile**" and the relationship between $\Phi(z) = P(Z \leq z)$ and $\Phi^{-1}(z)$ you should do the following two activities.

Activity 4.4

(a) Show that $x = \mu + \sigma$ is the $84th$ percentile for any $n(\mu; \sigma^2)$ distribution.

(b) Show that $x = \mu - \sigma$ is the $16th$ percentile for any $n(\mu; \sigma^2)$ distribution.

.....

.....

.....

.....

.....

.....

.....

.....

.....

Feedback  Feedback

From our first-year modules (descriptive statistics) you learned that a percentile value of a quantitative variable is a value that divides the ordered data into groups so that a certain percentage is above and another percentage is below. For example, the 95th percentile is the value below which 95% of the observations fall. (We mainly focused on the quartiles (ie the 25th, 50th and 75th percentiles) in first-year courses.)

For a theoretical distribution we simply translate "below which 95% of the observations fall" to $P(X \leq x) = 0.95$.

(a) $P(X \leq \mu + \sigma) = P\left(\frac{X-\mu}{\sigma} \leq \frac{\mu+\sigma-\mu}{\sigma}\right) = P(Z \leq 1) = \Phi(1) = 0.84$ (using table 1 Stoker)

(b) $P(X \leq \mu - \sigma) = P\left(\frac{X-\mu}{\sigma} \leq \frac{\mu-\sigma-\mu}{\sigma}\right) = P(Z \leq -1) = \Phi(-1)$ (cannot use table 1 Stoker directly)

now $\Phi(-1) = 1 - \Phi(1) = 1 - 0.84 = 0.16$

Activity 4.5

(a) Show that $\Phi^{-1}(0.05) = -1.645$. How do you interpret the value -1.645?

(b) Show that $\Phi^{-1}(0.15) = -1.036$. How do you interpret the value -1.036?

Feedback Feedback

(a) $\Phi^{-1}(0.05)$ means we have to find a z-value such that $P(Z \leq z) = 0.05$.

Now we already know that whenever the given probability is smaller than $\frac{1}{2} = 0.500$ it is *extra effort* to use table II of Stoker!

We also know that the final z-value will definitely be a *negative value*.

We find the value by working with the positive mirror image:

$P(Z \geq +z) = 0.05 \Longrightarrow P(Z \leq +z) = 1 - 0.05 = 0.95$. Since this given probability is now larger than 0.500 we can look up the value in table II of Stoker and find $+z = 1.645$. So, the value we were looking for is -1.645.

We interpret the value -1.645 as the $5th$ percentile for the $n(0;1)$ distribution.

(b) Similarly, $\Phi^{-1}(0.15)$ means we have to find a z-value such that $P(Z \leq z) = 0.15$.

Again we have to find the value by working with the positive mirror image:

$P(Z \geq +z) = 0.15 \Longrightarrow P(Z \leq +z) = 1 - 0.15 = 0.85$.

We can look up the value in table II of Stoker and find $+z = 1.036$. So, the value we were looking for is -1.036.

We interpret the value -1.036 as the $15th$ percentile for the $n(0;1)$ distribution.

It is easy to convert the α-th percentile for the $n(0;1)$ distribution to the corresponding α-th percentile for the $n(\mu;\sigma^2)$ distribution by the reverse standardisation. In a real-life example you will draw a scatter plot of the ordered observed sample values versus the nonstandardised percentiles of the normal distribution and call it a "**normal quantile plot**". This is at long last what you are going to do in the next activity!

Activity 4.6

Use your own generated sample from a $n(100;225)$ population and draw a "**normal quantile plot**". (You have to repeat activity 3.8 to have the default output. I have used the data of table 3.1 to produce the output in the feedback.)

Perform the steps **given in *Sall, Creighton and Lehman,*** chapter 7 page 152.

Feedback ⟶ Feedback

[*For beginners*: How to draw a "**normal quantile plot**:

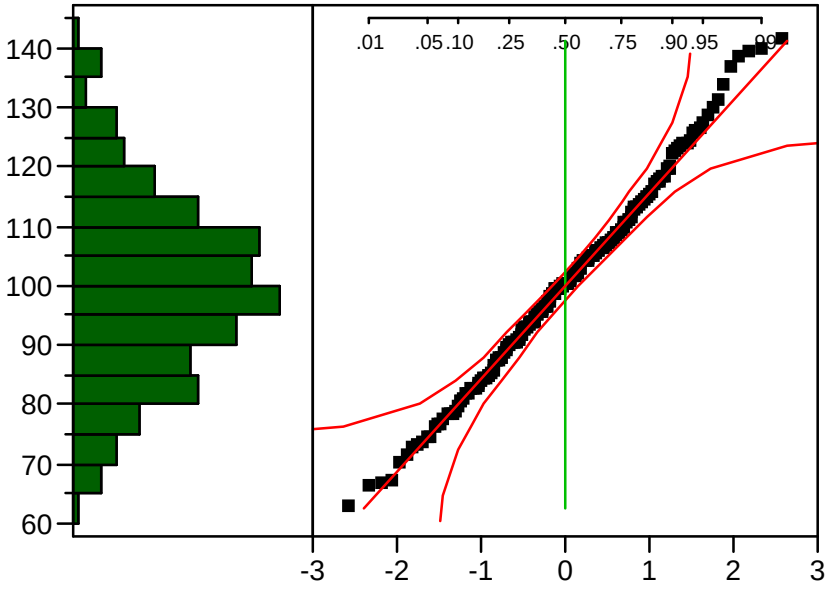
Have your own generated sample of size $n = 200$ from a $n(100;225)$ population open as the active data table.

Click on **Analyze**⟶**Distribution.**

In the dialog box **"Distribution"** select the column with the name normal and click on **Y, Columns** before you click on **OK**.

Click on the histogram bar (there is a red triangle before the variable name *normal* where you have to click. This opens a *submenu* where **Normal Quantile Plot** is the third of the possible choices which you select.]

To simplify your graphical output you can also *de-select the fourth option* on the same submenu which will then not produce the **Outlier Box Plot**. (It is usually included by the default setting). Doing this will show the following:



Normal Quantile Plot

As we would have expected, the normal quantile plot shows a perfect linear graph with no systematic deviation from the straight line.

4.3 Goodness-of-fit test for normality

The appropriate test statistic is the chi-square statistic

$$\chi^2 = \sum_{i=1}^{k} \frac{(\text{Observed frequency} \ - \ \text{Expected frequency})^2}{\text{Expected frequency}} \ .$$

By specifying the values of the parameters of the type of distribution (ie of the specified model under H_0) the probabilities of the intervals (or the categories into which the observed data were classified) become fully known. "Expected frequency" $\equiv$ "probability $\times$ sample size"

If we replace the unknown parameters by sample estimates and we still compute the expected frequencies for a specified type of distribution it feels like cheating, doesn't it? (We are almost tailoring the null hypothesis to make it more like the sample; and the fit then appears to be better than it really is.) We compensate for this "cheating" by reducing the number of degrees of freedom to $k - 1 - r$ and thus the critical region is shifted to the left (ie closer to zero) as compared with that of χ^2_{k-1}.

The **null hypothesis** must always **specify the distribution** that is being tested. In other words, you **know** that you must test for a normal distribution. (Or as you will see in the next section, for an exponential distribution or whatever the case might be.)

The typical table that you have to compile, before you can compute the value of the χ^2-test statistic, is usually

	These values are obtained (or given) by classifying the data ↓	*These values are obtained from the theoretical distribution* ↓
Intervals	Observed frequency	Expected frequency
interval 1	N_1	$n\pi_1$
interval 2	N_2	$n\pi_2$
$\vdots$	$\vdots$	$\vdots$
interval k	N_k	$n\pi_k$
Total	n	n

Always check if they are the same.

There is a crucial "connection" between these two columns. The one follows from the other.

So, what is the connection between the **intervals** and the **expected frequency**? One of the following two situations is possible:

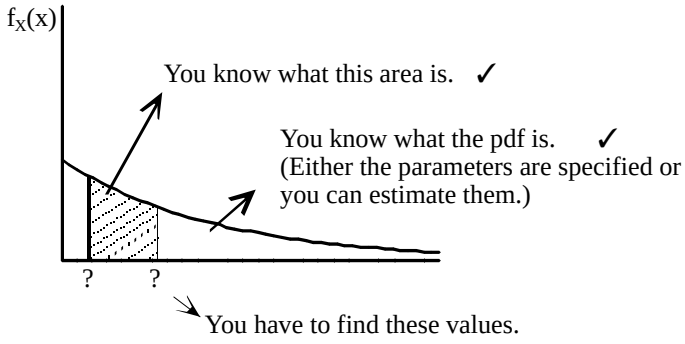


Figure 1:

Situation A:

The **expected frequencies are given** (in most cases with the wording "**equal** expected frequencies"), **BUT** you do not have the limits for the intervals and thus you have to compute them. (This of course would then lead to the classification of the raw data which gives you the observed frequencies.)

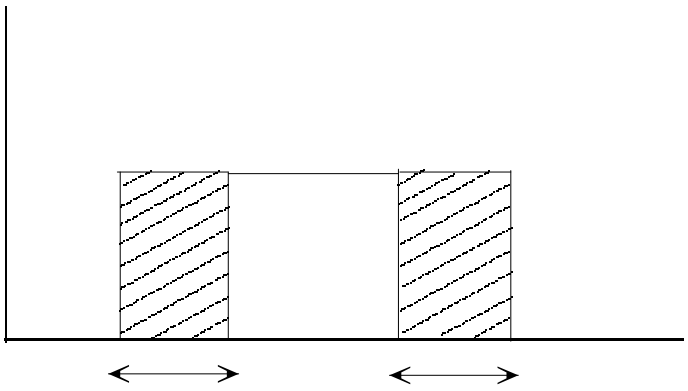
(This is the case with examples 4.3 and 4.4 and the "overview exercise" 4.1 question 6 of the study guide.)

HOW?

So, if there are k **equiprobable intervals**, the area for each interval under the pdf will be $\frac{1}{k}$.

[**Please note**: *Do not fall into the trap* of trying to create your own "*shortcut*" by simply dividing the RANGE ($R =$ largest$-$smallest value) into chunks of equal length to create the corresponding number of intervals! This will lead to the wrong limits.

Why not? If the areas are the same (ie they are *equiprobable*) it does not mean that the intervals must also be of the same length!! It all depends on the shape of the pdf. Only if the pdf has a *uniform distribution*, in other words it looks like this



will "equal probability" imply "equal interval length".]

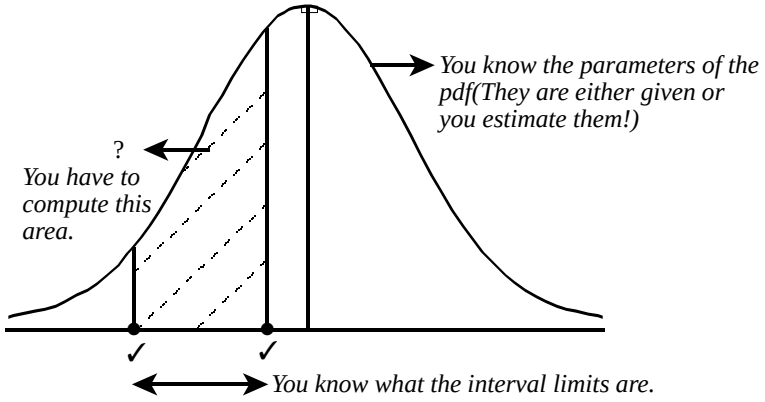


Figure 2:

Once you have computed the interval limits, you are able to classify the raw data into these intervals and thus obtain the observed frequencies. Some of the examples/exercises/examination questions falling in this section perform this task with the words: "Suppose that the observations are classified into these intervals, and we get the following:" but you must always keep in mind that in real-life situations you (or your computer) will have to do the classification/tallying yourself.

Situation B:

The **specific limits for the intervals are given** (and most often the data have been classified into these intervals or you have the original raw data and thus you are able to classify them into these intervals), **BUT** you do not have the expected frequencies and you have to compute them. (This is the case with example 4.5 and with exercise 4.1, questions 3 and 5 of the study guide.)

HOW?

Well, from the null hypothesis you know what the pdf is, and if you know what the intervals are, it means that you are able to compute the corresponding **areas under the curve**.

(Please note that exercise 4.1, question 4 of the study guide is the discrete equivalent of this situation!)

Activity 4.7

[:-) This is an optional activity!

You need only work through this activity if you feel unsure or unhappy about the class limits given in **example 4.5 of the study guide**.]

Refer to example 4.5 of the study guide.

Explain how to derive the class limits of the given frequency table.

Feedback ⟶ Feedback

Compiling a frequency table by hand (ie not using a computer) is rather cumbersome! It is not essential to order the values from small to large (like we luckily have in example 4.3 of the study guide) but it helps with the tallying and it also cross-checks that the smallest and the largest values are included.

Where do we start?
You start with the number of intervals. For this specific case it is already given to be ten classes. (You learned in first-year courses that there are formulae to guide you to roughly compute the number of classes.)

This means that **interval length** = $\frac{\text{(largest value } - \text{ smallest value)}}{\text{number of intervals}} = \frac{68-32}{10} = 3.6$ which we round to 4.

When designing a set of intervals for a data set you should take care of the following three points:

- There must be no doubt in which interval a specific observed value must be classified.
- No observation could ever be classified into more than one interval.
- All observations must be classified.

Please note that the *accuracy* of the recorded data determines the way we write down the interval limits. Since our data are recorded *correctly to one decimal place*, in other words it could include values such as 43.6; 56.2; 61.9, et cetera we need to make sure that our limits reflect this accuracy. What does this mean? Say for example we create a column with ten lines and we begin this process by first writing down the *lower bound* for each interval in each new line. It is easy to start with 30 in the first line and simply add four to get 34 for the next line, et cetera. So each lower value is simply the previous one plus four which means 30; 34; 38; 42 and so on, up to 66. **But, what must we write down for the upper value?** Because we assume that our data were measured as a real continuous variable we could (for starters!) use the following set of class intervals: 30 and under 34; 34 and under 38; 38 and under 42 but already bringing in the *one decimal place*. This leads to

Interval
30.0 and under 34.0
34.0 and under 38.0
38.0 and under 42.0
42.0 and under 46.0
46.0 and under 50.0
50.0 and under 54.0
54.0 and under 58.0
58.0 and under 62.0
62.0 and under 66.0
66.0 and under 70.0

Solely for the purpose of compiling a frequency table, the above set of intervals will work fine and we will experience no problems to classify values like 43.6; 56.2; 61.9, because (for example) 43.6 will fall in the fourth interval and 56.2 will fall in the $7th$ interval. Now we could refine the intervals by writing *30.0 and under 34.0* as $30.0 - 33.9$ and *34.0 and under 38.0* as $34.0 - 37.9$,et cetera. This means that we write down the upper class limit of one interval as close as possible to the lower limit of the following interval – chosen such that the intervals are still unambiguous.

Interval
30.0 – 33.9
34.0 – 37.9
38.0 – 41.9
42.0 – 45.9
46.0 – 49.9
50.0 – 53.9
54.0 – 57.9
58.0 – 61.9
62.0 – 65.9
66.0 – 69.9

Can we stop here? Unfortunately not! These intervals are not useful for a histogram. What we actually need for a graph *without any gaps* is where the upper limit of one class becomes the lower limit for the following class. We achieve this by finding a value halfway between the two values of the gap, in other words halfway between the possible values that the measured variable could take.
Halfway between 33.9 and 34.0 is 33.95.
So, our *final set of intervals* that have an interval length of four, that have no gaps and are unambiguous (meaning you do not wonder where you have to classify a value!) are the following:

Class intervals
29.95 – 33.95
33.95 – 37.95
37.95 – 41.95
41.95 – 45.95
45.95 – 49.95
49.95 – 53.95
53.95 – 57.95
57.95 – 61.95
61.95 – 65.95
65.95 – 69.95

What about JMP?

You must have been conditioned by now that we usually repeat what is done **manually** in the study guide by doing it **electronically on JMP in the workbook! How will you perform a goodness-of-fit test using JMP?** You will notice that there is not an option of a *chi-square test statistic* in

JMP, but only the option of the *Shapiro-Wilk W test.* Discussing the mathematics of this test falls beyond the scope of this module. It must however make intuitive sense to you and you must be able to interpret any output in this regard. Think of W as a measure of the "straightness of the normal quantile plot". To conceptualise this should be no problem because we have discussed these plots in detail! Think of the Shapiro-Wilk W test as something similar to the squared correlation coefficient between the ordered sample values (X_i^*) and the expected normal quantile values. It is difficult to explain the computation of the p-value if the sampling distribution of the test statistic is not explained but you have to accept the fact that a very small p-value (ie p-value $< \alpha$) rejects the null hypothesis that the sample comes from a normal population at the α-level of significance.

Activity 4.8

(a) Capture the raw data given in *example 4.3* of the study guide in a *JMP data table.*

(b) Save the file as *example4p3.jmp* for future use.

(c) Draw a histogram and superimpose a normal curve over the histogram. (Refer to activity 4.1 to refresh your memory.)

(d) Draw a normal quantile plot.

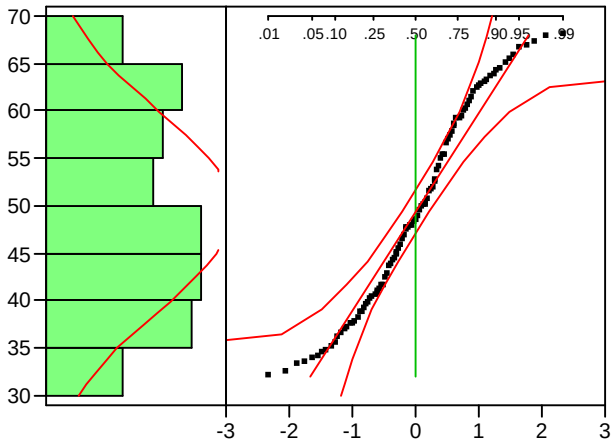
(e) Perform a Shapiro-Wilk W goodness-of-fit test for normality.

(f) What can you conclude from all the output?



Numbers (a) and (b) are not displayed. I called my variable X.

Numbers (c) and (d) produced the following output:



Normal Quantile Plot

From both these plots it *seems unlikely that the sample comes from a normal distribution*. The normal curve does not fit the histogram well. (The bars are too high in the tails and too low in the middle.) From the normal quantile plot we see *systematic deviation* from the straight line. The points on the far left are all above the line, then they are all below the line before they seem to cross the line and are all again above the line with the points on the far right all below the line.

(e) [*For beginners*: How to perform a Shapiro-Wilk W goodness-of-fit test for normality:

After you have performed the steps to produce the histogram, to superimpose a normal curve and to draw a normal quantile plot, you have to *right-click* on the **Fitted Normal** bar (scroll down if you do not see it immediately) to get the following submenu:

Diagnostic Plot
Quantile Plot
Density Curve
Goodness of Fit
Fix Parameters
Quantiles
Set Spec Limits for K Sigma
Spec Limits
Save Fitted Quantiles
Save Density Formula
Save Spec Limits
Remove Fit

where you select **Goodness of Fit**.

This will produce the Shapiro-Wilk W goodness-of-fit test output.]

Fitted Normal

Parameter Estimates

Type	Parameter	Estimate	Lower 95%	Upper 95%
Location	μ	49.457	47.383363	51.530637
Dispersion	s	10.450656	9.1757476	12.140269

Goodness-of-Fit Test

Shapiro-Wilk W Test

W	Prob<W
0.951064	0.0010*

Note: Ho = The data is from the Normal distribution. Small p-values reject Ho.

Since the p-value "**Prob**$<$**W**" is less than $\alpha\,(0.001 < 0.05)$, we reject the null hypothesis that the sample comes from a normal distribution.

Although this seems like a contradiction of the result of the chi-squared test in the study guide, we must point out that the χ^2-test was a borderline case.

 Now wasn't this fun to do? This is the kind of analysis an applied statistician does.

4.4 Goodness-of-fit tests in general

If you understand the goodness-of-fit test for a normal distribution, you will have no problem with this section!

For most distributions other than the normal distribution you will usually have the setup of "Situation A" as was explained above in the previous section. [In other words, the **expected frequencies are given** (in most cases with the wording "**equal** expected frequencies"), **BUT** you do not have the limits for the intervals and thus you have to compute them.] Falling under this category, we had to "work backwards" with *example 4.7 of the study guide to find the interval limits* under the assumption of an exponential pdf with parameter specified. (We are given that $\theta = 100$.)

Activity 4.9

Try to prove theorem 4.1 of the study guide for the special case where $k = 2$.

Feedback Feedback

If we look at the case $k = 2$, that is the binomial distribution, the frequencies are N_1 and $N_2 = n - N_1$ and the probabilities are π_1 and $\pi_2 = 1 - \pi_1$. [N_1 is the number of successes in n trials so that $N_1 \sim b(n;\ \pi_1)$ with $E(N_1) = n\pi_1$ and $var(N_1) = n\pi_1(1 - \pi_1)$. (See definition 1.14 and what follows.)]

We have to show that $Y^2 = \sum_{i=1}^{k} (N_i - n\pi_i)^2 / n\pi_i$ is approximately $\sim \chi^2_{k-1}$.

Now $Y^2 = (N_1 - n\pi_1)^2 / n\pi_1 + (N_2 - n\pi_2)^2 / n\pi_2 = \dfrac{(N_1 - n\pi_1)^2}{n\pi_1} + \dfrac{(n - N_1 - n(1 - \pi_1))^2}{n(1 - \pi_1)}$

$$= \frac{(N_1 - n\pi_1)^2}{n\pi_1} + \frac{(n\pi_1 - N_1)^2}{n(1 - \pi_1)} = \frac{(N_1 - n\pi_1)^2}{n}\left(\frac{1}{\pi_1} + \frac{1}{1 - \pi_1}\right)$$

$$= \frac{(N_1 - n\pi_1)^2}{n}\left(\frac{(1 - \pi_1) + \pi_1}{\pi_1(1 - \pi_1)}\right) = \frac{(N_1 - n\pi_1)^2}{n\pi_1(1 - \pi_1)}$$

Take the square root on both sides, that is $Y = \dfrac{N_1 - n\pi_1}{\sqrt{n\pi_1(1 - \pi_1)}}$.

Do you recognise Y as the variate which is approximately $n(0;\ 1)$ by the normal approximation to the binomial distribution? If we apply property (iv) of result 1.1 Y^2 is approximately χ^2_1 which is in accordance with theorem 4.1.

Activity 4.10

(a) Capture the raw data given in *example 4.7* of the study guide in a *JMP data table.*

(b) Save the file as *example4p7.jmp* for future use.

(c) Draw a histogram and superimpose a normal curve over the histogram. (If you like, draw a normal quantile plot and perform a Shapiro-Wilk W goodness-of-fit test for normality!)

(d) Draw a histogram and *superimpose an exponential curve* over the histogram.

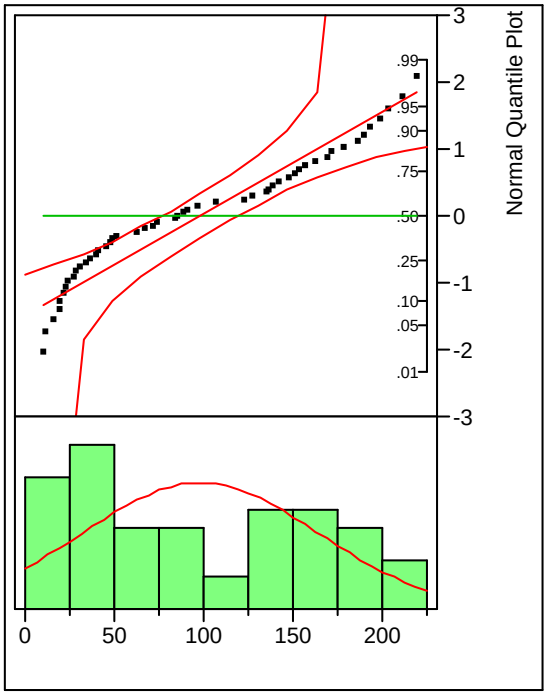
(e) Perform a Shapiro-Wilk W goodness-of-fit test for an *exponential distribution*.

(f) What can you conclude from all the output? How does the goodness-of-fit test for an *exponential distribution* differ from the study guide?

Feedback → Feedback

Numbers (a) and (b) are not displayed. I called my variable X.

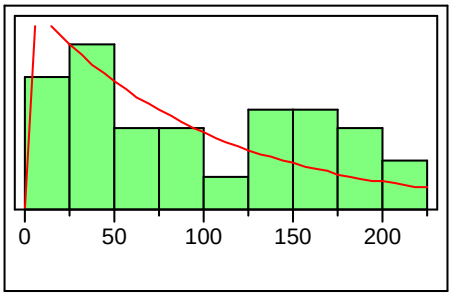
Number (c) produced the following output.



(c) This produced the following output:

This graph is obviously non-normal (the normal quantile plot even shows points outside the confidence band!) and we may conclude that the sample is not from a normal population without having to do the Shapiro-Wilk W goodness-of-fit test for normality!

(d)



Superimposing an exponential curve over the histogram seems to give a better fit but now it is still not so obvious whether the fit is good enough and we need a proper statistical test.

(e)

Fitted Exponential

Parameter Estimates

Type	Parameter	Estimate	Lower 95%	Upper 95%
Scale	s	97.977999	75.173365	131.019

Goodness-of-Fit Test

Kolmogorov's D

D	Prob>D
0.132697	0.1307

Note: Ho = The data is from the Exponential distribution. Small p-values reject Ho.

The output above is for the hypothesis test: H_0: The sample comes from an exponential distribution.

It differs from the study guide in the sense that the *theoretical parameter is not specified* and the computer estimates the unknown value as $\theta = 97.977999$. If the computer had used a chi-squared test it would have decreased the degrees of freedom by 1. However, JMP uses the Kolmogorov Smirnov test and we only need to interpret the p-value $= 0.1307$. Since this value $> \alpha = 0.05$, we cannot reject H_0 at the 5% level of significance. (Please note: Since this value $> \alpha = 0.10$, we cannot even reject H_0 at the 10% level of significance.) We assume that the data are from an exponential distribution.

4.5 Using the method of moments to test for normality

The following activity will enhance your understanding of why we could use the standardised moments of a sample to test for normality. Work through the questions without peeping at the feedback and try to get a feeling for the idea behind the test!

Activity 4.11

Refer to activity 3.7 (also see activity 3.8). This means the questions are about a random sample of size $n = 200$ from a $n\left(100;\ (15)^2\right)$ distribution.
Indicate whether the following statements are true or false:

(a) We could expect more or less 100 values to be larger than $\overline{X}$ and 100 values to be smaller than $\overline{X}$.

.....

.....

(b) About half of the differences $\left(X_i - \overline{X}\right)^3$ will be positive and the other half will be negative.

.....

.....

(c) The very large positive values will be cancelled out by the very large negative values.

.....

.....

(d) The magnitude of the numerical values of the $\left(X_i - \overline{X}\right)^3$ values will be a function of the scale of X. Hence $\Sigma\left(X_i - \overline{X}\right)^3$ will also be a function of σ.

.....

.....

(e) We can expect $\frac{(X_i - \overline{X})}{\sigma}$ to behave like a $n\,(0;\ 1)$ observation.

.....

.....

(f) We can expect $\frac{(X_i - \overline{X})^3}{\sigma^3}$ to behave like a cubed $n\,(0;\ 1)$ observation.

.....

.....

(g) We can expect $\Sigma\frac{(X_i - \overline{X})^3}{\sigma^3}$ to vary in the region of 0.

.....

.....

.....

Feedback Feedback

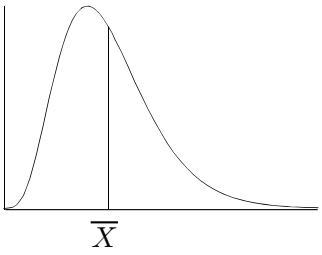
(a) True (but only if the sample comes from a normal population).

(b) True (this follows from (a) above).

Please note:

For any sample, regardless of the distribution from which it comes, it will always be true that $\Sigma\left(X_i - \overline{X}\right) = 0$.

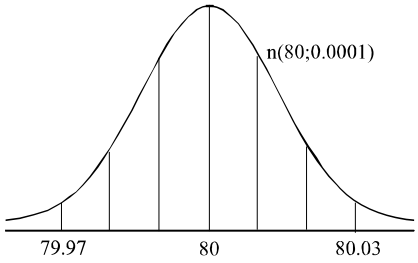
(c) True, but please note that this will only be true for a symmetrical distribution.



For example, for a positively skew distribution (illustrated above) the extreme positive $(X_i - \overline{X})^3$ values will be much larger than the extreme negative $(X_i - \overline{X})^3$ values.

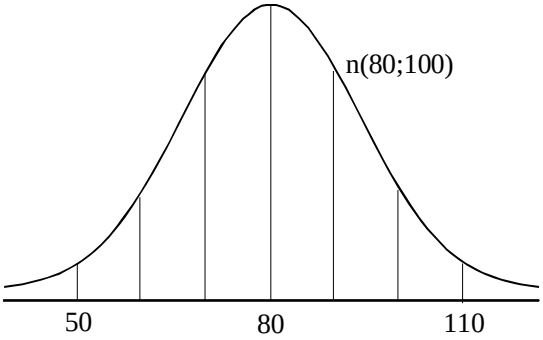
(d) True! The larger the spread of the original X_i-values, the larger the $\left(X_i - \overline{X}\right)^3$ values could become. If this does not make sense to you, imagine two contrasting normal populations from which you draw samples. In the sketch below we have two normal distributions which have the same mean value $(\mu = 80)$ but the variance of the one on the left is 0.0001 ($\Longrightarrow \sigma = 0.01$) whilst the variance of the one on the right is 100 ($\Longrightarrow \sigma = 10$).

(i)



What is the value of the most extreme positive $\left(X_i - \overline{X}\right)^3$ value? [Very small, e.g. $(80.03 - 80)^3 = (0.03)^3 = 0.000027$.]

(ii)



What is the value of the most extreme positive $\left(X_i - \overline{X}\right)^3$ value? [Very large, e.g. $(110 - 80)^3 = 30^3 = 27000$.]

(e) True! If we apply this to the illustrative examples above we get $\dfrac{(80.03 - 80)}{0.01} = 3$ and $\dfrac{(110 - 80)}{10} = 3$.

:-) Do you understand why we have to standardise?

(f) True.

(g) True, provided that the sample comes from a normal distribution.

Computing the standardised moments of a sample manually to test for normality is very tedious work – especially if we work with large samples.

This prompts many students to try and figure out shortcuts – shortcuts which are mostly incorrect!

Activity 4.12

How could you show that $\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^3 \neq \left[\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\right]^3$?

Feedback Feedback

(i) **Mathematically**

$$\text{LHS} \quad \sum_{i=1}^{n}\left(X_i-\overline{X}\right)^3 = \sum_{i=1}^{n}\left[X_i^3-3X_i^2\overline{X}+3X_i\overline{X}^2-\overline{X}^3\right]$$

$$= \sum_{i=1}^{n}X_i^3-3\overline{X}\sum_{i=1}^{n}X_i^2+3\overline{X}^2\sum_{i=1}^{n}X_i-\sum_{i=1}^{n}\overline{X}^3\Bigg]$$

$$= \sum_{i=1}^{n}X_i^3-3\overline{X}\sum_{i=1}^{n}X_i^2+3n\left(\overline{X}\right)^3-n\overline{X}^3$$

$$= \sum_{i=1}^{n}X_i^3-3\overline{X}\sum_{i=1}^{n}X_i^2-2n\left(\overline{X}\right)^3$$

$$\text{RHS} \quad \left[\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\right]^3 = \left[\sum_{i=1}^{n}X_i-\sum_{i=1}^{n}\overline{X}\right]^3$$

$$= \left[n\overline{X}-n\overline{X}\right]^3$$

$$= [0]^3$$

$$= 0$$

LHS $\neq$ RHS

(ii) **Use a numerical example**. (I have manipulated the values to ease computations and illustrate the point!)

Suppose we have a sample of size $n = 5$.

	X_i	$(X_i - \overline{X})$	$(X_i - \overline{X})^3$
	6	-3	-27
	7	-2	-8
	9	0	0
	10	$+1$	1
	13	$+4$	64
Total	45	0	30

$\overline{X} = \dfrac{\Sigma X_i}{5} = \dfrac{45}{5} = 9$

Clearly $\Sigma\left(X_i - \overline{X}\right) = 0$ **(ALWAYS THE CASE!)**

But $\Sigma\left(X_i - \overline{X}\right)^3 = 65 - 35 = 30$. [Some values are positive and some values are negative.]

Hence $\Sigma\left(X_i - \overline{X}\right)^3 \neq \left[\Sigma\left(X_i - \overline{X}\right)\right]^3$.

 Are you now convinced that it is not a good shortcut? Makes me see red in the examination and when I mark assignment questions!

It is important to keep in mind that if we want to test the hypothesis that a sample comes from a normal distribution and we want to make use of the method of moments we have to **perform two separate hypothesis tests**.

The first test is where we test the **skewness** of the sample. This test alone is not enough to guarantee normality because there are other distributions (for example the uniform distribution) that are also symmetrical about the mean.

The second test is for **kurtosis**. With this test we focus on another property of the normal distribution. Only if the sample passes both tests, may we conclude that the sample is most probably from a normal distribution.

The general framework to test for normality is the following: (For the sake of simplicity, we will only consider two-sided testing.)

A: Test for skewness

(a) Hypotheses: $H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$ (two-sided testing)

[Please note the Greek symbol!]

(b) Test statistic

From the given sample you compute $B_1 = \dfrac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^3}{\hat{\sigma}^3}$.

(c) Critical region

Critical values can be read from table A in the study guide. They are usually found from the knowledge of the distribution of the test statistic. We do not give you the distribution of B_1 (it falls outside the range of this module!), but it should feel correct to you that

- B_1 has a symmetrical distribution
- $E\left(B_1\right) = 0 = \beta_1$
- $var\left(B_1\right)$ is a function of n.

B: Test for kurtosis

(a) Hypotheses: $H_0 : \beta_2 = 3$

$H_1 : \beta_2 \neq 3$

(b) Test statistic

From the given sample you compute $B_2 = \dfrac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^4}{\hat{\sigma}^4}$.

(c) Critical region

Again we do not give the distribution of B_2 but you should know that

- B_2 does not have a symmetrical distribution
- $E\left(B_2\right) = 3$
- $var\left(B_2\right)$ is a function of n.

Activity 4.13

(i) What do you make of TABLE A of the study guide?

.....

.....

.....

(ii) Use the value $n = 40$ to derive any kind of probability statement from the values in the table.

.....

.....

.....

.....

Feedback  Feedback

(i)
- The variance of B_1 is clearly a function of n. The larger the value of n, the closer the critical value comes to zero. (Compare 0.711 if $n = 25$ with 0.127 if $n = 1\,000$.)
- Because B_1 has a symmetrical distribution about the value zero, the left-hand critical value corresponds to the right-hand critical value but has a negative sign.

(ii) For $n = 40$ we can make the following probability statements:

$$P\left[B_1 > 0.587\right] = 0.05$$
$$P\left[B_1 < -0.587\right] = 0.05$$
$$P\left[-0.587 < B_1 < 0.587\right] = 0.90$$

Thus, we call -0.587 and 0.587 the critical values and we will reject H_0 (two-sided at the 10% level) if $B_1 < -0.587$ or if $B_1 > 0.587$.

Activity 4.14

(i) What do you make of TABLE B of the study guide?

.....

.....

.....

.....

(ii) Use the value $n = 100$ to derive any kind of probability statement.

.....

.....

.....

.....

Feedback → Feedback

(i)
- The variance of B_2 is also clearly a function of n. The larger the value of n, the closer the critical values come to 3. (Compare the upper 5% value of 3.99 if $n = 50$ with the upper 5% value of 3.26 if $n = 1\,000$.)
- We can only use this test statistic if $n \geq 50$.
- B_2 is not symmetrical about the value 3.

(ii) For $n = 100$ we can make the following probability statements:

$P[B_2 > 3.77] = 0.05$

$P[B_2 < 2.35] = 0.05$

$P[2.35 < B_2 < 3.77] = 0.90$

Right, so we noted that computing the standardised moments of a sample manually is very tedious work (especially if we work with a large sample) and now our only hope is to employ JMP to do the calculations for us!

Activity 4.15

Use your own generated sample from a $n(100; 225)$ population and use JMP to test for normality by *using the method of moments*.
(Refer to activities 3.8 and 3.9 to have the default output. I used the data of table 3.1 to produce the feedback.)

Please note: From activity 3.9 we already know that JMP automatically gives a summary of some of the important statistics when we perform **Analyze**⟶**Distribution**.

Moments	
Mean	99.903554
Std Dev	15.714819
Std Err Mean	1.1112055
upper 95% Mean	102.0948
lower 95% Mean	97.712305
N	200

The output labelled **Moments** does not include B_1 and B_2 and we have to request it from JMP under **Display Options**⟶**More Moments.**

Feedback ⟶ Feedback

[*For beginners:* How to include more moments in the output:
Click on the histogram bar (there is a red triangle before the variable name *normal* where you have to click. This opens a *submenu* where **Display Options** is the first of the possible choices which you could select.] Tick the **More Moments** option.
This instruction results in the following output:

Moments	
Mean	99.903554
Std Dev	15.714819
Std Err Mean	1.1112055
upper 95% Mean	102.0948
lower 95% Mean	97.712305
N	200
Sum Wgt	200
Sum	19980.711
Variance	246.95553
Skewness	0.1744445
Kurtosis	-0.094571
CV	15.72999
N Missing	0

What do we do now?

"Skewness" and "kurtosis" are not treated as proper test statistics because there are no p-values included in the output. It also looks as if "kurtosis" has been standardised to have an expected value of 0 and not 3 . Whenever you are unsure of how a statistic has been computed or how you have to interpret it, you can go to the **Help Function of JMP**. For this example it is not much more of any assistance and we only get the following information from JMP:

"**Kurtosis** measures peakedness or heaviness of tails. It is based on the fourth moment about the mean and is computed as

$$\sum w_i^2 z_i^4 \frac{N(N+1)}{(N-1)(N-2)(N-3)} - 3\frac{(N-1)^2}{(N-2)(N-3)} \text{ , where } z_i = \frac{X_i - \overline{X}}{sd(X)}$$

and w_i is a weight term (= 1 for equally weighted items)."

This does not tell us how we could treat it as a test statistic. If all else fails go to another statistical package to compare information!

Using SPSS on the same data set produces the following results:

Descriptive Statistics

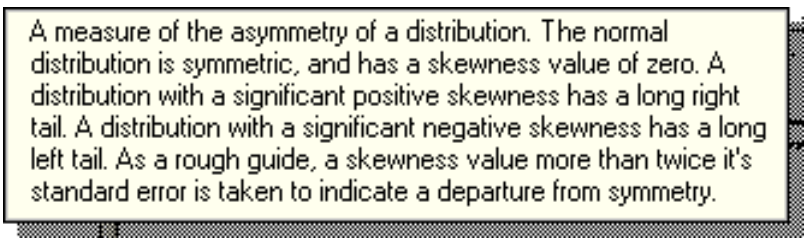
		normal	Valid N (listwise)
N	Statistic	200	200
Mean	Statistic	99.903553972171	
Std. Deviation	Statistic	15.714818787049	
Variance	Statistic	246.956	
Skewness	Statistic	.174	
	Std. Error	.172	
Kurtosis	Statistic	-.095	
	Std. Error	.342	

(Note that "normal" was the name of my variable.)

Do you notice that we obtain the same output as with JMP but with the big difference that some vital extra information are included (giving the "Std Errors" of the two statistics).

SPSS also has a help function which produces a bit more useful information:

Skewness =



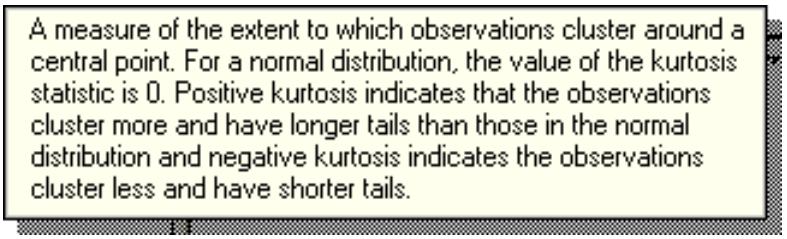
From the SPSS output and the printed explanation, we gather that we could standardise the skewness and compare it with the critical value of a $n(0;1)$ distribution.

$\Longrightarrow Z_{B_1} = \dfrac{B_1 - 0}{StdError} = \frac{0.174}{0.172} = 1.011\,6$

Performing a two-sided test and using $z_{\alpha/2} = 1.96$ (for $\alpha = 0.05$) we see that $-1.96 < 1.011\,6 < 1.96$.

We cannot reject the null hypothesis that the sample has the skewness of a normal population.

Kurtosis =



Similarly, using the SPSS output and the printed explanation, we gather that we could standardise the kurtosis and compare it with the critical value of a $n(0;1)$ distribution.

$\Longrightarrow Z_{B_2} = \frac{B_2 - 0}{StdError} = \frac{-0.095}{0.342} = -0.277\,78$

Performing a two-sided test and using $z_{\alpha/2} = 1.96$ (for $\alpha = 0.05$) we see that $-1.96 < -0.277\,78 < 1.96$.

We cannot reject the null hypothesis that the sample has the kurtosis of a normal population.

If you can figure out a way to produce the same output with JMP as with SPSS (regarding skewness and kurtosis) please let me know how!

Activity 4.16

Refer to activities 3.8 and 3.9 and use the data of table 3.1 to **manually** perform a *test for normality.* Use the *method of moments* and test two-sided at the 10% level of significance.

You make use of the following calculations:

$\sum_{i=1}^{200}(X_i - \overline{X})^2 = 49\,144.15$

$\sum_{i=1}^{200}(X_i - \overline{X})^3 = 133\,382.74$

$\sum_{i=1}^{200}(X_i - \overline{X})^4 = 34\,753\,109.98$

Feedback → Feedback

Using the "summary recipe" given just before activity 4.13, this test is easy as pie when the computational sums have been given!

A: Test for skewness

(a) Hypotheses: $H_0 : \beta_1 = 0$
$H_1 : \beta_1 \neq 0$ (two-sided testing)

(b) Test statistic

From the given sample we compute $B_1 = \dfrac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^3}{\hat{\sigma}^3}$.

$$\hat{\sigma}^3 = \left[\frac{1}{200}\sum_{i=1}^{200}(X_i - \overline{X})^2\right]^{\frac{3}{2}} = \left[\frac{1}{200}(49\,144.15)\right]^{\frac{3}{2}} = 3\,851.8$$

$$\begin{aligned} B_1 &= \frac{\frac{1}{200}\sum_{i=1}^{200}\left(X_i - \overline{X}\right)^3}{\hat{\sigma}^3} \\ &= \frac{\frac{1}{200}(133\,382.74)}{3\,851.8} \\ &= 0.173\,14 \end{aligned}$$

(c) Critical region

From TABLE A we find the upper 5% percentage point as 0.280. This means we will reject H_0 two-sided if $|B_1| > 0.280$.

Since $0.17314 < 0.280$ we cannot reject H_0. (Of course this is what we expected to happen because we used a normally generated sample!)

B: Test for kurtosis

(a) Hypotheses: $H_0 : \beta_2 = 3$
$H_1 : \beta_2 \neq 3$

(b) Test statistic

From the given sample we compute $B_2 = \dfrac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^4}{\hat{\sigma}^4}$.

where $\hat{\sigma}^4 = \left[\frac{1}{200}\sum_{i=1}^{200}(X_i - \overline{X})^2\right]^2 = \left[\frac{1}{200}(49\,144.15)\right]^2 = 60\,379.$

$$\begin{aligned} B_2 &= \frac{\frac{1}{200}\sum_{i=1}^{200}\left(X_i - \overline{X}\right)^4}{\hat{\sigma}^4} \\ &= \frac{\frac{1}{200}(34\,753\,109.98)}{60\,379} \\ &= 2.877\,9 \end{aligned}$$

(c) Critical region

From TABLE B we find the upper 5% percentage point as 3.57 and the lower 5% percentage point as 2.51. This means we will reject H_0 two-sided if $B_2 > 3.57$ or if $B_2 < 2.51$.

Since $2.51 < 2.8779 < 3.57$ we cannot reject H_0. (Of course this is again what we expected to happen because we used a normally generated sample!)

The sample passed both tests and hence we may conclude that it came from a normal population.

Now it is time to sit back and reflect on all the tests and graphical techniques we covered in this study unit.

Make yourself a mind map to organise and summarise the information.

STUDY UNIT 5

Statistical independence

5.1 The meaning of independence

In study unit 4 we discussed different tests for normality, because normality is an important assumption underlying the appropriateness of many statistical analyses. Another important assumption is that of "independent observations" which is also all too easily assumed. When you have completed this course and page through the study guide you will notice how many theorems and definitions contain lines containing either the word **independent** or the word **random**. So, many procedures we perform in the rest of the study guide are based on the assumption of independence.

Do we need a whole study unit to bring across this concept? Well, in the rest of the unit we expand on the idea of independence to include more than simply the independence of observations but also the independence of *two variables*. Then of course there is also the issue of how to measure INDEPENDENCE/DEPENDENCE.

Activity 5.1

Give the formal definition of the independence of two random variables X_1 and X_2.

.....

.....

.....

.....

.....

.....

Feedback → Feedback

Two random variables X_1 and X_2 are **independent** if their joint probability distribution function (pdf) is given by

$f_{X_1;X_2}(x_1;x_2) = f_{X_1}(x_1)\, f_{X_2}(x_2)$ for all x_1 and x_2.

Please note: This topic forms part of the syllabus of STA2603, where you will cover the work in more detail and will be given concrete examples.

Activity 5.2

Suppose that X_1 and X_2 are two discrete random variables with joint probabilities as given in the table below:

		X_2		
		1	2	$f_{X_1}(x_1)$
X_1	0	0.1	0.1	0.2
	1	0.26	0.14	0.4
	2	0.24	0.16	0.4
$f_{X_2}(x_2)$		0.6	0.4	1

Test to see whether X_1 and X_2 are independent.

.....

.....

.....

.....

.....

.....

.....



$P(X_1 = 2 \text{ and } X_2 = 1) = f_{X_1;X_2}(2;1) = 0.24$ and

$f_{X_1}(2) f_{X_2}(1) = (0.4)(0.6) = 0.24$ and we might hastily conclude that X_1 and X_2 are independent. **But**, if we select another combination of X_1 and X_2, for example

$X_1 = 0$ and $X_2 = 1$ we find that

$P(X_1 = 0 \text{ and } X_2 = 1) = 0.1$ and

$f_{X_1}(0) f_{X_2}(1) = 0.2 \times 0.6 = 0.12 \neq 0.1.$

So, we have found a combination of X_1 and X_2 where

$f_{X_1;X_2}(x_1;x_2) \neq f_{X_1}(x_1) f_{X_2}(x_2)$ and hence they are **not independent**!

5.2 Examples of dependence

From the five given examples of dependence in the study guide, it is clear that the outcome of one observation could have an influence on the outcome of any other observation and hence we have dependence. Much time and effort are spent on developing statistical analysis methods for these scenarios.

In the rest of unit 5 we will study the relationship of bivariate data sets, in other words, we will study the dependence/independence of **variables two at a time**. So, based on the **type of variables being measured**, we will use totally different techniques for different types to answer the question: Are the two variables related?

5.3 Contingency table analysis

When both the variables we are interested in are **nominal variables**, we usually summarise the data in a two-way classification table (called a cross-tabulation in many computer packages). Thus, when **two categorical variables** are observed and you count the number of experimental units that fall into each of the various intersections of the different category levels, you obtain a **contingency table**.

Note that the numbers in a contingency table always represent the **frequencies** (or counts) of the number of people/experimental units that fall into each category.

In this section we will cover **two different tests** for contingency tables:

1. The general $h \times k$ cross-tabulation ($h > 2; k > 2$)
2. The specific 2×2 cross-tabulation

Activity 5.3

A sample of 600 people was randomly selected and cross-classified according to **income** and **body condition**. The categories *Low income*; *Middle income* and *High income* are the income levels and the categories *Underweight*; *Average* and *Overweight* were used for body condition. The following results were obtained:

		Income		
		Low income	Middle income	High income
Body type (Condition)	Underweight	80	24	23
	Average	150	120	42
	Overweight	70	56	35

We want to test the hypothesis that the two factors **income** and **body condition** are independent.

Question 1

Refer to the setup above and *use the notation defined in the study guide (below example 5.1)*. Indicate whether the following statements are true or false:

(a) The cell frequencies $N_{ij}; i = 1; 2; 3$ and $j = 1; 2; 3$ may be considered to be random variables.

.....

(b) $N_{i\cdot} = \sum_{j=1}^{3} N_{ij}$ for $i = 1; 2; 3$ and $N_{\cdot j} = \sum_{i=1}^{3} N_{ij}$ for $j = 1; 2; 3$ are also random variables.

.....

(c) Testing the hypothesis that the two factors are **independent** implies testing

$H_0 : \pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}$ for $i = 1; 2; 3$ and $j = 1; 2; 3.$

.....

(d) The alternative hypothesis is that the probabilities π_{ij} are free.

.....

(e) $\pi_{i\cdot}$ and $\pi_{\cdot j}$ are unknown probabilities and are estimated from the contingency table as $\frac{N_{i\cdot}}{N_{\cdot\cdot}}$ and $\frac{N_{\cdot j}}{N_{\cdot\cdot}}$, respectively.

.....

(f) Under the alternative hypothesis, the maximum likelihood estimator of π_{ij} is $\hat{\pi}_{ij} = \frac{N_{ij}}{N_{\cdot\cdot}}$.

.....

(g) Under the null hypothesis of independence, the expected counts $e_{ij} = N_{\cdot\cdot}\hat{\pi}_{ij} = \frac{N_{i\cdot}N_{\cdot j}}{N_{\cdot\cdot}}$.

.....

(h) The test statistic, Y^2, defined in theorem 5.1 has the χ^2_ν distribution which has only one parameter, ν, that needs to be known before you can utilise the distribution.

.....

Question 2

Complete the table for the **expected frequencies**:

		Income			
		Low income	Middle income	High income	**Row totals** $(N_{i\cdot})$
Body type (Condition)	Underweight	.....	.....	21.17	127
	Average	.....	.....	.....	.....
	Overweight	80.50	53.67	.....	.....
Column totals	$(N_{\cdot j})$	300	.....	.....	600

Question 3

Compute the value for the test statistic Y^2.

.....

.....

.....

.....

Question 4

Use $\alpha = 0.05$ and look up the critical value.

.....

.....

.....

.....

Question 5

What is your final conclusion?

.....

.....

.....

.....

Feedback → Feedback

Question 1

(a) True.

(b) True. We have the case where $N_{\cdot\cdot} = 600$ is considered as a fixed grand total and then the marginal totals are random variables (as opposed to the case where $N_{\cdot j}$ is not a random variable, but a chosen sample size from population j).

(c) True. In STA1503 we defined two events A and B to be INDEPENDENT if $P(A \cap B) = P(A) \cdot P(B)$.

(d) True. "Free" means that they are not restricted to be the product $(\pi_{i\cdot})(\pi_{\cdot j})$.

(e) True. It can be proven that the maximum likelihood estimators of $\pi_{i\cdot}$ are $\hat{\pi}_{i\cdot} = \frac{N_{i\cdot}}{N_{\cdot\cdot}}$ and the maximum likelihood estimators of $\pi_{\cdot j}$ are $\hat{\pi}_{\cdot j} = \frac{N_{\cdot j}}{N_{\cdot\cdot}}$.

(f) True.

(g) True. $e_{ij} = N_{\cdot\cdot}\hat{\pi}_{ij} = N_{\cdot\cdot}\hat{\pi}_{i\cdot}\hat{\pi}_{\cdot j} = N_{\cdot\cdot}\left(\frac{N_{i\cdot}}{N_{\cdot\cdot}}\right)\left(\frac{N_{\cdot j}}{N_{\cdot\cdot}}\right) = \frac{N_{i\cdot}N_{\cdot j}}{N_{\cdot\cdot}}$

(h) True! The chi-square distribution, usually denoted as χ^2_ν , has only one parameter, ν, which for the contingency table application has $\nu = (h-1)(k-1)$.

Question 2

		Income			
		Low income	Middle income	High income	**Row totals** $(N_{i.})$
Body type	Underweight	63.5	42.33	21.17	127
(Condition)	Average	156.0	104.00	52.00	312
	Overweight	80.5	53.67	26.83	161
Column totals	$(N_{.j})$	300	200	100	600

Question 3

$$\begin{aligned} Y^2 &= \sum_{i=1}^{3}\sum_{j=1}^{3}\frac{(N_{ij}-e_{ij})^2}{e_{ij}} \\ &= \frac{(80-63.5)^2}{63.5}+\frac{(24-42.33)^2}{42.33}+\frac{(23-21.17)^2}{21.17}+\frac{(150-156)^2}{156} \\ &\quad +\frac{(120-104)^2}{104}+\frac{(42-52)^2}{52}+\frac{(70-80.5)^2}{80.5}+\frac{(56-53.67)^2}{53.67}+\frac{(35-26.83)^2}{26.83} \\ &= 20.96 \end{aligned}$$

Question 4

The degrees of freedom are

$$\begin{aligned} (3-1)(3-1) &= 4 \text{ and} \\ \chi^2_{0.05;4} &= 9.48773. \end{aligned}$$

Question 5

Since $20.96 > 9.48773$ we reject H_0. It seems as if *income* and *body type* are not independent factors.

The exact test for 2 x 2 tables

It is explained in this corresponding section of the study guide how to use **TABLE D**. An obvious "tip" for any assignment or examination question, is to use the hypergeometric distribution whenever $N = 12$ and when you have a 2×2 contingency table! (We do not supply you with tables for any other value of N.) The following activities will help you to understand this procedure.

Activity 5.4

Suppose you have the following 2×2 contingency tables, where $N = 12$:

(a)

1	3	4
3	5	8
4	8	$N = 12$

(b)

5	0	5
4	3	7
9	3	$N = 12$

(c)

4	2	6
3	3	6
7	5	$N = 12$

(d)

4	2	6
2	4	6
6	6	$N = 12$

Indicate how you will choose n and k.

Feedback → Feedback

(a)

1	3	4	$\longrightarrow k$
3	5	8	
4	8	$N = 12$	
$\uparrow$			
n			

(b)

5	0	5	$\longrightarrow n$
4	3	7	
9	3	$N = 12$	
	$\uparrow$		
	k		

(c)

4	2	6	doesn't matter
3	3	6	$\} \longrightarrow n$
7	5	$N = 12$	
	$\uparrow$		
	k		

(d)

4	2	6	
2	4	6	$\longrightarrow n$
6	6	$N = 12$	
	$\uparrow$		
	k		

doesn't matter - you might pick differently

Activity 5.5

Refer to activity 5.4.

Indicate the nature of the hypergeometric distribution that results from your choice of n and k for each one of the four 2×2 tables. Let X denote the random variable which has outcome x. Show this value.

.....

.....

.....

.....

.....

.....

.....

Feedback  Feedback

(a) $X \sim H(12; 4; 4)$ and we have the "observed outcome" of $X = 1$, that is x pinned by the cross-tabulation where $k = 4$ and $n = 4$ which results in $x = 1$.

(b) $X \sim H(12; 5; 3)$ and we have the "observed outcome" of $X = 0$, that is $x = 0$.

(c) $X \sim H(12; 6; 5)$ and we have the "observed outcome" of $X = 3$.

(d) $X \sim H(12; 6; 6)$ and we have the "observed outcome" of $X = 4$. (This of course is for the choice I have indicated in the FEEDBACK above!)



Now it is time to make friends with TABLE D and practise a bit how to use the table!

Activity 5.6

Refer to activities 5.4 and 5.5.

Part I: For each one of the four given 2×2 tables, use the observed value x and TABLE D to find the value of *both the probability expressions*

$P(X \leq x)$ and $P(X \geq x)$.

Part II: Optional exercise: Use TABLE D to show that for case (d) above the variable $X \sim H(12; 6; 6)$ has a *discrete probability distribution* and draw the probability function (similar to a bar chart).

.....

.....

.....

.....

Feedback → Feedback

Part I:

(a) $X \sim H\,(12; 4; 4)$ and $x = 1$:

$$\begin{array}{rcll} P\,(X \leq 1) & = & 0.594 & \text{(Read this directly from the table.)} \\ P\,(X \geq 1) & = & 1 - P\,(X \leq 0) & (X \text{ has a discrete distribution.}) \\ & = & 1 - 0.141 & \\ & = & 0.859 & \end{array}$$

(b) $X \sim H\,(12; 5; 3)$ and $x = 0$:

$$\begin{array}{rcll} \therefore P\,(X \leq 0) & = & 0.159 & \text{(Read this directly from the table.)} \\ P\,(X \geq 0) & = & 1 & \text{(This forms the whole sample space, that is all possibilities.)} \end{array}$$

(c) $X \sim H\,(12; 6; 5)$ and $x = 3$:

$$\begin{array}{rcll} P\,(X \leq 3) & = & 0.879 & \text{(Read this directly from the table.)} \\ P\,(X \geq 3) & = & 1 - P\,(X \leq 2) & \\ & = & 1 - 0.500 & \\ & = & 0.500 & \end{array}$$

(d) $X \sim H\,(12; 6; 6)$ and $x = 4$:

$$\begin{array}{rcll} P\,(X \leq 4) & = & 0.960 & \text{(Read this directly from the table.)} \\ P\,(X \geq 4) & = & 1 - P\,(X \leq 3) & \\ & = & 1 - 0.716 & \\ & = & 0.284 & \end{array}$$

Part II:

Optional exercise: If $X \sim H\,(12; 6; 6)$ (case (d) above) it means that the sample space of X is $A = \{0; 1; 2; 3; 4; 5; 6\}$ which has a discrete probability distribution where

$$P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5) + P(X = 6) = 1.$$

Using table D we find

$P(X = 0) = 0.001$

$P(X = 1) = 0.039$ which is $(0.040 - 0.001)$

$P(X = 2) = 0.244$ which is $(0.284 - 0.040)$

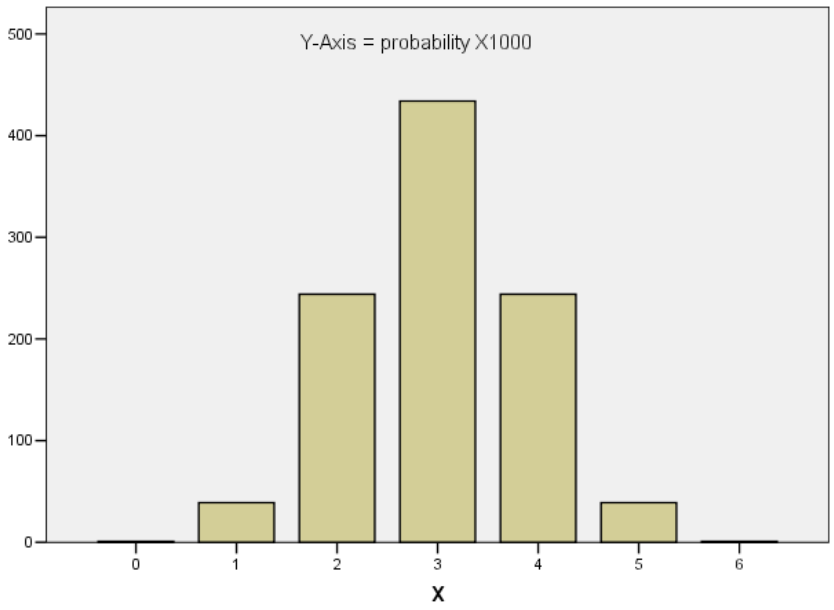
$P(X = 3) = 0.434$ which is $(0.716 - 0.282)$

$P(X = 4) = 0.244$ which is $(0.960 - 0.716)$

$P(X = 5) = 0.039$ which is $(0.999 - 0.960)$

$P(X = 6) = 0.001$ which is $(1.000 - 0.999)$

We may draw the following probability function:



Now it is time to apply these skills to a concrete example!

Activity 5.7

In a random sample of eight women, it was found that five smoked. In a random sample of four men, it was found that three smoked.

Test the null hypothesis that there is no association between gender and smoking against the alternative that more men than women smoke (meaning the proportion of men who smoke is greater than the proportion of women who smoke). Use $\alpha = 0.05$.

[**Hint:** Use any of the results of activities 5.4 to 5.6.]

Feedback  Feedback

Our first step is to compile a 2×2 contingency table!

		Nonsmokers	Smokers			
Gender	Male	1 ...x	3	4	$\longrightarrow$	k
	Female	3	5	8		
		4	8	$N = 12$		
		$\uparrow$				
		n				

This is exactly the 2×2 table given in activity 5.4 (a) and n; k and x are indicated as in this activity.

We have to test:
$H_0:$ Gender and smoking are independent factors.
$H_1:$ The proportion of male smokers is larger than the proportion of female smokers.

Please note that the word "larger" implies a one-sided alternative (or else we should have used the word "different"). The trickiest part of this problem is to translate the alternative hypothesis in terms of x (which is "fixed" by the choice of n and k).

If H_1 is true, we would expect the nonsmoking males to be few $\Longrightarrow$ we will reject H_0 in favour of H_1 if $P(X \leq x)$ becomes a very small probability. From activity 5.6(a), we already know that $P(X \leq 1) = 0.594$.

This is not a very small probability $\Longrightarrow$ we cannot reject H_0. Gender and smoking seem to be independent $\Longrightarrow$ whether you will smoke or not, is independent of whether you are male or female.

What about JMP?
You must have been conditioned by now that we usually repeat what is done **manually** in the study guide by doing it **electronically on JMP in the workbook! How will you create a contingency table using JMP? How will you perform a contingency analysis using JMP?**

Activity 5.8

Refer to **activity 5.3**.

(a) Manipulate JMP to create a data set that will result in the same contingency table as activity 5.3.

(b) Test the hypothesis that the two factors *income* and *body condition* are independent.

(c) Comment on the similarities (and the differences) between the manual test and the electronic version of JMP.



(a) In general, collected data will be captured in the following manner:

computer spreadsheet

	Variable 1	Variable 2	Variable 3	.....	Variable k
row 1					
row 2					
$\vdots$					
row n					

We need to rethink how a computer captures data and try to figure out how the spreadsheet can be manipulated to represent the sample of 600 people (actually needing 600 rows) *in fewer rows*. We will make use of a capturing technique called *weighting*.

A sample of 600 people were randomly selected and cross-classified according to **income** and **body condition**. The categories *Low income*; *Middle income* and *High income* were the "income" levels and the categories *Underweight*; *Average* and *Overweight* were used for "body condition".

For the first variable, which is **Income**, we can only have *three possible values*: *Low income*; *Middle income* and *High income*. Suppose we *code these values* as 1 = *Low income*; 2 = *Middle income* and 3 = *High income*. If we do the same for the second variable (**Body type**), we have 1 = *Underweight*; 2 = *Average* and 3 = *Overweight*. What does this mean? It means that should we have to enter the data, we will type 1s; 2s and 3s in the columns.

In the contingency table of activity 5.3, the *frequency of 80 in the first cell* (where Income = Low income and Body type = Underweight) means that if we have to type the data we would have to type **80 identical rows:**

income	body type
1	1

1. The same argument is true for the *frequency of 24 in the second cell* (where Income = Middle income and Body type = Underweight)**.** It means that if we have to type the data we would have to type **24 identical rows:**

income	body type
2	1

The *frequency of 35 in the last cell* (where Income = High income and Body type = Overweight) means that if we have to type the data we would have to type **35 identical rows:**

income	body type
3	3

We sidestep this process by *adding a third variable* which we call **count** (in which we type the frequencies) *and use only nine rows instead of 600 rows.* (Please note that this "count column" gives the frequency or the relative importance of each of these lines for the computer and in some computer packages it is called weighting!)

computer spreadsheet

	income	body type	count
row 1	1	1	80
row 2	2	1	24
row 3	3	1	23
row 4	1	2	150
row 5	2	2	120
row 6	3	2	42
row 7	1	3	70
row 8	2	3	56
row 9	3	3	35

(i) **Reproduce the above table in JMP.** (Revise activity 3.3 if your knowledge is a little rusty!)

Do you recall that we concluded activity 3.3 with: "*Question to ponder:* Are you comfortable with the 1s and 2s for gender and the numerical values 1 to 5 for the different choices of residence? We will come back to the issue of *changing the properties of a column* to give **labels** to the different numerical values. You also have to think about the **Data Type** and **Modeling Type** implications of having numerical values for these variables."



It is now time to learn how to "update and streamline" data by selecting the correct **Modeling Type** and to give **labels** to the different numerical values of a nominal variable.

[*For beginners*: How to select the correct **Modeling Type** and also **give labels** to the different numerical values of a nominal variable:

- To open the column information dialog box there are two routes:

  1] Double-click the column with the name *income* to automatically open the column information dialog box.

2] Highlight the first column. (The heading **income** changes to the colour blue which means that the computer now considers this column as active and will wait for further instructions from you.) Choose **Cols⟶Column Info (**the fifth option on the main menu will open the **Columns submenu** where you will find the option **Column Info** *fourth from the top*).

- Select **Nominal** under **Modeling Type**.
- The **Column Info** dialog box has a field called **Column Properties** at the bottom of the **Column Info** box which opens a pop-up menu:

  Formula

  Notes

  Range Check

  List Check

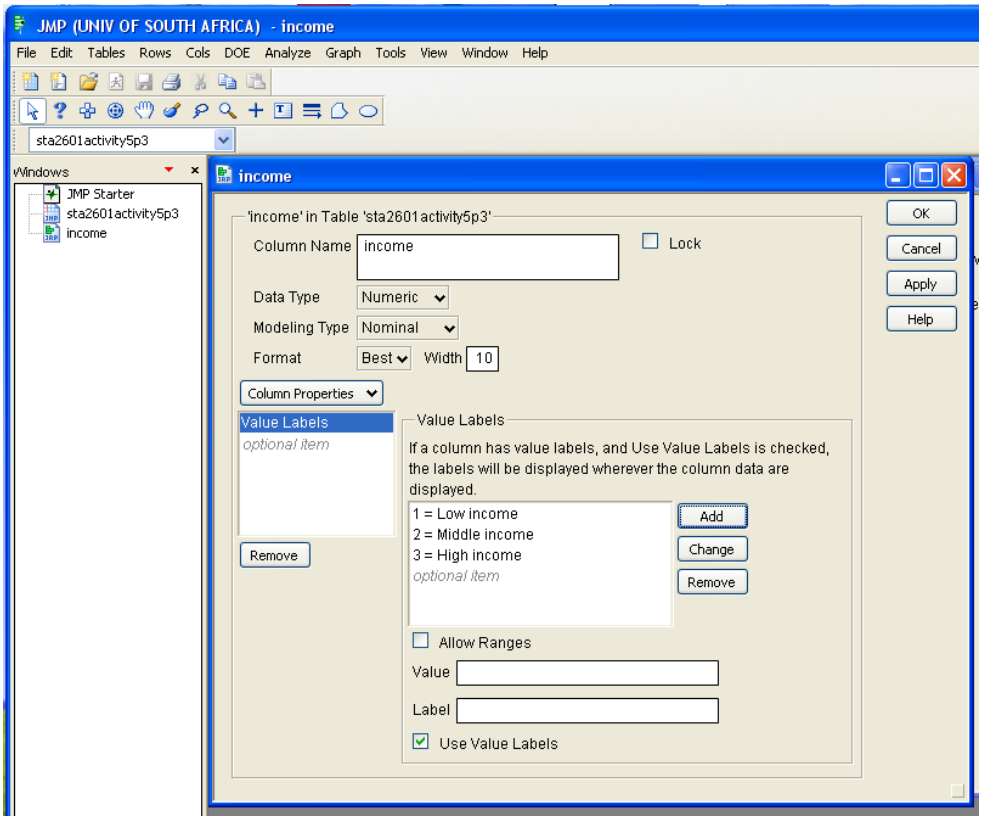
  Value Labels

  ↓

  etc. –

Choose **Column Properties ⟶Value Labels**

- Now you have to do your labelling: At the bottom of the "Value Labels" dialog box you will see two fields which you have to complete by typing in your **Value** (in the first box) and the **Label** (in the second box). Do them value by value; it is important that you click the "**Add**" button on the right to *activate the label*.



Repeat this process for the variable **body type.**

(b) [*For beginners*: How to test the hypothesis that the two factors *income* and *body type* are independent:

Select **Analyze** ⟶ **Fit Y by X.**

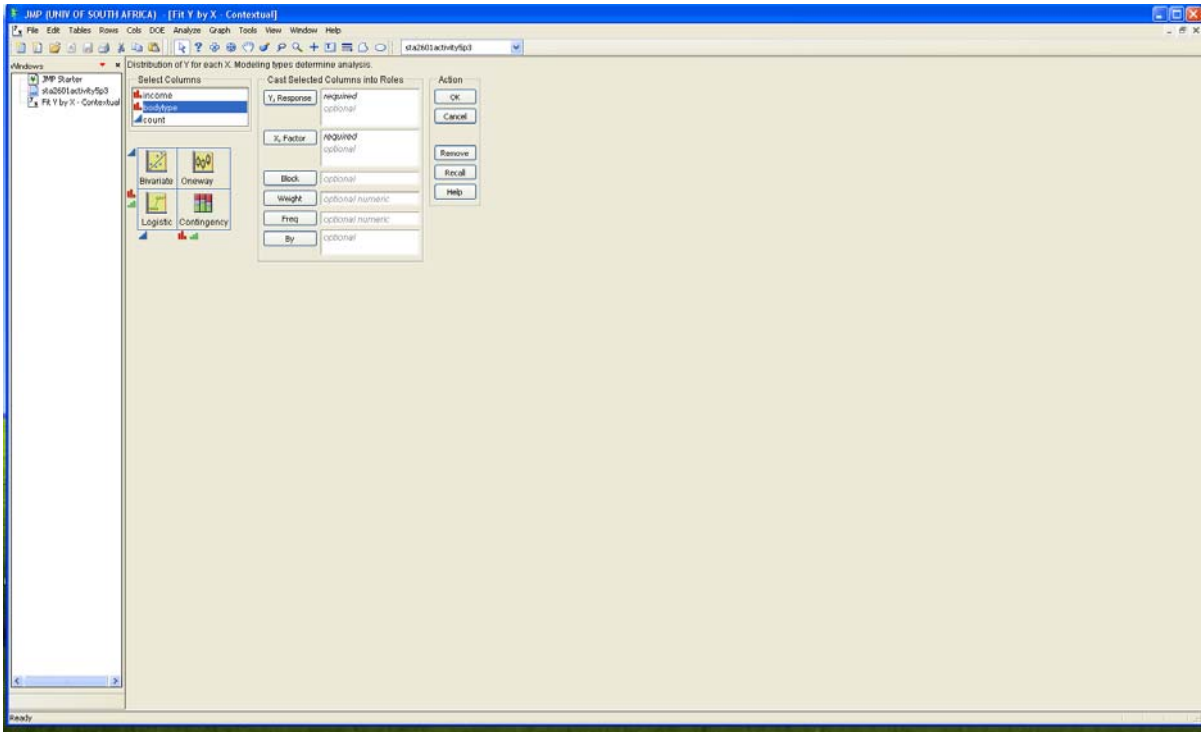
Now you get a report from JMP where the top line reads as follows:

"**Distribution of Y for each X. Modeling types determine analysis"**

You will notice that JMP uses three different icons for the three main modelling types: a *blue triangle* for a continuous variable; *a green histogram-type* icon for an ordinal variable and a *red bar chart* for a nominal variable.

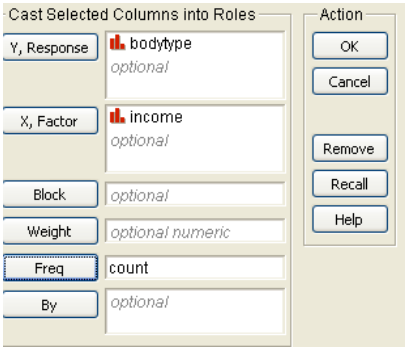
If you have selected nominal for the correct **Modeling Type**, JMP automatically uses the *icon of a red bar chart* to identify the two nominal variables in the variable box (*count* is not really a variable which will be analysed!!) and all you have to do is to select the correct "picture display" of the given four possible "**Y by X**" analyses in the bottom right-hand corner! This means that if Y = *a red bar chart* and *X = a red bar chart* you will have to do a contingency analysis. (Of course this

is what we already know to do because we study statistics!)



If we are trying to establish a connection between the two variables, would you think that *income* could influence *body type* or the other way around? In other words, if you have to select an "independent variable = X" and a "dependent variable = Y" how will you choose them? (It is not extremely critical for this analysis but for regression it is extremely important to make the correct choice!) Suppose that *income* influences *body type* which means we assume that "independent variable = X, Factor = *income*" and that "dependent variable = Y, Response = *body type*". Convey this choice to JMP.

It is very important that you also select *count* and move it into the "**freq**" box in the "**Cast Selected Columns into Roles**" dialog box.



If you click on OK JMP will compile a two-way frequency table, automatically supply the appropriate graphical summary and perform the chi-square goodness-of-fit test.

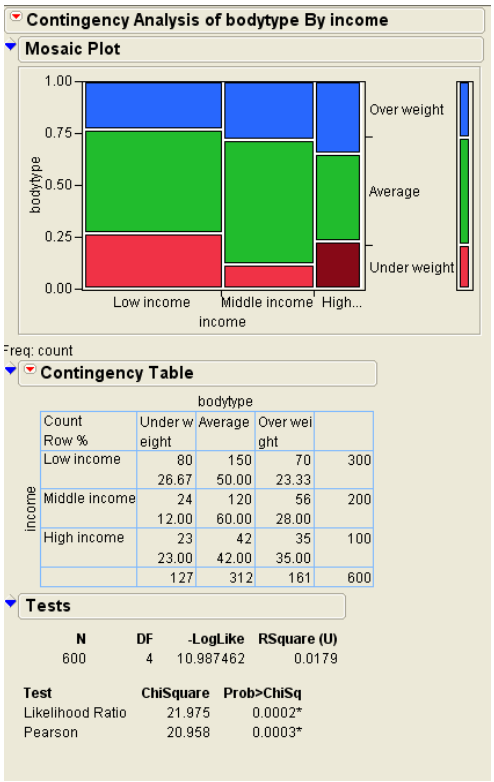
Discussion of the JMP output:

What JMP calls a "**Mosaic Plot**" is in effect a stacked bar chart but it is proportional to the size

of the subsamples. We can see that there were double the number of *Low income* respondents in the sample than the other two groups combined. (Also that there were double the number of *Middle income* respondents in the sample than *High income*.) The graph represents the "**Row %**" of the "**Contingency Table**". I have selected "*body type*" as the Y-variable and "*income*" as the X-variable. Changing the axes would result in a different mosaic plot and the graph will then represent the present "**Col %**" of the present "**Contingency Table**" which I did not select. (If you only want the "**Contingency Table**" without the percentages you can click inside the table and de-select the percentages.) For this analysis I have selected only the row percentages because this is in effect what we are comparing. (If *income* influences *body type*, the colour pattern of the mosaic plot will look distinctly different for the three groups of income.) Remember that the graph is only a visual aid and does not give statistical proof! Looking at the graph it seems as if *income* and *body type* are not independent factors because of the difference between (for example) middle income and high income in the mosaic plot. (Notice that low income and high income looks very alike in the mosaic plot!)

How does JMP compare with our manual test?

If you page back to activity 5.3 you will see that we rejected the null hypothesis of independence because the computed value of the chi-square test statistic was 20.96 which was greater than the critical value of 9.48773 (obtained from a table). JMP also computes the chi-square test statistic as 20.958 with a p-value of 0.0003 which is of course significant and leads to the same conclusion.



WORK THROUGH

Sall, Creighton and Lehman, **Chapter 12 Categorical Models**

Start reading on page 289 "*Two-Way Tables: Entering Count Data*" and read up to "*If You Have a Perfect Fit*".

Work through the instructions and steps under this heading, but you need *not continue with*

"If You Have a Perfect Fit".

Activity 5.9

Refer to **activity 5.7**.

(a) Manipulate JMP to create a data set that will result in the same 2×2 table as activity 5.7.

(b) Comment on the similarities (and the differences) between the manual test and the electronic version of JMP.

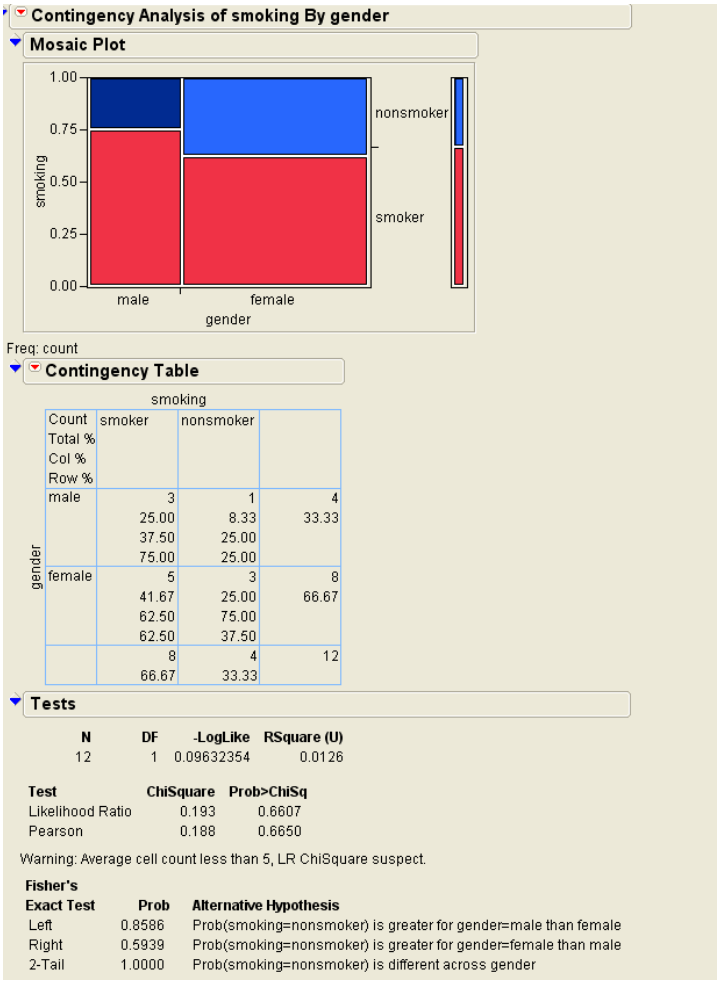
(c) Test the hypothesis that the two factors *gender* and *smoking* are independent against the *one-sided alternative* that there is a greater proportion of male smokers.

Feedback ⟶ Feedback

(a) Create a JMP data table to look like this:

gender	smoking	count
male	smoker	3
male	nonsmoker	1
female	smoker	5
female	nonsmoker	3

(b) Performing exactly the same analysis as in the previous activity, in other words select **Analyse ⟶ Fit Y by X,** the appropriate (useful) JMP output is the following:



Again we notice the **Mosaic Plot** (stacked bar chart proportional to the size of the subsamples). We can see that there were double the number of females in the sample and that the graph represents the **Row %** of the **Contingency Table**. I have selected "*smoking*" as the Y-variable and "*gender*" as the X-variable. Changing the axes would result in a different mosaic plot because the graph will then represent the **Col %** of the **Contingency Table**. (If you only want the **Contingency Table** without the percentages you can click inside the table and de-select the percentages or select maybe only the row percentages.)

(c) We get the same results as for the manual test but we have to understand how to interpret the output and pick the correct (in other words the appropriate) line of **Fisher's Exact Test**.

In activity 5.7 we computed manually (using a table!) that $P(x \leq 1) = 0.594$. This probability represented the p-value to reject H_0 in favour of H_1 : The proportion of **male smokers** is larger than the proportion of female smokers.

Where do we find this probability in the JMP table? The value 0.5939 is the value for the right-tailed test of the JMP Fisher's Exact Test output where JMP used the *complimentary* (different) wording but meaning exactly the same as our alternative hypothesis!

H_1 : The proportion of **female nonsmokers** is larger than the proportion of male nonsmokers.

We notice that all the p-values are very large and there is not any possibility of rejecting the null hypothesis of independence.

5.4 Correlation

Now we have the scenario where **two continuous variables** are simultaneously observed and we are interested in the question of whether or not they are related.

Correlation and independence

Activity 5.10

Give your own summary of theorems 5.2 and 5.3.

.....

.....

.....



X and Y independent $\overset{\text{no condition}}{\Longrightarrow}$ $\rho = 0$.

$\rho = 0$ $\overset{\text{condition: } (X;Y) \text{ has a bivariate normal distribution}}{\Longrightarrow}$ X and Y are independent.

Activity 5.11

Rewrite **definition 1.22** (see study unit 1) in terms of two random variables X and Y, in other words let X and Y be a pair of continuous random variables with a **bivariate normal distribution** and give $f_{X;Y}(x;y)$.

Let $E(X)=\mu_1;\ E(Y)=\mu_2;\ var(X)=\sigma_1^2;\ var(Y)=\sigma_2^2$ and $\rho=\dfrac{cov(X;Y)}{\sigma_1\sigma_2}$.

.....

.....

.....

.....

Feedback  Feedback

$$f_{X;Y}(x;y)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}e^{-1/2\left(1/(1-\rho)^2\right)\{((x-\mu_1)/\sigma_1)^2+((y-\mu_2)/\sigma_2)^2-2\rho((x-\mu_1)/\sigma_1)((y-\mu_2)/\sigma_2)\}}$$

This pdf is valid for

$-\infty<x<\infty;\ -\infty<y<\infty;\ \sigma_1>0;\ \sigma_2>0;\ -1<\rho<1.$

Although the definitions and mathematical properties of joint distributions are covered in detail in our modules on distribution theory (for example STA2603), you should have a basic "gut feeling" for the implications and the applications in STA2601.

Activity 5.12

If a pair of continuous random variables X and Y have a bivariate normal distribution, we use a special notation for this and write

$$(X;Y)\sim BVN\left(\mu_1;\mu_2;\sigma_1^2;\sigma_2^2;\rho\right).$$

(a) How many parameters define the joint distribution?

.....

.....

(b) What is the marginal distribution of X?

.....

.....

(c) What is the marginal distribution of Y?

.....

.....

(d) Define ρ.

.....

.....

Feedback → Feedback

(a) We need **five parameters** to define this distribution.

If we do not have the values of $E(X) = \mu_1;\ E(Y) = \mu_2;\ var(X) = \sigma_1^2;\ var(Y) = \sigma_2^2$ and $\rho =$ correlation coefficient of X and Y, we will not be able to "draw a graph" of the joint pdf of X and Y. This "graph" will look like a mountain in three-dimensional space. The total "volume" under this mountain $= 1$ and represents the total probability of the complete sample space. Mathematically speaking, we know that

$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X;Y}(x;y)\,dxdy = 1.$

(b) The marginal distribution of X is $\int_{-\infty}^{\infty} f_{X;Y}(x;y)\,dy = f_X(x)$ and this will be the pdf of a $n\left(\mu_1;\sigma_1^2\right)$ distribution.

(c) The marginal distribution of Y is $\int_{-\infty}^{\infty} f_{X;Y}(x;y)\,dx = f_Y(y)$ and this will be the pdf of a $n\left(\mu_2;\sigma_2^2\right)$ distribution.

(d) $\rho = \dfrac{cov(X;Y)}{\sqrt{var(X)\,var(Y)}}$.

The activities above created a rather lengthy "introduction" and "detour" to what we really want to do because we first focused on the theory! The hypothesis tests that we really aim at, concerns the relationship between two continuous variables, and these tests depend on *the assumption of*

bivariate normality. Can you see how we hammer on the importance of at least thinking about the theoretical assumptions? Throughout the rest of our study guide, an "assumption" implies a necessary theoretical condition that validates the derivation of the test statistic. The first actual test statistic is defined in *theorem 5.4*.
In the rest of this section, we will concentrate on the following two scenarios:

1. We have a single random sample (that is **one sample**) of paired observations $(X_i; Y_i)\ \ i = 1; 2...n$ from a bivariate normal population.
2. We have **two independent** random samples of paired observations $(X_i; Y_i)\ \ i = 1; 2...n$ from a bivariate normal population and $(V_j; W_j)\ \ j = 1; 2...m$ from an independent bivariate normal population.

Activity 5.13

Suppose you have a single random sample (in other words one sample) of paired observations $(X_i; Y_i)$ for $i = 1; 2...n$ from a bivariate normal population with correlation coefficient ρ.

Give a summary of the different hypothesis tests/confidence intervals you can perform. Also state the relevant test statistic.

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

Feedback → Feedback

1. We may use the test statistic $\mathbf{T} = \dfrac{\sqrt{n-2}R}{\sqrt{1-R^2}} \sim \mathbf{t}_{n-2}$

   to test $\mathbf{H}_0$: $\boldsymbol{\rho = 0}$ against $\begin{array}{ll} \diagup & H_1 : \rho > 0 \text{ or} \\ - & H_1 : \rho < 0 \text{ or} \\ \diagdown & H_1 : \rho \neq 0. \end{array}$

   Please note: We do not use this test statistic to derive a confidence interval for ρ.

2. (i) We may use the test statistic $\mathbf{Z} = \sqrt{n-3}\,(U - \eta) \sim n\,(0;1)$

   to test $\mathbf{H}_0 : \quad \rho = \rho_0$ against $\begin{array}{ll} \diagup & H_1 : \rho > \rho_0 \text{ or} \\ - & H_1 : \rho < \rho_0 \text{ or} \\ \diagdown & H_1 : \rho \neq \rho_0. \end{array}$

   (Please note that ρ_0 is now a value other than zero!)

   $U = \dfrac{1}{2} \ln \dfrac{1+R}{1-R}$ and $\eta = \dfrac{1}{2} \ln \dfrac{1+\rho}{1-\rho}$

   (ii) We use this test statistic to derive a confidence interval for ρ.

Activity 5.14

Suppose you have the following random sample of 12 paired observations $(X_i; Y_i)$ from a bivariate normal population, with correlation coefficient ρ.

X_i	7	9	10	10	8	11	13	12	13	12	15	16
Y_i	15	10	12	18	21	20	22	21	30	25	23	26

Complete the following table by filling in the blank cells and then answer the questions:

X_i	Y_i	X_iY_i	X_i^2	Y_i^2	$(X_i - X)$	$(X_i - X)^2$	$(Y_i - Y)$	$(Y_i - Y)^2$	$(X_i - X) \cdot (Y_i - Y)$
7	15	105	49	225	-4.33	18.75	-5.25	27.56	
9	10	90		100	-2.33	5.43	-10.25	105.06	23.88
10	12	120	100	144		1.77	-8.25	68.06	10.97
10	18		100	324	-1.33	1.77		5.06	2.99
8	21	168	64	441	-3.33	11.09	0.75	0.56	-2.50
11	20		121		-0.33	0.11	-0.25	0.06	0.08
13	22	286	169	484	1.67		1.75	3.06	
12	21	252		441	0.67	0.45		0.56	0.50
13	30	390	169		1.67	2.79	9.75	95.06	16.28
12	25	300	144	625		0.45	4.75	22.56	3.18
15	23	345	225	529	3.67	13.47	2.75	7.56	10.09
16	26	416	256	676	4.67	21.81	5.75	33.06	26.85
136	**243**	**2872**	**1622**	**5289**	**0**	**80.67**	**0**	**368.25**	**118.00**

(a) Use the different total values in the table to show that

$$\sum (X_i - X)^2 = \sum X_i^2 - \frac{\left(\sum X_i\right)^2}{n}.$$

.....

.....

.....

(b) Use the different total values in the table to show that

$$\sum (X_i - X)(Y_i - Y) = \sum X_i Y_i - \frac{\left(\sum X_i\right)\left(\sum Y_i\right)}{n}.$$

.....

.....

.....

(c) Compute R, the sample correlation coefficient.

.....

.....

.....

.....

Feedback → Feedback

X_i	Y_i	X_iY_i	X_i^2	Y_i^2	$(X_i - X)$	$(X_i - X)^2$	$(Y_i - Y)$	$(Y_i - Y)^2$	$(X_i - X) \cdot (Y_i - Y)$
7	15	105	49	225	-4.33	18.75	-5.25	27.56	22.73
9	10	90	81	100	-2.33	5.43	-10.25	105.06	23.88
10	12	120	100	144	-1.33	1.77	-8.25	68.06	10.97
10	18	180	100	324	-1.33	1.77	-2.25	5.06	2.99
8	21	168	64	441	-3.33	11.09	0.75	0.56	-2.50
11	20	220	121	400	-0.33	0.11	-0.25	0.06	0.08
13	22	286	169	484	1.67	2.79	1.75	3.06	2.92
12	21	252	144	441	0.67	0.45	0.75	0.56	0.50
13	30	390	169	900	1.67	2.79	9.75	95.06	16.28
12	25	300	144	625	0.67	0.45	4.75	22.56	3.18
15	23	345	225	529	3.67	13.47	2.75	7.56	10.09
16	26	416	256	676	4.67	21.81	5.75	33.06	26.85
136	**243**	**2872**	**1622**	**5289**	**0**	**80.67**	**0**	**368.25**	**118.00**

(a) From the table (seventh column), we see that $\sum \left(X_i - \overline{X}\right)^2 = 80.67$.

RHS: $$\sum X_i^2 - \frac{\left(\sum X_i\right)^2}{n} = 1622 - \frac{(136)^2}{12}$$
$$= 1622 - 1541.33$$
$$= 80.67$$

Thus $$\sum (X_i - X)^2 = \sum X_i^2 - \frac{\left(\sum X_i\right)^2}{n}$$

Please note: Some of you might think that this is a trivial result and that you do not need to know it, because you will in any case do all the computations on a smart calculator. Very well, but sometimes we are mean enough not to give you the original data, but only some of the totals!

(b) From the table (tenth column), we see that $\sum (X_i - X)(Y_i - Y) = 118$.

RHS: $$\sum X_iY_i - \frac{\left(\sum X_i\right)\left(\sum Y_i\right)}{n} = 2872 - \frac{(136)(243)}{12}$$
$$= 118$$

Thus $$\sum (X_i - X)(Y_i - Y) = \sum X_iY_i - \frac{\left(\sum X_i\right)\left(\sum Y_i\right)}{n}$$

(c) $$R = \frac{\sum (X_i - X)(Y_i - Y)}{\sqrt{\sum (X_i - X)^2 \sum (Y_i - Y)^2}}$$

$$= \frac{118}{\sqrt{(80.67)(368.25)}}$$

$$= 0.685$$

Activity 5.15

Refer to the data of the **previous activity (5.14)**.

Test $H_0 : \rho = 0$ against
$H_1 : \rho > 0$ at the 1% level of significance.

Feedback → Feedback

Since we are testing $H_0 : \rho = 0,$ we may use the test statistic

$$T = \frac{\sqrt{n-2}\,(R)}{\sqrt{1-R^2}} \sim t_{n-2}. \qquad \text{(theorem 5.4)}$$

For the data of activity 5.14, we have already computed $R = 0.685.$

$$\begin{aligned} \therefore T &= \frac{\sqrt{12-2}\,(0.685)}{\sqrt{1-(0.685)^2}} \\ &= \frac{2.166}{0.7285} \\ &= 2.9734 \end{aligned}$$

We will reject the null hypothesis and accept the one-sided alternative if $T \geq t_{0.01;10} = 2.764$ (table III Stoker).

Since $2.9734 > 2.764$, we reject H_0 and conclude that there is a significant positive correlation between X and Y.

Activity 5.16

Refer to the data of activity 5.14.

Test $H_0 : \rho = 0.6$ against
$H_1 : \rho \neq 0.6$ at the 5% level of significance.

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

Feedback  Feedback

Since we are testing $H_0 : \rho = \rho_0$ (with $\rho_0 = 0.6 \neq 0$), we have to use the test statistic defined in **theorem 5.5**.

Using the value $R = 0.685$ (from **activity 5.14**), we find that

$$\begin{aligned} U &= \frac{1}{2}\ln\frac{1+R}{1-R} \\ &= \frac{1}{2}\ln\frac{1+0.685}{1-0.685} \\ &= \frac{1}{2}\ln 5.349 \\ &= 0.8385. \end{aligned}$$

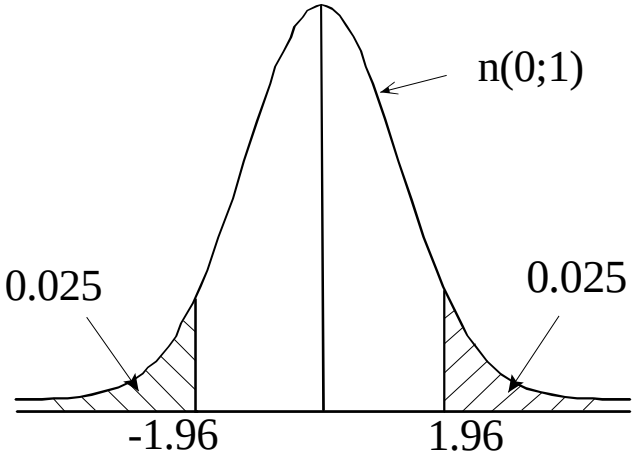
We also compute

$$\begin{aligned} \eta &= \frac{1}{2}\ln\frac{1+\rho_0}{1-\rho_0} \\ &= \frac{1}{2}\ln\frac{1+0.6}{1-0.6} \\ &= \frac{1}{2}\ln 4.0 \\ &= 0.6931. \end{aligned}$$

Although we have a smallish sample $(n = 12)$ we still assume that $Z = \sqrt{n-3}\,(U-\eta) \sim n\,(0;1)$ is the valid test statistic. So,

$$\begin{aligned} Z &= \sqrt{12-3}\,(0.8385 - 0.6931) \\ &= 3\,(0.1454) \\ &= 0.4362. \end{aligned}$$

We reject the null hypothesis if $Z \geq z_{\alpha/2}$ or $Z \leq -z_{\alpha/2}$.



Since $-1.96 < 0.4362 < 1.96$ we cannot reject H_0.

Activity 5.17

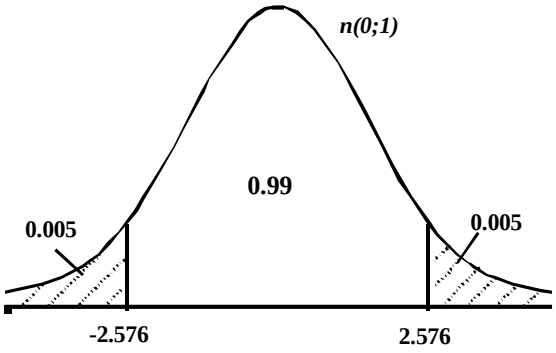
Refer to the data of activity 5.14.

Derive a 99% confidence interval for ρ.

Feedback Feedback

We have to use the test statistic in **theorem 5.5** to derive a confidence interval for ρ.

The first step is to find the 99% bounds for the $n(0;1)$ distribution.



$P\left[-z_{\alpha/2} \leq Z \leq z_{\alpha/2}\right] = 1-\alpha.$

From table II (Stoker), we find that $z_{0.005} = 2.576$

$\therefore P\left[-2.576 \leq Z \leq 2.576\right] = 0.99.$

The second step is to transform back to η (which is a function of the unknown ρ).

$\therefore P\left[-2.576 \leq \sqrt{n-3}\,(U-\eta) \leq 2.576\right] = 0.99$

$$\therefore P\left[U - \frac{2.576}{\sqrt{n-3}} \leq \eta \leq U + \frac{2.576}{\sqrt{n-3}}\right] = 0.99$$

Using the value of U that was computed in activity 5.16 and substituting $\sqrt{n-3} = \sqrt{9} = 3$ we find that

$$P\left[0.8385 - \frac{2.576}{3} \leq \eta \leq 0.8385 + \frac{2.576}{3}\right] = 0.99.$$

$\therefore$ The 99% confidence interval for η is

$$-0.0202 \leq \eta \leq 1.6972.$$

Now the third (and final) step is to convert this in terms of ρ by using the function

$$\rho = \frac{e^{\eta} - e^{-\eta}}{e^{\eta} + e^{-\eta}}.$$

For the lower bound, we get

$$\begin{aligned}\rho &= \frac{e^{-0.0202} - e^{-(-0.0202)}}{e^{-0.0202} + e^{-(-0.0202)}} \\ &= \frac{-0.04}{2.000} \\ &= -0.02.\end{aligned}$$

For the upper bound, we get

$$\begin{aligned}\rho &= \frac{e^{1.6972} - e^{-1.6972}}{e^{1.6972} + e^{-1.6972}} \\ &= \frac{5.275}{5.6418} \\ &= 0.935.\end{aligned}$$

Thus, we are 99% confident that the true (but unknown) value of ρ is between -0.02 and 0.935, that is $-0.02 \leq \rho \leq 0.935$.

Please note: This is a very wide confidence interval, because we have chosen a very large probability and also because we have such a small sample.

The last subsection of section 5.4 is to test whether *two correlation coefficients* for two independent bivariate normal populations are the same, in other words, we wish to test

$$H_0 : \rho_1 = \rho_2.$$

We know that whenever we want to test a null hypothesis, we need a test statistic that will be "connected" to the null hypothesis and of which the distribution will be known.

☺ Think of the test statistic that will test this hypothesis as the "*two-sample extension*" of theorem 5.5.

Activity 5.18

Indicate whether the following statements are true or false: (Do not simply say "true" or "false", but try to substantiate your conclusion!!)

(a) If $Z = \sqrt{n-3}\,(U-\eta) \sim n\,(0;1)$, then $U - \eta \sim n\left(0; \frac{1}{n-3}\right)$.

.....

.....

.....

(b) If (in general) $X \sim n\left(\mu_1; \sigma_1^2\right)$ and $Y \sim n\left(\mu_2; \sigma_2^2\right)$ then $X - Y \sim n(\mu_1 - \mu_2; \sigma_1^2 + \sigma_2^2)$.

.....

.....

.....

(c) If $Z_1 = \sqrt{n_1-3}\,(U_1-\eta_1) \sim n\,(0;1)$, then $U_1 - \eta_1 \sim n\left(0; \frac{1}{n_1-3}\right)$.

.....

.....

.....

(d) If $Z_2 = \sqrt{n_2-3}\,(U_2-\eta_2) \sim n\,(0;1)$, then $U_2 - \eta_2 \sim n\left(0; \frac{1}{n_2-3}\right)$.

.....

.....

.....

(e) If $U_1 - \eta_1 \sim n\left(0; \frac{1}{n_1-3}\right)$ independent of $U_2 - \eta_2 \sim n\left(0; \frac{1}{n_2-3}\right)$, then

$(U_1 - \eta_1) - (U_2 - \eta_2) \sim n\left(0; \frac{1}{n_1-3} + \frac{1}{n_2-3}\right)$.

.....

.....

.....

Feedback Feedback

(a) True.

$U - \eta = \frac{1}{\sqrt{n-3}} \cdot Z$

Now we have to apply the following very important theorem (from first-year statistics!):

If (in general) $Y = aX + k$

then $E(Y) = aE(X) + k$

$var(Y) = a^2 var(X)$

$$\Longrightarrow \quad var(U-\eta) = \left(\frac{1}{\sqrt{n-3}}\right)^2 var(Z)$$

$$= \frac{1}{n-3}(1) \qquad \text{since } Z \sim n(0;1)$$

and $E(U-\eta) = \frac{1}{\sqrt{n-3}} E(Z) = 0.$

(b) Only true if X and Y are **independent**!

Again, we apply a very important theorem from first-year statistics!

If (in general) $V = aX + bY$

then $E(V) = aE(X) + bE(Y)$

and $var(V) = a^2 var(X) + b^2 var(Y) + 2ab\, cov(X;Y)$

$X - Y = 1(X) + (-1)Y$

$\Longrightarrow \quad var(X-Y) = 1^2 var(X) + (-1)^2 var(Y) + 2(1)(-1)\ cov(X;Y)$

$= \sigma_1^2 + \sigma_2^2 - 2cov(X;Y).$

If X and Y are independent, $Cov(X;Y) = 0$ and the last term will fall away.

(c) This follows directly from (a), which was true!

(We have only changed the statement by giving subscripts!)

So, if $Z_1 = \sqrt{n_1 - 3}\,(U_1 - \eta_1) \sim n(0;1)$, then $U_1 - \eta_1 \sim n\left(0; \frac{1}{n_1 - 3}\right)$.

(d) True!

(This follows from (a) and (c).)

(e) True!

(This follows from the feedback to question (b) and from (c) and (d) above.)

Activity 5.19

Refer to (e) of the previous activity

If you know that $(U_1 - \eta_1) - (U_2 - \eta_2) \sim n\left(0; \frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}\right)$, show how you will standardise this variable to obtain a $n(0;1)$ variable.

.....

.....

.....

.....

Feedback Feedback

Let $X = (U_1 - \eta_1) - (U_2 - \eta_2) \sim n\left(0; \frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}\right)$

$\Longrightarrow E(X) = 0$ and $var(X) = \frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}$

So $Z = \frac{X - E(X)}{\sqrt{var(X)}} = \frac{X - 0}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}$

that is, $Z = \frac{(U_1 - \eta_1) - (U_2 - \eta_2)}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}} \sim n(0;1)$.

Activity 5.20

In a random sample of 43 paired observations, a correlation coefficient was computed as $R = 0.7$ and in a second random sample of 73 paired observations, a correlation coefficient of $R = 0.65$ was computed. Assume that both samples are from bivariate normal populations and that the two populations are independent. Test $H_0 : \rho_1 = \rho_2$ against $H_1 : \rho_1 > \rho_2$ at the 5% level of significance.

.....

.....

Feedback → Feedback

We use the test statistic $Z = \dfrac{U_1 - U_2}{\sqrt{\dfrac{1}{n_1 - 3} + \dfrac{1}{n_2 - 3}}} \sim n\,(0;1)$ to test $H_0 : \rho_1 = \rho_2$ against $H_1 : \rho_1 > \rho_2$.

$$\begin{aligned} U_1 &= \frac{1}{2}\ln\frac{1+R_1}{1-R_1} \qquad \text{and} \quad U_2 = \frac{1}{2}\ln\frac{1+R_2}{1-R_2} \\ &= \frac{1}{2}\ln\frac{1+0.7}{1-0.7} \qquad \vdots \qquad = \frac{1}{2}\ln\frac{1+0.65}{1-0.65} \\ &= 0.867 \qquad \vdots \qquad = 0.775 \end{aligned}$$

$$\begin{aligned} \therefore Z &= \frac{0.867 - 0.775}{\sqrt{\dfrac{1}{43-3} + \dfrac{1}{73-3}}} \\ &= \frac{0.092}{0.198} \\ &= 0.464 \end{aligned}$$

We will reject H_0 if $Z \geq z_{0.05} = 1.645$.

Since $0.464 < 1.645$ we cannot reject H_0.

What about JMP?

We conclude this section by doing **electronically with JMP** what was done **manually** in the study guide! This is primarily for academic purposes. Redoing something like activity 5.14 (including a very small sample!) on a computer does not really make you appreciate enough how powerful and fast a computer is and how it takes the arduous labour out of computations, but it provides a basis for comparison. (We are spoiled by fancy calculators and computers and we want to laugh at the sight of a calculation table as in activity 5.14!) In real-life situations where you have big data sets, it is almost impossible to do a correlation/regression analysis by hand.

I know that you would not be able to resist playing around with all the different options in JMP. If you understand theoretically what buttons you press and why you do it, it is fine! However, you should keep in mind that regression (which we discuss in study unit 8) is a powerful technique and becomes very complicated when there are more than one predictor variable. As you progress with your studies (STA2601, the forerunner for STA3701; STA3701, the forerunner for STA4801 and STA4804, et cetera) you will learn how to employ JMP in an increasingly complicated manner to aid you in your regression analysis.

Activity 5.21

Refer to **activity 5.14**.

(a) *Capture the raw data* in a JMP data table.

(b) *Compute R*. (Compare this with the feedback of activity 5.14.)

(c) Test the hypothesis that the two variables are uncorrelated against the alternative that they are positively correlated.

(d) Is it possible to test the hypothesis that the two variables have a correlation coefficient of $\rho = 0.6$ against the alternative that $\rho \neq 0.6$?

(e) Comment on the similarities (and the differences) between the manual test and the electronic version of JMP.

Feedback  Feedback

(a) *The data* table in JMP must have two columns (name them X and Y respectively) and 12 rows. Save it as *sta2601act5p14.jmp.*

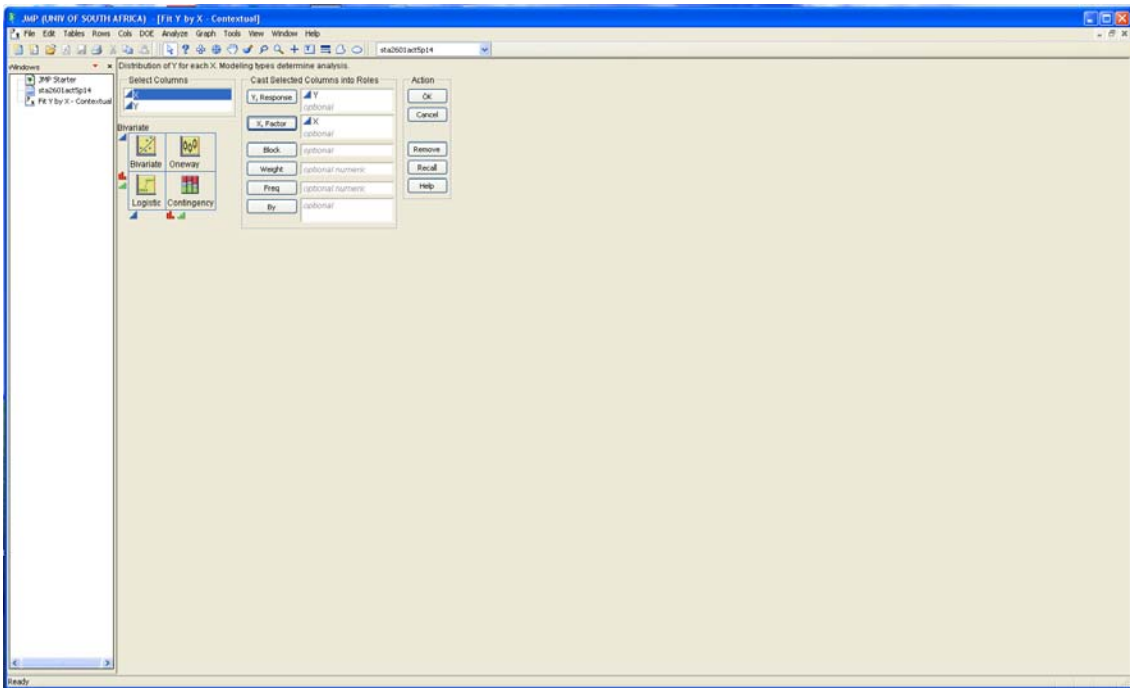
(b) [*For beginners:* How to c*ompute R*.

FIRST POSSIBILITY:

Select **Analyze** $\longrightarrow$ **Fit Y by X.** (Yes, this looks the same as the initial steps for the contingency table analysis of activity 5.8.)

Now you get a report from JMP where the top line reads as follows:

"**Distribution of Y for each X. Modeling types determine analysis"**



The only difference between this analysis and the contingency table analysis is that we now have two of the same icons: **two** *blue triangles* for the two continuous variables X and Y.

JMP automatically uses the *icon of a blue triangle* to identify the two continuous variables in the variable box and all you have to do is to select the correct "picture display" of the given four possible "**Y by X**" analyses (which is of course in the top left-hand corner)! This means that if Y=*a blue triangle* and *X=a blue triangle* you will have to do a **Bivariate** analysis (and this is what we already know to do because we study statistics).

If you click on the **OK** button you immediately get the "**Bivariate Fit of Y By X"** output which is a *scatter plot*.

Click on the **Bivariate Fit of Y By X** bar (there is a red triangle where you have to click) and open the submenu:

Show Points
Fit Mean
Fit Line
Fit Polynomial
↓
Script

where you select **Fit Line.**]

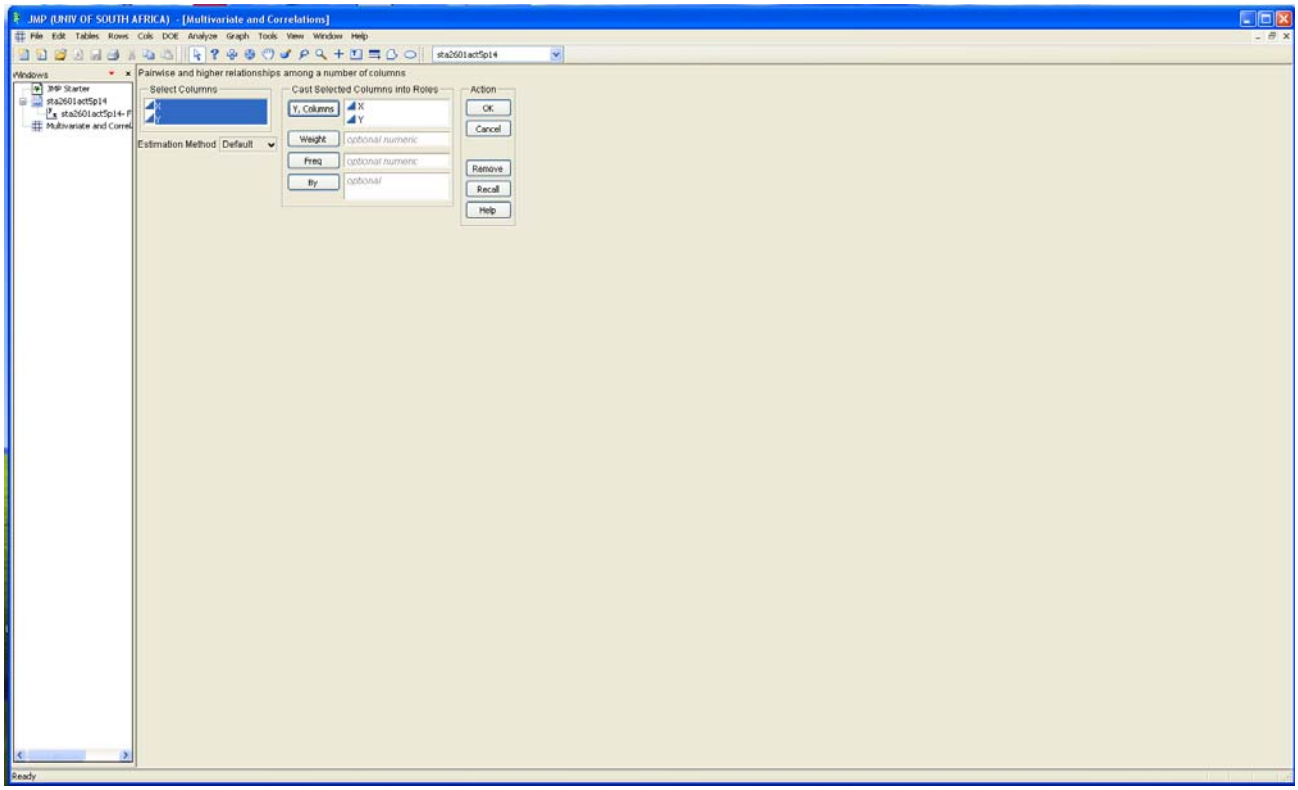
By doing this, we are actually busy doing a regression analysis (which will only be dealt with in detail in study unit 8). You will notice the **Summary of Fit** gives **RSquare** $= 0.468735$ in the first line. In activity 5.14 we computed $R = 0.685 \Longrightarrow R^2 = (0.685)^2 = 0.469$ (rounded to three decimal places). This is about all that we can use from this output!

Summary of Fit	
RSquare	0.468735
RSquare Adj	0.415608
Root Mean Square Error	4.423103
Mean of Response	20.25
Observations (or Sum Wgts)	12

SECOND POSSIBILITY:

Select **Analyse** $\longrightarrow$ **Multivariate Methods**$\longrightarrow$ **Multivariate.**

Make sure you move both variables, X and Y, into the **Y, Columns** box, and click **OK**. (For a proper multivariate analysis you will move more than two variables into this box.)



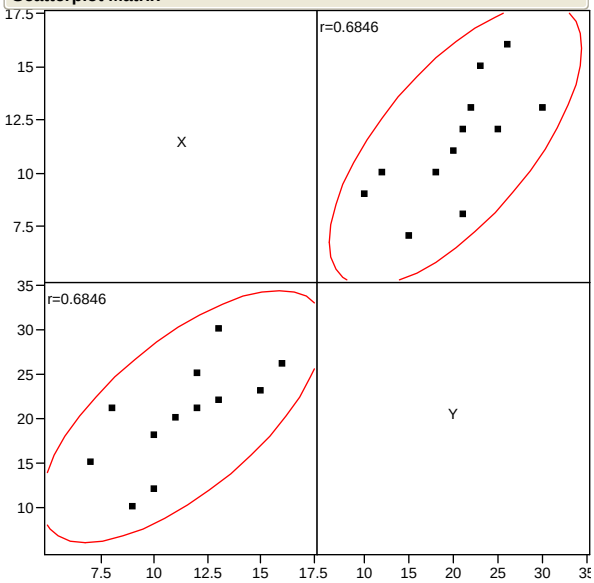
The output shows a 2×2 correlation matrix and a 2×2 scatter plot matrix. (Should you have entered five variables it would show a 5×5 matrix.) From the output we get the direct value of $R_{X,Y} = 0.6846$ as well as the appropriate graphical display.

Multivariate

Correlations

	X	Y
X	1.0000	0.6846
Y	0.6846	1.0000

Scatterplot Matrix



(c) JMP does not include a hypothesis test to test that the two variables are uncorrelated against the alternative that they are positively correlated.

(d) JMP does not include a hypothesis test to test the hypothesis that the two variables have a correlation coefficient of $\rho = 0.6$ against the alternative that $\rho \neq 0.6$.

(e) If we choose the **Analyze** $\longrightarrow$ **Fit Y by X** option to obtain a fitted regression line, JMP gives the **t Ratio** as 2.97 for the slope (under the output **Parameter Estimates)** which is in effect a test for $H_0 : \rho = 0$. We will come back to this point when we discuss regression. (If you go back to activity 5.14 you will see that we computed $T = 2.9734$ and rejected the null hypothesis $H_0 : \rho = 0.)$

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	3.6714876	5.725519	0.64	0.5358
X	1.4628099	0.49247	2.97	0.0140*

The manual procedure allows for more (as well as direct) possible tests of ρ than the electronic version using JMP.

You should condition yourself to feel in control when you use JMP. Only apply procedures (and interpret the output!) of which you understand the theory.

 Go back to the learning outcomes of this study unit in the study guide. Can you tick them off with confidence?

STUDY UNIT 6

Inference on variances

In study unit 4 we discussed different tests for normality, because normality is an important assumption underlying the appropriateness of many statistical analyses. You will notice in the next study unit, which focuses on various tests of means, that another important assumption will be the one of "equal variances". So, in this study unit we will develop such tests but we will also discuss inference on variances as methods standing on their own and not necessarily being the prerequisite for other tests.

6.1 One-sample problem

There could be real-life situations when the *variability* of a variable is our main focus. We most often don't know what the values of the parameters of a distribution are, and we could estimate the variance of an unknown population by computing a *confidence interval for the variance*.

Do you recall that in statistics, *inference* means we have as evidence the limited information contained in the outcome of a sample and we want to conclude something about the unknown population from which the sample was drawn? We could also be interested in *testing a hypothesis* about a variance. Testing a hypothesis or computing a confidence interval is called *inference on a variance*.

Activity 6.1

Give your own interpretation of the application of theorem 6.1 of the study guide.

.....

.....

.....

Feedback  Feedback

Theorem 6.1 has two distinct applications:

Either we will *know* what the value of the population mean, μ, is (and thus use the (a) part which has a χ^2_n distribution) or we will *not know* what the value of the population mean, μ, is (and thus use the (b) part which has a χ^2_{n-1} distribution).

Activity 6.2

Using the notation of the study guide, what does the following mean?

(a) $\chi^2_{0.975;\,4}$

.....

.....

(b) $\chi^2_{0.025;\,7}$

.....

.....

(c) $\chi^2_{0.90;\,8}$

.....

.....

(d) $\chi^2_{p;\,\upsilon}$

.....

.....

Feedback → Feedback

(a) $\chi^2_{0.975;\,4}$ will be a *numerical value* (read from table IV Stoker) that is associated with a χ^2-variable with $\upsilon = 4$ degrees of freedom such that the probability to the right of this value is 0.975 (implying that probability to the left of this value is 0.025).

$$P(\chi^2 \geq \chi^2_{0.975;\,4}) = 0.975$$

(b) $\chi^2_{0.025;\,7}$ will be a *numerical value* (read from table IV Stoker) that is associated with a χ^2-variable with $\upsilon = 7$ degrees of freedom such that the probability to the right of this value is 0.025 (implying that probability to the left of this value is 0.975).

$$P(\chi^2 \geq \chi^2_{0.025;\,7}) = 0.025$$

(c) $\chi^2_{0.90;\,8}$ will be a *numerical value* (read from table IV Stoker) that is associated with a χ^2-variable with $\upsilon = 8$ degrees of freedom such that the probability to the right of this value is 0.90 (implying that probability to the left of this value is 0.10).

$$P(\chi^2 \geq \chi^2_{0.90;\,8}) = 0.90$$

(d) $\chi^2_{p;\,\upsilon}$ will be a *numerical value* that is associated with a χ^2-variable with υ degrees of freedom such that the probability to the right of this value is p (implying that probability to the left of this value is $1 - p$).

$$P(\chi^2 \geq \chi^2_{p;\,\upsilon}) = p$$

Activity 6.3

Let $X_1,\ X_2,\ ...,\ X_7$ be independent $n\left(8;\ 9\right)$ variables. Are the following statements *correct or incorrect*?

(a) $\sum_{i=1}^{7} \frac{\left(X_i - 8\right)^2}{9}$ has a χ_7^2 distribution.

.....

(b) $\sum_{i=1}^{7} \frac{\left(X_i - \overline{X}\right)^2}{9}$ has a χ_6^2 distribution.

.....

(c) $E\left[\sum_{i=1}^{7} \frac{\left(X_i - \overline{X}\right)^2}{9}\right] = 6$

.....

(d) $Var\left[\sum_{i=1}^{7} \frac{\left(X_i - 8\right)^2}{9}\right] = 12$

.....

(e) $\frac{\left(X_1 - 8\right)^2}{9}$ has a X_1^2 distribution.

.....

(f) If $a < X < b \Rightarrow \frac{1}{a} > \frac{1}{X} > \frac{1}{b}$ provided that a and b are positive.

.....

Feedback → Feedback

(a) **Correct**. From theorem 6.1 (a) (which = **result 1.2** of study unit 1) it follows that

$$\sum_{i=1}^{n} \frac{(X_i-\mu)^2}{\sigma^2} \sim \chi^2_n \text{ and thus } \sum_{i=1}^{7} \frac{(X_i-8)^2}{3^2} = \sum_{i=1}^{7} \frac{(X_i-8)^2}{9} \sim \chi^2_7.$$

(b) **Correct**. From theorem 6.1 (b) (which = **result 1.3** of study unit 1) it follows that

$$\sum_{i=1}^{n} \frac{(X_i-\overline{X})^2}{\sigma^2} \sim \chi^2_{n-1} \text{ and thus } \sum_{i=1}^{7} \frac{(X_i-\overline{X})^2}{3^2} = \sum_{i=1}^{7} \frac{(X_i-\overline{X})^2}{9} \sim \chi^2_{7-1}.$$

(c) **Correct**. From **property (i) of result 1.1** it follows that

$$E\left(\chi^2_n\right) = n.$$

From question (b) above it holds that $E\left(\chi^2_6\right) = 6.$

(d) **Incorrect**. From **property (i) of result 1.1** it follows that

$$Var\left(\chi^2_n\right) = 2n.$$

From question (a) above it holds that $Var\left(\chi^2_7\right) = 2\,(7) = 14.$

(e) **Correct.** From **property (iv) of result 1.1** it follows that if $n = 1$ then $\left(\frac{X-\mu}{\sigma}\right)^2 \sim \chi^2_1.$

This means that $\frac{(X_1-8)^2}{9}$ does indeed have a χ^2_1 distribution.

(f) **Correct.**

Activity 6.4

Refer to activity 6.3 above. Suppose we let $U_1 = \sum_{i=1}^{7} \frac{(X_i - 8)^2}{9}$ and $U_2 = \sum_{i=1}^{7} \left[\frac{X_i - \overline{X}}{3}\right]^2$.

Use table IV of Stoker, to find the following $\chi^2_{\alpha,f}$ values:

(a) $P(\chi^2_{1-\alpha/2,\,f} < U_1 < \chi^2_{\alpha/2,\,f}) = 0.95$

.....

.....

(b) $P(\chi^2_{1-\alpha/2,\,f} < U_2 < \chi^2_{\alpha/2,\,f}) = 0.95$

.....

.....

(c) $P(U_1 > \chi^2_{1-\alpha,\,f}) = 0.95$

.....

.....

(d) $P(U_2 > \chi^2_{1-\alpha,\,f}) = 0.95$

.....

.....

(e) $P(U_1 < \chi^2_{\alpha,\,f}) = 0.95$

.....

.....

(f) $P(U_2 < \chi^2_{\alpha,\,f}) = 0.95$

.....

.....

Feedback → Feedback

(a) $U_1 = \sum_{i=1}^{7} \frac{(X_i - 8)^2}{9}$ has a χ^2_7 distribution $\Longrightarrow P(\chi^2_{0.975;7} < U_1 < \chi^2_{0.025;7}) = P(1.68987 < U_1 < 16.0128) = 0.95.$

(b) $U_2 = \sum_{i=1}^{7} \left[\frac{X_i - \overline{X}}{3}\right]^2$ has a χ^2_6 distribution $\Longrightarrow P(\chi^2_{0.975;6} < U_2 < \chi^2_{0.025;6}) = P(1.237347 < U_2 < 14.4494) = 0.95.$

(c) $P(U_1 > \chi^2_{0.95;\,7}) = P(U_1 > 2.16735) = 0.95$

(d) $P(U_2 > \chi^2_{0.95;6}) = P(U_2 > 1.63539) = 0.95$

(e) $P(U_1 < \chi^2_{0.05;7}) = P(U_1 < 14.0671) = 0.95$

(f) $P(U_2 < \chi^2_{0.05;6}) = P(U_2 < 12.5916) = 0.95$

Activity 6.5

In a random sample of size $n = 15$ is was found that $\sum_{i=1}^{15} X_i = 51$ and $\sum_{i=1}^{15} X_i^2 = 210.$

Assume that $\mu = 3$ for the population from which the sample was drawn.

(a) Compute $\sum_{i=1}^{15} (X_i - \mu)^2$.

.....

.....

.....

.....

(b) Would you accept $H_0 : \sigma^2 = 2$ against $H_0 : \sigma^2 \neq 2$ for this population if you assume normality? Use $\alpha = 0.05$.

.....

.....

.....

.....

.....

.....

.....

.....

Feedback ⟶ Feedback

(a) $\sum_{i=1}^{15}(X_i - \mu)^2 = \sum_{i=1}^{15} {X_i}^2 - 2\mu \sum_{i=1}^{15} X_i + n\mu^2 = 210 - 2(3)(51) + 15(3)^2 = 210 - 171.0 = 39.0$

(b) To test $H_0 : \sigma^2 = 2$ against $H_0 : \sigma^2 \neq 2$ we use the test statistic

$U = \dfrac{\sum_{i=1}^{n}(X_i - \mu)^2}{c} \sim \chi^2_n$ (assuming normality).

We reject H_0 two-sided if $U \leq \chi^2_{1-\alpha/2;\, f}$ or if $U \geq \chi^2_{\alpha/2;\, f}$.

$U = \dfrac{\sum_{i=1}^{15}(X_i - \mu)^2}{c} = \dfrac{39}{2} = 19.5 \sim \chi^2_{15}$

We reject H_0 two-sided at the 5% level of significance if $U \leq \chi^2_{0.975;15} = 6.26214$ or if $U \geq \chi^2_{0.025;15} = 27.4884$.

Since $6.26214 \leq U \leq 27.4884$ we cannot reject H_0.

Activity 6.6

Consider the following random sample of size $n = 20$ drawn from a normal population with unknown mean and unknown variance.

14	15	22	17	16	17	23	25	19	18
22	17	18	21	15	22	20	21	18	20

(a) Test the hypothesis that the variance is at most 7. Use $\alpha = 0.05$.

(b) Derive a 95% one-sided confidence interval in accordance with the hypothesis test.

Feedback Feedback

(a) $\sum_{i=1}^{n} X_i = 380 \Longrightarrow \overline{X} = \frac{380}{20} = 19$ and $\sum_{i=1}^{20}(X_i - \overline{X})^2 = \sum_{i=1}^{20}(X_i - 19)^2 = 170$

To test $H_0 : \sigma^2 = 7$ against $H_0 : \sigma^2 < 7$ we use the test statistic

$U = \dfrac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{c} \sim \chi^2_{n-1}$ and we reject H_0 left-sided if $U \leq \chi^2_{1-\alpha;\, f}$.

$$U = \frac{\sum_{i=1}^{20}(X_i - \overline{X})^2}{7} = \frac{170}{7} = 24.2857 \sim \chi^2_{19}$$

We reject H_0 left-sided at the 5% level of significance if $U \leq \chi^2_{0.95;\,19} = 10.1170$.

Since $U > 10.1170$ we cannot reject H_0.

(b) A 95% one-sided confidence interval in accordance with the alternative hypothesis is the interval

$$[0; \frac{\sum_{i=1}^{20}(X_i - \overline{X})^2}{\chi^2_{1-\alpha;\,n-1}}] = [0; \frac{170}{\chi^2_{0.95;\,19}}] = [0; \frac{170}{10.1170}] = [0; 16.8034].$$

Since this interval includes the value 7, we cannot reject H_0. (Of course this must never contradict the conclusion of question (a).)

What about JMP?

How will you create a confidence interval or perform a hypothesis test for a single sample variance using JMP? There is not a separate analysis in JMP to do this!

6.2 Two independent samples

You will find exactly the same heading of "Two independent samples" in the following study unit when we deal with the topic of *inference on means*. There is a definite connection between this subsection and its counterpart in the following study unit, because the hypothesis test we derive in this section is the *key to checking the theoretical assumption* $\sigma_1^2 = \sigma_2^2$ we will need in the following study unit when we discuss the difference between two means. So, already make a mental note of this connection. Another reason why you have to make this connection, is that JMP does not treat the "testing of variances" as a main option but only as part of the subtests when the main issue is the testing of means. In other words, you have to "manipulate" JMP to obtain variance tests.

Activity 6.7

Let $X_1;\ X_2;\ ...;\ X_{10}$ be a random sample from a $n\,(100;\ 100)$ distribution and let $Y_1;\ Y_2;\ ...;\ Y_{12}$ be a random sample from an independent $n\,(50;\ 81)$ distribution.

Suppose we define

$$\begin{aligned} V_1 &= \sum_{i=1}^{10} \left[(X_i - 100)/10\right]^2 \\ V_2 &= \sum_{i=1}^{12} \left[(Y_i - 50)/9\right]^2 \\ V_3 &= \sum_{i=1}^{10} \left[\left(X_i - \overline{X}\right)/10\right]^2 \\ V_4 &= \sum_{i=1}^{12} \left[\left(Y_i - \overline{Y}\right)/9\right]^2. \end{aligned}$$

Indicate whether the following statements are *correct* or *incorrect*.

(a) If $U_1 = \dfrac{V_1/10}{V_2/12} \Longrightarrow U_1 \sim F_{10;\,12}$

.....

(b) $\dfrac{1}{U_1} \sim F_{10;\,12}$

.....

(c) $P\left(U_1 < 2.75\right) = 0.95$

.....

(d) $P\left(\dfrac{1}{U_1} < 2.91\right) = 0.95$

.....

(e) If $U_2 = \dfrac{V_3/9}{V_4/11} \Longrightarrow U_2 \sim F_{9;\,11}$

.....

(f) $P\left(\dfrac{1}{U_2} > 3.105\right) = 0.05$

.....

(g) If the observed value of the test statistic $U_1 = 2.1$ when testing the null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$ against $H_1 : \sigma_1^2 \neq \sigma_2^2$, we cannot reject H_0 at the 5% level.

.....

(h) Testing $H_1 : \sigma_1^2 > \sigma_2^2$ is the same as testing $H_1 : \dfrac{\sigma_2^2}{\sigma_1^2} > 1$.

.....

Feedback ⟶ Feedback

(a) Correct. $V_1 \sim \chi^2_{10}$ and $V_2 \sim \chi^2_{12}$ and hence $U_1 = \dfrac{V_1/10}{V_2/12} \sim F_{10;\,12}$.

(b) Incorrect. If we change the roles of V_1 and V_2, we have $\dfrac{1}{U_1} \sim F_{12;10}$.

(c) Correct. Since $U_1 \sim F_{10;12}$ and thus $\nu_1 = 10$ and $\nu_2 = 12$, we see from table V (Stoker) that $P\left(U_1 < 2.75\right) = 0.95$

(d) Correct. Since $\dfrac{1}{U_1} \sim F_{12;10}$ and now $\nu_1 = 12$ and $\nu_2 = 10$, we see from table V (Stoker) that $P\left(\dfrac{1}{U_1} < 2.91\right) = 0.95$.

(e) Correct. $V_3 \sim \chi^2_9$ and $V_4 \sim \chi^2_{11}$ and hence $U_2 = \dfrac{V_3/9}{V_4/11} \sim F_{9;11}$.

(f) Correct. Since $\dfrac{1}{U_2} \sim F_{11;9} \Longrightarrow P\left(\dfrac{1}{U_2} > 3.105\right) = 0.05$.

(The value for $\nu_1 = 11$ and $\nu_2 = 9$ is not tabulated and we interpolate between $\nu_1 = 10$ and $\nu_2 = 9$; $\nu_1 = 12$ and $\nu_2 = 9$.)

(g) Correct. From table VI (Stoker) we read $F_{0.025;\,10;12} = 3.37$. The lower point is $F_{0{,}975;\,10;12}$, which is equal to $\dfrac{1}{F_{0{,}025;\,12;10}} = \dfrac{1}{3.62} = 0.276\,24$ and since $0.276\,24 < U_1 < 3.37$ we cannot reject H_0.

(h) Incorrect. Testing $H_1 : \sigma_1^2 > \sigma_2^2$ is the same as testing $H_1 : \frac{\sigma_2^2}{\sigma_1^2} < 1$, because

$$\begin{aligned} & \sigma_1^2 > \sigma_2^2 \\ \Longrightarrow \quad & \tfrac{\sigma_1^2}{\sigma_1^2} > \tfrac{\sigma_2^2}{\sigma_1^2} \\ \Longrightarrow \quad & 1 > \tfrac{\sigma_2^2}{\sigma_1^2}. \end{aligned}$$

The activity above assesses whether you understand the theory behind the test and now you are ready to apply it to a real-life situation!

Activity 6.8

The birth weights of 15 rats bred in laboratory A were:

230; 233; 226; 232; 235; 233; 229; 231; 236; 230; 234; 235; 234; 232; 234 grams. (Call this a sample from population A.)

The birth weights of 12 rats bred from a different cross-breed in laboratory B were:

231; 232; 229; 232; 231; 224; 232; 231; 225; 226; 235; 233 grams. (Call this a sample from population B.)

The researchers from laboratory B claim that the birth weights of their new cross-breed are **more homogeneous** than the birth weights from laboratory A.

(a) Express the claim as a proper hypothesis and test it at the 5% level of significance.

(b) On what assumptions is your test statistic based?

(c) Derive the appropriate 95% one-sided confidence interval for $\frac{\sigma_B^2}{\sigma_A^2}$ in accordance with the hypothesis to be tested.

(d) Derive an upper 95% bound for the ratio $\frac{\sigma_B}{\sigma_A}$ of the standard deviations.

Feedback Feedback

(a) “More homogeneous” means the same as “more similar” and this implies **less variability**. The null hypothesis is always the case including the "="sign, that is $H_0 : \sigma_A^2 = \sigma_B^2$ which implies we test against the alternative hypothesis $H_1 : \sigma_B^2 < \sigma_A^2$.

(b) We may use the test statistic $F = \frac{\sigma_2^2}{\sigma_1^2} \cdot \frac{S_1^2}{S_2^2} \sim F_{n_1-1;\, n_2-1}$ if we consider the two samples

from the different laboratories as independent samples from normal populations. For this application $n_1 = 15$ and $n_2 = 12$ (and “1” $=$ “A” and “2” $=$ “B”).

We need to compute S_A^2 and S_B^2.

$$\overline{X}_A = \frac{3484}{15} = 232.2667 \text{ and } S_A^2 = \frac{\sum_{i=1}^{15} (X_A - 232.2667)^2}{(15-1)} = \frac{100.9333}{14} = 7.2095$$

$$\overline{X}_B = \frac{2761}{12} = 230.0833 \text{ and } S_B^2 = \frac{\sum_{i=1}^{16} (X_B - 230.0833)^2}{(12-1)} = \frac{126.9167}{11} = 11.5378$$

Under $H_0 : \sigma_A^2 = \sigma_B^2$ the F-statistic becomes $F = \frac{S_A^2}{S_B^2} = \frac{7.2095}{11.5378} = 0.624\,86.$

The critical region to test H_0 against $H_1 : \sigma_B^2 < \sigma_A^2$ is where $F \geq F_{\alpha;\, n_1-1;\, n_2-1} \equiv F_{0.05;\, 14;\, 11} = 2.74$ (interpolated as $2.72 + \frac{1}{3}(2.79 - 2.72)$).

Since $F = 0.62489 < 2.74$ H_0 cannot be rejected.

(Please note that you could also take $F = \frac{S_B^2}{S_A^2} = \frac{11.5378}{7.2095} = 1.600\,4$ but then the

rejection region becomes $F \leq F_{0.95;\, 11;\, 14} = \frac{1}{F_{0.05;\, 14;\, 11}} = \frac{1}{2.74} = 0.364\,96$).

(c) The appropriate one-sided confidence interval for $\frac{\sigma_B^2}{\sigma_A^2}$ can be derived as:

$$P\left[\frac{\sigma_B^2}{\sigma_A^2} \cdot \frac{S_A^2}{S_B^2} < F_{\alpha;\, n_1-1;\, n_2-1}\right] = 1 - \alpha$$

$$\Rightarrow P\left[\frac{\sigma_B^2}{\sigma_A^2} < \frac{S_B^2}{S_A^2} F_{0.05;\, 14;\, 11}\right] = 0.95$$

$$\Rightarrow P\left[\frac{\sigma_B^2}{\sigma_A^2} < (1.600\,4)\,(2.74)\right] = 0.95$$

$$\Rightarrow P\left[\frac{\sigma_B^2}{\sigma_A^2} < 4.385\,1\right] = 0.95$$

So, based on the sample outcomes, a 95% one-sided confidence interval for $\frac{\sigma_B^2}{\sigma_A^2}$ is $(0;\ 4.385\,1)$.

Since this interval includes the value 1 $\Rightarrow \frac{\sigma_B^2}{\sigma_A^2} = 1 \Rightarrow \sigma_B^2 = \sigma_A^2$.

Only if the upper value for this interval is < 1 may we conclude that $\sigma_B^2 < \sigma_A^2$.

(d) An upper 95% bound for the ratio $\frac{\sigma_B^2}{\sigma_A^2}$ of the variances is $(0;\ 4.384\,51)$, hence an upper 95% bound for the ratio $\frac{\sigma_B}{\sigma_A}$ of the **standard deviations** is $\sqrt{4.384\,51} = 2.094\,0$.

What about JMP?

How will you create a confidence interval or perform a hypothesis test for two independent sample variances using JMP? We briefly discussed the situation of computers and testing for equal variances (in general) in the last section of study unit 6. We will now redo the activity above (where we have *two independent samples*) by using JMP.

Activity 6.9

Refer to **activity 6.8**.

(a) Manipulate JMP to create a data table that will represent the two independent samples and result in the same output as we obtained manually. (Name and save this data table as *act6p8.jmp* as we will come back to this data set when we test for means in the next study unit.)

(b) Test the hypothesis that the birth weights of the new cross-breed from laboratory B are **more homogeneous** than the birth weights from laboratory A.

(c) Comment on the similarities (and the differences) between the manual test and the electronic version of JMP.

Feedback → Feedback

(a) Your JMP table should look like the snapshot below:

27/27 Rows

	group	weights
1	a	230
2	a	233
3	a	226
4	a	232
5	a	235
6	a	233
7	a	229
8	a	231
9	a	236
10	a	230
11	a	234
12	a	235
13	a	234
14	a	232
15	a	234
16	b	231
17	b	232
18	b	229
19	b	232
20	b	231
21	b	224
22	b	232
23	b	231
24	b	225
25	b	226
26	b	235
27	b	233

Please note: You should not be tempted to type the two sets of weights as two separate columns. (This is the most common mistake that beginners make!) You need to think like a computer! There are different weights for different rats but still, there is simply the one continuous variable called "weights". To distinguish between the two sets of weights for the two laboratories, we need to add an extra column to the data table. The first column represents the two different groups and is called a **grouping variable**. I always use the variable name "*group*" to fulfil this function. Do you agree that we should declare "*group*" as a nominal variable before we continue? Select **Nominal** under **Modeling Type**. (See activity 5.8 to refresh your memory.)

(b) Aha, did you get the feeling from (a) above that this could be an application of **Analyze** ⟶ **Fit Y by X**? (Yes, this could lead to the same initial steps as for the contingency table analysis of activity 5.8 as well as the initial steps for the correlation/regression analysis! It is simply the *combination of type of variables* that is going to change.)

(b) [*For beginners:* How to perform a test for equality of variances:

Select **Analyze** ⟶ **Fit Y by X.**

Again you get the (familiar by now) **Report from JMP** where the top line reads as follows:

"**Distribution of Y for each X. Modeling types determine analysis"**

We know that JMP uses three different icons for the three main modelling types: a *blue triangle* for a continuous variable; *a green histogram-type* icon for an ordinal variable and a *red bar chart* for a nominal variable.

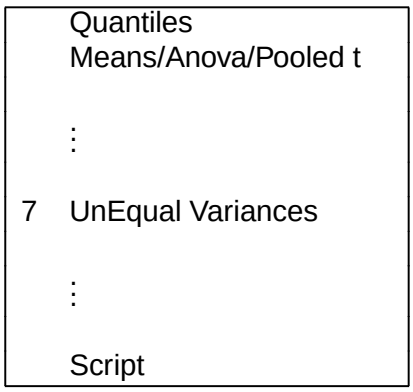
If you have selected nominal as the correct **Modeling Type** for the variable "*group"*, JMP automatically uses the *icon of a red bar chart* to identify the one nominal variable in the variable box and the *icon of a blue triangle* to identify the one continuous variable "*weights".*

(It was easier in study unit 5 to grasp that we were trying to establish "a connection between the two variables", where we had to select an "independent variable = X" and a "dependent variable = Y")

How will you choose X and Y for this situation? Suppose that *group* influences *weight* which means we assume that "independent variable = X, Factor = *group*" and that "dependent variable = Y, Response = *weight*". **Convey this choice to JMP in the relevant boxes**.

All you have to do is to look at the "picture display" of the given four possible "**Y by X**" analyses, and see that for this application it will be the display in the top right-hand corner called **Oneway**. Actually, "Oneway" refers to the appropriate analysis if we want to *test for means* and we will come back in detail to the testing of means. If you click on OK JMP will perform a **Oneway Analysis of weights By group.**

If you click on the red triangle in the **Oneway Analysis of weights By group** bar in the output, it opens a submenu with 16 choices.



Select the seventh one $\longrightarrow$ **UnEqual Variances**.

The output will have the heading **Tests that the Variances are Equal** and it will show five different test statistics:

Test	F Ratio	DFNum	DFDen	p-Value
O'Brien[.5]	0.8033	1	25	0.3786
Brown-Forsythe	0.1825	1	25	0.6729
Levene	0.8544	1	25	0.3641
Bartlett	0.6607	1	.	0.4163
F Test 2-sided	1.6004	11	14	0.4033

We see in the last line that the **F Ratio** for the **F Test 2-sided** $= 1.6004$ (we computed the value $1.600\,4$ manually in activity 6.8) and it has **DFNum=11** (11 degrees of freedom for the *numerator*) and **DFDen=14** (14 degrees of freedom for the *denominator*) which means an $F_{11;\,14}$ distribution (we also used this to obtain our manual critical value in activity 6.8). The p-value of 0.4033 is the two-sided exceedance probability and even if we halve this (for a one-sided test) $\frac{0.4033}{2} = 0.201\,65 > 0.05 \Longrightarrow$ we cannot reject H_0.

(c) JMP does not give any confidence intervals for σ^2 but only the necessary output to test for the "subhypothesis test" $H_0 : \sigma_A^2 = \sigma_B^2$ against the alternative hypothesis $H_1 : \sigma_B^2 \neq \sigma_A^2$ since the "main-hypothesis test" is actually $H_0 : \mu_A = \mu_B$ against the alternative hypothesis $H_1 : \mu_A \neq \mu_B$ which we will come back to!

6.3 Paired observations

From the given examples of dependence in study unit 5 of the study guide, it is clear that when the outcome of one observation could have an influence on the outcome of any other observation, we have a form of dependence and we need to use a different technique for our inference. This is the reason why we had to derive the complicated test (see result 6.1) developed from theorems 6.2 and 6.3 to test $H_0 : \sigma_1^2 = \sigma_2^2$ in the case of *paired observations*!

Although the emphasis in STA2601 is not on the theoretical derivations, we do like you to understand why and where we apply specific techniques and that these methods are not simply fabricated! The best applied statisticians are those who properly understand the theory. We do not want you to learn the proofs of theorems 6.2 and 6.3 parrot-fashion. In fact, we will never ask you to prove them in any examination or in an assignment!

Activity 6.10

Give your own interpretation of theorem 6.2 in the study guide.

.....

.....

.....

.....

.....

.....

Feedback → Feedback

There is no way on earth that you could have logically figured out that you had to "create" two new variables from the existing paired observations $(X_{1j}; X_{2j})$ in order to prove that $\sigma_{X1}^2 = \sigma_{X2}^2$. It is a very clever mathematical trick!

However, try to have a clear picture of this mathematical manipulation!

If we have paired observations $(X_{1j}; X_{2j})$ it means (think like a computer) that we type the observations as two separate columns having the same number of rows, where the first column shows the values for X_{1j} and the second column shows the values for X_{2j}. Now we add a third column where we type in the corresponding value $(X_{1j} - X_{2j})$ for each pair and add a fourth column where we type in the corresponding value $(X_{1j} + X_{2j})$ for each pair.

Forget for the moment about the first two columns and "zoom in" on column 3 $= Y_1$ and column 4 $= Y_2$. If we apply the work covered in study unit 5, we know that there is a test for $H_0 : \rho_{Y_1Y_2} = 0$.

Now theorem 6.2 states that if $H_0 : \rho_{Y_1Y_2} = 0$ is true it actually proves that $\sigma^2_{X1} = \sigma^2_{X2}$ is true! Sounds impossible but that is the magic of mathematics!

 In Afrikaans we could say: "Kul jou hier, kul jou daar...!]

Activity 6.11

Give your own interpretation of theorem 6.3 in the study guide.

.....

.....

.....

.....

.....

.....

Feedback  Feedback

This theorem builds on theorem 6.2 which states that if $H_0 : \rho_{Y_1Y_2} = 0$ is true for the artificially created pairs it actually proves that $\sigma^2_{X1} = \sigma^2_{X2}$ is true for the original paired values. So, it **uses the test statistic which we normally use to test** $H_0 : \rho_{Y_1Y_2} = 0$ but then "converts it back" to the pairs so that we only have to use the original first column values for X_{1j} and the second column values for X_{2j}. In other words, what we have to compute are three statistics $U_{11}; U_{12}$ and U_{22} which are expressed in terms of the original paired observations.

$$T = \frac{\sqrt{n-2}\, R}{\sqrt{1-R^2}} \longrightarrow \cdots \longrightarrow \cdots = \frac{\sqrt{n-2}\,(U_{11} - U_{22})}{2\sqrt{U_{11}U_{22} - U_{12}^2}}$$

Activity 6.12

Mineral compositions were determined for 12 rocks taken from an area in England with many excellently preserved plant fossils. The percentage of siderite ($FeCO_3$) concentrations was determined using two different analysis techniques. Do you think that technique B leads to more variable measurements than technique A? Test at the 10% level of significance.

Rock	1	2	3	4	5	6	7	8	9	10	11	12
% siderite method A	62.2	58.3	74.7	75.0	64.9	59.0	74.3	81.1	89.5	58.3	74.7	63.5
% siderite method B	61.8	57.2	76.1	74.2	63.8	59.7	73.2	80.1	89.2	57.8	74.9	62.7

Feedback  Feedback

We have to test $H_0 : \sigma_A^2 = \sigma_B^2$ against $H_0 : \sigma_A^2 < \sigma_B^2$ (technique B leads to *more variable measurements* than technique A).

To compute the test statistic T we have to compute the three statistics $U_{11}; U_{12}$ and U_{22} in terms of the original paired observations, which we do in tabular format.

Rock	% siderite method A	% siderite method B	$A-\overline{A}$	$B-\overline{B}$	$(A-\overline{A})^2$	$(B-\overline{B})^2$	$(A-\overline{A})(B-\overline{B})$
1	62.2	61.8	−7.425	−7.425	55.1306	55.1306	55.1306
2	58.3	57.2	−11.325	−12.025	128.2556	144.6006	136.1831
3	74.7	76.1	5.075	6.875	25.7556	47.2656	34.8906
4	75.0	74.2	5.375	4.975	28.8906	24.7506	26.7406
5	64.9	63.8	−4.725	−5.425	22.3256	29.4306	25.6331
6	59.0	59.7	−10.625	−9.525	112.8906	90.7256	101.2031
7	74.3	73.2	4.675	3.975	21.8556	15.8006	18.5831
8	81.1	80.1	11.475	10.875	131.6756	118.2656	124.7906
9	89.5	89.2	19.875	19.975	395.0156	399.0006	397.0031
10	58.3	57.8	−11.325	−11.425	128.2556	130.5306	129.3881
11	74.7	74.9	5.075	5.675	25.7556	32.2056	28.8006
12	63.5	62.7	−6.125	−6.525	37.5156	42.5756	39.9656
Total	835.5	830.7	0	0	1113.3222	1130.2822	1118.3122

We will reject H_0 at the 10% level of significance, if $T < -t_{0.10;\,10} = -1.372$.

$$T = \frac{\sqrt{n-2}\,(U_{11}-U_{22})}{2\sqrt{U_{11}U_{22}-U_{12}^2}} = \frac{\sqrt{12-2}\,(1113.3222-1130.2822)}{2\sqrt{(1113.3222)(1130.2822)-1118.3122^2}} = \frac{-53.6322}{176.0237} - 0.304\,69$$

Since $-0.304\,69 \not< -1.372$ we cannot reject $H_0 : \sigma_A^2 = \sigma_B^2$ and conclude that technique B does not lead to *more variable measurements* than technique A.

6.4 More than two independent samples

What we said in the introduction to 6.2 holds true for this section as well except that we now work with *more than two* independent samples. You will find exactly the same heading of "More than two independent samples" in the following study unit when we deal with the topic of *inference on means*. Again, there is a definite connection between this subsection and its counterpart in the following study unit, because the hypotesis test we derive in this section is the *key to checking the theoretical assumption* $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \cdots = \sigma_k^2$ we will need in study unit 7 when we discuss the difference between more than two means. So, already make a mental note of the connection between the two sections. Once again, as was the case in section 6.2, JMP does not treat the "testing of more than two variances" as a main topic but only as part of the subtests when the main issue is the testing of more than two means, called ANOVA. In other words, you have to "manipulate" JMP to obtain the corresponding variance test.

Activity 6.13

The following table gives the rate of flow of blood through the kidneys for three distinct age groups:

	Rate of flow for 20-year-old adults (X_{1i})	Rate of flow for 40-year-old adults (X_{2i})	Rate of flow for 60-year-old adults (X_{3i})
$i = 1$	542	440	339
$i = 2$	544	443	340
$i = 3$	545	444	342
$i = 4$	548	446	343
$i = 5$	549	448	344
$i = 6$	550	450	350
$i = 7$	550	451	352
$i = 8$	551	454	355
$i = 9$	554	455	356
$i = 10$	557	459	359

Assume that the data above represent three independent random samples of size $n = 10$ from $n\left(\mu_j;\ \sigma_j^2\right)$ distributions $(j = 1;\ 2$ and $3)$.

Test $H_0: \ \sigma_1^2 = \sigma_2^2 = \sigma_3^2$ against
$H_1: \ \sigma_p^2 \neq \sigma_q^2$ (for at least one $p \neq q$)

at the 5% level of significance.

Feedback → Feedback

The test statistic is $U = \dfrac{\max\limits_j S_j^2}{\min\limits_j S_j^2}$.

The necessary computations are done in the table below:

	X_{1i} (20)	X_{2i} (40)	X_{3i} (60)	$(X_{1i}-549)^2$	$(X_{2i}-449)^2$	$(X_{3i}-348)^2$
$i=1$	542	440	339	49	81	81
$i=2$	544	443	340	25	36	64
$i=3$	545	444	342	16	25	36
$i=4$	548	446	343	1	9	25
$i=5$	549	448	344	0	1	16
$i=6$	550	450	350	1	1	4
$i=7$	550	451	352	1	4	16
$i=8$	551	454	355	4	25	49
$i=9$	554	455	356	25	36	64
$i=10$	557	459	359	64	100	121
Total	5490	4490	3480	186	318	476

From the table we may compute $S_j^2 = \dfrac{\sum \left(X_{ji} - \overline{X}_j\right)^2}{(n-1)}$ for $j = 1, 2,$ and 3.

$$S_1^2 = \frac{186}{9} = 20.6667;$$

$$S_2^2 = \frac{318}{9} = 35.3333;$$

$$S_3^2 = \frac{476}{9} = 52.8889$$

$$U = \frac{\max\limits_j S_j^2}{\min\limits_j S_j^2} = \frac{52.8889}{20.6667} = 2.5591$$

To use table E we assume that $k = 3$ and $v = 9 \Longrightarrow$ the upper 5% critical value for this problem is 5.34.

Since $2.559 < 5.34$ H_0 cannot be rejected and we may assume that $\sigma_1^2 = \sigma_2^2 = \sigma_3^2$.

Activity 6.14

Refer to **activity 6.13**.

(a) Manipulate JMP to create a data table that will represent the three independent samples and result in the same output as we obtained manually. (Name and save this data table as *act6p13.jmp* as we will come back to this data set when we perform an ANOVA test for means in the next study unit.)

(b) Test the hypothesis that $\sigma_1^2 = \sigma_2^2 = \sigma_3^2$ against the alternative $\sigma_p^2 \neq \sigma_q^2$ (for at least one $p \neq q$).

(c) Comment on the similarities (and the differences) between the manual test and the electronic version of JMP.

Feedback → Feedback

(a) Your JMP table should look like the snapshot below:

	group	rateflow
1	a	542
2	a	544
3	a	545
4	a	548
5	a	549
6	a	550
7	a	550
8	a	551
9	a	554
10	a	557
11	b	440
12	b	443
13	b	444
14	b	446
15	b	448
16	b	450
17	b	451
18	b	454
19	b	455
20	b	459
21	c	339
22	c	340
23	c	342
24	c	343
25	c	344
26	c	350
27	c	352
28	c	355
29	c	356
30	c	359

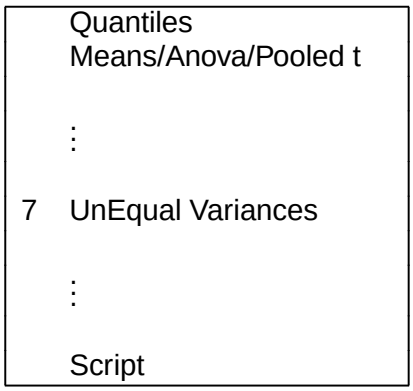
Please note: Similar to *activity 6.9* we need to add an extra column to the data table. Again I used the variable name "*group*" for the first column to represent the three different groups. Make sure to declare "*group*" as a nominal variable before you continue. (Select **Nominal** under **Modeling Type**.) All the rate flow values should be typed in the second column below one another, implying we have one continuous variable called "*rateflow".*

(b) Click on **Analyze** ⟶ **Fit Y by X.**

Complete the "**Distribution of Y for each X. Modeling types determine analysis"** dialog box by choosing "independent variable = X, Factor = *group*" and "dependent variable =Y, Response = *rateflow*".

If you click on OK JMP will perform a **Oneway Analysis of rateflow By group**

If you click on the red triangle in the **Oneway Analysis of rateflow By group** bar in the output, it opens a submenu with 16 choices.



Select the seventh one ⟶ **UnEqual Variances.**

The output will have the heading **Tests that the Variances are Equal** and it will show *four* different test statistics:

Test	F Ratio	DFNum	DFDen	Prob > F
O'Brien[.5]	1.8999	2	27	0.1691
Brown-Forsythe	2.5682	2	27	0.0953
Levene	2.7112	2	27	0.0845
Bartlett	0.9160	2	.	0.4001

If you compare this with the JMP output of activity 6.9 you will notice that now the "fifth line" (**F Ratio** for the **F Test 2-sided)** is *not included* because it is not a valid test anymore!

(c) Although JMP does not give the U-test of our study guide, it does provide appropriate alternative tests to test for the "sub-hypothesis test" $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2$ against the alternative hypothesis $H_1 : \sigma_p^2 \neq \sigma_q^2$ (for at least one pair $p \neq q$). Keep in mind that the "main-hypothesis test" is actually $H_0 : \mu_1 = \mu_2 = \mu_3$ against the alternative hypothesis $H_1 : \mu_p \neq \mu_q$ (for at least one pair $p \neq q$) which we will come back to!

All you need to be able to do is to make the correct decision: since none of the p-values are less than α we cannot reject the null hypothesis and may conclude that $\sigma_1^2 = \sigma_2^2 = \sigma_3^2$.

Go back to the learning outcomes of this study unit in the study guide. Can you tick them off with confidence?

STUDY UNIT 7

Inference on means

In study unit 6 we treated the inference on variances as methods standing on their own and not necessarily as the prerequisite for other tests. Now we focus on various tests for means and we refer back to the previous study unit because an important assumption will be the one of "equal variances".

7.1 One-sample problem

In many real-life situations *the mean* of a variable is our main focus. We most often don't know what the values of the parameters of a distribution are, and we could estimate the mean of an unknown population by computing a *confidence interval for the mean*. We could also be interested in *testing a hypothesis* about a mean.

You will notice that these two topics also appear in the textbook in **Chapter 7: Univariate Distributions: One Variable, One Sample,** in a subsection called *Statistical Inference on the Mean***.**

Activity 7.1

Give your own interpretation of the **application** of theorem 7.1 of the study guide.

……

……

……

……

……

……

……

Feedback → Feedback

Theorem 7.1 has two distinct **applications**:
Either we will *know* what the value of the population variance, σ^2, is (and thus use the (a) part where $\sqrt{n}(\overline{X}-\mu)$ has a normal distribution) or we will *not know* what the value of the population variance, σ^2, is (and thus use the (d) part where $\sqrt{n}(\overline{X}-\mu)$ has a t-distribution).

Activity 7.2

Refer to the data of **activity 6.6** where we considered the following random sample of size $n = 20$ drawn from a normal population with unknown mean and unknown variance.

14	15	22	17	16	17	23	25	19	18
22	17	18	21	15	22	20	21	18	20

(a) Test the hypothesis that the mean is at most 20. Use $\alpha = 0.05$.

(b) Derive a 95% one-sided confidence interval in accordance with the hypothesis test.

(c) Compute a 95% two-sided confidence interval.

Feedback → Feedback

(a) From activity 6.6 we already know that $\sum_{i=1}^{n} X_i = 380 \Longrightarrow \overline{X} = \frac{380}{20} = 19$ and

$$\frac{\sum_{i=1}^{20}(X_i - \overline{X})^2}{n-1} = \frac{170}{19} = 8.947\,4.$$

To test $H_0 : \mu = 20$ against $H_0 : \mu < 20$ we use the test statistic

$T = \frac{(\overline{X} - \mu_0)}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$ and we reject H_0 left-sided if $T \leq -t_{\alpha;n-1}$.

$T = \frac{(19-20)}{\frac{\sqrt{8.9474}}{\sqrt{20}}} = -1.495\,1 \sim t_{19}$

We reject H_0 left-sided at the 5% level of significance if $T \leq -t_{0.05;19} = -1.729$.

Since $T > -1.729$ we cannot reject H_0.

(b) A 95% one-sided confidence interval in accordance with the alternative hypothesis is the interval

$$\frac{(\overline{X} - \mu)}{\frac{S}{\sqrt{n}}} > -t_{0.05;19} \Longrightarrow \mu < \overline{X} + \left(\frac{S}{\sqrt{n}}\right)(t_{0.05;19}) \Longrightarrow \mu < 19 + \left(\frac{2.991\,2}{\sqrt{20}}\right)(1.729) = 20.1564.$$

(In other words the interval $(-\infty; 20.1564)$.)

Since this upper boundary includes the value $\mu = 20$ we cannot reject H_0. (Of course this must never contradict the conclusion of question (a).)

(c) A 95% two-sided confidence interval is computed as $\overline{X} - \left(\frac{S}{\sqrt{n}}\right)(t_{0.025;19}) < \mu < \overline{X} + \left(\frac{S}{\sqrt{n}}\right)(t_{0.025;19})$

$$\Longrightarrow 19 - \left(\frac{2.991\,2}{\sqrt{20}}\right)(2.093) < \mu < 19 + \left(\frac{2.991\,2}{\sqrt{20}}\right)(2.093) \Longrightarrow 17.6001 < \mu < 20.3999.$$

In other words a 95% two-sided confidence interval for the unknown mean is $(17.6001;\ 20.3999)$.

WORK THROUGH

Sall, Creighton and Lehman, **Chapter 7 Univariate Distributions: One Variable, One Sample**

Start reading on page 135 "*Statistical Inference on the Mean*" and read up to "*Case Study: The Earth's Ecliptic*".

Activity 7.3

Follow all the instructions and steps for this case study "*The Earth's Ecliptic"* and try to reproduce the output of **figure 7.13** on page 141 of the textbook.

.....

Feedback → Feedback


The next activity will really thrill the inquisitive student!

Activity 7.4

Refer to the data of activity 7.2 above and **use JMP** to perform the one-sided hypothesis test for a single sample **mean** and to create a confidence interval.

Feedback → Feedback

(a) Performing a hypothesis test for the mean of a single sample:

For beginners: Your data table must have 20 rows with one column containing the sample values. (I called the variable X.)

Select **Analyze** $\longrightarrow$ **Distribution.**

Move the variable X to the **Y, Columns** box and click on **OK**.

This gives the standard default output **Histogram + Boxplot**

Quantiles

Moments

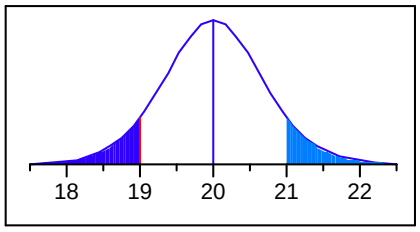
Click on the histogram bar (there is a red triangle before the variable name *X* where you have to click. This opens a *submenu* with 16 choices. Select **Test Mean** which is the eighth of the possible choices.

This opens the **JMP: Test Mean** dialog box which has three possible fields to complete. You have to complete the first field which asks you to **Specify Hypothesized Mean.** Type 20. Ignore the other two.

(When you know what the value of σ^2 is, you are able to complete the second field which asks you to **Enter True Standard Deviation to do z-test rather than t-test,** in other words when we want to apply theorem 7.1 (a) of our study guide. The third field gives you the opportunity to select the choice of performing a nonparametric test we dealt with in our first-year courses.)

Click on **OK** to get the following output:

Test Mean=value	
Hypothesized Value	20
Actual Estimate	19
df	19
Std Dev	2.99122
	t Test
Test Statistic	-1.4951
Prob > |t|	0.1513
Prob > t	0.9243
Prob < t	0.0757



Of course this must check with our manual computations! The only difference is that a computer never gives a critical value but always all the possible p-values from which *you have to select the appropriate one*.

If the instruction of the exercise was to perform a **left-sided test**, you only have to report that $T = -1.4951$ with $P(T < -1.4951) = 0.0757 > \alpha \Longrightarrow H_0 : \mu = 20$ cannot be rejected.

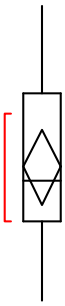
(b) Obtaining confidence intervals:

The selection of **Analyze** $\longrightarrow$ **Distribution** $\longrightarrow$ **OK** automatically gives the standard default output **Moments** which includes a two-sided confidence interval.

Moments	
Mean	19
Std Dev	2.9912152
Std Err Mean	0.6688561
upper 95% Mean	20.399932
lower 95% Mean	17.600068
N	20

From this we find the 95% confidence interval $(17.600068;\ 20.399932)$.

Please note that a visual representation of the confidence interval is the **diamond shaped** diagram within the box of the **Quantile Box Plot.** The horizontal line is the *median* and for a perfectly symmetrical distribution it will cut the diamond in two!



To obtain a 95% **one-sided confidence interval** you will have to click on the histogram bar (there is a red triangle before the variable name *X* where you have to click) and open the *submenu* where **Confidence Interval** is the tenth of the possible choices which you have to select. This opens the **JMP: Confidence Intervals** dialog box which gives you the choice of any value for $(1-\alpha)$ and which provides three possible choices for the interval. Click the **One-sided upper limit.**

One-sided Confidence Interval

Parameter	Estimate	Lower CI	Upper CI	1-Alpha
Mean	19	.	20.15654	0.950
Std Dev	2.991215	.	4.099192	

From the output we may conclude that a 95% one-sided confidence interval (in accordance with the left-sided alternative hypothesis) is the interval $(-\infty; 20.15654)$.

Do you think that any intelligent person who manages the JMP software will be able to make such a statement? I doubt it! You have to study statistics to know how to really interpret the output!

DO NOT delete your output! You will need it for the next activity.

Activity 7.5

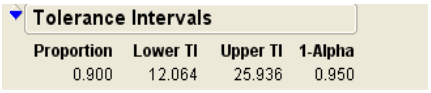
(a) Refer to the data of activity 7.2 as well as activity 7.4 above and **use JMP** to create a two-sided **Tolerance Interval.**

[The **Tolerance Interval** option from a **Distribution report** displays a dialog where you have to select the confidence $(1-a)=0.95$ and the proportion $(p)=0.90$. Also select the appropriate option for a two-sided tolerance interval from the dialog box.]

(b) How does this compare with the two-sided confidence interval we computed in activities 7.2 and 7.4?

Feedback ⟶ Feedback

To obtain a 95% **Tolerance Interval** you will have to click on the histogram bar of the output for activity 7.4 and open the *submenu* where **Tolerance Interval** is the 12th of the possible choices which you have to select. This opens the **JMP: Tolerance Intervals** dialog box which gives you the choice of any value for $(1-\alpha)$ (select 0.95) and the proportion (p) (select 0.90). Also select the appropriate option for a two-sided tolerance interval from the dialog box.


Tolerance Intervals

Proportion	Lower TI	Upper TI	1-Alpha
0.900	12.064	25.936	0.950

From this we find the 95% **tolerance interval** to be $(12.064;\ 25.936)$ which means we are 95% confident that at least 90% of the population data (from which the sample was drawn) lie between 12.064 and 25.93 as compared to the 95% confidence interval of $(17.600068;\ 20.399932)$.

Does the difference make sense to you? The 95% confidence interval for the mean, that is $(17.600068;\ 20.399932)$ implies we are 95% confident that the unknown mean of the population from which the sample was drawn lies between 17.600068 and 20.399932. The two intervals are totally different in concept.

7.2 The power of the test and the noncentral t-distribution

Activity 7.6

Suppose that a random sample of size $n = 16$ is drawn from a normal population of which we know the variance is $\sigma^2 = 36$. We test at the 5% level of significance $H_0 : \mu = 12$ against the alternative $H_1 : \mu < 12$.

What is the power of the test if the true unknown population mean is in fact $\mu = 9$?

Feedback → Feedback

We know that $Z_0 = \dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim n(0;1)$

$\Longrightarrow$ if $H_0 : \mu = 12$ is true, that $\overline{X} \sim n\left(12; \dfrac{36}{16}\right)$.

We reject H_0 at the 5% level of significance if $Z_0 \leq -1.645$

$$\begin{aligned} \Longrightarrow \frac{\overline{X} - 12}{6/\sqrt{16}} &\leq -1.645 \\ \Longrightarrow \overline{X} &\leq 12 - 1.645(6/\sqrt{16}) \\ \Longrightarrow \overline{X} &\leq 9.532\,5 \end{aligned}$$

(which implies we do not reject H_0 if $\overline{X} > 9.532\,5$).

Now

$$\begin{aligned} \beta = P(\text{type II error}) &= P(H_0 \text{ is not rejected} \mid H_1 \text{ is true}) \\ &= P(\overline{X} > 9.532\,5 \mid \overline{X} \sim n(9; \tfrac{36}{16})) \\ &= P\left(\frac{\overline{X} - 9}{6/\sqrt{16}} > \frac{9.532\,5 - 9}{6/\sqrt{16}}\right) \quad \text{(standardise using } \mu = 9 \text{ and not } \mu = 12\text{)} \\ &= P(Z > 0.355\,) \\ &= 0.3613 \qquad \left(\frac{0.3632 + 0.3594}{2}\right) \end{aligned}$$

Hence the power of the test is $1 - \beta = 1 - 0.3613 = 0.638\,7$.

After doing the previous activity manually, you will appreciate even more what a computer can do in the blink of an eye in the next activity!

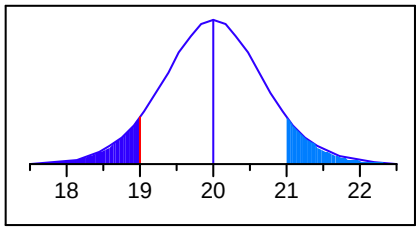
Activity 7.7

This is a JMP **power animation** exercise. (See *Sall, Creighton and Lehman,* **chapter 7 Univariate Distributions: One Variable, One Sample** page 148 for this **animation** exercise.) You can either do it for the "*Case Study: The Earth's Ecliptic*" to reproduce figure 7.17 of the textbook, or you can do it for the data set of activity 7.4 to reproduce the figure below in the feedback. You need to activate (for the data set of your choice) the **JMP: Test Mean** dialog where you have to **Specify a Hypothesized Mean.**

Test Mean=value

Hypothesized Value	20
Actual Estimate	19
df	19
Std Dev	2.99122

	t Test
Test Statistic	-1.4951
Prob > |t|	0.1513
Prob > t	0.9243
Prob < t	0.0757



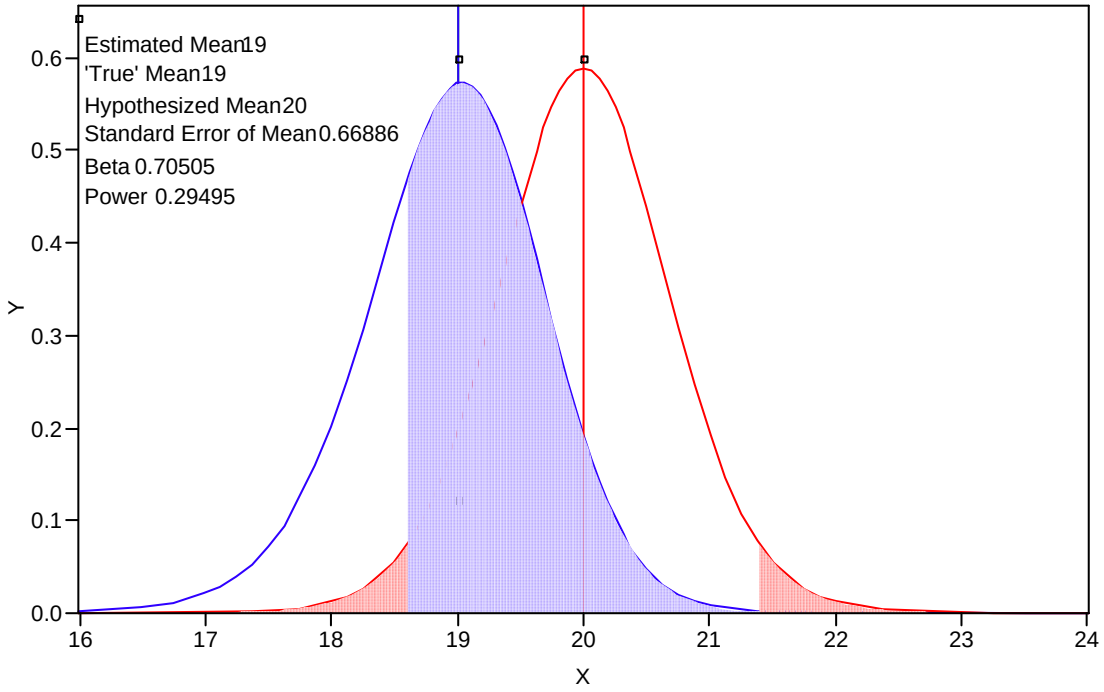
If you now click on the red triangle before the **Test Mean=value** in the output, you get a small submenu of

PValue animation
Power animation
Remove Test

Select the **Power animation.**

Feedback → Feedback

Is this not amazing!



This is of course the default output for two-sided testing where $n = 20$ and $\alpha = 0.05$. Keep in mind that animation exercises are interactive where you are supposed to change the parameters to see what happens. So, **play around with the interactive buttons!** Change Two sided to Low Side and to High Side. Also change the sample size to $n = 15$ and see what happens. Change the sample size back to $n = 20$ but now change Alpha to $\alpha = 0.01$ and see what happens.

I hope that this enhanced your understanding of the concept *power of a test*. If you feel unsure about the p-value for a hypothesis test you can try to do the PValue animation as well.

Activity 7.8

$H_0 : \mu = 20$ is tested against the alternative $H_1 : \mu \neq 20$ using a sample of size $n = 25$. Find the power of the test if we know that the population mean is $20 - \sqrt{0.64\sigma^2}$ at the level

(a) $\alpha = 0.05$

(b) $\alpha = 0.01$.

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

Feedback Feedback

H_0 is tested against H_1 with the test statistic $T = \dfrac{\sqrt{n}(\overline{X} - \mu_0)}{S} \sim t_{n-1}$.

To find the power we need to compute $\phi = \dfrac{1}{\sqrt{2}}\delta$ where $\upsilon = n - 1 = 24$ and

$$\begin{aligned} \delta &= \frac{\sqrt{n}(\mu - \mu_0)}{\sigma} \\ &= \frac{\sqrt{25}[(20 - \sqrt{0.64\sigma^2}) - 20]}{\sigma} \\ &= \frac{5(-0.8\sigma)}{\sigma} \\ &= -4.0. \end{aligned}$$

Hence $\phi = \dfrac{1}{\sqrt{2}}\delta = \dfrac{-4}{\sqrt{2}} = -2.828\,4.$

From table F we read off the power

(a) at the level $\alpha = 0.05$ as 0.96 (take $\phi = 2.8$ and use $v = 20$)

(b) at the level $\alpha = 0.01$ as 0.86 (take $\phi = 2.8$ and use $v = 20$).

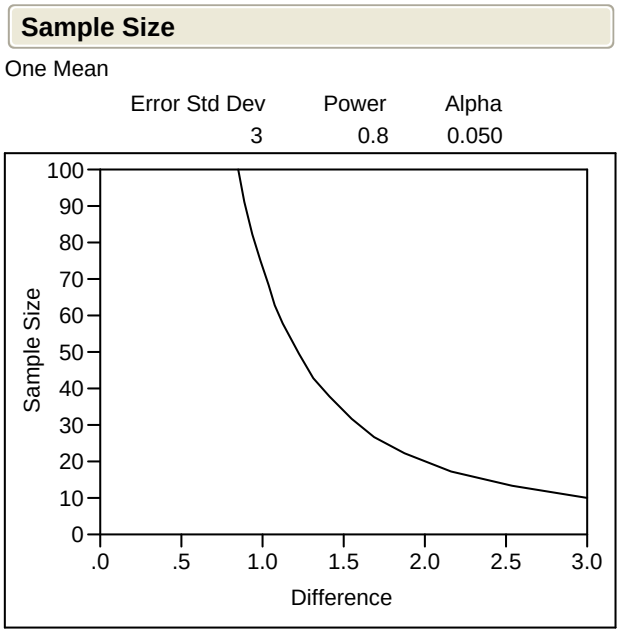
Activity 7.9



This is an optional activity because computing the desired sample size for a given power falls beyond the scope of this module although we do emphasise the relation between the power and the sample size (see the second note towards the end of the section on power in the study guide).

JMP computes a graph showing the desired sample size for a given power. You will find this under DOE (Design of Experiments).

Feedback  Feedback



7.3 Two-sample problem: independent samples

We had the same setup of "two independent samples" in the previous study unit when we dealt with the topic of *inference on variances*. There is a definite connection between this subsection and its counterpart in the previous study unit, because the hypothesis test we derive in this section is based on *the theoretical assumption of equal variances*. If this assumption is not met, we have an alternative route to follow (see section 7.5). So, performing the test manually, we strictly need to always first test for the equality of the variances. You will see that JMP also treats the "testing of variances" as part of the subtests when the main issue is the testing of means.

Activity 7.10

Indicate whether the following statements are correct or incorrect and try to rectify the incorrect statements to make them true.

(a) $S^2_{pooled} = \dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} = \dfrac{\Sigma(X_{1i}-\overline{X}_1)^2 + \Sigma(X_{2i}-\overline{X}_2)^2}{n_1+n_2-2}$

.....

.....

(b) If we derive a confidence interval for $(\mu_1-\mu_2)$ we use $SE = \sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}$ but if we test $H_0 : \mu_1 = \mu_2$ we use $SE = \sqrt{S^2_{pooled}\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}$.

.....

.....

(c) In a one-tailed test for the difference between two population means $(\mu_1-\mu_2)$, if the null hypothesis is rejected when the alternative hypothesis, $H_1 : \mu_1 < \mu_2$ is false, a type I error is committed.

.....

.....

Feedback → Feedback

(a) Correct. With a little algebraic manipulation it follows from the definitions of $S_1^2 = \frac{\Sigma(X_{1i} - \overline{X}_1)^2}{n_1 - 1}$ and $S_2^2 = \frac{\Sigma(X_{2i} - \overline{X}_2)^2}{n_2 - 1}$ that $(n_1 - 1)S_1^2 = \Sigma(X_{1i} - \overline{X}_1)^2$ and that $(n_2 - 1)S_2^2 = \Sigma(X_{2i} - \overline{X}_2)^2$.

(b) **Incorrect**. We use $SE = \sqrt{s_{pooled}^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$ for both the hypothesis test and the confidence interval!

(c) Correct.

You will find that in most of the exercises on this section, whether they are for an assignment or the examination**,** the information you have to work with will either be:

- raw data for two samples, or
- summarised data given in a table format as

	Population	
	1	2
Sample size	n_1	n_2
Sample mean	$\overline{x}_1$	$\overline{x}_2$
Sample variance	S_1^2	S_2^2

There could be "variations" on the theme of summarised data where **computed sums** are given instead of sample statistics, for example Σx_{1i} instead of $\overline{x}_1$ or Σx_{1i}^2 and Σx_{1i} instead of S_1^2.

In the case of raw data, you must try to have at least a *Scientific Pocket Calculator* with *Statistical Functions* that will enable you to compute the **sample statistics**.

Activity 7.11

Refer to **exercise 7.7 of the study guide**. Were you able to derive the t-statistic? If yes, check your solution against the feedback below; if no, make sure you understand the feedback below.

Feedback → Feedback

$H_0 : \mu_1 = 2\mu_2 \Longrightarrow \mu_1 - 2\mu_2 = 0$ is tested against the alternative $H_1 : \mu_1 > 2\mu_2 \Longrightarrow \mu_1 - 2\mu_2 > 0.$

We base the derivation on the following three assumptions:

1. We have two **independent** populations from which we drew small random samples.
2. Both populations have **normal** distributions.
3. Both populations have **the same variance**, that is $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

From this we may say that $\overline{X}_1 \sim n\left(\mu_1; \frac{\sigma^2}{n_1}\right)$

and $\overline{X}_2 \sim n\left(\mu_2; \frac{\sigma^2}{n_2}\right) \Longrightarrow 2\,\overline{X}_2 \sim n\left(2\mu_2; \frac{4\sigma^2}{n_2}\right)$. [Yes, in general, if

$Y = 2X \Longrightarrow var(Y) = 2^2 var(X) = 4var(X).]$

From the assumption of independence, it now follows that

$\overline{X}_1 - 2\,\overline{X}_2 \sim n\left([\mu_1 - 2\mu_2]; \left[\frac{\sigma^2}{n_1} + \frac{4\sigma^2}{n_2}\right]\right)$. [Yes, in general, if $Z = X - Y \Longrightarrow var(Z) = var(X) + var(Y)$ and not a minus sign!]

We may write this as $\overline{X}_1 - 2\,\overline{X}_2 \sim n\left(\mu_1 - 2\mu_2;\ \sigma^2\left(\frac{1}{n_1} + \frac{4}{n_2}\right)\right)$.

If we standardise in the normal manner we may write

$$Z = \frac{(\overline{X}_1 - 2\,\overline{X}_2) - (\mu_1 - 2\mu_2)}{\sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{4}{n_2}\right)}} = \frac{(\overline{X}_1 - 2\,\overline{X}_2) - (\mu_1 - 2\mu_2)}{\sigma\sqrt{\left(\frac{1}{n_1} + \frac{4}{n_2}\right)}} \sim n(0;1).$$

From theorem 6.1 (b) we have that $\sum_{j=1}^{n_1} \frac{\left(X_{1j} - \overline{X}_1\right)^2}{\sigma^2}$ has a $\chi^2_{n_1-1}$ distribution and that $\sum_{j=1}^{n_2} \frac{\left(X_{2j} - \overline{X}_2\right)^2}{\sigma^2}$ has a $\chi^2_{n_2-1}$ distribution.

Because S_1^2 and S_2^2 are independent $\sum_{j=1}^{n_1} \frac{\left(X_{1j} - \overline{X}_1\right)^2}{\sigma^2} + \sum_{j=1}^{n_2} \frac{\left(X_{2j} - \overline{X}_2\right)^2}{\sigma^2}$ has a $\chi^2_{(n_1-1)+(n_2-1)}$ distribution.

We may write this as $\frac{1}{\sigma^2}[\sum_{j=1}^{n_1}\left(X_{1j} - \overline{X}_1\right)^2 + \sum_{j=1}^{n_2}\left(X_{2j} - \overline{X}_2\right)^2] \sim \chi^2_{n_1+n_2-2}.$

A t-distribution with υ degrees of freedom is defined as $T = \frac{n(0;1)}{\sqrt{\chi^2_\upsilon/\upsilon}}$.

If we apply this to the variables defined above we get

$$T = \frac{\frac{(\overline{X}_1 - 2\,\overline{X}_2) - (\mu_1 - 2\mu_2)}{\sigma\sqrt{\left(\frac{1}{n_1} + \frac{4}{n_2}\right)}}}{\sqrt{\frac{1}{\sigma^2}[\sum_{j=1}^{n_1}\left(X_{1j} - \overline{X}_1\right)^2 + \sum_{j=1}^{n_2}\left(X_{2j} - \overline{X}_2\right)^2]/(n_2 + n_2 - 2)}}$$

which simplifies to $T = \frac{(\overline{X}_1 - 2\,\overline{X}_2) - (\mu_1 - 2\mu_2)}{\sqrt{\frac{[\sum_{j=1}^{n_1}\left(X_{1j} - \overline{X}_1\right)^2 + \sum_{j=1}^{n_2}\left(X_{2j} - \overline{X}_2\right)^2]}{(n_2 + n_2 - 2)}}\sqrt{(\left(\frac{1}{n_1} + \frac{4}{n_2}\right)}} \sim t_{n_2+n_2-2}.$



The activities above assessed whether you understand the theory behind the testing for the difference between means and now you are ready to apply it to a real-life situation!

Activity 7.12

(Refer to activity 6.9 of the previous study unit.) The birth weights of 15 rats bred in laboratory A were:

230; 233; 226; 232; 235; 233; 229; 231; 236; 230; 234; 235; 234; 232; 234 grams. (Call this a sample from population A.)

The birth weights of 12 rats bred from a different cross-breed in laboratory B were:

231; 232; 229; 232; 231; 224; 232; 231; 225; 226; 235; 233 grams. (Call this a sample from population B.)

(a) Test $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$ at the 5% level of significance.

(b) On what assumptions do you base your test statistic?

(c) Derive a 95% two-sided confidence interval for $(\mu_1 - \mu_2)$.

Feedback → Feedback

Let us **start with the (b) part**: we already know from activity 6.9 of the previous study unit, that we could not reject $H_0 : \sigma_B^2 = \sigma_A^2$.

So, based on the following three assumptions:

1. we have two **independent** populations (from which we drew small random samples)
2. both populations have **normal** distributions
3. both populations have **the same variance**, that is $\sigma_1^2 = \sigma_2^2 = \sigma^2$

we may use the **test statistic** $T = \dfrac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_{pooled}^2}\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$ which has a t-distribution with $\upsilon = (n_1 + n_2 - 2)$ degrees of freedom.

$S_{pooled}^2 = \dfrac{\Sigma(X_{1i} - \overline{X}_1)^2 + \Sigma(X_{2i} - \overline{X}_2)^2}{n_1 + n_2 - 2}$ is the pooled estimate of the **unknown common variance** assuming that $\sigma_1^2 = \sigma_2^2$.

(a) From activity 6.9 of the previous study unit we also know that

$$\overline{X}_A = \frac{3\,484}{15} = 232.2667 \text{ and } S_A^2 = \frac{\sum_{i=1}^{15} (X_A - 232.2667)^2}{(15-1)} = \frac{100.9433}{14} = 7.2095;$$

$$\overline{X}_B = \frac{2\,761}{12} = 230.0833 \text{ and } S_B^2 = \frac{\sum_{i=1}^{12} (X_B - 230.0833)^2}{(12-1)} = \frac{126.9167}{11} = 11.5378$$

$$\Longrightarrow S_{pooled}^2 = \frac{100.9433 + 126.9167}{15 + 12 - 2} = \frac{227.86}{25} = 9.114\,4$$

$$\Longrightarrow T = \frac{(232.2667 - 230.0833) - (0)}{\sqrt{9.114\,4}\sqrt{\dfrac{1}{15} + \dfrac{1}{12}}} = \frac{2.1834}{(3.019)(0.387\,3)} = 1.867\,3.$$

The critical region to test H_0 against H_1 is where $|T| \geq t_{0.025;\,15+12-2} = 2.060.$

Since $1.867\,3 < 2.060$ H_0 cannot be rejected.

(c) The 95% two–sided confidence interval for $(\mu_1 - \mu_2)$ can be derived as:

$(\overline{X}_1 - \overline{X}_2) \pm (t_{0.025;\,15+12-2})\sqrt{s^2_{pooled}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 2.1834 \pm (2.060)(3.019)(0.387\,30) = 2.1834 \pm 2.408\,7$

$(-0.225\,3;\ 4.592\,1).$

Since this interval includes the **value zero** $\Longrightarrow (\mu_1 - \mu_2) = 0 \Longrightarrow \mu_1 = \mu_2$ (which confirms that H_0 cannot be rejected).

What about JMP?

How will you create a confidence interval or perform a hypothesis test for the difference between two independent sample means using JMP? Before we redo the activity above (where we have *two independent samples*) by making use of JMP, you first need to work through the corresponding section of the textbook to understand all the possible output.

WORK THROUGH

Sall, Creighton and Lehman, **chapter 8 The Difference Between Two Means**

Start reading on page 168 "*Two Independent Groups*" and read up to "*Equal or Unequal Variances?*".

Activity 7.13

Work through the example explained in this section of the textbook; follow all the instructions and steps for the *Htwt12.jmp* data set and try to finally reproduce the output of **figure 8.4** on page 173.

Feedback → Feedback

Activity 7.14

Refer to activity 7.12 and also to **activities 6.8 and 6.9** of the previous study unit.

(a) Open the data table *act6p8.jmp* which you created and saved in activity 6.9.

(b) Test $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$ at the 5% level of significance.

(c) Derive a 95% two-sided confidence interval for $(\mu_1 - \mu_2)$.

(d) Comment on the similarities (and the differences) between the manual test and the electronic version of JMP.

Feedback → Feedback

(a) From activity 6.9 you know that the data table should look like the snapshot below:

27/27 Rows	group	weights
1	a	230
2	a	233
3	a	226
4	a	232
5	a	235
6	a	233
7	a	229
8	a	231
9	a	236
10	a	230
11	a	234
12	a	235
13	a	234
14	a	232
15	a	234
16	b	231
17	b	232
18	b	229
19	b	232
20	b	231
21	b	224
22	b	232
23	b	231
24	b	225
25	b	226
26	b	235
27	b	233

(b) [*For beginners:* How to perform a test for equality of means:

Select **Analyze** ⟶ **Fit Y by X.**

It makes sense that *group* influences *weight* which means we assume that "Independent variable = X, Factor = *group*" and that "Dependent variable = Y, Response = *weights*". **Convey this choice to JMP in the relevant boxes**.

If you click on OK JMP will perform a **Oneway Analysis of weights By group**.

Do you notice that the initial output contains only a very basic graph showing the individual observations of two separate groups?

Before we now request a test for means, we should strictly speaking first verify whether the populations variances are equal. (In other words if this was the first time you worked with this

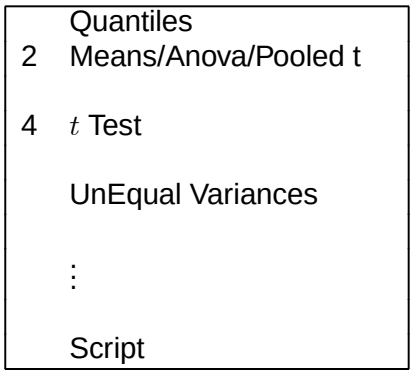
data set, you should have selected to do **Tests that the Variances are Equal** (we will not repeat this because we emphasized it in the activities of the previous study unit.)

[To refresh your memory: Click on the red triangle in the

Oneway Analysis of weights By group bar in the output. Select the seventh one $\longrightarrow$ **UnEqual Variances,** which will produce five different test statistics to **Test that the Variances are Equal.** The p-values of all the different tests to test for the "subhypothesis test" $H_0 : \sigma_A^2 = \sigma_B^2$ against the alternative hypothesis $H_1 : \sigma_B^2 \neq \sigma_A^2 > 0.05 \Longrightarrow$ we cannot reject H_0.]

Now we can focus our attention on the **test for the difference between the means** because we know that we may use the test statistic where we pool the variances (this knowledge influences the next step). JMP calls this the **Pooled t** statistic.

If you click on the red triangle in the **Oneway Analysis of weights By group** bar in the output, it again opens the submenu with the 16 choices.

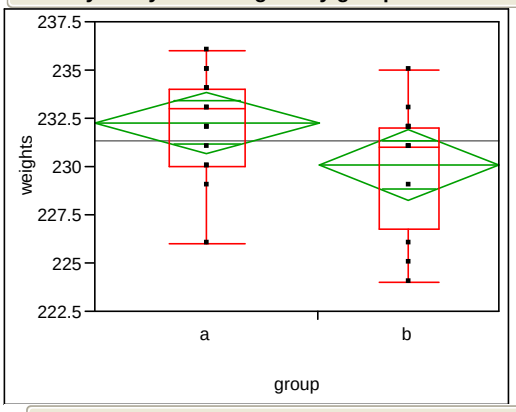


Select the **second** one $\longrightarrow$ **Means/Anova/Pooled t.**

Notice what happens to the output:

Apart from the reports on all the different computations, the initial graph is enriched with much additional information. We will come back to the discussion thereof.

Oneway Analysis of weights By group



Quantiles

Level	Minimum	10%	25%	Median	75%	90%	Maximum
a	226	227.8	230	233	234	235.4	236
b	224	224.3	226.75	231	232	234.4	235

Oneway Anova

Summary of Fit

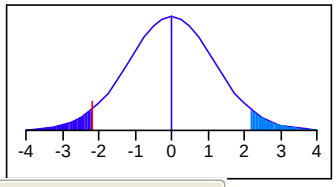
Rsquare	0.122404
Adj Rsquare	0.0873
Root Mean Square Error	3.01894
Mean of Response	231.2963
Observations (or Sum Wgts)	27

t Test

b-a

Assuming equal variances

Difference	-2.1833	t Ratio	-1.86732
Std Err Dif	1.1692	DF	25
Upper CL Dif	0.2247	Prob > |t|	0.0736
Lower CL Dif	-4.5914	Prob > t	0.9632
Confidence	0.95	Prob < t	0.0368*



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
group	1	31.77963	31.7796	3.4869	0.0736
Error	25	227.85000	9.1140		
C. Total	26	259.62963			

Means for Oneway Anova

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
a	15	232.267	0.77949	230.66	233.87
b	12	230.083	0.87149	228.29	231.88

Std Error uses a pooled estimate of error variance

What can we conclude from the output under the **t Test?** (We should get the same values for the calculations as those we obtained manually in activity 7.12.) JMP shows that the difference is taken as $b - a$ (which means there will be a difference in sign to our calculations because we worked with $a - b$).

- The value of the t-statistic is -1.86732 with 25 degrees of freedom and the two-sided p-value is 0.0736 which means that $H_0 : \mu_b = \mu_a$ cannot be rejected at the 5% level of significance. (Do you notice that JMP never makes any conclusions?)
- The 95% two–sided confidence interval for $(\mu_b - \mu_a)$ is $(-4.5914;\ 0.2247)$. [We manually derived the 95% two–sided confidence interval for $(\mu_1 - \mu_2)$ as $(-0.225\,3;\ 4.592\,1)$ which is the swapped interval with a slight difference in calculations due to maybe rounding errors.]

 Please note that if we cannot assume equal variances we have to pick a different test statistic which we will deal with in section 7.5.

7.4 Paired observations

Similar to the situation in study unit 6.3 where we had to derive the complicated test to test $H_0 : \sigma_1^2 = \sigma_2^2$ in the case of *paired observations,* we also have to derive a different test for $H_0 : \mu_1 = \mu_2$ when we have paired observations. Luckily this test simplifies to the one-sample setup of section 7.1 except that we have to "convert" the null hypothesis about μ_1 and μ_2 to an appropriate null hypothesis for μ_D.

Activity 7.15

In activity 6.12 of the previous study unit, you tested for the equality of variances for the following set of **paired observations**:

Mineral compositions were determined for 12 rocks taken from an area in England with many excellently preserved plant fossils. The percentage of siderite ($FeCO_3$) concentrations was determined using two different analysis techniques. Do you think that technique B leads to more variable measurements than technique A? Test at the 10% level of significance.

Rock	1	2	3	4	5	6	7	8	9	10	11	12
% siderite method A	62.2	58.3	74.7	75.0	64.9	59.0	74.3	81.1	89.5	58.3	74.7	63.5
% siderite method B	61.8	57.2	76.1	74.2	63.8	59.7	73.2	80.1	89.2	57.8	74.9	62.7

Use the same data set to test whether $H_0 : \mu_1 = \mu_2$ is true against a two-sided alternative at the 5% level of significance.

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

Feedback  Feedback

We have dependent (paired) observations and need to work with the differences.

(a) **Hypotheses**

$$H_0 : \mu_D = 0 \qquad\qquad H_1 : \mu_D \neq 0$$

where $d_i =$ % siderite using method A $-$ % siderite using method B.

$$d_i : \quad 0.4 \quad 1.1 \quad -1.4 \quad 0.8 \quad 1.1 \quad -0.7 \quad 1.1 \quad 1.0 \quad 0.3 \quad 0.5 \quad -0.2 \quad 0.8$$

(b) The **test statistic** is

$$T = \frac{\bar{d} - 0}{S_d/\sqrt{n}}$$

where

$$\begin{aligned} \bar{d} &= \frac{\sum d_i}{12} = \frac{4.8}{12} = 0.4 \ ; \\ S_d^2 &= \frac{\sum d_i^2 - n\bar{d}^2}{(n-1)} = \frac{8.9 - 12(0.4)^2}{11} = 0.634\,55 \ ; \\ S_d &= \sqrt{0.634\,55} = 0.796\,59. \end{aligned}$$

Therefore,

$$\begin{aligned} T &= \frac{0.4 - 0}{0.796\,59/\sqrt{12}} \\ &= 1.739\,5\ . \end{aligned}$$

We reject H_0 at the 5% level of significance if $|T| \geq t_{0.025;\,11} = 2.201$ and since $1.739\,5 < 2.201$ we cannot reject H_0.

What about JMP?

How will you create a confidence interval or perform a hypothesis test concerning the mean for paired observations using JMP? We will now redo the activity above (where we have *two dependent samples*) by using JMP.

Activity 7.16

Refer to **activity 7.15**.

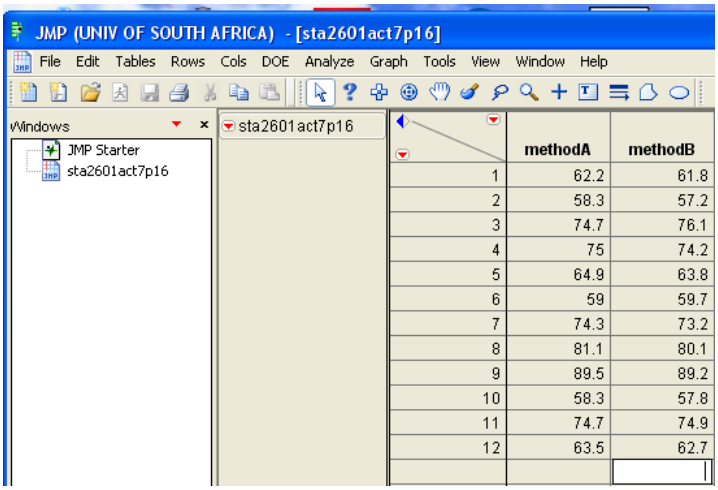
(a) Manipulate JMP to create a data table that will represent the paired observations (that is two dependent samples) and will result in the same output as we obtained manually.

(b) Test whether $H_0 : \mu_1 = \mu_2$ is true against a two-sided alternative at the 5% level of significance.

(c) Comment on the similarities (and the differences) between the manual test and the electronic version of JMP.

Feedback → Feedback

(a) Your JMP table should look like the snapshot below:



	methodA	methodB
1	62.2	61.8
2	58.3	57.2
3	74.7	76.1
4	75	74.2
5	64.9	63.8
6	59	59.7
7	74.3	73.2
8	81.1	80.1
9	89.5	89.2
10	58.3	57.8
11	74.7	74.9
12	63.5	62.7

In other words, we have two columns next to each other with 12 rows of values.

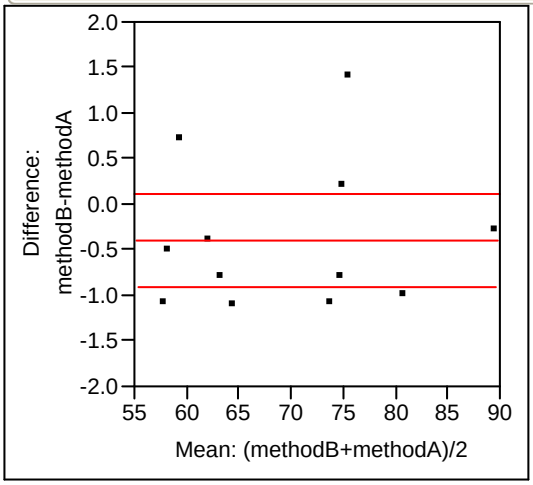
Select **Analyze** ⟶ **Matched Pairs**

Select both columns and move them to the Y, Paired Response block.

If you click on OK JMP will perform a **Matched Pairs Analysis.**

Matched Pairs

Difference: methodB-methodA



methodB	69.225	t-Ratio	-1.73948
methodA	69.625	DF	11
Mean Difference	-0.4	Prob > |t|	0.1098
Std Error	0.22995	Prob > t	0.9451
Upper95%	0.10613	Prob < t	0.0549
Lower95%	-0.9061		
N	12		
Correlation	0.99692		

What can we conclude from the output under the **Matched Pairs?** (We should get the same values for the calculations as those we obtained manually in activity 7.15.) JMP shows that the difference is taken as Method$B-$Method$A$ (which means there will be a difference in sign from our calculations because we worked with Method$A-$Method$B$).

- The value of the t-statistic is -1.73948 with 11 degrees of freedom and the two-sided p-value is 0.1098 which means that $H_0 : \mu_B = \mu_A$ cannot be rejected at the 5% level of significance.
- The 95% two–sided confidence interval for $(\mu_B - \mu_A)$ is $(-0.9061;\ 0.10613)$. [We did not derive the 95% two–sided confidence interval for $(\mu_1 - \mu_2)$ manually.] Note that this interval includes zero $\Longrightarrow \mu_D = 0 \Longrightarrow (\mu_B - \mu_A) = 0 \Longrightarrow \mu_B = \mu_A$.

7.5 Independent samples with unequal variances

Activity 7.17

A counselling psychologist wants to test the hypothesis that the average spatial perception score of engineering students is higher than the average spatial perception score of business management students. She draws a random sample of 16 third-year engineering students and an independent random sample of 20 third-year business management students and applies a spatial perception measurement scale (developed by a renowned psychometrist) to the 36 students. She obtained the following scores which she captured on an Excel spreadsheet:

Engineers	**Business management**
92	84
85	73
101	61
119	57
102	70
107	48
108	67
100	39
96	73
121	37
114	84
98	47
96	89
116	40
104	74
101	86
	71
	97
	48
	104

(a) Test $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 > \mu_2$ at the 5% level of significance.

(b) On what assumptions do you base your test statistic?

(c) Derive a 90% two-sided confidence interval for $(\mu_1 - \mu_2)$.

(d) What one-sided confidence intervals can you conclude from the confidence interval derived in (c)?

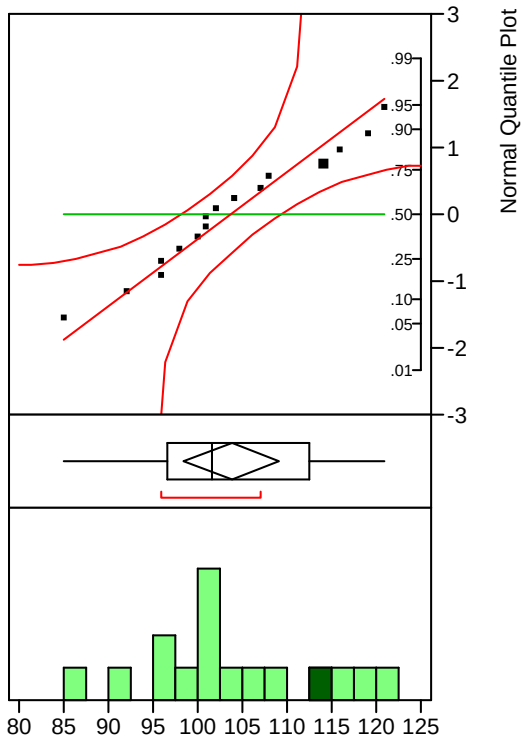
You may make use of the following calculations:

$$\Sigma X_{1i} = 1\,660 \qquad \Sigma X_{1i}^2 = 173\,714$$
$$\Sigma X_{2i} = 1\,349 \qquad \Sigma X_{2i}^2 = 98\,495$$

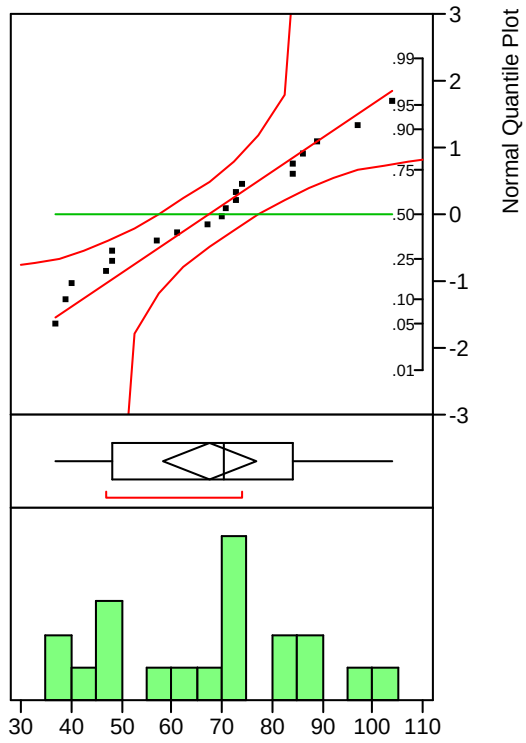
Feedback Feedback

(a) Suppose we call the engineers the "first sample" and the business management students the "second sample".

Regarding the assumptions, the very first test we perform should be to check for normality (before we even test $H_0 : \sigma_1^2 = \sigma_2^2$ against the alternative hypothesis $H_1 : \sigma_1^2 \neq \sigma_2^2$.)



Histogram: box plot and normal quantile plot for engineers



Histogram: box plot and normal quantile plot for business management

From the graphs above it seems reasonable to assume normality.



Are you able to reproduce the graphs?

Now we may use the test statistic $F = \dfrac{\sigma_2^2}{\sigma_1^2} \cdot \dfrac{S_1^2}{S_2^2} \sim F_{n_1-1;\, n_2-1}$ to test $H_0 : \sigma_1^2 = \sigma_2^2$

against the alternative hypothesis $H_1 : \sigma_1^2 \neq \sigma_2^2$.

For this application $n_1 = 16$ and $n_2 = 20 \implies$ we reject H_0 if $F \geq F_{0.025;\, 15;\, 19} = 2.62$ or if

$F \leq F_{0.975;\,15;\,19} = \dfrac{1}{F_{0.025;\,19;\,15}} \simeq \dfrac{1}{2.76} = 0.3623$ (Take the value $v_1 = 20$ and $v_2 = 15$.)

We need to compute S_1^2 and S_2^2 where

$$\overline{X}_1 = \frac{1\,660}{16} = 103.75 \text{ and } S_1^2 = \frac{\sum_{i=1}^{16} X_{1i}^2 - 16(\overline{X}_1)^2}{(16-1)} = \frac{173\,714 - 16(103.75)^2}{15} = 99.2667$$

$$\overline{X}_2 = \frac{1\,349}{20} = 67.45 \text{ and } S_2^2 = \frac{\sum_{i=1}^{20} X_{2i}^2 - 20(\overline{X}_2)^2}{(20-1)} = \frac{98\,495 - 20(67.45)^2}{19} = 394.9974$$

$F = \dfrac{\sigma_2^2}{\sigma_1^2} \cdot \dfrac{S_1^2}{S_2^2} = 1 \cdot \dfrac{99.2667}{394.9974} = 0.251\,31$ and since $0.251\,31 < 0.3623$ we reject H_0.

[If we swap the roles of "1" and "2" it will be just as valid to compute $F = \dfrac{\sigma_1^2}{\sigma_2^2} \cdot \dfrac{S_2^2}{S_1^2} = \frac{394.9974}{99.2667} = 3.979\,2 > 2.62$.]

(b) From (a) above it follows that we have to use the t-statistic for independent samples with *unequal variances* if we want to test $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 > \mu_2$.

$$\begin{aligned}
\therefore T &= \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{(\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2})}} \\
&= \frac{(103.75 - 67.45) - 0}{\sqrt{(\dfrac{99.2667}{16} + \dfrac{394.9974}{20})}} \qquad \text{because } \mu_1 = \mu_2 \implies (\mu_1 - \mu_2) = 0 \\
&= \frac{36.3}{\sqrt{25.954}} = 7.125\,3
\end{aligned}$$

The formula for the degrees of freedom is horrific to compute!

$$\begin{aligned}
\upsilon &= \frac{\left[\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}\right]^2}{\dfrac{S_1^4}{n_1^2(n_1-1)} + \dfrac{S_2^4}{n_2^2(n_2-1)}} \\
&= \frac{\left[\dfrac{99.2667}{16} + \dfrac{394.9974}{20}\right]^2}{\dfrac{(99.2667)^2}{16^2(16-1)} + \dfrac{(394.9974)^2}{20^2(20-1)}} \\
&= \frac{673.6121}{23.0954} = 29.1665
\end{aligned}$$

The critical region to test H_0 against H_1 is where $T \geq t_{0.05;\,\upsilon} = t_{0.05;\,29.166} \approx 1.699$.(No need to even interpolate since the numerical value of T is already $>>$any tabulated value!) We reject H_0 and assume that the average spatial perception of engineering students is higher than the average spatial perception of business management students.

(c) A $(1-a)100\%$ two-sided confidence interval for $(\mu_1 - \mu_2)$ will be given by $(\overline{X}_1 - \overline{X}_2) \pm t_{\alpha/2;\,29.1665}\sqrt{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)}$.

$\Longrightarrow$A 90% two-sided confidence interval for $(\mu_1 - \mu_2)$ will be given by

$$\begin{aligned}(\overline{X}_1 - \overline{X}_2) \pm t_{0.05;\,29}\sqrt{(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2})} &= (103.75 - 67.45) \pm (1.699)\sqrt{(\frac{99.2667}{16} + \frac{394.9974}{20})} \\ &= 36.3 \pm 8.655\,6 \\ &= (27.6444;\ 44.9556).\end{aligned}$$

(d) We may use the upper limit of a $(1-a)100\%$ two-sided confidence interval to obtain a $(1-\frac{a}{2})100\%$ one-sided confidence interval of the form $(-\infty; UpperL)$ and vice versa we may use the lower limit of a $(1-a)100\%$ two-sided confidence interval to obtain a $(1-\frac{a}{2})100\%$ one-sided confidence interval of the form $(LowerL; \infty)$. Since we are interested in $(\mu_1 - \mu_2) >$ "something" we will use the form $(LowerL; \infty)$.

$\Longrightarrow$A 95% one-sided confidence interval for $(\mu_1 - \mu_2)$ will be given by $(27.6444;\ \infty)$.

$\Longrightarrow$ We are 95% confident that $(\mu_1 - \mu_2) > 27.6444$.

$\Longrightarrow$ We are 95% confident that $\mu_1 > \mu_2 + 27.6444 \Longrightarrow$ the engineering students score on average 27.6444 points higher on spatial perception than business management students.

 Why did I give you this question? It guides you to manipulate JMP for a one-sided interval!

 Now we are challenged to draw the same conclusions from JMP!

Activity 7.18

Refer to activity 7.17 above (compare this with activity 7.14).

(a) Create a data table (similar to the one you created in activity 7.14) for the data given in activity 7.17 (save it as *act7p17.jmp*).

(b) Test $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 > \mu_2$ at the 5% level of significance.

(c) Derive an appropriate 95% one-sided confidence interval for $(\mu_1 - \mu_2)$.

(d) Comment on the similarities (and the differences) between the manual test and the electronic version of JMP.

Feedback ⟶ Feedback

(a) Similar to the data table you created in activity 7.14 the data table should have two columns where the first one (called *group*) acts as the grouping variable and the second one (called *score*) contains the spatial perception scores.

(b) Similar to the initial steps you performed in activity 7.14 you will do the same here.

Select **Analyze** $\longrightarrow$ **Fit Y by X.**

Convey the choice of "Independent variable = X, Factor = *group*" and that "Dependent variable = Y, Response = *score*" to JMP in the relevant boxes.

If you click on OK JMP will perform a **Oneway Analysis of score By group**.

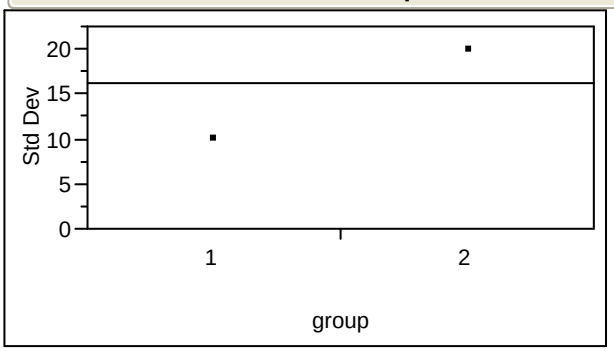
Once again (as in activity 7.14) the initial output contains only a very basic graph showing the individual observations of two separate groups.

Even though we know from the manual test performed in activity 7.17 that $\sigma_1^2 \neq \sigma_2^2$ we should strictly speaking also test whether the population variances are equal. (In other words consider it as the first time you work with this data set and now select to do a **Tests that the Variances are Equal.** Click on the red triangle in the **Oneway Analysis of score By group** bar in the output.

Select the seventh one $\longrightarrow$ **UnEqual Variances,** which will produce five different test statistics to **Test that the Variances are Equal.**

Your output should look like this:

Tests that the Variances are Equal



Level	Count	Std Dev	MeanAbsDif to Mean	MeanAbsDif to Median
1	16	9.96327	7.84375	7.62500
2	20	19.87454	16.30500	16.05000

Test	F Ratio	DFNum	DFDen	p-Value
O'Brien[.5]	7.5938	1	34	0.0093*
Brown-Forsythe	6.7486	1	34	0.0138*
Levene	8.0360	1	34	0.0077*
Bartlett	6.8783	1	.	0.0087*
F Test 2-sided	3.9792	19	15	0.0092*

Welch Anova testing Means Equal, allowing Std Devs Not Equal

F Ratio	DFNum	DFDen	Prob > F
50.7701	1	29.166	<.0001*

t Test
7.1253

The p-values (of all the different tests to test for the "subhypothesis test" $H_0 : \sigma_A^2 = \sigma_B^2$ against the alternative hypothesis $H_1 : \sigma_B^2 \neq \sigma_A^2$) are all $< 0.05 \Longrightarrow$ we reject H_0 and cannot assume equal variances.

What do you also notice about the output? JMP automatically computed the Welch t-statistic which one has to use when the variances are unequal. However, the output does not give a p-value for the "main test" neither any confidence intervals, so we have to select the appropriate test from the menu. This is of course the same route if you directly want to focus your attention on the **test for the difference between the means** when you know that you may not use the test statistic where we pool the variances. You therefore have to select option 4 of the submenu.

If you click on the red triangle in the **Oneway Analysis of score By group** bar in the output, it opens the submenu with the 16 choices.

Quantiles
2 Means/Anova/Pooled t
4 t Test
UnEqual Variances
⋮
Script

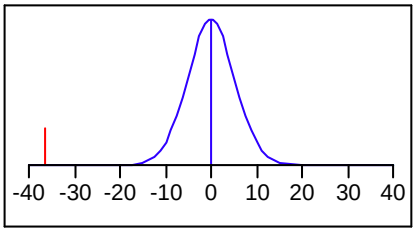
Select the **fourth** one $\longrightarrow$ t **Test.**

Notice what happens to the output:

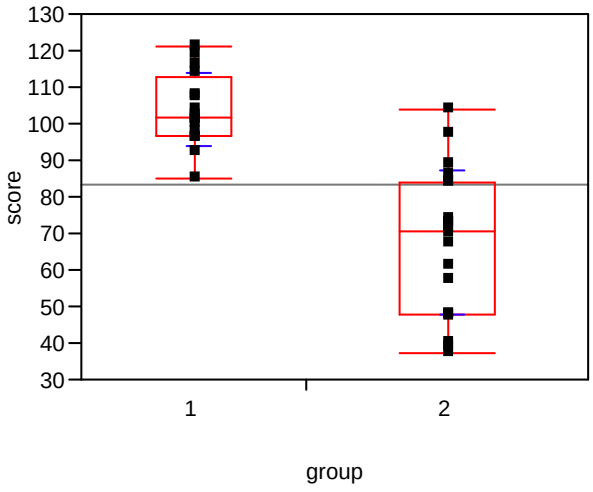
t Test

2-1
Assuming unequal variances

Difference	-36.300	t Ratio	-7.12532
Std Err Dif	5.095	DF	29.16644
Upper CL Dif	-25.883	Prob > |t|	<.0001*
Lower CL Dif	-46.717	Prob > t	1.0000
Confidence	0.95	Prob < t	<.0001*



In contrast to what happened in activity 7.14 where we selected option 2, this initial graph is not enriched with a box plot (or a means diamond plot). Since it is very informative to have the box plots, we may obtain this by clicking on the first option of the submenu which is **Quantiles.** (In other words, click again on the red triangle in the **Oneway Analysis of scores By group**.)



This picture visualises what we confirmed with the statistical tests: The two groups differ in all respects – neither the spread nor the means are the same.

What can we conclude from the output under the **t Test**? (We should get the same values for the calculations as those we obtained manually in activity 7.17.) JMP shows that the difference is taken as $2 - 1$ (which means there will be a difference in sign from our calculations because we worked with $1 - 2$).

- The value of the t-statistic is -7.12532 with 29.1664 degrees of freedom and the two-sided p-value is 0.0001 which means that $H_0 : \mu_2 = \mu_1$ is rejected at **any** level of significance.
- The 95% two–sided confidence interval for $(\mu_2 - \mu_1)$ is $(-46.717;\ 25.883)$. [We did not manually derive the 95% two–sided confidence interval.]

(c) Please note that we have to manipulate JMP to obtain a 95% one-sided interval by computing a 90% two–sided confidence interval for $(\mu_2 - \mu_1)$.

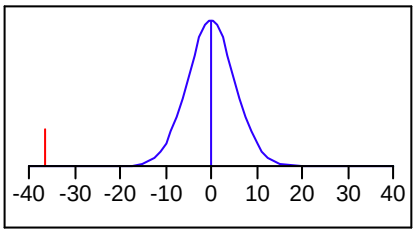
Click on the red triangle in the **Oneway Analysis of score By group** bar in the output and select **Set α Level.** Change the α-level to $\alpha = 0.10$.

t Test

2-1

Assuming unequal variances

Difference	-36.300	t Ratio	-7.12532
Std Err Dif	5.095	DF	29.16644
Upper CL Dif	-27.645	Prob > |t|	<.0001*
Lower CL Dif	-44.955	Prob > t	1.0000
Confidence	0.9	Prob < t	<.0001*



Thus a 90% two-sided confidence interval is $(-44.955;\ -27.645)$ from which we may conclude $P[(\mu_2 - \mu_1) < -27.645) = 0.95 \implies$ we are 95% confident that $\mu_2 < \mu_1 - 27.645 \implies$ that the business management students score on average 27.645 points lower on spatial perception than engineering students.

This is exactly the same conclusion we drew with the manual computation, that is that the engineering students score on average 27.6454 points higher on spatial perception than business management students.

7.6 More than two independent samples

The Anova test we do in this section has the same three assumptions (extended to more than two samples) which we needed to do the pooled t-test:

1. We have more than two **independent** populations (from which we draw random samples).
2. All the populations have **normal** distributions.
3. All populations have **the same variance**, that is $\sigma_1^2 = \ \sigma_2^2 = ... = \sigma_k^2$.

Activity 7.19

In **activity 6.13** of the previous study unit, you already tested for the equality of **variances** for three independent samples (which we luckily repeated with JMP in activity 6.14 – so we have the electronic data set).

Now we have to test for the equality of **means** for three independent samples.

 Yes, we first have to do it manually before we repeat it on JMP in the next activity! In a typical examination question, we will give you the table with the raw data but we will usually also give you some of the computations.

The following table gives the rate of flow of blood through the kidneys for three distinct age groups:

	Rate of flow for 20-year-old adults (X_{1i})	Rate of flow for 40-year-old adults (X_{2i})	Rate of flow for 60-year-old adults (X_{3i})
$i = 1$	542	440	339
$i = 2$	544	443	340
$i = 3$	545	444	342
$i = 4$	548	446	343
$i = 5$	549	448	344
$i = 6$	550	450	350
$i = 7$	550	451	352
$i = 8$	551	454	355
$i = 9$	554	455	356
$i = 10$	557	459	359

You may assume that

$$\begin{aligned}
\Sigma X_{1i} &= 5490 \\
\Sigma X_{2i} &= 4490 \\
\Sigma X_{3i} &= 3480 \\
\sum \left(X_{1i} - \overline{X}_1\right)^2 &= 186 \\
\sum \left(X_{2i} - \overline{X}_2\right)^2 &= 318 \\
\sum \left(X_{3i} - \overline{X}_3\right)^2 &= 476
\end{aligned}$$

(a) Test $H_0 : \mu_1 = \mu_2 = \mu_3$ at the 5% level of significance against $H_1 : \mu_p \neq \mu_q$ (for at least one pair $p \neq q$).

(b) On what assumptions do you base your test statistic?

(c) Perform multiple comparisons.

Feedback Feedback

(a) For the ANOVA test we need the following calculations:

$$
\begin{aligned}
MSE &= \frac{SSE}{(kn-k)} \quad \text{(definition 7.2 in the study guide)} \\
&= \frac{\sum_{i=1}^{3}\sum_{j=1}^{10}(X_{ij}-\overline{X}_i)^2}{3(10)-3} \\
&= \frac{186+318+476}{27} \\
&= \frac{980.0}{27} \\
&= 36.2963
\end{aligned}
$$

We have $k = 3$ random samples of size $n = 10$ each.

	$\sum_{j=1}^{10} X_{ij}$	$\overline{X}_i$	$(\overline{X}_i - \overline{X})^2$
$i = 1$	5 490	549	$(549 - 448.6667)^2 = 10066.7711$
$i = 2$	4 490	449	$(449 - 448.6667)^2 = 0.111\,1$
$i = 3$	3 480	348	$(348 - 448.6667)^2 = 10\ 133.7845$
Total$(\sum_{i=1}^{3})$	13 460	$\overline{X} = 448.6667$	20 200.6667

The test statistic is $F = \dfrac{MSTr}{MSE} \sim F_{k-1;kn-1}$ under H_0. (See theorem 7.7 in the study guide.)

H_0 is rejected if $F \geqslant F_{\alpha;k-1;kn-k} = F_{0.05;2;27} = 3.34$ (Stoker tables).

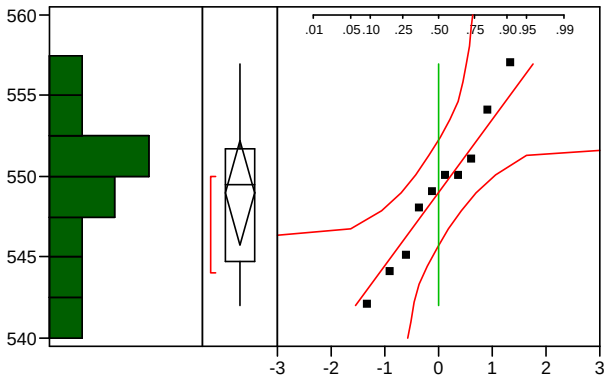
$$\begin{aligned} F &= \frac{MSTr}{MSE} = \frac{n\sum_{i=1}^{k}(\overline{X}_i - \overline{X})^2/(k-1)}{\sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij} - \overline{X}_i)^2/(kn-k)} \\ &= \frac{10(20\,200.6667)/2}{36.2963} \\ &= \frac{101\,003.3335}{36.2963} \\ &= 2782.7446 \end{aligned}$$

H_0 is rejected at the 5% level of significance since $2782.7446 \ggg 3.34$.

(b) (i) INDEPENDENCE: The one group cannot influence the other groups.

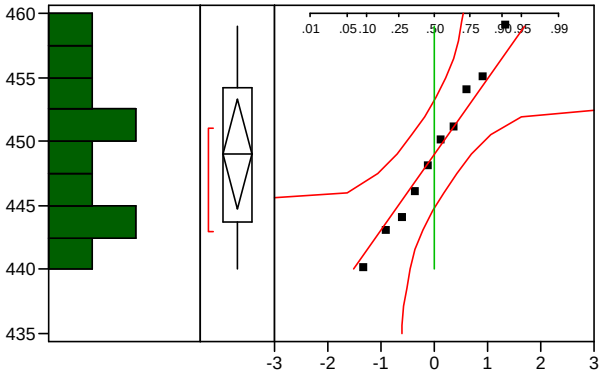
(ii) NORMALITY: Most clinical variables have a normal distribution (if we not simply want to assume it, we can easily test with JMP doing a normal quantile plot – **see graphs below**).

(iii) EQUAL POPULATION VARIANCES: This was formally tested in activities 6.13 and 6.14 of the previous study unit where $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2$ could not be rejected.



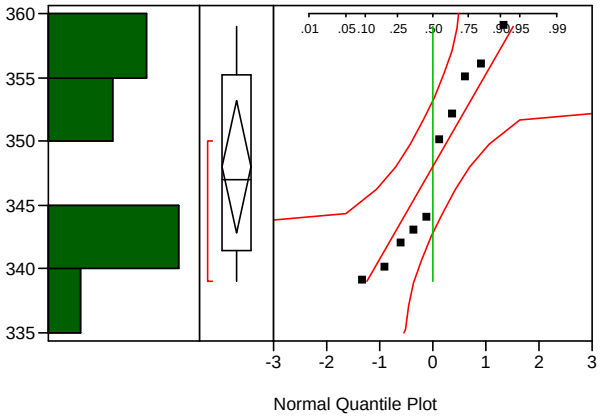
20-year-old adults (X_{1i})

Assume normality.



40-year-old adults (X_{2i})

Assume normality.



60-year-old adults (X_{3i})

Assume normality.

(c) We may compute, for each pair of means $\overline{X}_p$ and $\overline{X}_q$, a T-statistic

$$T_{pq} = \frac{\overline{X}_p - \overline{X}_q}{S\sqrt{\frac{1}{n} + \frac{1}{n}}} = \frac{\sqrt{n}(\overline{X}_p - \overline{X}_q)}{\sqrt{2}S}$$

and reject $H_0(p;q) : \mu_p = \mu_q$ in favour of $H_1 : \mu_p \neq \mu_q$

if $|T_{pq}|$ exceeds $\sqrt{(k-1)F_{\alpha;k-1;kn-k}}$.

So, for this problem we will reject H_0 if $|T_{pq}| \geqslant \sqrt{2(3.34)} = 2.5846$

where $T_{pq} = \sqrt{\frac{10}{2}}\frac{1}{\sqrt{36.2963}}(\overline{X}_p - \overline{X}_q) = 0.371\,15(\overline{X}_p - \overline{X}_q)$

$\Longrightarrow$ reject H_0 if

$$\begin{aligned} & \left|0.371\,15(\overline{X}_p - \overline{X}_q)\right| \geqslant 2.5846 \\ \Longrightarrow \quad & \left|(\overline{X}_p - \overline{X}_q)\right| \geqslant \frac{2.5846}{0.371\,15} \\ \Longrightarrow \quad & \left|(\overline{X}_p - \overline{X}_q)\right| \geqslant 6.9638 \end{aligned}$$

$\overline{X}_1 - \overline{X}_2 = 100; \quad \overline{X}_2 - \overline{X}_3 = 101; \quad \overline{X}_1 - \overline{X}_3 = 201 \Longrightarrow$ all the differences are significant $\Longrightarrow$ reject all $H_0(p;q) : \mu_p = \mu_q$ in favour of $H_1 : \mu_p \neq \mu_q$.

Activity 7.20

Repeat activity 7.19 by using JMP.

Feedback ⟶ Feedback

(a) Refer to the output of activity 6.14 to refresh your memory on the data set *act6p13.jmp* which you have to use.

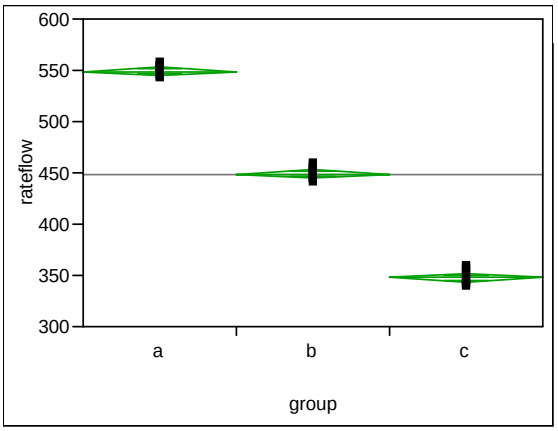
Click on **Analyze** $\longrightarrow$ **Fit Y by X.**

Complete the "**Distribution of Y for each X. Modeling types determine analysis"** dialog box by choosing "Independent variable = X, Factor = *group*" and "Dependent variable = Y, Response = *rateflow*".

Click on OK to perform a **Oneway Analysis of rateflow By group**.

If you click on the red triangle in the **Oneway Analysis of rateflow By group** bar in the output, it opens the submenu with 16 choices.

Select the **second** one $\longrightarrow$ **Means/Anova.**



Oneway Anova

Summary of Fit

Rsquare	0.995172
Adj Rsquare	0.994814
Root Mean Square Error	6.024641
Mean of Response	448.6667
Observations (or Sum Wgts)	30

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
group	2	202006.67	101003	2782.745	<.0001*
Error	27	980.00	36		
C. Total	29	202986.67			

Means for Oneway Anova

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
a	10	549.000	1.9052	545.09	552.91
b	10	449.000	1.9052	445.09	452.91
c	10	348.000	1.9052	344.09	351.91

Std Error uses a pooled estimate of error variance

The outputs **Analysis of Variance** and **Means for Oneway Anova** give the same calculations (and values) for what we obtained in the previous activity when we computed it manually. The only difference is the absence of a critical value and the inclusion of a p-value. Since $0.0001 << 0.01$ we reject the null hypothesis. The first report (**Summary of Fit**) computes additional measures which will be dealt with in third-year STA3701.

(b) We should have, strictly speaking, first verified whether the population variances are equal. (In other words if this was the first time you worked with this data set you should have selected to do a **Tests that the Variances are Equal** (we will not repeat this because we emphasised it in activities 6.14 and 6.15 of the previous study unit).

[To refresh your memory: Click on the red triangle in the

Oneway Analysis of weights By group bar in the output. Select the sixth one $\longrightarrow$ **UnEqual Variances,** which will produce the test statistics to **Test that the Variances are Equal.** The p-values of all the different tests are $> 0.05 \Longrightarrow$ the "subhypothesis test" $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2$ could not be rejected.]

Secondly, you can test for normality by doing a **normal quantile plot** for each sample. (This was already done under (b) of the previous activity and will not be repeated here – **see graphs in activity 7.19 above**.)

(c) **Now for the multiple comparisons:**

Click on the red triangle in the **Oneway Analysis of weights By group** bar in the output.

Select the fourth one ⟶ **Compare Means,** which will open another menu from which you select **All Pairs, Tukey HSD.** Do not select the first one **Each Pair, Student's t,** because it makes no adjustment for multiple tests. Tukey's HSD protects the overall error rate.

Means Comparisons

Comparisons for all pairs using Tukey-Kramer HSD

q*	Alpha
2.47942	0.05

Abs(Dif)-LSD

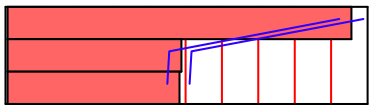
	a	b	c
a	-6.68	93.32	194.32
b	93.32	-6.68	94.32
c	194.32	94.32	-6.68

Positive values show pairs of means that are significantly different.

Level		Mean
a	A	549.00000
b	B	449.00000
c	C	348.00000

Levels not connected by same letter are significantly different.

Level	- Level	Difference	Lower CL	Upper CL
a	c	201.0000	194.3197	207.6803
b	c	101.0000	94.3197	107.6803
a	b	100.0000	93.3197	106.6803



From this we also conclude that all pairs of means are significantly different.



It is time to draw yourself a mind map of this study unit!

STUDY UNIT 8

Regression

8.1 Correlation and regression

In section 5.4 of study unit 5 we explained the scenario where we have **two continuous variables** (which are simultaneously observed) and where we are interested in the question of whether or not they are correlated. In South Africa, politicians tell us that "poverty is correlated with crime" and we formalised the idea of correlation in a statistical sense. On the other hand, sometimes the two variables, X and Y, are related in a particular way. It may be that **the value of Y depends on the value of X**. In other words, the value of X in some way explains the value of Y, which makes it a *causal process*. Preferably more than three observations for the same value of X should be performed in order to estimate the random component of a true regression experiment. So, if one of the two variables can be classified as the dependent variable Y and the other as the independent variable X, and if the data show a straight line pattern, it is possible to describe the relationship by estimating the *equation of a straight line*. This technique is known as **linear regression** and will be covered in this study unit.

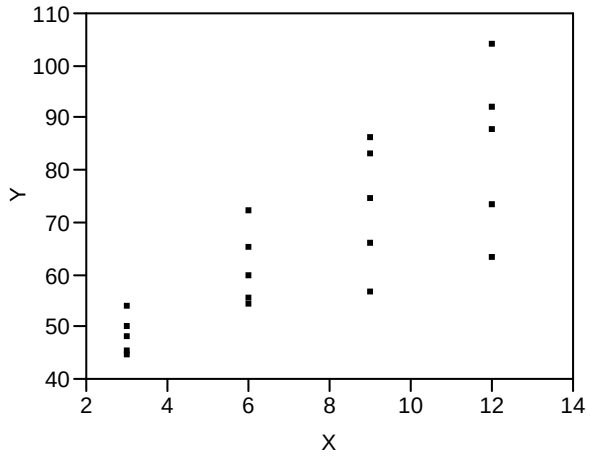
8.2 The simple linear regression model

Activity 8.1

Consider the following bivariate data set:

X	Y
3	49.9
3	53.8
3	44.1
3	45.2
3	47.8
6	71.96
6	65.02
6	59.45
6	55.24
6	53.91
9	85.79
9	82.9
9	74.14
9	65.49
9	56.33
12	103.7
12	91.8
12	87.46
12	73.22
12	63.06

Plotting X against Y for this data set results in the following scatter plot:



Would you say that the simple linear regression model is appropriate?

.....

.....

.....

.....

Feedback  Feedback

No, the simple linear regression model is not appropriate because the variance seems to increase as X increases.

Activity 8.2

Refer to activity 8.1 above.

(a) Create your own JMP data table for the given data.

(b) Create $\frac{Y}{X}$ and $\frac{1}{X}$ from the existing columns (save it as *act8p1.jmp*).

(c) Plot $\frac{Y}{X}$ versus $\frac{1}{X}$ and comment.

Feedback ⟶ Feedback

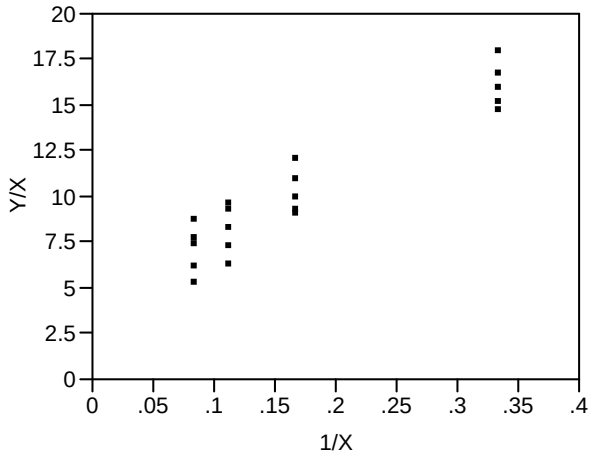
(a) + (b) Your data table should have the four columns:

X	Y	1/X	Y/X
3	49.9	0.33333333	16.6333333
3	53.8	0.33333333	17.9333333
3	44.1	0.33333333	14.7
3	45.2	0.33333333	15.0666667
3	47.8	0.33333333	15.9333333
6	71.96	0.16666667	11.9933333
6	65.02	0.16666667	10.8366667
6	59.45	0.16666667	9.90833333
6	55.24	0.16666667	9.20666667
6	53.91	0.16666667	8.985
9	85.79	0.11111111	9.53222222
9	82.9	0.11111111	9.21111111
9	74.14	0.11111111	8.23777778
9	65.49	0.11111111	7.27666667
9	56.33	0.11111111	6.25888889
12	103.7	0.08333333	8.64166667
12	91.8	0.08333333	7.65
12	87.46	0.08333333	7.28833333
12	73.22	0.08333333	6.10166667
12	63.06	0.08333333	5.255

[*For beginners*: How to create a new variable from an existing one.

- Click on the **Cols** button and then select **New Column.**
- The **New Column** dialog box allows you to rename the column. Replace *Column3* with **1/X**.
- Next choose **Column Properties** ⟶**Formula.**
- Selecting **Formula** shows another dialog box with *Table Columns* in the first column (that is, the current variables in the data set, for example X and Y); *mathematical operators* in the second column and (grouped) *Functions* in the third column. Do not select any of the functions but go directly to the **red rectangle** that says no formula.
- If you double-click this red rectangle it becomes black because it goes into edit mode and you can now type your own formula.
- (It is always safer to use the existing variables displayed in *Table Columns* as well as the *mathematical operators* in the second column when you create a formula.) **Type 1** and then select ÷ from the *mathematical operators* (this immediately creates a fraction $\frac{1}{\square}$ where you can select and move X as the denominator.) Close the formula Editor by clicking OK. Also click **OK** in the **New Column** dialog box.
- Repeat the whole process to create Y/X.]

(c) Select **Analyze** $\longrightarrow$ **Fit Y/X by 1/X.**



Now the simple linear regression model seems to be appropriate because the variance remains constant and the points follow a linear trend.

8.3 Estimation



This section will make no sense if you did not master the section on *maximum likelihood estimation* in study unit 2! Apart from the theoretical derivations and motivations, this section should make intuitive sense to you as well.

To draw the best fitting line through the points on a scatter diagram by taking a ruler and simply guessing how it would fit "right through the middle of the points", would probably result in a different line each time we repeat the exercise! There is an exact mathematical procedure called the *least-squares principle* that will give the same line every time! This method relies on finding a line such that the sum of the squared differences between the data points and the line is minimised. This method of *least-squares estimation* and the *maximum likelihood estimation* method lead to the same formulas to estimate β_0 and β_1. The best fitting line to a sample scatter plot is indicated by $\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X$.

$\widehat{\beta}_0$ *is a constant:* This is the value at which the straight line **intersects** or cuts the *y*-axis.

$\widehat{\beta}_1$ *is a constant:* $\widehat{\beta}_1$ is the **slope** of the line. It indicates the change in Y for every change of one unit in X.

In section 5.4 we observed that perfect correlation (that is $R = 1$) is practically nonexistent in real-life situations. The same applies to perfect regression because there will usually be some scatter around the regression line. The closer the observed data points are scattered around the line, the better our predictions of Y based solely on X will be. How do we estimate this random component or "closeness around the line"? We will define a measure of variation, similar to the ordinary standard deviation, but instead of summing the squared deviations from the mean, it uses the squared deviations between the observed value Y_i and the calculated (predicted) value $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$ (for each i). This variance around the line (which is theoretically the variance σ^2 of the random variable E) is denoted as $\widehat{\sigma}^2 = S^2$.

(**Please note** that after you have worked through activities 8.3-8.7 you will be directed to the textbook where there is an animation exercise "**Seeing least squares**" which you can do. This is optional and for fun!)

Activity 8.3

Indicate whether the following statements are correct or incorrect and try to rectify the incorrect statements to make them true.

(a) $S^2 = \dfrac{\sum \left(Y_i - \widehat{Y}_i\right)^2}{n-2}$ estimates the variance around the line.

.....

.....

(b) The **least-squares estimators** for β_0 and β_1 are:

$$\widehat{\beta}_1 = \frac{n\sum XY - (\sum X)(\sum Y)}{n(\sum X^2) - (\sum X)^2}$$

$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X}.$$

.....

.....

(c) $\widehat{\beta}_1 = R\left(\frac{S_x}{S_y}\right) = \frac{S_{xy}}{S_x^2}$

.....

.....

Feedback → Feedback

(a) Correct.

(b) Correct. *Maximum likelihood estimation* and *least-squares estimation* result in the same estimates.

(c) Correct. With a little algebraic manipulation it can be shown that this is the **alternative computational formula for** $\widehat{\beta}_1$. Notice that the necessary sums and sums of squares required to compute $\widehat{\beta}_1$ are the same as for the computation of the correlation coefficient R. For this reason $\widehat{\beta}_1$ can be expressed as a function of R.



The same principle reflected in theorem 6.1 with $\overline{X}$, the sample mean, being a random variable, applies to the observed value of $\widehat{\beta}_1$, the **slope** of the regression line. In other words $\widehat{\beta}_1$ will vary from sample to sample – thus it is a **random variable**. A random variable is of no use if you do not know the value of its mean and variance.

It was shown in theorem 8.1 that $E(\widehat{\beta}_1) = \beta_1$ and

$$Var(\widehat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}.$$

Of course, the same is true for $\widehat{\beta}_0$, the **intersect** of the regression line. In other words $\widehat{\beta}_0$ will vary from sample to sample – thus it is also a **random variable** which is of no use unless you know the value of its mean and variance.

It was also shown in theorem 8.1 that $E(\widehat{\beta}_0) = \beta_0$ and

$$Var(\widehat{\beta}_0) = \sigma^2\left[\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\right].$$

Activity 8.4

(a) If you assume that the **random variable** $\widehat{\beta}_1$ has a normal distribution, write down the standardised form of $\widehat{\beta}_1$.

.....

.....

.....

.....

(b) If you assume that the **random variable** $\widehat{\beta}_0$ has a normal distribution, write down the standardised form of $\widehat{\beta}_0$.

.....

.....

.....

.....

Feedback Feedback

(a) $Z_{\widehat{\beta}_1} = \dfrac{\widehat{\beta}_1 - E(\widehat{\beta}_1)}{\sqrt{var(\widehat{\beta}_1)}} = \dfrac{\widehat{\beta}_1 - \beta_1}{\sqrt{\dfrac{\sigma^2}{\sum\limits_{i=1}^{n}\left(X_i - \overline{X}\right)^2}}}$

(b) $Z_{\widehat{\beta}_0} = \dfrac{\widehat{\beta}_0 - E(\widehat{\beta}_0)}{\sqrt{var(\widehat{\beta}_0)}} = \dfrac{\widehat{\beta}_0 - \beta_0}{\sqrt{\sigma^2\left[\dfrac{1}{n} + \dfrac{\overline{X}^2}{\sum\limits_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\right]}}$



The problem with $Z_{\widehat{\beta}_1} = \dfrac{\widehat{\beta}_1 - \beta_1}{\sqrt{\dfrac{\sigma^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}}}$ being a "workable" variable, is that we need to **know the value of** σ^2.

In real life this is usually not the case! The alternative is to replace σ^2 with its unbiased estimator $S^2 = \dfrac{\sum\left(Y_i - \widehat{Y}_i\right)^2}{n-2}$. **This, however, changes the nature of the theoretical distribution.** So, for inference about the **slope** of the regression line, $Z_{\widehat{\beta}_1} = \dfrac{\widehat{\beta}_1 - \beta_1}{\sqrt{\dfrac{\sigma^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}}}$ **is replaced by** $T_1 = \dfrac{\widehat{\beta}_1 - \beta_1}{\sqrt{\dfrac{S^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}}}$ which has a t-distribution with $\upsilon = n-2$ degrees of freedom.

8.4 Inference on the coefficients

Activity 8.5

How will you change $Z_{\widehat{\beta}_0} = \dfrac{\widehat{\beta}_0 - E(\widehat{\beta}_0)}{\sqrt{var(\widehat{\beta}_0)}} = \dfrac{\widehat{\beta}_0 - \beta_0}{\sqrt{\sigma^2\left[\dfrac{1}{n} + \dfrac{\overline{X}^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\right]}}$ to be a "workable" variable, if we do not **know the value of** σ^2?

.....

.....

.....

.....

.....

Feedback → Feedback

Replace σ^2 with its unbiased estimator $S^2 = \dfrac{\sum \left(Y_i - \widehat{Y}_i\right)^2}{n-2}$. **This, however, changes the nature of the theoretical distribution.** So, for inference about the **intercept** of the regression line, $Z_{\widehat{\beta}_0} = \dfrac{\widehat{\beta}_0 - \beta_0}{\sqrt{\sigma^2\left[\dfrac{1}{n} + \dfrac{\overline{X}^2}{\sum\limits_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\right]}}$ **is replaced by** $T_0 = \dfrac{\widehat{\beta}_0 - \beta_0}{\sqrt{S^2\left[\dfrac{1}{n} + \dfrac{\overline{X}^2}{\sum\limits_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\right]}}$ which has a t-distribution with $\upsilon = n-2$ degrees of freedom.

Using the formula $S^2 = \dfrac{\sum \left(Y_i - \widehat{Y}_i\right)^2}{n-2}$ implies that you first need to compute $\widehat{\beta}_1$ and $\widehat{\beta}_0$ to obtain an equation for the regression line and then calculate the (predicted) value $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$ (for each i). This enables you to compute the squared deviations between the observed value Y_i and the predicted value $\widehat{Y}_i$. There is a less laborious method to compute S^2.

Activity 8.6

Show that an alternative formula to compute $S^2 = \dfrac{\sum\limits_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2}{n-2}$ is to compute

$$S^2 = \frac{\sum\limits_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 - \widehat{\beta}_1^2 \sum\limits_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{(n-2)}.$$

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

Feedback → Feedback

$$\sum_{i=1}^{n}\left(Y_i-\widehat{Y}_i\right)^2=\sum_{i=1}^{n}\left(Y_i-[\widehat{\beta}_0+\widehat{\beta}_1X_i]\right)^2=\sum_{i=1}^{n}\left(Y_i-[(\overline{Y}-\widehat{\beta}_1\overline{X})+\widehat{\beta}_1X_i]\right)^2$$

$$=\sum_{i=1}^{n}\left(Y_i-\overline{Y}+\widehat{\beta}_1\overline{X}-\widehat{\beta}_1X_i]\right)^2=\sum_{i=1}^{n}\left([Y_i-\overline{Y}]-[\widehat{\beta}_1X_i-\widehat{\beta}_1\overline{X}]\right)^2 \text{ (regroup terms)}$$

$$=\sum_{i=1}^{n}\left([Y_i-\overline{Y}]^2-2[Y_i-\overline{Y}]\widehat{\beta}_1[X_i-\overline{X}]+\widehat{\beta}_1^2[X_i-\overline{X}]^2\right)=\sum_{i=1}^{n}[Y_i-\overline{Y}]^2-\sum_{i=1}^{n}2[Y_i-\overline{Y}]\widehat{\beta}_1[X_i-\overline{X}]+\sum_{i=1}^{n}\widehat{\beta}_1^2[X_i-\overline{X}]^2$$

But

$$\begin{aligned}\sum_{i=1}^{n}2[Y_i-\overline{Y}]\widehat{\beta}_1[X_i-\overline{X}] &= 2\widehat{\beta}_1\sum_{i=1}^{n}[Y_i-\overline{Y}][X_i-\overline{X}]\\ &= 2\widehat{\beta}_1\frac{\sum_{i=1}^{n}[Y_i-\overline{Y}][X_i-\overline{X}]}{\sum_{i=1}^{n}[X_i-\overline{X}]^2}\cdot\frac{\sum_{i=1}^{n}[X_i-\overline{X}]^2}{1} \quad \text{(trick manipulation!)}\\ &= 2\widehat{\beta}_1(\widehat{\beta}_1)\sum_{i=1}^{n}[X_i-\overline{X}]^2\\ &= 2\widehat{\beta}_1^2\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2\end{aligned}$$

Hence, $\sum_{i=1}^{n}\left(Y_i-\widehat{Y}_i\right)^2=\sum_{i=1}^{n}[Y_i-\overline{Y}]^2-2\widehat{\beta}_1^2\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2+\sum_{i=1}^{n}\widehat{\beta}_1^2[X_i-\overline{X}]^2=\sum_{i=1}^{n}[Y_i-\overline{Y}]^2-\widehat{\beta}_1^2\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$.

You will notice that this is similar to what we derive in section 8.6 of the study guide as $(n-2)S^2=S_{22}-\frac{S_{12}^2}{S_{11}}$. (See activity 8.11 when you have studied section 8.6.)


We will not expect you to reproduce the above proof in the examination. However, it helps to understand that alternative formulae are not merely fabricated! Activities 8.3 to 8.6 assessed whether you understand the theory behind regression and now you are ready to apply it comprehensively to a real-life situation!

Activity 8.7

A random sample was taken of 15 car owners, using the same insurance company and driving cars in the same price class. The table below lists their driving experience (in years) and their monthly car insurance premiums (in rand).

Driving experience	Monthly premium
34	175
4	280
12	231
39	180
6	275
2	282
21	217
17	220
45	148
31	185
26	200
9	250
15	241
24	220
30	196

(a) Can we consider this as bivariate data on *two numerical variables*?

(b) Can we consider this as an ideal causal regression experiment?

(c) Draw a scatter diagram of the data.

(d) Calculate the regression equation.

(Insurance premium is the dependent variable.)

(e) What is the estimated variance around the line? Use this to derive values for the **standard error** of $\widehat{\beta}_0$ and the **standard error** of $\widehat{\beta}_1$.

(f) Derive a 95% confidence interval for β_0, the intercept of the population regression equation.

(g) Derive a 95% confidence interval for β_1, the slope of the population regression equation.

You may make use of the following calculation table:

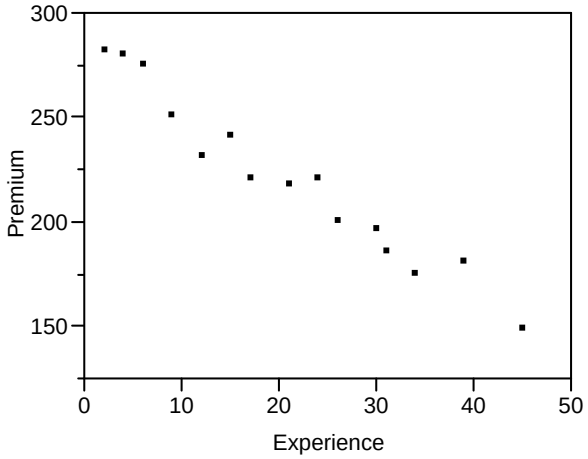
Experience	Premium	$(X-\overline{X})^2$	$(X-\overline{X})$	$(Y-\overline{Y})$	$(X-\overline{X})(Y-\overline{Y})$	$(Y-\overline{Y})^2$
34	175	169	13	−45	−585	2 025
4	280	289	−17	60	−1 020	3 600
12	231	81	−9	11	−99	121
39	180	324	18	−40	−720	1 600
6	275	225	−15	55	−825	3 025
2	282	361	−19	62	−1 178	3 844
21	217	0	0	−3	0	9
17	220	16	−4	0	0	0
45	148	576	24	−72	−1 728	5 184
31	185	100	10	−35	−350	1 225
26	200	25	5	−20	−100	400
9	250	144	−12	30	−360	900
15	241	36	−6	21	−126	441
24	220	9	3	0	0	0
30	196	81	9	−24	−216	576
315	**3 300**	**2 436**	**DO NOT ADD**		**−7 307**	**22 950**

Feedback → Feedback

(a) Yes, both variables are scale variables. The one is discrete and the other one is continuous.

(b) No, we did not select a fixed number of values for a control variable (years of driving experience) to observe repeated values of the response variable (insurance premium) at each of these values. Hence this is really a *passive, noncausal regression study*.

(c)



From the scatter diagram we can see that linear regression is applicable.

(d) To calculate the regression equation we use the totals supplied in the table:

$$\widehat{\beta}_1 = \frac{\sum (X-\overline{X})(Y-\overline{Y})}{\sum (X-\overline{X})^2}$$

$$= \frac{-7307}{2436} = -3.0$$

and $\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X}$

$$= \frac{3300}{15} - (-3.0)\frac{315}{15}$$

$$= 220 + 63$$

$$= 283\,.$$

So, the estimated regression equation is

$$\widehat{Y} = 283 - 3X\,.$$

(e) $S = \sqrt{\dfrac{\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2}{n-2}}$ or $\sqrt{\dfrac{\sum_{i=1}^{n}(Y_i-\overline{Y})^2 - \widehat{\beta}_1^2\sum_{i=1}^{n}(X_i-\overline{X})^2}{(n-2)}}$

$$= \sqrt{\frac{22950 - (-3)^2(2436)}{13}}$$

$$= \sqrt{\frac{1026}{13}} = \sqrt{78.9231}$$

$$= 8.8839$$

Using the value of S^2, we may find expressions for the estimated variances of $\widehat{\beta}_0$ and $\widehat{\beta}_1$:

$$\widehat{Var(\widehat{\beta}_0)} = S^2 \left[\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2} \right] = 78.9231 \left(\frac{1}{15} + \frac{(21)^2}{2436} \right) = 19.5493 = S^2_{\widehat{\beta}_0}$$

$\sqrt{19.5493} = 4.421\,5$ is called the **standard error** of $\widehat{\beta}_0 \Longrightarrow SE(\widehat{\beta}_0) = 4.421\,5$

and

$$\widehat{Var(\widehat{\beta}_1)} = \frac{S^2}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2} = \frac{78.9231}{2436} = 0.0324 = S^2_{\widehat{\beta}_1} .$$

$\sqrt{0.0324} = 0.18$ is called the **standard error** of $\widehat{\beta}_1 \Longrightarrow SE(\widehat{\beta}_1) = 0.18.$

(f) A 95% confidence interval for $\widehat{\beta}_0$ is given by

$$\begin{aligned} & \widehat{\beta}_0 \pm t_{0.025;\,13} SE(\widehat{\beta}_0) \\ = \;& 283 \pm (2.160)\,(4.421\,5) \\ = \;& 283 \pm 9.550\,4 \\ = \;& (273.4496; 292.5504)\,. \end{aligned}$$

(g) A 95% confidence interval for $\widehat{\beta}_1$ is given by

$$\begin{aligned} & \widehat{\beta}_1 \pm t_{0.025;\,13} SE(\widehat{\beta}_1) \\ = \;& -3 \pm (2.160)\,(0.18) \\ = \;& -3 \pm 0.3888 \\ = \;& (-3.3888; -2.6112)\,. \end{aligned}$$

What about JMP?

Before you work through the corresponding section of the textbook, you need to understand all the possible outputs.

The typical computer output (for most statistical packages) will include an ANOVA table (almost exactly the same as we obtained in the last section of the previous study unit where we tested for the difference between means).

Source of variation Source	Degrees of freedom DF	Sum of Squares SS	Mean Square MS	F-Ratio F
Model	$(r-1)$	$\widehat{\beta}_1 \sum \left(X-\overline{X}\right)^2$	MSR	$F=\frac{MSR}{MSE}$
Error	$(n-r)$	$\sum \left(Y_i-\widehat{Y}_i\right)^2$	MSE	
C. Total	$(n-1)$	$\sum_{i=1}^{n} \left(Y_i-\widehat{Y}_i\right)^2$		

This output actually measures "how good" the regression is! If the regression equation fits the data perfectly (that is if $R=\pm 1)$, the predicted value, $\widehat{Y}_i=\widehat{\beta}_0+\widehat{\beta}_1 X_i$, will be exactly equal to the observed value Y_i. This implies that

$$S^2=\frac{\sum \left(Y_i-\ \widehat{Y}_i\right)^2}{n-2}=0\,.$$

So, can we use S as a measure of how good the regression is? We cannot use S^2 as such because it depends on the scale of Y_i (the dependent variable) and will be expressed *in the same units*. It needs to be standardised in a special manner. From activity 8.6 we know that

$$\sum_{i=1}^{n}\left(Y_i-\widehat{Y}_i\right)^2=\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2-\widehat{\beta}_1^2\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2 \qquad (1)$$

which means that

$$\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2=\sum_{i=1}^{n}\left(Y_i-\widehat{Y}_i\right)^2+\widehat{\beta}_1^2\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2\,. \qquad (2)$$

But in the second equation, the term on the left of the expression is the total variation in Y, in other words $\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2=$ total variation in Y. Equation 2 is the classical breakdown of the total variation in the Y-values into two other variations:

SST	=	SSE	$+$	SSR
↓		↓		↓
Total variation		sum of squares for error		sum of squares for regression

For perfect regression $SSE=0 \Longrightarrow SST=0+SSR$ which means that $SST=SSR$.

Thinking in terms of the F-statistic as we did in the ANOVA situation (see study unit 7), this would also mean that $\frac{SSR}{SST}=1$. So,

$$\frac{SSR}{SST} = \frac{\widehat{\beta}_1^2 \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2}$$

is the **proportion of the total variation in** Y **that is explained by the regression equation** which measures how "good" the regression is.

It can also be shown (with a little algebraic manipulation) that

$$\frac{\widehat{\beta}_1^2 \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2} = \frac{SSR}{SST} = R^2.$$

R^2 is called the **coefficient of determination** and measures how good the regression is, in other words it is a measure of goodness of fit whilst R (without the square) measures how well the linear relationship between X and Y is.

Another important spin-off from the classical breakdown of variance, that is

$$SST = \underbrace{SSE}_{\substack{\downarrow \\ \text{sum of squares} \\ \text{for error}}} + \underbrace{SSR}_{\substack{\downarrow \\ \text{sum of squares} \\ \text{for regression}}}$$

is to create an F-test where the ratio $\frac{SSR/df}{SSE/df} = \frac{MSR}{MSE}$ will have an F-distribution which can be used to test whether the regression model (in this application *the regression line*) accounts for a significant portion of the variation in the Y-variable (this is similar to theorem 7.7 where we derived $F = \frac{MSTreatment}{MSError} \sim F_{\upsilon_1;\, \upsilon_2}$).

Using this approach as a hypothesis test for the overall regression model (that includes both $\widehat{\beta}_0$ and $\widehat{\beta}_1$) the typical computer output will start off with an ANOVA table almost exactly the same as in section 7.6.

Source of variation Source	Degrees of freedom df	Sum of Squares SS	Mean Sum of Squares MS	F-value F
Regression	$(r-1)$	$\widehat{\beta}_1^2 \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$	MSR	$F = \frac{MSR}{MSE}$
Error	$(n-r)$	$\sum_{i=1}^{n} \left(Y_i - \widehat{Y}_i\right)^2$	MSE	
Total	$(n-1)$			

 **YOU WILL NOT BE EXAMINED ON THE THEORETICAL EXPLANATION OF THE ANOVA TABLE OF THE REGRESSION OUTPUT. You will only be examined on the theory that is covered in the study guide.**

WORK THROUGH

Sall, Creighton and Lehman, **chapter 10 Fitting Curves Through Points: Regression**

Start reading on page 235 "*Overview*" and read up to (including) "*Confidence Intervals on the Estimates*".

Please note that you can try on your own to open and run *demoLeastSquares.jsl* as explained in the textbook on page 237 (with a reference to page 57) doing the interactive animation exercise "**Seeing Least Squares**". This is optional and for fun! However, you should be able to do the proper regression analysis which is explained under "**Fitting a Line and Testing the Slope**".

 Even without having to laboriously compute the calculation table in activity 8.7, producing the answers still required a lot of effort and thus you will appreciate what a computer can do in the blink of an eye in the next activity !

Activity 8.8

Refer to **activity 8.7**.

(a) Create a JMP data table that will represent the paired observations.

(b) Perform a regression analysis and interpret your output.

(c) Comment on the similarities (and the differences) between the manual test and the electronic version of JMP.

Feedback → Feedback

(a) I have called the data set *act8p7.jmp* and named the first column *Experience* and the second column *Premium.*

(b) [*For beginners:* How to *fit a regression line.*

Select **Analyze** ⟶ **Fit Y by X.** (Yes we have done this many times before!)

The problem stated that *Experience* influences *Premium* which means we assume that "independent variable = X, Factor = *Experience*" and that "dependent variable = Y, Response = *Premium*". Convey this choice to JMP.

If you click on the **OK** button you immediately get the "**Bivariate Fit of Premium By Experience"** output which is a *scatter plot*.

Click on the **Bivariate Fit of Premium By Experience** bar (there is a red triangle where you have to click) and open the submenu

Show Points
Fit Mean
Fit Line
Fit Polynomial

↓
Script

where you select **Fit Line.**]

This results in the following output:

Linear Fit

Premium = 282.99138 - 2.9995895 Experience

Summary of Fit

RSquare	0.955033
RSquare Adj	0.951574
Root Mean Square Error	8.909803
Mean of Response	220
Observations (or Sum Wgts)	15

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	21918.000	21918.0	276.0990
Error	13	1032.000	79.4	Prob > F
C. Total	14	22950.000		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	282.99138	4.434372	63.82	<.0001*
Experience	-2.999589	0.180522	-16.62	<.0001*

From the **Linear Fit** we conclude $Y = 282.99138 - 2.9995895(X)$ which we round to $\widehat{Y} = 283 - 3X$ (the same as for the manual computations).

From the first line of the **Parameter Estimates** report we conclude that $\widehat{\beta}_0 = 282.99138$ and the **standard error** of $\widehat{\beta}_0$, which we denote as $SE(\widehat{\beta}_0) = 4.434372$.

To test $H_0 : \beta_0 = 0$ the computed value of the test statistic

$$T_0 = \frac{\widehat{\beta}_0 - \beta_0}{\sqrt{S^2 \left[\frac{1}{n} + \frac{\overline{X}^2}{\sum\limits_{i=1}^{n} \left(X_i - \overline{X}\right)^2} \right]}} = 63.82 \Longrightarrow \text{reject } H_0.$$

(We did not compute T_0 manually in activity 8.7 and you can cross-check if you get $T_0 = 63.82$.)

In the last line from the **Parameter Estimates** report we conclude that $\widehat{\beta}_1 = -2.999589$ and the **standard error** of $\widehat{\beta}_1$, which we denote as $SE(\widehat{\beta}_1) = 0.180522$.

To test $H_0 : \beta_1 = 0$ the computed value of the test statistic $T_1 = \dfrac{\widehat{\beta}_1 - \beta_1}{\sqrt{\dfrac{S^2}{\sum\limits_{i=1}^{n} \left(X_i - \overline{X}\right)^2}}} = -16.62 \Longrightarrow$ reject H_0.

(We did not compute T_1 manually in activity 8.7 and you can cross-check if you get $T_1 = -16.62.$)

The default **Parameter Estimates** report does not show the *confidence intervals* for the estimated parameters but they can be activated by doing the following:

Put your mouse pointer inside the **Parameter Estimates** report block on *Intercept* and right-click to open the submenu Table Style . Columns ⋮ ⋮ Make Into Matrix.

Select **Columns** and $\surd$ tick **Lower 95%** and **Upper 95%**. This action will produce the output shown below.

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%
Intercept	282.99138	4.434372	63.82	<.0001*	273.4115	292.57126
Experience	-2.999589	0.180522	-16.62	<.0001*	-3.389583	-2.609596

From this output a 95% confidence interval for $\widehat{\beta}_0$ is given by $(273.4115; 292.57126)$ (we manually obtained $(273.4496; 292.5504)$).

From this output a 95% confidence interval for $\widehat{\beta}_1$ is given by $(-3.389583; -2.609596)$ (we manually obtained $(-3.3888; -2.6112)$).

 It is a good feeling to be able to interpret **any electronic regression analysis output** even if you are not familiar with the particular statistical package. Redoing the **regression analysis** for the data of activity 8.7 by **using the statistical package SPSS,** the computer output includes the following ANOVA table:

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	21918.000	1	21918.000	276.099	.000[a]
	Residual	1032.000	13	79.385		
	Total	22950.000	14			

a. Predictors: (Constant), Driving experience

b. Dependent Variable: Monthly car insurance

SPSS also presents the other computations in the following manner:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	282.991	4.434		63.818	.000	273.412	292.571
	Driving experience	-3.000	.181	-.977	-16.616	.000	-3.390	-2.610

a. Dependent Variable: Monthly car insurance

 Does every little bit of output make sense to you? If it does – you have mastered simple linear regression.

Please also note the following:

- In more advanced regression analysis, *two or more* independent variables are used to predict the dependent variable. We call this multiple regression analysis. You will learn more about this in advanced modules.
- When we began our discussion of correlation in study unit 5 we emphasised that we must have paired observations of *interval-scaled measurements*. However, for the true and ideal causal regression model it implies that the values of the independent variable X are fixed and can be predetermined. This implies that the dependent variable must be continuous and of interval scale and that Y varies with the same amount (measured by σ) when X is either a large or a small value. When the *variation is the same for all x-values*, we use the term *homoscedasticity.* Many computer packages provide diagnostic tools to check the validity of homoscedasticity. We only touched on this issue in activity 8.2 and it belongs to more advanced modules.

8.5 Inference on the regression line

Activity 8.9

Refer to example 8.2(c) of the study guide. Compute the confidence limits for a **future observation** for the data set of example 8.2.

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

Feedback → Feedback

We have to compute $(\widehat{\beta}_0 + \widehat{\beta}_1 X_i) \pm t_{\alpha/2;(n-2)}\ S\sqrt{1 + \frac{1}{n} + \frac{(X_i - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}}$ for the different values of X.

For this application $t_{\alpha/2;(n-2)} = 2.228$.

X	$\widehat{\beta}_0 + \widehat{\beta}_1 X$	$3\sqrt{1 + \frac{1}{12} + \frac{(X-3)^2}{60}}$	Lower limit	Upper limit
0	15	3.331 7	7.577	22.423
1	18	3.217 1	10.8323	25.1677
2	21	3.146 4	13.9898	28.0102
3	24	3.122 5	17.0431	30.9569
4	27	3.146 4	19.9898	34.0102
5	30	3.217 1	22.8323	37.1677
6	33	3.331 7	25.577	40.423

Activity 8.10

(a) Copy the data of example 8.2(c) of the study guide as a JMP data table.

(b) Fit the regression line.

(c) Draw the confidence bands around the line. (In other words both the ***confidence limits for a future observation*** and the narrower ***confidence limits for the regression line.)***

Feedback → Feedback

(a) I have called the data set *act8p10.jmp* and named the first column *Fertilizer* and the second column *Yield.*

(b) [*For beginners:* How to *fit a regression line.*

Select **Analyze** ⟶ **Fit Y by X.** (Yes we have done this many times before!)

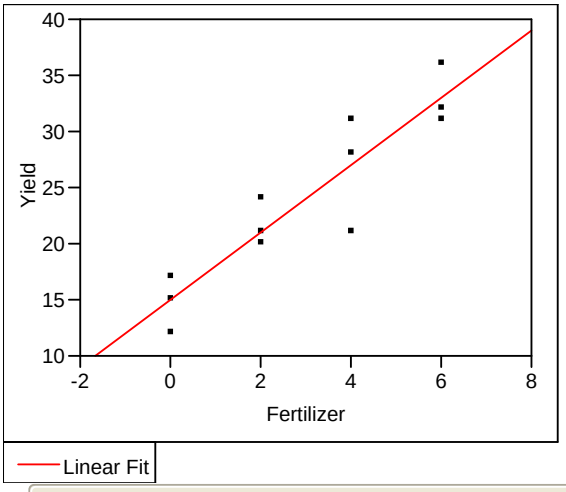
Convey to JMP "independent variable = X, Factor = *Fertilizer*" and that "dependent variable = Y, Response = *Yield*".

Click on the **Bivariate Fit of Yield By Fertilizer** bar (there is a red triangle where you have to click) and open the submenu

Show Points
Fit Mean
Fit Line
Fit Polynomial

↓
Script

where you select **Fit Line.**]



Linear Fit

Yield = 15 + 3 Fertilizer

Summary of Fit

RSquare	0.857143
RSquare Adj	0.842857
Root Mean Square Error	3
Mean of Response	24
Observations (or Sum Wgts)	12

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	540.00000	540.000	60.0000
Error	10	90.00000	9.000	Prob > F
C. Total	11	630.00000		<.0001*

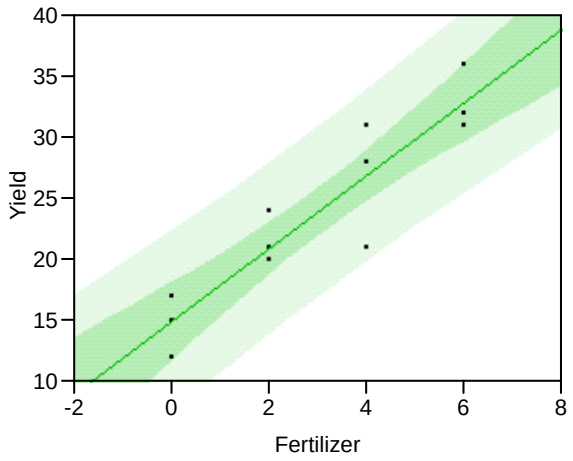
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	15	1.449138	10.35	<.0001*
Fertilizer	3	0.387298	7.75	<.0001*

Click on the red triangle in the **Linear Fit** bar in the output and now you have a choice between "**Curves**" and "**Shaded**" for your confidence bands. I prefer the shaded area for a better picture and thus I have selected

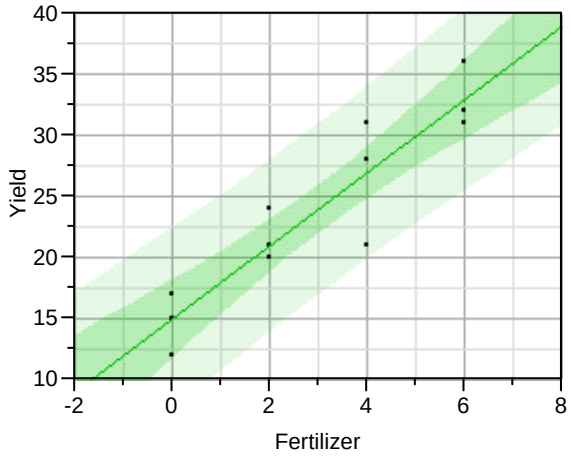
√ **Line of Fit**
Confid Curves Fit
Confid Curves Indiv
⋮
⋮
√ **Confid Shaded Fit**
√ **Confid Shaded Indiv**
Remove Fit.

The wider band is for "**Confid Shaded Indiv**" which is our "***confidence limits for a future observation***" and the narrower band is for "**Confid Shaded Fit**" which is our "***confidence limits for the regression line***".



In order to compare the manually computed values with the computer output you can add gridlines to your graph for ease of reading off the values.

For example, if $X = 2$ we read from the graph that the confidence limits for a future observation is ± 14 up to ± 28 (which is of course what we computed in activity 8.9 as $(13.9898; 28.0102)$) whilst we read from the graph that the confidence limits for the regression line if $X = 2$ is ± 19 up to ± 23 (which we have to compare with the tabulated interval computed in example 8.2(b) in the study guide, we obtained $(18.89; 23.11)$).



8.6 Relationship between tests for correlation and regression

Activity 8.11

In this section of the study guide, in other words section 8.6, we derive that $(n-2)S^2 = S_{22} - \frac{S_{12}^2}{S_{11}}$.

Use the definition of S_{22}, S_{12} and S_{11} as given in the study guide and show that the alternative computational formula for S^2 which was derived in activity 8.6 equals this expression.

(You have to show that $(n-2)S^2 = S_{22} - \frac{S_{12}^2}{S_{11}} = \sum_{i=1}^{n}[Y_i - \overline{Y}]^2 - \widehat{\beta}_1^2 \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$.)

Feedback → Feedback

In activity 8.6 we proved that $(n-2)S^2 = \sum_{i=1}^{n}[Y_i - \overline{Y}]^2 - \widehat{\beta}_1^2 \sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$.

If we now substitute the definition of $\widehat{\beta}_1$ into this equation we get

$$\begin{aligned}
(n-2)S^2 &= \sum_{i=1}^{n}[Y_i - \overline{Y}]^2 - \widehat{\beta}_1^2 \sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \\
&= \sum_{i=1}^{n}[Y_i - \overline{Y}]^2 - \left[\frac{\sum_{i=1}^{n}[Y_i - \overline{Y}][X_i - \overline{X}]}{\sum_{i=1}^{n}[X_i - \overline{X}]^2}\right]^2 \sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \\
&= \sum_{i=1}^{n}[Y_i - \overline{Y}]^2 - \frac{\left[\sum_{i=1}^{n}[Y_i - \overline{Y}][X_i - \overline{X}]\right]^2}{\sum_{i=1}^{n}[X_i - \overline{X}]^2} \\
&= S_{22} - \frac{S_{12}^2}{S_{11}}.
\end{aligned}$$

8.7 Simple linear regression in matrix notation

You will not be examined on this section but it is very important when you intend to continue with statistics at third-year level (or higher up) to grasp the matrix notation of linear models.



It is time to draw yourself a mind map of this study unit!

STUDY UNIT 9

Typical examination questions

9.1 Worksheet

QUESTION 1

Are the following statements true or false? Give the correct statement if the given statement is false.

(a) If $X_1,\ X_2,\ ...,\ X_{10}$ is a random sample from a $n\left(10;\ 9\right)$ distribution, then $Var\left(X_{10}\right) = 3.$

(b) $E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E\left[X_i\right]$

(c) $Var\left(aX\right) = aVar\left(X\right)$

(d) The maximum likelihood estimator of a parameter θ is always unbiased.

(e) The probability $P\left(H_1 \text{ accepted} | H_1 \text{ is true}\right) = 1 - \beta$ is called the exceedance probability.

(f) If the correlation coefficient between X and Y, $\rho = 0$, then X and Y are always independent.

QUESTION 2

Assume that the pdf of a $n\left(\mu;\ \sigma^2\right)$ variate is given by

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2/\sigma^2}.$$

Let $X_1,\ X_2,\ ...,\ X_n$ be a random sample from a $n\left(0;\ \sigma^2\right)$ distribution.

(a) Show that the maximum likelihood estimator for σ^2 is given by $T = \sum_{i=1}^{n} X_i^2/n$.

(b) Show that T is an unbiased estimator for σ^2.

QUESTION 3

The most used chemical method of determining the iron status of the body, is to measure the ferritin in a blood sample of the patient. A physician took a random sample of 42 presumably healthy females between the ages of 20 and 35 years, took a blood sample from each and found their ferritin (X_i) measurements to be the following:

(Note that the sample values are arranged in order of magnitude to ease classification into intervals.)

10	15	17	21	27	30	35	37	40	42	45	46	47
50	51	54	55	57	57	58	59	59	61	61	61	62
64	65	66	70	72	75	80	82	83	86	90	94	96
105	110	125										

You may make use of any of the following calculations:

$$\sum_{i=1}^{42} X_i = 2\,520; \qquad \sum_{i=1}^{42} X_i^2 = 178\,562; \qquad \sum_{i=1}^{42} \left(X_i - \bar{X}\right)^2 = 27\,362$$

Consider the **following six equiprobable** class intervals:

$$\begin{aligned} X &\leq 35.34 \\ 35.34 &\leq X \leq 48.97 \\ 48.97 &\leq X \leq 60.00 \\ 60.00 &\leq X \leq 71.03 \\ 71.03 &\leq X \leq 84.66 \\ X &\geq 84.66 \end{aligned}$$

(a) **Clearly show** how this table can be derived under the assumption of normality. **You need only do it for one interval: THE FOURTH INTERVAL** $(60.00 - 71.03)$.

(b) Use the given six equiprobable intervals to classify the sample values.

(c) Use a **chi-square goodness-of-fit test** to test at the 5% level of significance whether the ferritin measurements have a normal distribution.

(d) Calculate a **two-sided 95% confidence interval for** μ_X (the population **mean**) where X represents the ferritin measurement of females between the ages of 20 and 35 years.

(e) Derive and calculate a two-sided 95% confidence interval for σ_X^2 (the population **variance**) where X represents the ferritin measurement of females between the ages of 20 and 35 years.

(f) Suppose that the same physician repeats this whole "experiment" with an independent second sample of 30 presumably healthy females above 40 years of age and finds that

$$\sum_{i=1}^{30} Y_i = 1\,905 \qquad \text{and} \qquad \sum_{i=1}^{30} \left(Y_i - \bar{Y}\right)^2 = 20\,365.25.$$

[Y_i represents the ferritin measurement of the i-th patient of the second sample.]

(i) Test the hypothesis $H_0 : \mu_X = \mu_Y$ against
$H_1 : \mu_X < \mu_Y$

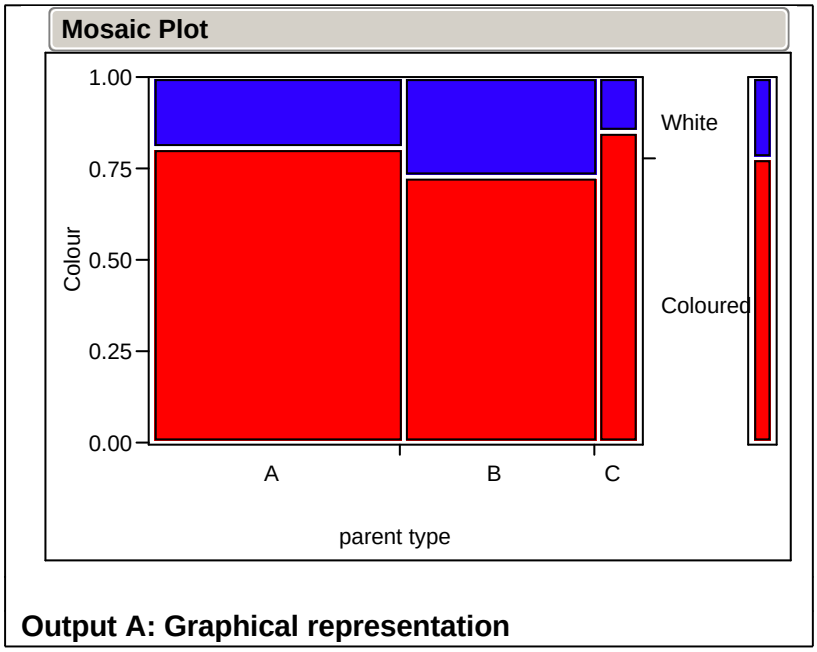
at the 1% level of significance.

Can you conclude that older women have a larger mean ferritin score than females in the age group 20-35 years?

(ii) Under what assumptions is the test statistic that you use valid?

QUESTION 4

Breeding experiments were carried out on Rhode Island Red Fowl in a study of colour mutation. The number of chicks with **white down** (where down is defined as the first covering of a young bird) versus the number of chicks with **coloured down,** was observed for three different classification types of Rhode Island Red Fowl parents. It was observed that for type A parents, there were 50 chicks with white down versus 210 chicks with coloured down. For type B parents, there were 54 chicks with white down versus 146 chicks with coloured down and for type C parents, there were six chicks with white down versus 34 chicks with coloured down.
The breeder wants to test whether an association exists between the type of parent and the colour of the down.



Output A: Graphical representation

Contingency Table

		Colour		
	Count Total % Col % Row %	Coloured	White	
parent type	A	210 42.00 53.85 80.77	50 10.00 45.45 19.23	260 52.00
	B	146 29.20 37.44 73.00	54 10.80 49.09 27.00	200 40.00
	C	34 6.80 8.72 85.00	6 1.20 5.45 15.00	40 8.00
		390 78.00	110 22.00	500

Output B: Contingency table

Tests

Source	DF	-LogLike	RSquare (U)
Model	2	2.61036	0.0099
Error	497	260.84362	
C. Total	499	263.45398	
N	500		

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	5.221	0.0735
Pearson	5.218	0.0736

Output C: Computations

Do these results indicate a significant association between the type of parent and the colour of the down of the chicks?

Substantiate your answer by paying attention to the following detail:

(a) How do you interpret the **Mosaic Plot of Output A**?

(b) State the appropriate null and alternative hypotheses for this test.

(c) What test statistic is used to test these hypotheses and what is the value of the test statistic?

(d) Looking at the row percentages in Output B, can you draw any conclusions?

(e) What is your final conclusion?

QUESTION 5

The correlation coefficient between a psychological test score and academic performance for a random sample of 147 first-year students is 0.40. The corresponding coefficient for a random sample of 125 second-year students is 0.59. Do a hypothesis test to see whether there is a significant difference between these two coefficients. Use $\alpha = 0.05$ and assume bivariate normality.

QUESTION 6

Four teaching methods were applied to four comparable classes of students. At the end of the course the same test was applied to each class. The following results were obtained:

I: $\overline{X}_1 = 6.2 \quad S_1^2 = 4.7 \quad n = 5$

II: $\overline{X}_2 = 8.4 \quad S_2^2 = 2.8 \quad n = 5$

III: $\overline{X}_3 = 7.0 \quad S_3^2 = 2.5 \quad n = 5$

IV: $\overline{X}_4 = 2.8 \quad S_4^2 = 3.2 \quad n = 5$

Consider the data as random samples from $n\left(\mu_i;\ \sigma_i^2\right)$ populations.

Test $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$ against

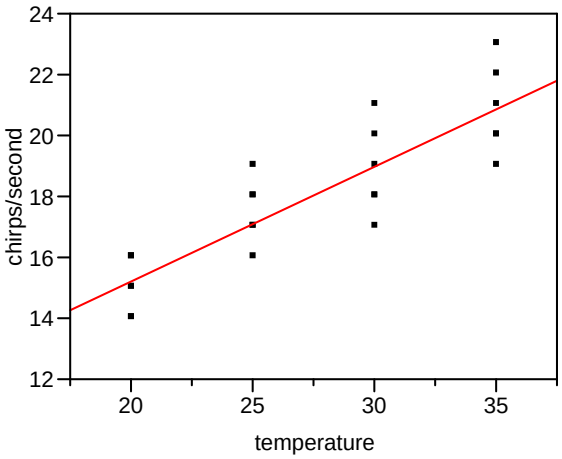
$H_1 : \sigma_p^2 \neq \sigma_q^2$ for at least one $p \neq q$.

Use $\alpha = 0.05$.

QUESTION 7

An entomologist is studying the chirping behaviour of the common brown ground cricket. He counted the chirps per second (response variable, Y) at four fixed temperature (^{0}C) settings (control variable, X) and reported the data which were captured in the following table:

temperature	chirps/second
20	16
20	16
20	15
20	14
20	14
20	15
25	16
25	17
25	18
25	19
25	18
25	17
30	19
30	18
30	20
30	21
30	18
30	17
35	21
35	20
35	22
35	23
35	19
35	20



Output A: Graphical representation

Linear Fit

chirps/second = 7.6833333 + 0.3766667 temperature

Summary of Fit

RSquare	0.765757
RSquare Adj	0.75511
Root Mean Square Error	1.216366
Mean of Response	18.04167
Observations (or Sum Wgts)	24

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	106.40833	106.408	71.9196
Error	22	32.55000	1.480	Prob > F
C. Total	23	138.95833		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	7.6833333	1.246404	6.16	<.0001*
temperature	0.3766667	0.044415	8.48	<.0001*

Output B: Computations

The entomologist wants to predict the number of chirps/second from the temperature.
Use the output of the JMP analysis and answer the following questions: (**Substantiate your answer where necessary.**)

(a) Would you say that simple linear regression is a suitable model?

(b) What is the equation of the regression line?

(c) What is the expected number of chirps/second if the temperature is 30^0C? How could you compute this value in a totally different way (in other words by not making use of the regression line)?

(d) What test statistic is used to test the null hypothesis that the **intercept is zero** and what is the value of the test statistic? Is this test significant at the 1% level of significance?

(e) What test statistic is used to test the null hypothesis: $H_0 : \beta_1 = 0$ and what is the value of the test statistic? Is this test significant at the 1% level of significance?

(f) What is the value of $S^2 = \dfrac{\Sigma \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)^2}{(n-2)}$?

QUESTION 8

State whether each of the following statements is TRUE or FALSE. If the statement is FALSE, write down the correct statement (that is, show how you change the statement to make it a TRUE statement):

(a) Any function $T = T\left(X_1,\ X_2,\ ...,\ X_n\right)$ of a random sample $X_1,\ X_2,\ ...,\ X_n$ is called a statistic if it can be computed without using unknown parameters.

(b) The most efficient estimator of two or more unbiased estimators is that estimator obtained by least squares estimation.

(c) You cannot determine a maximum likelihood estimator if you do not know the formula for the probability density function of the specific distribution.

(d) The significance level of a test is usually denoted by $\beta = P\left(H_0 \text{ is rejected} \mid H_0 \text{ is true}\right)$.

(e) A type II error is committed if we accept H_0 when H_0 is in fact false.

(f) The power of a test is found by computing $1 - P\left(\text{type I error}\right)$.

(g) If we want to perform k simultaneous inference tests, with the overall significance level equal to at most α, each test is performed at level of significance α.

QUESTION 9

Let $X_1,\ X_2,\ ...,\ X_n$ be a random sample from a distribution with the following probability density function:

$$\begin{aligned} P(X = x) &= \theta(1-\theta)^{x-1} \text{ for } x > 1 \\ &= 0 \text{ elsewhere.} \end{aligned}$$

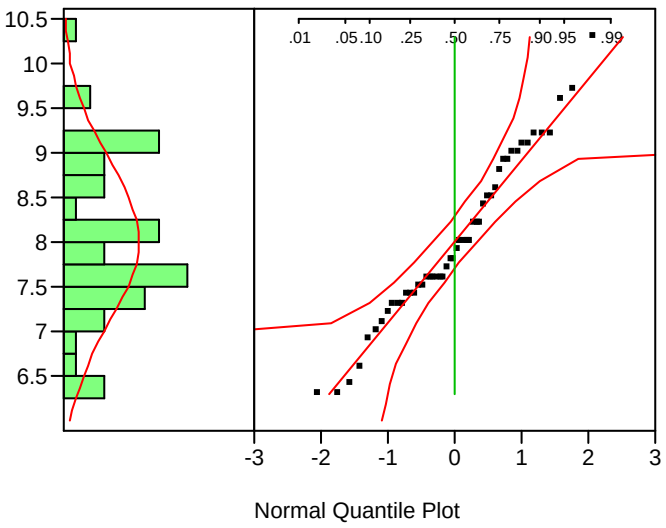
Find the maximum likelihood estimator for θ.

QUESTION 10

A manufacturer of rubber gloves (for household purposes) selected a random sample of 50 females and measured the palm width of the right hand of each female.

The table below shows the 50 observations (in the first column) with five extra columns of calculations where **the last row shows the total of each column**:

X_i	X_i^2	X_i^3	$(X_i-\overline{X})$	$(X_i-\overline{X})^3$	$(X_i-\overline{X})^4$
palmsize	X**2	X**3	X-mean	(X-mean)**3	(X-mean)**4
7.6	57.76	438.976	-0.406	-0.06692342	0.027170907
8.5	72.25	614.125	0.494	0.120553784	0.059553569
7.1	50.41	357.911	-0.906	-0.74367742	0.673771739
7.7	59.29	456.533	-0.306	-0.02865262	0.0087677
7.8	60.84	474.552	-0.206	-0.00874182	0.001800814
7.6	57.76	438.976	-0.406	-0.06692342	0.027170907
8.9	79.21	704.969	0.894	0.714516984	0.638778184
7.5	56.25	421.875	-0.506	-0.12955422	0.065554433
8	64	512	-0.006	-2.16E-07	1.296E-09
9.2	84.64	778.688	1.194	1.702209384	2.032438004
10.3	106.09	1092.727	2.294	12.07202818	27.69323265
9.2	84.64	778.688	1.194	1.702209384	2.032438004
8	64	512	-0.006	-2.16E-07	1.296E-09
8.8	77.44	681.472	0.794	0.500566184	0.39744955
7.6	57.76	438.976	-0.406	-0.06692342	0.027170907
8.2	67.24	551.368	0.194	0.007301384	0.001416468
9.1	82.81	753.571	1.094	1.309338584	1.432416411
6.3	39.69	250.047	-1.706	-4.96520382	8.47063771
7.3	53.29	389.017	-0.706	-0.35189582	0.248438446
6.9	47.61	328.509	-1.106	-1.35289902	1.496306312
8.4	70.56	592.704	0.394	0.061162984	0.024098216
9	81	729	0.994	0.982107784	0.976215137
7	49	343	-1.006	-1.01810822	1.024216865
6.3	39.69	250.047	-1.706	-4.96520382	8.47063771
9	81	729	0.994	0.982107784	0.976215137
8.9	79.21	704.969	0.894	0.714516984	0.638778184
9.1	82.81	753.571	1.094	1.309338584	1.432416411
7.4	54.76	405.224	-0.606	-0.22254502	0.13486228
7.6	57.76	438.976	-0.406	-0.06692342	0.027170907
7.8	60.84	474.552	-0.206	-0.00874182	0.001800814
9.7	94.09	912.673	1.694	4.861163384	8.234810772
8.2	67.24	551.368	0.194	0.007301384	0.001416468
7.3	53.29	389.017	-0.706	-0.35189582	0.248438446
7.2	51.84	373.248	-0.806	-0.52360662	0.422026932
7.6	57.76	438.976	-0.406	-0.06692342	0.027170907
7.9	62.41	493.039	-0.106	-0.00119102	0.000126248
8.6	73.96	636.056	0.594	0.209584584	0.124493243
9.2	84.64	778.688	1.194	1.702209384	2.032438004
8.2	67.24	551.368	0.194	0.007301384	0.001416468
8	64	512	-0.006	-2.16E-07	1.296E-09
7.6	57.76	438.976	-0.406	-0.06692342	0.027170907
7.4	54.76	405.224	-0.606	-0.22254502	0.13486228
7.5	56.25	421.875	-0.506	-0.12955422	0.065554433
7.3	53.29	389.017	-0.706	-0.35189582	0.248438446
8	64	512	-0.006	-2.16E-07	1.296E-09
6.4	40.96	262.144	-1.606	-4.14225302	6.652458344
7.4	54.76	405.224	-0.606	-0.22254502	0.13486228
8.5	72.25	614.125	0.494	0.120553784	0.059553569
9.6	92.16	884.736	1.594	4.050092584	6.455847579
6.6	43.56	287.496	-1.406	-2.77943142	3.907880571
400.3	3245.83	26653.273		10.2144816	87.81989029



Output A: Graphical representation of hand palm sizes

Moments

Mean	8.006
Std Dev	0.9150466
Std Err Mean	0.1294071
upper 95% Mean	8.2660534
lower 95% Mean	7.7459466
N	50

Fitted Normal

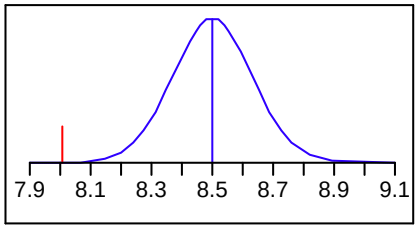
Parameter Estimates

Type	Parameter	Estimate	Lower 95%	Upper 95%
Location	μ	8.006	7.7459466	8.2660534
Dispersion	s	0.9150466	0.7643695	1.14027

Test Mean=value

Hypothesized Value	8.5
Actual Estimate	8.006
df	49
Std Dev	0.91505

	t Test
Test Statistic	-3.8174
Prob > |t|	0.0004*
Prob > t	0.9998
Prob < t	0.0002*



Output B: Moments and t-test of hand palm sizes

(a) What can you conclude about normality from the **JMP graphical output in A**?

(b) Use the **method of moments to test for normality.** (Use any of the given calculations, but **clearly show all your computations**. Perform both tests two-sided at the 10% level of significance.)

(c) Is there any reason to reject the null hypothesis $H_0 : \mu = 8.5$? Test two-sided at the 5% level of significance. Show how you use and interpret **the JMP output in B**.

(d) Give a 95% confidence interval for the unknown mean μ. Does this interval confirm your conclusion in question (c)?

(e) Show how to manually compute a 95% confidence interval for the unknown **variance** of the population. Does this confirm the 95% confidence interval for the unknown **standard deviation** of the population which is given as $(0.7643695; 1.14027)$ in the JMP output B? [Hint: You need to interpolate the critical values for a fair comparison.]

QUESTION 11

A soft drink company sells two kinds of cool drink: regular and sugar-free. A random sample consisting of six males and six females were asked what their cool drink preference was. Four males preferred the regular cool drink whilst only two females preferred the regular cool drink. The rest of them preferred the sugar-free cool drink. Is it true that females on average **drink significantly more sugar-free cool drink than men**? Test at the 5% level of significance and show how you use an exact test.

QUESTION 12

Ten judges of a beauty pageant have to rate the contestants according to looks and personality. A statistician has the feeling that the rating of looks could influence the rating of personality and that there could be a positive correlation between the two ratings. She decided to analyse the ratings for one particular contestant and obtained the following:

Looks X_i	2.5	3.0	5.0	4.0	3.5	2.0	3.3	5.2	2.8	5.8
Personality Y_i	5.0	4.7	3.0	4.2	4.5	3.0	5.9	5.5	3.2	6.1

(a) Compute R, the correlation coefficient between the ratings of looks and personality.

(b) Test $H_0 : \rho = 0$ against $H_1 : \rho > 0$ at the 5% level of significance.

QUESTION 13

Two independent random samples from $n\left(\mu_1;\ \sigma_1^2\right)$ and $n\left(\mu_2;\ \sigma_2^2\right)$ distributions respectively, yielded the following statistics:

Sample 1: $n_1 = 10$ $\qquad \sum_{i=1}^{10} X_{1i} = 190$ $\qquad \sum_{i=1}^{10} X_{1i}^2 = 5\,374$

Sample 2: $n_2 = 16$ $\qquad \sum_{i=1}^{16} X_{2i} = 408$ $\qquad \sum_{i=1}^{16} X_{2i}^2 = 12\,564$

Test $H_0 : \sigma_1^2 = \sigma_2^2$ against

$H_1 : \sigma_1^2 > \sigma_2^2$ (5% level of significance).

QUESTION 14

A random sample $X_1,\ X_2,\ ...,\ X_{16}$ from a normal population yielded the following statistics:

$$\sum_{i=1}^{16} X_i \ = \ 96 \quad ; \qquad \sum_{i=1}^{16} X_i^2 \ = \ 4\,500.$$

(a) Derive a *95% confidence interval* for the unknown mean μ if you assume that $\sigma = 15.$

(b) Derive a *95% confidence interval* for the unknown mean μ if you assume that σ is unknown.

QUESTION 15

In order to test the usefulness of various fuel mixtures, the following experiment was performed: Twenty identical cars were randomly divided into four groups consisting of five cars each. Group 1 used petrol as fuel, while the other three groups used fuel mixtures A, B and C, respectively. The test consisted of measuring (under identical conditions) how far each car could travel on five litres of fuel. The results were as follows:

	Distance in km					$\overline{x}_i$	$\sum_{j=1}^{5}(x_{ij}-\overline{x}_i)^2$
Group 1 (Petrol)	55	55	54	55	56	$\overline{x}_1 = 55$	2
Group 2 (A)	41	39	40	39	41	$\overline{x}_2 = 40$	4
Group 3 (B)	44	46	45	45	45	$\overline{x}_3 = 45$	2
Group 4 (C)	42	39	39	41	39	$\overline{x}_4 = 40$	8
						$\overline{x} = 45$	

(Please note that the data are exactly the same as the **data for question 5**.)

(a) Test $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$ against $H_1 : \sigma_p^2 \neq \sigma_q^2$ for at least one pair $p \neq q$. Use $\alpha = 0.05$.

(b) Test at the 5% level whether the means for the four fuel mixtures differ significantly.

(c) What assumptions do you have to make to perform this statistical test?

QUESTION 16

In order to determine the effect of the amount of fertilizer (of a specific composition) on the growth of young walnut trees, 12 trees of the same age and all 30 cm high, were chosen and randomly divided into four groups of three each. After one season of growth the increased heights of the trees were measured.

	Fertilizer (grams/tree)	Increase in height (cm)			
	x_i	y_i	$x_i - \overline{x}$	$(x_i - \overline{x})^2$	$y_i\,(x_i - \overline{x})$
Group 1	1	14	-3	9	-42
	1	11	-3	9	-33
	1	16	-3	9	-48
Group 2	3	23	-1	1	-23
	3	19	-1	1	-19
	3	20	-1	1	-20
Group 3	5	20	1	1	20
	5	30	1	1	30
	5	27	1	1	27
Group 4	7	35	3	9	105
	7	31	3	9	93
	7	30	3	9	90
Total	48	276	0	60	180

Consider the simple linear regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ where the ε_is are independent $n\left(0;\ \sigma^2\right)$ random variables.

(a) Compute the regression line of y on x.

(b) Find a 90% confidence interval for the slope of the regression line computed in (a).

(c) What is the expected growth (increase in height) for $x = 6$?

(d) What is the standard error for the expected growth you computed in (c), in other words when $x = 6$?

QUESTION 17

An examination consists of two papers. The results for ten candidates are shown below.

Paper X	87	53	59	80	70	73	66	61	52	49
Paper Y	91	59	63	84	67	77	68	65	53	59

(a) Give the formulae used to calculate the estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ for an estimated regression line.

(b) The regression line calculated for the paper Y mark on the paper X mark is given by

$$Y = 9.001 + 0.917X$$

An eleventh candidate scored 63 in paper X, but missed the second paper because of illness. What would you estimate her mark in paper Y to be?

(c) $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ can be tested by the test statistic

$$T = \frac{\widehat{\beta}_1}{S/d} \qquad \text{where}$$

$$d^2 = \sum (x_i - \overline{x})^2 \text{ and}$$

$$S^2 = \frac{1}{n-2} \sum \left(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i \right)^2 .$$

Give the critical region where H_0 will be rejected if $\alpha = 0.05$.

(d) The calculated value for the test statistic is $T = 10.267$. What can you conclude?

(e) Test $H_0 : \beta_1 = 1$ against $H_1 : \beta_1 \neq 1$ at the 5% level of significance.

Use $d^2 = 1\,420$ and $S = 3.3656$. Give your final conclusion.

QUESTION 18

Complete the following statements in your answer book (that is, *give the missing words* and do not waste time rewriting everything).

(a) The hypothesis "The average blood platelet count of male cancer patients differs from the average blood platelet count of healthy men" requires a .....-tailed test of significance.

(b) An appropriate null hypothesis for the above test could be ..... .

(c) If you reject the null hypothesis on the basis of sample data, when in fact no difference exists, you have made a ..... error.

(d) If you accept the null hypothesis on the basis of sample data when in fact a true difference exists, you have made a ..... error.

(e) If the observed difference is "statistically highly significant" it means that the p-value will be very ..... .

(f) The *efficiency* of an estimator is a function of its ..... .

(g) When multiple measurements or observations are made on each of the individuals in a sample the assumption of ..... is violated.

QUESTION 19

(a) Write down, in general terms, the method of obtaining a least squares estimator.

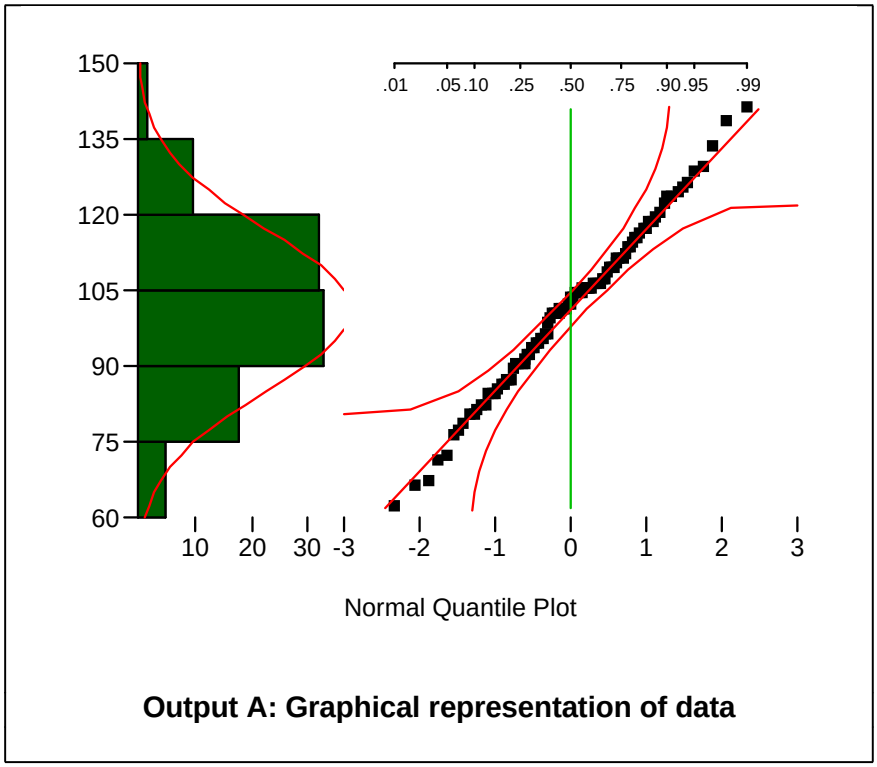
(b) Let $X_1;\ X_2;\ ...;\ X_n$ be independent random variables from a distribution with expected value θ. Find the least squares estimator for θ.

QUESTION 20

The following 100 observations have been gathered in an experiment:

Table A: Data

95	78	108	123	109	90	106	102	105	87
86	107	80	91	96	119	111	102	107	101
87	104	81	105	118	106	94	110	101	99
92	82	92	106	72	67	84	77	104	84
126	105	76	86	115	111	89	125	115	87
94	93	123	104	112	103	109	84	105	101
62	122	114	93	104	133	101	116	90	85
96	103	138	98	141	128	95	113	100	80
66	100	71	117	106	90	120	100	100	105
129	117	82	118	113	109	106	124	111	90



Output A: Graphical representation of data

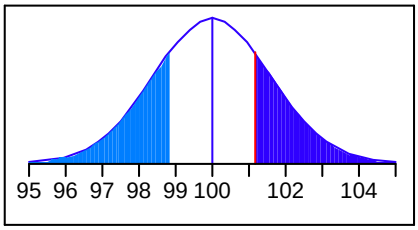
Moments

Mean	101.17
Std Dev	16.001929
Std Err Mean	1.6001929
upper 95% Mean	104.34513
lower 95% Mean	97.99487
N	100

Test Mean=value

Hypothesized Value	100
Actual Estimate	101.17
df	99
Std Dev	16.0019

	t Test
Test Statistic	0.7312
Prob > |t|	0.4664
Prob > t	0.2332
Prob < t	0.7668



Output B: Moments and t-test for data

You may assume that $\Sigma X_i = 10\ 117$ and $\Sigma(X_i - \overline{X})^2 = 25350.11$

Table B

Interval	Observed frequency, N_i	Expected frequency, $\hat{e}_i = n\hat{\pi}_i$
$X < 75$	5	5.94
$75 \leq X < 90$	18	.....
$90 \leq X < 105$	33	.....
$105 \leq X < 120$	32	27.27
$120 \leq X < 135$	10	.....
$X \geq 135$	2	1.43
Totals	100	100

(a) What can you conclude about normality from the **JMP graphical output in A**?

(b) Apply a **manual goodness-of-fit test** to test the null hypothesis that the sample comes from a $n\left(100; (16)^2\right)$ distribution.

Copy and complete the frequency tabulation (given in table B above) in your examination script.

[Hint: The intervals are **not equiprobable** because they are of the same length.]

Use $\alpha = 0.05$ and **clearly show all your computations**.

(c) Is there any reason to reject the null hypothesis $H_0 : \mu = 100$ against the alternative $H_1 : \mu > 100$? (Show how you use and interpret **the JMP output in B.)**

(d) Write down the 95% confidence interval for the mean. What can you conclude from this interval?

QUESTION 21

In an experiment to test whether pedigreed dogs are more susceptible to an influenza virus than cross-bred dogs, six pedigreed dogs and six cross-bred dogs were exposed to the virus. [All 12 the dogs were more or less of the same age, the same sex, the same weight and in the same physical condition.] The experiment yielded the following results:

	Type of breed		
	Pedigreed	Cross-bred	
Contracted influenza	5	2	7
Did not contract influenza	1	4	5
	6	6	12

Test the hypothesis that the two types of breeds are equally susceptible, against the alternative that pedigreed dogs are more susceptible than cross-breds. Test at the 5% level of significance.

QUESTION 22

In a random sample of 39 observations the sample correlation coefficient is $r = -0.35$.

Test $H_0 : \rho = -0.2$ against

$H_1 : \rho \neq -0.2$ at the 5% level.

QUESTION 23

(a) Let $X_1; X_2; \ldots; X_n$ be a random sample from a normal distribution with **unknown mean** μ. Use the distribution of $U = \sum_{i=1}^{n} \frac{\left(X_i - \bar{X}\right)^2}{\sigma^2}$ to show that a $100\,(1-\alpha)\,\%$ two-sided confidence interval for σ^2 is given by

$$\left[\frac{\sum \left(X_i - \bar{X}\right)^2}{\chi^2_{\frac{\alpha}{2};n-1}} \quad ; \quad \frac{\sum \left(X_i - \bar{X}\right)^2}{\chi^2_{1-\frac{\alpha}{2};n-1}} \right].$$

(b) In a sample of 20 from a $n\left(\mu;\sigma^2\right)$ distribution with μ unknown, we obtained $\sum_{i=1}^{20} X_i = 60$ and $\sum_{i=1}^{20} X_i^2 = 380.$

(i) Construct a 90% two-sided confidence interval for σ^2.

(ii) Use this interval to test the hypothesis that $\sigma^2 = 16$.

QUESTION 24

Let $X_1;\ X_2;\ ...;\ X_n$ be a random sample of size $n = 11$ from a normal population.
Let $Y_1;\ Y_2;\ ...;\ Y_m$ be a random sample of size $m = 16$ from a normal population independent of the first sample. Assume the following statistics:

$$\sum_{i=1}^{11} X_i = 115.5 \quad ; \quad \sum_{i=1}^{11} (X_i - \overline{X})^2 = 110$$

$$\sum_{j=1}^{16} Y_j = 208 \quad ; \quad \sum_{j=1}^{16} (Y_j - \overline{Y})^2 = 160.$$

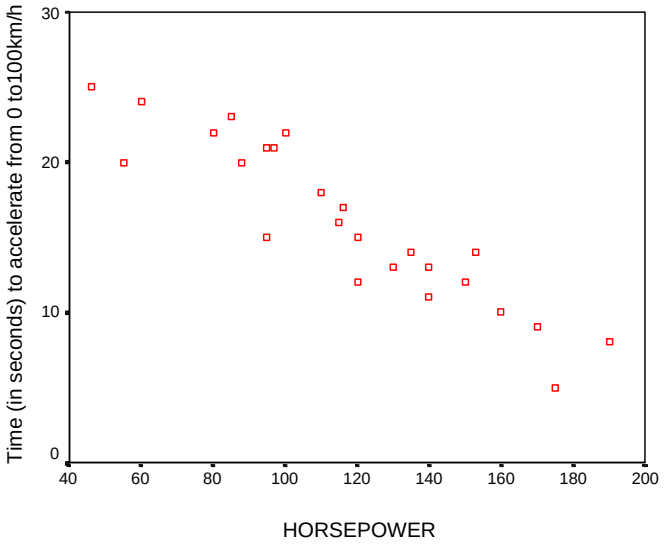
We want to test $H_0 : \mu_X = \mu_Y$ against
$H_1 : \mu_X \neq \mu_Y$ at the 5% level of significance.

(a) **Equality of variances** will influence the choice of the test statistic. Use $\alpha = 0.05$ and clearly show how you test for this assumption of equal variances.

(b) Perform the appropriate hypothesis test for the **difference between two means**.

QUESTION 25

The graph below represents a scatter diagram of a random sample of 25 cars where x depicts the horsepower and y depicts the time in seconds to accelerate from 0 to 100 km/h.



You may make use of any of the following results:

$$\sum_{i=1}^{25} X_i = 2\,925; \qquad \sum_{i=1}^{25} (X_i - \overline{X})^2 = 34\,384$$

$$\sum_{i=1}^{25} Y_i = 400; \qquad \sum_{i=1}^{25} (Y_i - \overline{Y})^2 = 708$$

$$\sum_{i=1}^{25} (X_i - \overline{X})(Y_i - \overline{Y}) = -4\,539$$

(a) Briefly discuss the applicability of simple linear regression for this data set.

(b) Compute a regression line where acceleration time is a function of the horsepower of the vehicle.

(c) Assume that $\sum_{i=1}^{25} (y_i - \widehat{y}_i)^2 = 108.89$ for the regression line computed in (b).

(i) What is the predicted time to accelerate from 0 to 100 km/h for a car with a horsepower of 200?

(ii) What is the standard error of this estimate?

Test $H_0 : \beta_1 = 0$ against

$H_1 : \beta_1 < 0$ at the 1% level of significance.

QUESTION 26

Indicate whether the following statements are true or false. Give the correct statement if the given statement is false.

(a) The statistic T is called an unbiased estimator for the parameter θ if $E(T) = \theta$.

(b) The significance level of a test is denoted by

$$\alpha \equiv P[H_0 \text{ is accepted} \mid H_0 \text{ is true}].$$

(c) The most efficient of several unbiased estimators is the one obtained by the method of least squares.

(d) If you reject the null hypothesis $\sigma_1^2 = \sigma_2^2$ on the basis of sample data, when in fact no difference exists between the variances of two populations, you have made a type II error.

(e) If X and Y have a bivariate normal distribution and $\rho = 0$ then the two variables X and Y are independent.

(f) If the sample size n increases, then the power of the test decreases.

(g) Repeated measurements on the same individual, $(X_i;\ Y_i)$ for $i = 1, 2, ..., n$ cannot be considered as independent observations.

QUESTION 27

Let $X_1;\ X_2;\ ...;\ X_n$ be a random sample from an exponential distribution with the following probability density function:

$$f_x(x;\ \theta) = \frac{1}{\theta} e^{-\frac{x}{\theta}}, \quad x > 0$$

Find the maximum likelihood estimator for θ.

QUESTION 28

From a sample of 75 independent observations the following statistics were computed:

$$\overline{X} = 64; \qquad \sum_{i=1}^{75} \left(X_i - \overline{X}\right)^2 = 1\,875;$$

$$\sum_{i=1}^{75} \left(X_i - \overline{X}\right)^3 = -7\,968 \qquad \text{and} \qquad \sum_{i=1}^{75} \left(X_i - \overline{X}\right)^4 = 187\,500.$$

(a) Test for both **skewness** and **kurtosis** to decide whether this is a sample from a normal distribution. Use the 10% level (two-sided) and follow the INSTRUCTIONS given under point 3 on page 1 of this paper.

(b) Discuss the possibility of testing $H_0 : \mu = 60$ against $H_1 : \mu \neq 60$ for the population from which this sample was drawn. (Do not perform a statistical test.)

QUESTION 29

In an investigation to determine whether a relationship exists between income and preferred Sunday newspaper, a random sample of 200 Sunday newspaper readers were chosen, and the following was recorded:

		Sunday newspaper			Total
		A	B	C	
Income	Below R10 000	8	30	12	50
	R10 000-R20 000	23	50	27	100
	Above R20 000	9	20	21	50
Total		40	100	60	200

Based on this data, would you think that the newspaper that a person reads is independent of his or her income? Test at the 5% level.

QUESTION 30

(a) Let $X_1;\ X_2;\ ...;\ X_n$ be a random sample from a normal distribution with **known mean** μ. Use the distribution of $U = \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2}$ to show that a $100\,(1-\alpha)\,\%$ two-sided confidence interval for σ^2 is given by

$$\left[\frac{\Sigma\,(X_i - \mu)^2}{\chi^2_{\frac{\alpha}{2};\,n}} \quad ; \quad \frac{\Sigma\,(X_i - \mu)^2}{\chi^2_{1-\frac{\alpha}{2};\,n}} \right].$$

(b) In a sample of size $n = 20$ from a $n\left(\mu;\ \sigma^2\right)$ distribution with $\mu = 4$, we obtained

$$\sum_{i=1}^{20} X_i = 60 \quad \text{and} \quad \sum_{i=1}^{20} X_i^2 = 380.$$

(i) Construct a 90% two-sided confidence interval for σ^2.

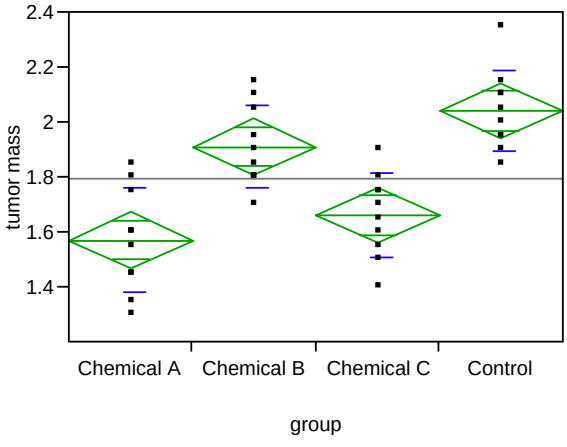
(ii) How can you use this interval to test the hypothesis that $\sigma^2 = 9$?

QUESTION 31

A company manufacturing medicine is screening various chemicals for possible use against cancer. They have three possible chemicals which they wish to test. Forty mice are selected, and cancer cells are implanted into each. The mice are then divided at random (for example by lottery) into four groups of 10 mice each; three groups are treated with the three chemicals and the other group serves as a control group which receives no treatment. After a fixed period the tumours in the mice are removed and weighed. The mass (in grams) of the tumours were found to be as follows:

Chemical A:	1.60;	1.45;	1.85;	1.30;	1.80;	1.65;	1.55;	1.75;	1.45;	1.35
Chemical B:	1.70;	2.05;	1.80;	2.15;	1.80;	1.95;	2.10;	1.85;	1.90;	1.80
Chemical C:	1.70;	1.75;	1.50;	1.40;	1.90;	1.65;	1.55;	1.60;	1.80;	1.75
Control:	1.90;	2.05;	2.35;	1.85;	2.10;	1.95;	2.10;	2.15;	2.00;	1.95

Study the following JMP output and answer the questions given below:



Output A: Graphical representation

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
group	3	1.4210000	0.473667	18.7797	<.0001*
Error	36	0.9080000	0.025222		
C. Total	39	2.3290000			

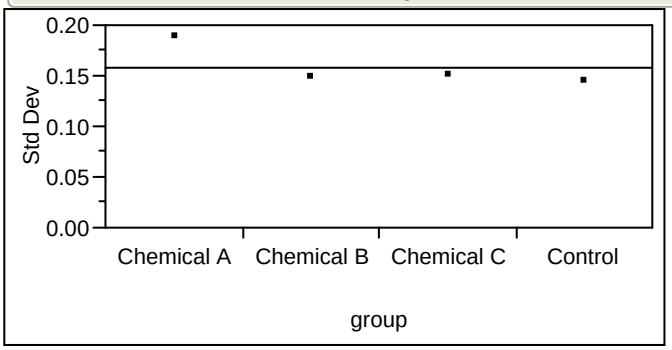
Means for Oneway Anova

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
Chemical A	10	1.57000	0.05022	1.4681	1.6719
Chemical B	10	1.91000	0.05022	1.8081	2.0119
Chemical C	10	1.66000	0.05022	1.5581	1.7619
Control	10	2.04000	0.05022	1.9381	2.1419

Std Error uses a pooled estimate of error variance

Output B: Computations for ANOVA

Tests that the Variances are Equal



Level	Count	Std Dev	MeanAbsDif to Mean	MeanAbsDif to Median
Chemical A	10	0.1873796	0.1500000	0.1500000
Chemical B	10	0.1486981	0.1220000	0.1200000
Chemical C	10	0.1505545	0.1200000	0.1200000
Control	10	0.1449138	0.1100000	0.1100000

Test	F Ratio	DFNum	DFDen	Prob > F
O'Brien[.5]	0.4682	3	36	0.7063
Brown-Forsythe	0.3711	3	36	0.7743
Levene	0.3937	3	36	0.7583
Bartlett	0.2606	3	.	0.8538

Welch Anova testing Means Equal, allowing Std Devs Not Equal

F Ratio	DFNum	DFDen	Prob > F
17.4327	3	19.926	<.0001*

Output C: Tests that the variances are equal

Means Comparisons

Comparisons for all pairs using Tukey-Kramer HSD

q*	Alpha
2.69323	0.05

Abs(Dif)-LSD

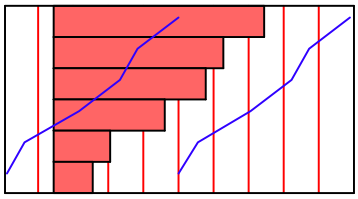
	Control	Chemical B	Chemical C	Chemical A
Control	-0.19128	-0.06128	0.18872	0.27872
Chemical B	-0.06128	-0.19128	0.05872	0.14872
Chemical C	0.18872	0.05872	-0.19128	-0.10128
Chemical A	0.27872	0.14872	-0.10128	-0.19128

Positive values show pairs of means that are significantly different.

Level		Mean
Control	A	2.0400000
Chemical B	A	1.9100000
Chemical C	B	1.6600000
Chemical A	B	1.5700000

Levels not connected by same letter are significantly different.

Level	- Level	Difference	Lower CL	Upper CL
Control	Chemical A	0.4700000	0.278716	0.6612845
Control	Chemical C	0.3800000	0.188716	0.5712845
Chemical B	Chemical A	0.3400000	0.148716	0.5312845
Chemical B	Chemical C	0.2500000	0.058716	0.4412845
Control	Chemical B	0.1300000	-0.061284	0.3212845
Chemical C	Chemical A	0.0900000	-0.101284	0.2812845



Output D: Multiple comparisons

(a) Do you think it is reasonable to assume that the four groups may be considered as *independent groups?*

(b) What additional graphical technique(s) would you need to decide that the four groups may be considered as *coming from normal populations?*

(c) Do you think it is reasonable to assume that the four groups have *equal population variances*? (Substantiate your answer.)

(d) **Do these results indicate that the tumours respond differently to the treatments**? (In other words can you conclude that the **mean tumour masses** for the four different groups differ significantly?)

Substantiate your answer by paying attention to the following detail:

(i) State the appropriate null and alternative hypotheses for this test.

(ii) What test statistic is used to test these hypotheses?

(iii) What is the value of the test statistic?

(e) Discuss the results of the multiple comparisons in Output D on all pairs of means by only interpreting the **confidence intervals** given in the last few lines.

QUESTION 32

The following data show the improvement (gain in reading speed) of six students and the number of weeks they attended a speed reading course:

Number of weeks	3	4	2	3	5	7
Speed gain (words/min)	65	90	22	81	132	180

(a) Fit a least squares regression line from which we can predict the gain in reading speed from the number of weeks a person has been attending the course.

(b) The 95% confidence interval for β_1 is given by (22.18; 38.07). What can one deduce from that?

(c) Predict the gain in reading speed for a person who has been in the program for six weeks.

(d) When testing the null hypothesis $H_0 : \beta_1 = 0$, the t-statistic was calculated as $t = 10.53$. Use this to show that an estimate, S^2, for $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2$ is approximately $2\,095.257$.

(e) Give a 95% confidence interval for the expected gain in speed for a future observation of $x = 6$.

9.2 Worksheet: Solutions

QUESTION 1

(a) False. $Var\left(X_{10}\right)=9$.

(b) True.

(c) False. $Var\left(aX\right)=a^{2}Var\left(X\right)$.

(d) False. The maximum likelihood estimators are more efficient than other estimators "but **they are not necessarily unbiased**".

(e) False. $1-\beta$ is the power of test, not the exceedance probability.

(f) False. If X and Y do not have a bivariate distribution then $\rho=0,$ does not necessarily imply independence.

QUESTION 2

(a) The maximum likelihood is

$$\begin{aligned} f_X(x) &= \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}(\,x-\mu)\,^2/\sigma^2} \\ &= \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\,X\,^2/\sigma^2} \text{ since } \mu=0 \\ L(\sigma) &= \prod_{i=1}^{n} f\left(X_i;\,\sigma\right) \\ &= \prod_{i=1}^{n}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\,X_i\,^2/\sigma^2} \\ &= \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\,X_1\,^2/\sigma^2}\times\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\,X_2\,^2/\sigma^2}\times\cdots\times\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\,X_n\,^2/\sigma^2} \\ &= \frac{1}{\left(\sigma\sqrt{2\pi}\right)^n}e^{-\frac{1}{2}\sum\,X_i\,^2/\sigma^2} \end{aligned}$$

$\therefore$

$$\begin{aligned}\log L(\sigma) &= -n\log\sigma - \frac{n}{2}\log 2\pi - \frac{1}{2}\sum {X_i}^2/\sigma^2 \\ \frac{d\log L(\sigma)}{d\sigma} &= \frac{-n}{\sigma} + \sum_{i=1}^{n} {X_i}^2 \times \frac{1}{\sigma^3} \\ &= \frac{-n\sigma^2 + \sum_{i=1}^{n} {X_i}^2}{\sigma^3}\end{aligned}$$

Setting $\frac{d\log L(\sigma)}{d\sigma} = 0$

$$\begin{aligned}\therefore \quad 0 &= \frac{-n\sigma^2 + \sum_{i=1}^{n} X_i^2}{\sigma^3} \\ 0 &= -n\sigma^2 + \sum_{i=1}^{n} X_i^2 \\ n\sigma^2 &= \sum_{i=1}^{n} X_i^2 \\ \Longrightarrow \widehat{\sigma}^2 &= \frac{\sum_{i=1}^{n} X_i^2}{n}\end{aligned}$$

(b)

$$\begin{aligned}E(T) &= E\left(\frac{\sum_{i=1}^{n} X_i^2}{n}\right) \\ &= \frac{1}{n}\left(\sum_{i=1}^{n} E\left(X_i^2\right)\right)\end{aligned}$$

$$\begin{aligned}\text{Now } E\left(X_i^2\right) &= Var(X_i) + E((Xi))^2 \\ &= \sigma^2 \text{ since } \mu = 0\end{aligned}$$

$$\begin{aligned}E(T) &= \frac{1}{n}\left(\sum_{i=1}^{n} \sigma^2\right) \\ &= \frac{1}{n} n\sigma^2 \\ &= \sigma^2\end{aligned}$$

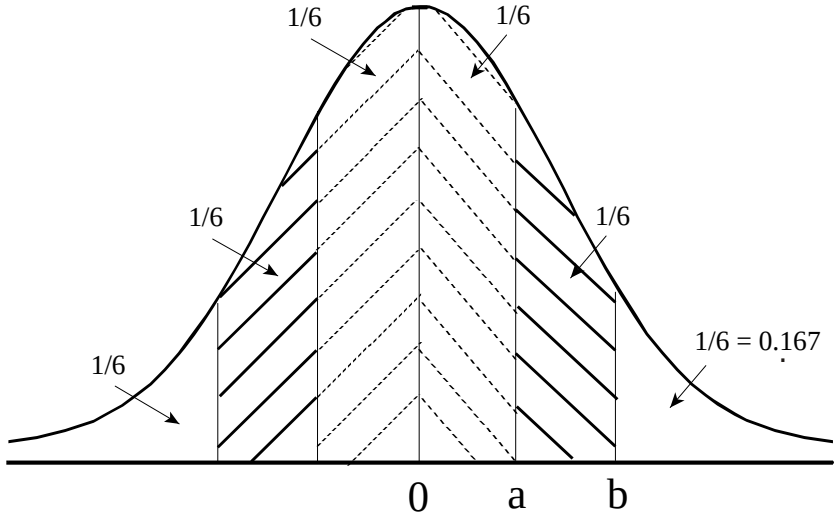


Figure 3:

QUESTION 3

(a) Since μ and σ are unknown, we estimate them.

$$\begin{aligned}\overline{X} &= \frac{\sum_{i=1}^{n} X_i}{n} \\ &= \frac{2\,340}{42} \\ &= 60\end{aligned}$$

$$\begin{aligned}\sigma &= \sqrt{\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}{n}} \\ &= \sqrt{\frac{27\,362}{42}} \\ &= \sqrt{651.4762} \\ &\approx 25.524\end{aligned}$$

If we have to use six classes of equal intervals then the probability of each interval is $\pi_i = \frac{1}{6}$ for each interval $\Rightarrow n\pi_i = 7$.

The fourth interval is where $0 \leq Z \leq a$.

From the sketch above the value "a" is found from table II as

$$\Phi(0.432) = P(Z \leq a) = 0.5 + 0.167 = 0.667.$$

Thus $a = 0.432.$

Since $0 \leq Z \leq a$. That is,

$$\begin{aligned}
0 &\leq Z &&\leq 0.432 \\
0 &\leq \frac{X-60}{25.524} &&\leq 0.432 \\
0 &\leq X-60 &&\leq 0.432 \times 25.524 \\
0 &\leq X-60 &&\leq 11.026368 \\
\Rightarrow 60 &\leq X &&\leq 60 + 11.026368 \\
60 &\leq X &&\leq 71.03.
\end{aligned}$$

(b)

Equal probability intervals	Expected frequency	Count marks	Observed frequency
$-\infty < X \leq 35.34$	7	𝍸 ||	7
$35.34 < X \leq 48.97$	7	𝍸 |	6
$48.97 < X \leq 60$	7	𝍸 ||||	9
$60 < X \leq 71.03$	7	𝍸 |||	8
$71.03 < X \leq 84.66$	7	𝍸	5
$84.66 < X \leq \infty$	7	𝍸 ||	7
Total	42		

(c)

H_0 : The sample comes from a normal distribution.
H_1 : The sample does not come from a normal distribution.

$$\begin{aligned}
Y^2 &= \sum_{k=1}^{k} \frac{(Observed - Expected)^2}{Expected} \\
&= \frac{(7-7)^2}{7} + \frac{(6-7)^2}{7} + \frac{(9-7)^2}{7} + \frac{(8-7)^2}{7} + \frac{(5-7)^2}{7} + \frac{(7-7)^2}{7} \\
&= 0 + \frac{1}{7} + \frac{4}{7} + \frac{1}{7} + \frac{4}{7} \\
&= \frac{10}{7} \\
&\approx 1.4286
\end{aligned}$$

From table IV (Stoker) $\chi^2_{\alpha;k-1} = \chi^2_{0.05;5} = 11.0705$ and $\chi^2_{\alpha;k-r-1} = \chi^2_{0.05;3} = 7.18473.$

Since $1.4286 < 7.18473$, we cannot reject H_0 at the 5% level of significance. We may conclude that the sample comes from a normal distribution.

(d) $\overline{X} = 60$, $t_{\alpha/2;(n-1)} = t_{0.025;41} = 2.02$.

$$\begin{aligned} s_x^2 &= \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{n-1} \\ &= \frac{27\,362}{41} \\ &= 667.3659 \\ \Longrightarrow \quad s &= \sqrt{667.3659} \approx 25.8334 \end{aligned}$$

The 95% confidence interval for μ is

$$\begin{aligned} &\overline{X} \pm t_{\alpha/2;(n-1)} \times \frac{S}{\sqrt{n}} \\ &60 \pm 2.02 \times \frac{25.8334}{\sqrt{42}} \\ &60 \pm 8.0521 \\ &(51.9479; 68.0521). \end{aligned}$$

(e) The confidence interval for σ^2 is $\left[\dfrac{\Sigma\left(X_i - \overline{X}\right)^2}{\chi^2_{\frac{1}{2}\alpha;n-1}}; \quad \dfrac{\Sigma\left(X_i - \overline{X}\right)^2}{\chi^2_{1-\frac{1}{2}\alpha;n-1}}\right]$

$\chi^2_{0.025;41} = 60.5946$ and $\chi^2_{0.975;41} = 25.2255$.

The 95% confidence interval is

$$\begin{aligned} &\left[\frac{27\,362}{60.5496}; \quad \frac{27\,362}{25.2255}\right] \\ &\left[451.894; \quad 1\,084.696\right]. \end{aligned}$$

(f) (i) $\overline{Y} = 1\,905/30 = 63.5$,

$$\begin{aligned} s_y^2 &= \frac{\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2}{n-1} \\ &= \frac{20\,365.25}{29} \\ &= 702.25 \end{aligned}$$

$H_0 : \sigma_1^2 = \sigma_2^2$ against $H_1 : \sigma_1^2 \neq \sigma_2^2$

The test statistic is

$$\begin{aligned} F &= \frac{\sigma_2^2}{\sigma_1^2} \cdot \frac{s_1^2}{s_2^2} \\ &= 1 \cdot \frac{667.3659}{702.25} \\ &\approx 0.9503. \end{aligned}$$

The critical values are $F_{0.05;41,29} = 1.80$.

Since $0.9503 < 1.80$, we cannot reject H_0 at the 5% level of significance and conclude that the variances are equal.

$$\begin{aligned} S_p^2 &= \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} \\ &= \frac{41(667.3659) + 29(702.25)}{42+30-2} \\ &= \frac{47\,727.2519}{70} \\ &\approx 681.8179 \end{aligned}$$

$\Longrightarrow S_p = \sqrt{681.8179} \approx 26.1116$

The test statistic is

$$\begin{aligned} T &= \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \\ &= \frac{(60-63.5)-0}{26.1116\sqrt{\frac{1}{42}+\frac{1}{30}}} \\ &= \frac{-3.5}{6.2419} \\ &\approx -0.5607. \end{aligned}$$

Test is one-tailed. The critical value is $t_{\alpha;(n_1+n_2-2)} = t_{0.01;70} = 2.384$. Reject H_0 if $T < -2.384$.

Since $-0.5607 > -2.384$, we do not reject H_0 at the 1% level and conclude that the claim is false. The older women and females in the age group 20-35 years have the same mean ferritin score.

(ii) It is valid if both populations are normal. Thus we assume second sample is normal.

QUESTION 4

(a) The mosaic output shows that the proportion of coloured-down chicks was almost four times the proportion of whites in parent type A and parent type C. There were almost three times the number of coloured-down chicks as compared to whites in parent type B. However, the proportions seem to be almost the same as evidenced by the horizontal lines (alignments). The hypothesis of no association might not be rejected.

(b) H_0: There is no association between type of parent and the colour of the down of the chicks.

H_1: There is an association between type of parent and the colour of the down of the chicks.

(c) The test statistic is $Y^2 = \sum_{k=1}^{k} \frac{(Observed - Expected)^2}{Expected}$ and the value is $Y^2 = 5.218.$

(d) Yes, the row percentage seems to be similar. For coloured-down chicks it is 81%, 73% and 85% for parent type A, B, and C, respectively. One might expect the null hypothesis not to be rejected.

(e) The critical value is $\chi^2_{0.05;2} = 5.99147$. Since $5.218 < 5.99147$, we do not reject H_0 at the 5% level of significance and conclude that there is no association between parent type and colour down of chicks. Alternatively the p-value $= 0.0736 < 0.05$, so we also reject H_0.

QUESTION 5

$n_1 = 147 \qquad R_1 = 0.4 \qquad n_2 = 125 \qquad R_1 = 0.59$

We want to test $H_0 : \rho_1 = \rho_2$ against $H_1 : \rho_1 \neq \rho_2$.

$U_1 = \frac{1}{2}\log_e \frac{1+0.4}{1-0.4} \approx 0.4236$

$U_2 = \frac{1}{2}\log_e \frac{1+0.59}{1-0.59} \approx 0.6777$

The test statistic is

$$\begin{aligned} z &= \frac{U_1 - U_2}{\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}} \\ &= \frac{0.4236 - 0.6777}{\sqrt{\frac{1}{147-3} + \frac{1}{125-3}}} \\ &= \frac{0.2541}{\sqrt{0.0151}} \\ &\approx -2.0675. \end{aligned}$$

The test is a two-tailed test. $\alpha = 0.05$, $\alpha/2 = 0.025$, and $Z_{0.025} = 1.96$. We reject H_0 if $Z > 1.96$ or $Z < -1.96$.

Since -2.0675 < -1.96, we reject H_0 at the 5% level and conclude that the correlations are significantly different, that is, $\rho_1 \neq \rho_2$.

QUESTION 6

Test $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$ against $H_1 : \sigma_p^2 \neq \sigma_q^2$ for at least one $p \neq q$.

The test statistic is

$$\begin{aligned} U &= \frac{\max\limits_i S_i^2}{\min\limits_i S_i^2} \\ &= \frac{4.7}{2.5} \\ &= 1.88. \end{aligned}$$

The critical value is 20.6. H_0 is rejected if $U > 20.6$.

Since $1.88 < 20.6$, we do not reject H_0 at the 5% level and conclude that the variances of the four populations are equal.

QUESTION 7

(a) Yes, there is a strong positive relationship.

(b) The equation of the regression line is

$$\widehat{Chirps/\sec ond} = 7.6833 + 0.3767 \textit{Temperature}$$

(c) The expected number of chirps/second is

$$\begin{aligned} \widehat{Chirps/\sec ond} &= 7.6833 + 0.3767(30) \\ &= 18.9843. \end{aligned}$$

We can estimate by taking repeated measurements of chirps/second when temperature is 30°C and estimate the mean response.

(d) The test statistic is

$$T_0 = \frac{\hat{\beta}_0 - \beta_0}{S\sqrt{\dfrac{1}{n} + \dfrac{\overline{X}^2}{\sum\limits_{i=1}^{n}(X_i - \overline{X})^2}}}.$$

The value of the test statistic is $T_0 = 6.16$. Yes, the p-value < 0.0001. Thus, since p-value < 0.01, we reject H_0 and conclude that it is significant.

(e) The test statistic is

$$T = \frac{\hat{\beta}_1 - \beta_1}{\frac{S}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}}}.$$

The value of the test statistic is $T = 8.48$. Yes, the p-value $<< 0.0001, \Longrightarrow$ p-value < 0.01, we reject H_0 and conclude that $\beta_1$is significantly different from zero.

(f) Value of $S^2 = \dfrac{\Sigma\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)^2}{(n-2)} = 1.48.$

QUESTION 8

(a) True.

(b) False. The most efficient estimator of two or more unbiased estimators is the one with the smallest variance.

(c) True.

(d) False. The significance level of a test is usually denoted by $\alpha = P\left(H_0 \text{ is rejected} \mid H_0 \text{ is true}\right).$

(e) True.

(f) False. The power of a test is found by computing $1 - P\left(\text{type II error}\right)$.

(g) False. Each test is performed at level $\dfrac{\alpha}{k}$.

QUESTION 9

$f_X(x) = \theta\left(1-\theta\right)^{X-1}$ for $x > 1$. The maximum likelihood estimator is

$$\begin{aligned}
L(\theta) &= \prod_{i=1}^{n} f\left(X_i;\ \theta\right) \\
&= \prod_{i=1}^{n} \theta\left(1-\theta\right)^{X_i - 1} \\
&= \theta\left(1-\theta\right)^{X_1 - 1} \times \theta\left(1-\theta\right)^{X_2 - 1} \times \cdots \times \theta\left(1-\theta\right)^{X_n - 1} \\
&= \theta^n (1-\theta)^{X_1 + X_2 + \ldots + X_n - n} \\
&= \theta^n (1-\theta)^{\sum X_i - n}
\end{aligned}$$

$\therefore$

$$\begin{aligned}
\log L(\theta) &= n\log\theta + \left[\sum_{i=1}^{n} Xi - n\right]\log(1-\theta) \\
\frac{d\log L(\theta)}{d\theta} &= \frac{n}{\theta} + \left[\sum_{i=1}^{n} Xi - n\right] \times \frac{-1}{1-\theta} \\
&= \frac{n(1-\theta) - \theta\left[\sum_{i=1}^{n} Xi - n\right]}{\theta(1-\theta)}.
\end{aligned}$$

Setting $\frac{d\log L(\theta)}{d\theta} = 0$

$$\begin{aligned}
\therefore \quad \frac{n(1-\theta) - \theta\left[\sum_{i=1}^{n} Xi - n\right]}{\theta(1-\theta)} &= 0 \\
n(1-\theta) - \theta\left[\sum_{i=1}^{n} X_i - n\right] &= 0 \\
n - n\theta - \theta\sum_{i=1}^{n} X_i + n\theta &= 0 \\
n - \theta\sum_{i=1}^{n} X_i &= 0 \\
\theta\sum_{i=1}^{n} X_i &= n \\
\widehat{\theta} &= \frac{n}{\sum_{i=1}^{n} X_i} \text{ or} \\
\Longrightarrow \quad \widehat{\theta} &= \frac{1}{\overline{X}}
\end{aligned}$$

QUESTION 10

(a) From the histogram, the data seem to be normally distributed. The normal quantile plot seems to follow a straight line (with no serious deviations) – Note: Comment subjective.

(b) **Test for skewness:**

We have to test $H_0 : \beta_1 = 0$ against
$H_1 : \beta_1 \neq 0.$

We will reject H_0 if $|B_1| > 0.534$ (in other words if $B_1 < -0.534$ or if $B_1 > 0.534$ (using table A)).

The value of the test statistic is $B_1 = \dfrac{\frac{1}{n}\Sigma\left(X_i - \overline{X}\right)^3}{\left[\frac{1}{n}\Sigma\left(X_i - \overline{X}\right)^2\right]^{\frac{3}{2}}}$.

We do not have $\Sigma\left(X_i-\overline{X}\right)^2$ but it can be derived from the given information.

$$\Sigma\left(X_i-\overline{X}\right)^2=\Sigma X_i^2-\frac{(\Sigma X_i)^2}{n} \quad =3\,245.83-\frac{(400.3)^2}{50}$$

$$=41.0282$$

$\Sigma\left(X_i-\overline{X}\right)^3=10.2145$

$$\therefore B_1=\frac{\frac{1}{50}(10.2145)}{\left[\frac{1}{50}(41.0282)\right]^{\frac{3}{2}}}=\frac{0.20429}{0.74331}\approx 0.2748$$

Since $0.2748>0.534$ we do not reject H_0 at the 10% level.

Test for kurtosis

We have to test $H_0:\beta_2=3$ against
$H_1:\beta_2\neq 3.$

$\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^4 = 87.81989.$ We will reject H_0 at the 10% level of significance (two-sided) if $B_2>3.99$ or if $B_2<2.15$ (from table B).

The value of the test statistic is: $B_2=\dfrac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^4}{\left[\frac{1}{n}\Sigma\left(X_i-\overline{X}\right)^2\right]^2}=\dfrac{\frac{1}{50}(87.81989)}{\left[\frac{1}{50}(41.0282)\right]^2}=\dfrac{1.7564}{(0.820564)^2}\approx$ $2.6085.$

Since $2.15<2.6085<3.99,$ we do not reject H_0 at the 10% level.

The sample **passed both tests** and hence we conclude that the sample is from a normal population.

(c) Yes. $H_0:\mu=8.5$ against $H_1:\mu\neq 8.5$.

Method I: Since p-value $=0.0004<0.05$, we reject H_0 at the 0.05 level and conclude that $\mu\neq 8.5.$

Method II: $t_{cal}=\frac{\overline{x}-\mu}{\frac{s}{\sqrt{n}}}=\frac{8.006-8.5}{\frac{0.91505}{\sqrt{50}}}\approx -3.8174.$ $t_{\frac{\alpha}{2};(n-1)}=t_{0.025;49}\approx 2.012.$ Since $|-3.8174|>2.012,$ the null hypothesis is rejected at the 0.05 level, thus $\mu\neq 8.5.$

(d) The 95% confidence interval for μ is 7.7459 to 8.2661. Yes, 8.5 is not contained in the interval.

(e) The confidence interval for the variance is $\left[\dfrac{\Sigma\left(X_i-\overline{X}\right)^2}{\chi^2_{\frac{1}{2}\alpha;n-1}};\quad \dfrac{\Sigma\left(X_i-\overline{X}\right)^2}{\chi^2_{1-\frac{1}{2}\alpha;n-1}}\right]$

$\chi^2_{0.025;49} = 70.2124$ and $\chi^2_{0.975;49} = 31.565$.

The 95% confidence interval for σ^2 is

$$\left[\frac{41.0282}{70.2124}; \quad \frac{41.0282}{31.565}\right]$$
$$[0.58434; \quad 1.2998]\,.$$

Now the confidence interval for σ is

$$\left[\sqrt{0.58434}; \quad \sqrt{1.2998}\right]$$
$$[0.76442; \quad 1.14009]\,.$$

Yes. There is a slight difference due to rounding up.

QUESTION 11

We have to test:

H_0 : There is no association between kind of cool drink and gender.
H_1 : Females tend to drink more sugar-free cool drink than men.

This is an exact 2×2 test and for fixed values of N, n and k the variable x (under H_0) will follow a hypergeometric distribution.

	Gender				
	Male	Female			
Type of cool drink		x			
		$\uparrow$			
Regular	4	2	6	$\longleftarrow$	n
Sugar-free	2	4	6		
	6	6	12	$\longrightarrow$	N
		$\uparrow$			
		k			

Females "drink significantly more sugar-free cool drink than men". If it is true we would expect regular female drinkers to be few. $\Longrightarrow$ We will reject H_0 in favour of H_1 if $P(X \leq x)$ becomes a very small probability. Using table D we find that for $x = 2$ and $P(X \leq 2) = 0.284$ (from table D)
$> 0.05 = \alpha$.

[If the students chooses $x = 4$, then we compute $P(X \geq 4) = 1 - P(X \leq 3) = 1 - 0.716 = 0.284$. Since $0.284 > 0.05$, we do not reject H_0 and conclude that there is no association between kind of cool drink and gender.

QUESTION 12

(a) $n = 10 \qquad \sum x = 37.1 \quad \sum x^2 = 151.91$
$\sum xy = 172.56 \quad \sum y = 45.1 \quad \sum y^2 = 215.49$

$$\begin{aligned} R &= \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sqrt{\left(\sum x^2 - \frac{\sum x}{n}\right)\left(\sum y^2 - \frac{\sum y}{n}\right)}} \\ &= \frac{172.56 - \frac{(37.1)(45.1)}{10}}{\sqrt{\left(151.91 - \frac{(37.1)^2}{10}\right)\left(215.49 - \frac{(45.1)^2}{10}\right)}} \\ &= \frac{5.239}{\sqrt{(14.269)(12.089)}} \\ &= \frac{5.239}{\sqrt{172.497941}} \\ &= \frac{5.239}{13.13385} \\ &\approx 0.3989 \end{aligned}$$

(b) $H_0 : \rho = 0$ against $H_1 : \rho > 0$

The critical value is $t_{\alpha;n-2} = t_{0.05;8} = 1.86$. Reject H_0 if $T > 1.86$.

$$\begin{aligned} T &= \frac{R\sqrt{n-2}}{\sqrt{1-R^2}} \\ &= \frac{0.3989\sqrt{8}}{\sqrt{1-0.3989^2}} \\ &= \frac{1.12825958}{\sqrt{0.84087879}} \\ &\approx 1.2304 \end{aligned}$$

Since $1.2304 < 1.86$, we do not reject H_0 and conclude that the correlation is not significant, that is, $\rho = 0$.

QUESTION 13

We need to test: $H_0 : \sigma_1^2 = \sigma_2^2$ against $H_1 : \sigma_1^2 > \sigma_2^2$.

$$\begin{aligned} S_1^2 &= \frac{1}{n_1 - 1}\left(\Sigma X_{1i}^2 - \frac{(\Sigma X_{1i})^2}{n}\right) \\ &= \frac{1}{9}\left(5\,374 - \frac{(190)^2}{10}\right) \\ &= \frac{1}{9}\left(1\,764\right) \\ &= 196 \end{aligned}$$

$$\begin{aligned} S_2^2 &= \frac{1}{n_2 - 1}\left(\Sigma X_{2i}^2 - \frac{(\Sigma X_{2i})^2}{n}\right) \\ &= \frac{1}{15}\left(12\,564 - \frac{(408)^2}{16}\right) \\ &= \frac{1}{15}\left(1\,466.4\right) \\ &= 97.76 \end{aligned}$$

The test statistic is

$$\begin{aligned} F &= \frac{\sigma_2^2}{\sigma_1^2} \cdot \frac{S_1^2}{S_2^2} \\ &= 1 \times \frac{196}{97.76} \\ &\approx 2.0049. \end{aligned}$$

The critical value is $F_{\alpha;n_1-1;n_2-1} = F_{0.05;9,15} = 2.59$. Reject H_0 if $F > 2.59$.

Since $2.0049 < 2.59$, we do not reject H_0 and conclude that $\sigma_1^2 = \sigma_2^2$, that is the variances are equal.

QUESTION 14

(a) $\sum_{i=1}^{16} X_i = 96; \qquad \sum_{i=1}^{16} X_i^2 = 4\,500; \qquad \overline{X} = 96/16 = 6; \qquad Z_{\alpha/2} = Z_{0.025} = 1.96$

The 95% confidence interval for the unknown mean μ when σ is known is

$$\begin{aligned}
&\overline{X} \pm Z_{\alpha/2;(n-1)} \times \frac{\sigma}{\sqrt{n}} \\
&6 \pm 1.96 \times \frac{15}{\sqrt{16}} \\
&6 \pm 7.35 \\
&(-1.35; 13.35).
\end{aligned}$$

(b) If σ is unknown we use s. Thus,

$$\begin{aligned}
S^2 &= \frac{1}{n-1}\left(\Sigma X_i^2 - \frac{(\Sigma X_i)^2}{n}\right) \\
&= \frac{1}{15}\left(4\,500 - \frac{(96)^2}{16}\right) \\
&= \frac{1}{15}(3\,924) \\
&= 261.6 \\
\Longrightarrow \quad & s = \sqrt{261.6} \approx 16.1741
\end{aligned}$$

$t_{\alpha/2;(n-1)} = t_{0.025;15} = 2.131.$

The 95% confidence interval for μ is

$$\begin{aligned}
&\overline{X} \pm t_{\alpha/2;(n-1)} \times \frac{S}{\sqrt{n}} \\
&6 \pm 2.131 \times \frac{16.1741}{\sqrt{16}} \\
&6 \pm 8.6168 \\
&(-2.6168; 14.6168).
\end{aligned}$$

QUESTION 15

(a)

$$\begin{aligned} S_1^2 &= \frac{\sum_{j=1}^{5}\left(X_{1j}-\overline{X}_1\right)^2}{n_1-1} & S_2^2 &= \frac{\sum_{j=1}^{5}\left(X_{2j}-\overline{X}_2\right)^2}{n_2-1} \\ &= \frac{2}{4} & &= \frac{4}{4} \\ &= 0.5 & &= 1 \end{aligned}$$

$$\begin{aligned} S_3^2 &= \frac{\sum_{j=1}^{5}\left(X_{3j}-\overline{X}_3\right)^2}{n_3-1} & S_4^2 &= \frac{\sum_{j=1}^{5}\left(X_{4j}-\overline{X}_4\right)^2}{n4-1} \\ &= \frac{2}{4} & &= \frac{8}{4} \\ &= 0.5 & &= 2 \end{aligned}$$

Testing $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$ against $H_1 : \sigma_p^2 \neq \sigma_q^2$ for at least one $p \neq q$

Test statistic is

$$\begin{aligned} U &= \frac{\max\limits_i S_i^2}{\min\limits_i S_i^2} \\ &= \frac{2}{0.5} \\ &= 4. \end{aligned}$$

The critical value is 20.6. H_0 is rejected if $U > 20.6$.

Since $4 < 20.6$, we do not reject H_0 at the 5% level and conclude that the variances of the four populations are the same.

(b) Testing $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ against $H_1 : \mu_p \neq \mu_p$ for at least one pair $p \neq q$

$$\begin{aligned} SSE &= \sum_i (X_{ij}-\overline{X}_i)^2 \\ &= 2+4+2+8 \\ &= 16 \end{aligned}$$

$$\begin{aligned} SST_r &= n\sum_{i=1}^{k}(X_i - \overline{X})^2 \\ &= 5[(55-45)^2 + (40-45)^2 + (45-45)^2 + (40-45)^2] \\ &= 5[100 + 25 + 0 + 25] \\ &= 5[150] \\ &= 750 \end{aligned}$$

$$MSE = \frac{SSE}{kn-k} = \frac{16}{16} = 1$$

$$MSTr = \frac{n\sum_{i=1}^{4}(\overline{X}_i - \overline{X})^2}{k-1} = \frac{750}{3} = 250$$

$$F = \frac{MSTr}{MSE} = \frac{250}{1} = 250$$

$$\begin{aligned} SST &= \Sigma X_i^2 - \frac{(\Sigma X)^2}{n} \\ &= 41\,266 - \frac{(900)^2}{20} \\ &= 766 \end{aligned}$$

The ANOVA table is

Source of variation	Sum of squares	Degrees of freedom	Mean square	F
Treatments	750	3	250	250
Error	16	16	1	
Total	766	19		

The critical value is $F_{0.05;3;16} = 3.24$. Reject H_0 if $F > 3.24$.

Since $250 > 3.24$, we reject H_0 at the 5% level and conclude that there is sufficient evidence to indicate a difference in means.

(c) The assumptions are:

- independence
- normality, and
- equal variances.

QUESTION 16

(a) $$\begin{aligned} \widehat{\beta}_1 &= \frac{\sum Y_i(X_i-\overline{X})}{\sum(X_i-\overline{X})^2} = \frac{180}{60} = 3 \\ \widehat{\beta}_0 &= \overline{Y}-\widehat{\beta}_1\overline{X} = 23-3(4) = 11 \end{aligned}$$

$\therefore$The estimated regression line is

$$\widehat{Y} = 11+3X.$$

(b) The confidence interval for the slope β_1 is

$$\widehat{\beta}_1 \pm t_{\alpha/2;n-2} \times \frac{S}{d}$$

$t_{\alpha/2;n-2} = t_{0.025;10} = 2.228 \qquad \widehat{\beta}_1 = 3.$

$$\begin{aligned} S^2 &= \frac{1}{n-2}\sum(Y_i-\widehat{Y}_i)^2 \\ &= \frac{1}{10}\left(0^2+(-3)^2+2^2+3^2+\cdots+(-2)^2\right) \\ &= \frac{1}{10}\left(9+4+9+1+36+16+1+9+1+4\right) \\ &= \frac{1}{10}(90) \\ &= 9 \\ \Longrightarrow\ & s=\sqrt{9}=3. \end{aligned}$$

Thus, the 95% confidence interval for $\widehat{\beta}_1$ is

$$\begin{aligned} &\widehat{\beta}_1 \pm t_{\alpha/2;n-2} \times \frac{S}{d} \\ &3 \pm 2.228 \times \frac{3}{\sqrt{60}} \\ &3 \pm 0.1114 \\ &(2.8886; 3.1114)\,. \end{aligned}$$

(c) The expected growth increase in height for $x=6$ is

$$\begin{aligned} \widehat{Y} &= 11+3(6) \\ &= 29. \end{aligned}$$

(d) The standard error for the expected growth when $x = 6$ is

$$\begin{aligned} S\sqrt{\frac{1}{n} + \frac{\left(X - \overline{X}\right)^2}{\sum\left(X_i - \overline{X}\right)^2}} &= 3\sqrt{\frac{1}{12} + \frac{(6-4)^2}{60}} \\ &= 3\sqrt{\frac{1}{12} + \frac{4}{60}} \\ &= 3\sqrt{0.15} \\ &\approx 1.1619. \end{aligned}$$

QUESTION 17

(a) The formulae are $\widehat{\beta}_1 = \frac{\sum Y_i(X_i - \overline{X})}{\sum(X_i - \overline{X})^2}$ $\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1\overline{X}.$

(b) The regression line is

$$\widehat{Y} = 9.001 + 0.917X.$$

The mark for paper Y is

$$\begin{aligned} \widehat{Y} &= 9.001 + 0.917(63) \\ &= 66.772 \\ &\approx 68. \end{aligned}$$

(c) The critical value is $t_{\frac{\alpha}{2};(n-2)} = t_{0.025;8} = 2.306$. The critical region is reject H_0 if $t > 2.306$ or $t < -2.306$.

(d) Since $T = 10.267 > 2.306$, we reject H_0 at the 5% level of significance and conclude that $\beta_1 \neq 0$. Thus, the slope is significant.

(e) We have to test $H_0 : \beta_1 = 1$ against $H_1 : \beta_1 \neq 1$. The test statistic is

$$\begin{aligned} T_{1,c} &= \frac{\widehat{\beta}_1 - c}{\frac{s}{\overline{d}}} \\ &= \frac{0.917 - 1}{\frac{3.3656}{1\,420}} \\ &= \frac{-0.083}{0.00237014} \\ &\approx -35.019. \end{aligned}$$

Since $-35.019 < -2.306$, we reject H_0 at the 5% level of significance and conclude that $\beta_1 \neq 1$.

QUESTION 18

(a) The hypothesis "The average blood platelet count of male cancer patients differs from the average blood platelet count of healthy men" requires a **two**-tailed test of significance.

(b) An appropriate null hypothesis for the above test could be $H_0 : \mu_1 = \mu_2$.

(c) If you reject the null hypothesis on the basis of sample data, when in fact no difference exists, you have made a **type I** error.

(d) If you accept the null hypothesis on the basis of sample data when in fact a true difference exists, you have made a **type II** error.

(e) If the observed difference is "statistically highly significant" it means that the p-value will be very **small**.

(f) The *efficiency* of an estimator is a function of its **variance**.

(g) When multiple measurements or observations are made on each of the individuals in a sample the assumption of **independence** is violated.

QUESTION 19

(a) Let $Q\left(X;\theta_1...\theta_j...,\theta_k\right) = \sum_{i}^{n}\left(X_i - E(X_i)\right)^2$.

Compute $\frac{dQ}{d\theta_j}; j = 1,....,k.$

Set $\frac{dQ}{d\theta_j} = 0; j = 1,....,k$ and solve to obtain $\widehat{\theta}_1,..\widehat{\theta}_j..,\widehat{\theta}_k.$

(b)

$$\begin{aligned} Q(X;\theta) &= \sum_i^n (X_i - E(X_i))^2 \\ &= \sum_i^n (X_i - \theta)^2 \\ &= \sum_i^n \left(X_i^2 - 2X_i\theta + \theta^2\right) \\ &= \sum_i^n X_i^2 - 2\theta \sum_i^n X_i + \sum_i^n \theta^2 \\ &= \sum_i^n X_i^2 - 2\theta \sum_i^n X_i + n\theta^2 \end{aligned}$$

$$\frac{dQ}{d\theta} = -2\sum_i^n X_i + 2n\theta$$

Set $\frac{dQ}{d\theta} = 0$ then

$$\begin{aligned} -2\sum_i^n X_i + 2n\theta &= 0 \\ \implies \sum_i^n X_i &= n\theta \\ \implies \widehat{\theta} &= \frac{\sum_i^n X_i}{n} \text{ (the least squares estimator for } \theta). \end{aligned}$$

QUESTION 20

(a) From the histogram of the **JMP graphical output in (A)** we conclude that the distribution on "looks" fairly symmetrical since the normal curve superimposed covers the histogram symmetrically (this is subjective).

From the normal quantile plot of the **JMP graphical output in (A)** we conclude that the distribution could be normal because there is no systematic deviation.

(b) We have to test:

$H_0 : X \sim n(100, 16^2)$ against
$H_1 : X$ does not have a $n(100, 16^2)$ distribution.

Although the intervals have the same length it does not imply that they will have the same probabilities! (This is similar to example 4.5 of the study guide.) **We have to compute the probability for each interval in order to find the expected frequency.** *[Three of the interval "Expected Frequency" values are given.]*

Now

$$\begin{aligned} P[X < 75] &= P\left[\frac{X-\mu}{\sigma} \leq \frac{75-100}{16}\right] = P[Z \leq -1.5625] = P[Z \geq 1.5625] \\ &= 1 - 0.9406 = \underline{0.0594} \text{ (given)} \end{aligned}$$

$$\begin{aligned} P[75 \leq X < 90] &= P\left[\frac{75-100}{16} \leq \frac{X-\mu}{\sigma} \leq \frac{90-100}{16}\right] = P[-1.5625 \leq Z \leq -0.625] \\ &= P[0.625 \leq Z \leq 1.5625] = 0.9406 - 0.7340 = \underline{0.2066} \end{aligned}$$

$$\begin{aligned} P[90 \leq X < 105] &= P\left[\frac{90-100}{16} \leq \frac{X-\mu}{\sigma} \leq \frac{105-100}{16}\right] = P[-0.625 \leq Z \leq 0.3125] \\ &= (0.6217 - 0.5) + (0.7340 - 0.5) = \underline{0.3557} \end{aligned}$$

$$\begin{aligned} P[105 \leq X < 120] &= P\left[\frac{105-100}{16} \leq \frac{X-\mu}{\sigma} \leq \frac{120-100}{16}\right] = P[0.3125 \leq Z \leq 1.25] \\ &= 0.8944 - 0.6217 = \underline{0.2727} \text{ (given)} \end{aligned}$$

$$\begin{aligned} P[125 \leq X < 135] &= P\left[\frac{120-100}{16} \leq \frac{X-\mu}{\sigma} \leq \frac{135-100}{16}\right] = P[1.25 \leq Z \leq 2.1875] \\ &= 0.9857 - 0.8944 = \underline{0.0913} \end{aligned}$$

$$\begin{aligned} P[X \geq 135] &= P\left[\frac{X-\mu}{\sigma} \geq \frac{135-100}{16}\right] = P[Z \geq 2.1875] \\ &= 1 - 0.9857 = \underline{0.0143} \text{ (given)} \end{aligned}$$

$0.0594 + 0.2066 + 0.3557 + 0.2727 + 0.0913 + 0.0143 = 1.$

Expected frequency = probability $\times$ 100

Table B				
Interval	Observed frequency, N_i	Expected frequency, $\hat{e}_i = n\hat{\pi}_i$		$(N_i - \hat{e}_i)^2$
$X < 75$	5	5.94		$(-0.94)^2 = 0.8836$
$75 \leq X < 90$	18	20.66		$(-2.66)^2 = 7.0756$
$90 \leq X < 25.5245$	33	35.57		$(-2.57)^2 = 6.6049$
$25.5245 \leq X < 120$	32	27.27		$(4.73)^2 = 22.373$
$120 \leq X < 135$	10	9.13		$(0.87)^2 = 0.7569$
$X \geq 135$	2	1.43		$(0.57)^2 = 0.3249$
Totals	100	100		

$$\begin{aligned} Y^2 &= \sum_{k=1}^{6} \frac{(Observed - Expected)^2}{Expected} \\ &= \frac{(5-5.94)^2}{5.94} + \frac{(18-20.66)^2}{20.66} + \frac{(33-35.57)^2}{35.57} + \cdots + \frac{(10-9.13)^2}{9.13} + \frac{(2-1.43)^2}{1.43} \\ &= \frac{0.8836}{5.94} + \frac{7.0756}{20.66} + \frac{6.6049}{35.57} + \frac{22.373}{27.27} + \frac{0.7569}{9.13} + \frac{0.3249}{1.43} \\ &\approx 1.8074 \end{aligned}$$

From table IV we see that $\chi^2_{0.05;(6-1)} = 11.0705$. This implies we reject H_0 if $Y^2 \geq 11.0705$.

Since $1.8074 < 11.0705$, we cannot reject H_0 at the 5% level of significance. Assume normality.

(c) We have to test the null hypothesis $H_0 : \mu = 100$ against the alternative $H_1 : \mu > 100$ (given).
From **the JMP output in B** the value of the t-test statistic is $t = 0.7312$ and p-value (prob$> t = 0.4664$). Cannot reject H_0.

(d) The 95% confidence interval for the mean is $(97.99487; 104.34513)$. Since this interval includes the value 100 we conclude we cannot reject $H_0 : \mu = 100$.

QUESTION 21

We have to test:

H_0 : No association between kind of breed and contraction of influenza (independent factors).
H_1 : Pedigreed dogs are more susceptible than cross-breeds.

This is an exact 2×2 test and for fixed values of N, n and k the variable x (under H_0) will follow a hypergeometric distribution.

	Type of breed				
	Pedigreed	Cross-bred			
Contracted influenza	5	2	7		
Did not contract influenza	$\mathbf{1 = x}$	4	5	$\longleftarrow$	k
	6	6	12	$\longrightarrow$	N
	$\uparrow$				
	n				

Please note that the words *"more susceptible"* imply a one-sided alternative (or else we should have used the word "different"). The trickiest part of this problem is to translate the alternative hypothesis in terms of x (which is "fixed" by the choice of n and k).

If H_1 is true, we would expect the pedigreed dogs to be few$\Longrightarrow$we will reject H_0 in favour of H_1 if $P(X \leq x)$ becomes a very small probability. Using table D we find that for $x = 1$ and $P(X \leq 1) = 0.121$.

This is not a very small probability $\Longrightarrow$ we cannot reject $H_{0.}$

[If students choose $x = 4$, then we compute $P(X \geq 4) = 1 - P(X \leq 3) = 1 - 0.879 = 0.121$. Since $0.121 > 0.05$, we do not reject H_0 and conclude that kind of breed and contraction of influenza seem to be independent $\Longrightarrow$ whether dog is pedigree or not, is independent of whether it will contract influenza.

QUESTION 22

Since we are testing $H_0 : \rho = \rho_0$ (with $\rho = -0.2 \neq 0$), we have to use the test statistic defined in **theorem 5.5**.

$z = \sqrt{n-3}(U - \eta) \sim n(0;1)$ is the valid test statistic.

$$\begin{aligned} U &= \frac{1}{2}\log_e \frac{1+R}{1-R} \\ &= \frac{1}{2}\log_e \frac{1+(-0.35)}{1-(-0.35)} \\ &= \frac{1}{2}\log_e 0.48148 \\ &\approx -0.3654 \end{aligned}$$

We also compute

$$\begin{aligned} U &= \frac{1}{2}\log_e \frac{1+\rho}{1-\rho} \\ &= \frac{1}{2}\log_e \frac{1+(-0.2)}{1-(-0.2)} \\ &= \frac{1}{2}\log_e 0.66667 \\ &\approx -0.2057. \end{aligned}$$

So, the test statistic is

$$\begin{aligned} z &= \sqrt{n-3}(U-\eta) \\ &= \sqrt{39-3}(-0.3654-(-0.2027)) \\ &= 6\times -0.1627 \\ &= -0.9762. \end{aligned}$$

We reject H_0 if $Z < -Z_{\alpha/2} = -1.96$ or greater than 1.96. Since $-1.96 < -0.9762 < 1.96$, we do not reject H_0 and conclude that $\rho = -0.2$ at the 5% level.

QUESTION 23

(a) From **theorem 6.1** we know that $U = \dfrac{\Sigma\left(X_i-\overline{X}\right)^2}{\sigma^2}$ then $U \sim \chi^2_{n-1}$.

$$\begin{aligned} 1-\alpha &= P\left(\chi^2_{1-\frac{1}{2}\alpha;n-1} < U < \chi^2_{\frac{1}{2}\alpha;n-1}\right) \\ &= P\left[\chi^2_{1-\frac{1}{2}\alpha;n-1} < \frac{\Sigma\left(X_i-\overline{X}\right)^2}{\sigma^2} < \chi^2_{\frac{1}{2}\alpha;n-1}\right] \\ &= P\left[\frac{1}{\chi^2_{1-\frac{1}{2}\alpha;n-1}} > \frac{\sigma^2}{\Sigma\left(X_i-\overline{X}\right)^2} > \frac{1}{\chi^2_{\frac{1}{2}\alpha;n-1}}\right] \\ &= P\left[\frac{\Sigma\left(X_i-\overline{X}\right)^2}{\chi^2_{\frac{1}{2}\alpha;n-1}} < \sigma^2 < \frac{\Sigma\left(X_i-\overline{X}\right)^2}{\chi^2_{1-\frac{1}{2}\alpha;n-1}}\right] \end{aligned}$$

Therefore $\left[\dfrac{\Sigma\left(X_i-\overline{X}\right)^2}{\chi^2_{\frac{1}{2}\alpha;n-1}};\quad \dfrac{\Sigma\left(X_i-\overline{X}\right)^2}{\chi^2_{1-\frac{1}{2}\alpha;n-1}}\right]$ is a $100\,(1-\alpha)\,\%$ two-sided confidence interval for σ^2.

(b) (i)

$$\begin{aligned} \Sigma\left(x_i-\overline{x}\right)^2 &= \Sigma x_i^2 - n\overline{x}^2 \\ &= \Sigma x_i^2 - \frac{\left(\Sigma x_i\right)^2}{n} \\ &= 380 - \frac{(60)^2}{20} \\ &= 380-180 \\ &= 200 \end{aligned}$$

The 90% two-sided confidence interval for σ^2 now becomes $\left(\dfrac{200}{\chi^2_{\frac{1}{2}\alpha;n-1}};\quad \dfrac{200}{\chi^2_{1-\frac{1}{2}\alpha;n-1}}\right)$ that is

$$\left(\frac{200}{30.144};\quad \frac{200}{10.117}\right) = (6.635;\quad 19.769)\,.$$

(ii) Since this interval includes the value 16 we cannot reject the null hypothesis $H_0 : \sigma^2 = 16$.

QUESTION 24

(a) We first need to test whether populations have the **same variance**, in other words test

$H_0 : \sigma_X^2 = \sigma_Y^2$ against $H_1 : \sigma_X^2 \neq \sigma_Y^2$.

The test statistic is

$$\begin{aligned} F &= \frac{s_1^2}{s_2^2} \\ &= \frac{\Sigma\left(X_i - \overline{X}\right)^2/(n_1-1)}{\Sigma\left(Y_i - \overline{Y}\right)^2/(n_2-1)} \\ &= \frac{110/10}{160/15} \\ &\approx 1.03. \end{aligned}$$

Reject H_0 in favour of H_1 if $F \leq F_{1-\alpha/2;n_1-1;n_2-1}$ or if $F \geq F_{\alpha/2;n_1-1;n_2-1}$. The rejection region is $F \leq F_{0.975;10;15} = \dfrac{1}{F_{0.025;15;10}} = \dfrac{1}{3.52} = 0.284$ or if $F \geq F_{0.025;10;15} = 3.06$. Since $0.284 < 1.03 < 3.06$, we cannot reject H_0 at the 5% level of significance and conclude that $\sigma_X^2 = \sigma_Y^2$.

(b) We wish to test $H_0 : \mu_X = \mu_Y$. Since the variances are equal.

$$\begin{aligned} S_p^2 &= \frac{\Sigma\left(X_i - \overline{X}\right)^2 + \Sigma\left(Y_i - \overline{Y}\right)^2}{n_1 + n_2 - 2} \\ &= \frac{110 + 160}{11 + 16 - 2} \\ &= \frac{270}{25} \\ &= 10.8 \end{aligned}$$

$\Longrightarrow S_{pooled} = \sqrt{10.8} \approx 3.286.$

The test statistic is

$$\begin{aligned} T &= \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \\ &= \frac{(10.5 - 13) - 0}{3.286\sqrt{\frac{1}{11} + \frac{1}{16}}} \\ &= \frac{-2.5}{3.286\sqrt{0.1534}} \\ &\approx -1.9425. \end{aligned}$$

Test is two-tailed. The critical value is $t_{\alpha/2;(n_1+n_2-2)} = t_{0.025;25} = 2.06$. Reject H_0 if $|T| > 2.06$.

Since $-2.06 < -1.9425 < 2.06 \Longrightarrow$ cannot reject H_0 at the 5% level and conclude that $\mu_X = \mu_Y$.

QUESTION 25

(a) From the graph we see that the points are scattered around a straight line (that is, not showing curvature or heteroscedacity).

(b) $\widehat{\beta}_1 = \frac{\sum(X_i-\overline{X})(Y_i-\overline{Y})}{\sum(X_i-\overline{X})^2} = \frac{-4\,539}{34\,384} \approx -0.13201$

$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1\overline{X} = \frac{400}{25} - (-0.13201)(\frac{2\,925}{25}) = 16 + 15.445 = 31.445$

So, the estimated regression equation is

$$\widehat{Y} = 31.445 - 0.132X.$$

(c) Given that $\sum_{i=1}^{25}(y_i - \widehat{y}_i)^2 = 108.89 \Longrightarrow S^2 = \frac{\sum_{i=1}^{25}(y_i - \widehat{y}_i)^2}{n-2} = \frac{108.89}{25-2} \approx 4.734.$

(i) If $x = 200$, the predicted time is $\widehat{Y} = 31.445 - 0.132(200) = 5.045$ seconds.

(ii) Since $x = 200$ is not an observed point of the data set, it must be considered a "future observation" $\Longrightarrow$ the error of the estimate is (see section 8.5 of the study guide)

$$\begin{aligned} S\sqrt{1 + \frac{1}{n} + \frac{(X - \overline{X})^2}{\sum(X_i - \overline{X})^2}} &= \sqrt{\frac{108.89}{25-2}}\sqrt{1 + \frac{1}{25} + \frac{(200-117)^2}{34\,384}} \\ &= \sqrt{4.734}\sqrt{1.240} \\ &= \sqrt{5.87} \\ &\approx 2.42. \end{aligned}$$

(d) We want to test $H_0 : \beta_1 = 0$ against
$H_1 : \beta_1 < 0.$

The test statistic is

$$\begin{aligned} T &= \frac{\hat{\beta}_1 - \beta_1}{\frac{S}{d}} \\ &= \frac{-0.132 - 0}{\frac{\sqrt{4.734}}{\sqrt{34\,384}}} \\ &\approx -11.250. \end{aligned}$$

The critical value is $-t_{(\alpha;n-2)} = -t_{0.01;23} = -2.5$. Reject H_0 if $T \leq -2.5$.

Since $-11.25 < -2.5$, we reject H_0 at the 1% level. This means that the regression line is significant to explain the variability in y. (Only when $\beta_1 = 0$, does it imply that regression is meaningless.)

QUESTION 26

(a) True.

(b) False. The significance level of a test is denoted by

$$\alpha \equiv P\,[H_0 \text{ is rejected} \mid \ H_0 \text{ is true}].$$

(c) False. The most efficient of several unbiased estimators is the one with the smallest variance.

(d) False. If you reject H_0 (when in fact it is true) you have a type I error (or type II error is when you fail to reject H_0 when in actual fact it is false).

(e) True.

(f) False. If the sample size n increases, then the power of the test increases.

(g) True.

QUESTION 27

$f_x(x;\ \theta) = \frac{1}{\theta}e^{-\frac{x}{\theta}}, \quad x > 0$. The maximum likelihood estimator is

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} f(X_i;\ \theta) \\ &= \prod_{i=1}^{n} \frac{1}{\theta}e^{-\frac{x}{\theta}} \\ &= \frac{1}{\theta}e^{-\frac{X_1}{\theta}} \times \frac{1}{\theta}e^{-\frac{X_2}{\theta}} \times \cdots \times \frac{1}{\theta}e^{-\frac{X_n}{\theta}} \\ &= \frac{1}{\theta^n}e^{-\frac{\sum_{i=1}^{n} X_i}{\theta}} \end{aligned}$$

$\therefore$

$$\begin{aligned} \log L(x;\theta) &= -n\log\theta - \sum_{i=1}^{n} X_i/\theta \\ \frac{d\log L(x;\theta)}{d\theta} &= \frac{-n}{\theta} + \sum_{i=1}^{n} X_i \times \frac{1}{\theta^2} \\ &= \frac{-n\theta + \sum_{i=1}^{n} X_i}{\theta^2}. \end{aligned}$$

Setting $\frac{d\log L(x;\theta)}{d\theta} = 0$

$$\begin{aligned} \therefore \quad \frac{-n\theta + \sum_{i=1}^{n} X_i}{\theta^2} &= 0 \\ -n\theta + \sum_{i=1}^{n} X_i &= 0 \\ n\theta &= \sum_{i=1}^{n} X_i \\ \widehat{\theta} &= \frac{\sum_{i=1}^{n} X_i}{n} \text{ or} \\ \Longrightarrow \quad \widehat{\theta} &= \overline{X}. \end{aligned}$$

QUESTION 28

(a) **Test for skewness:**

We have to test $H_0 : \beta_1 = 0$ against
$H_1 : \beta_1 \neq 0.$

The test statistic is $B_1 = \dfrac{\frac{1}{n}\Sigma\left(X_i - \overline{X}\right)^3}{\left[\frac{1}{n}\Sigma\left(X_i - \overline{X}\right)^2\right]^{\frac{3}{2}}}.$

Observed value for B_1 is $B_1 = \dfrac{\frac{1}{75}(-7\,968)}{\left[\frac{1}{75}(1\,875)\right]^{\frac{3}{2}}} = \dfrac{-106.24}{[25]^{\frac{3}{2}}} = 0.84992.$

The critical value can be found by interpolation (at the 10% level) $0.459 + \frac{1}{2}(0.432 - 0.459) = 0.4455.$
We will reject H_0 at the 10% level if $|B_1| >$ tabulated percentage point $= 0.4455.$

Since $0.84992 > 0.4455,$ we reject H_0 in favour of H_1. We conclude that the distribution is not symmetric.

Test for kurtosis

We have to test $H_0 : \beta_2 = 3$ against
$H_1 : \beta_2 \neq 3.$

Test statistic is $B_2 = \dfrac{\frac{1}{n}\sum\limits_{i=1}^{n}\left(X_i - \overline{X}\right)^4}{\left[\frac{1}{n}\Sigma\left(X_i - \overline{X}\right)^2\right]^2}$. We will reject H_0 at the 10% level of significance (two-sided) if $B_2 <$ lower 5% point or if 3.99 or if $B_2 >$ upper 5% point in the table.

The observed value of the B_2 is $B_2 = \dfrac{\frac{1}{75}(187\,500)}{\left[\frac{1}{75}(1\,875)\right]^2} = \dfrac{2\,500}{625} = 4.$

Critical values:

The lower 5% point is 2.27.

The upper 5% point is 3.87.

Since $4 > 3.87,$ we reject H_0 at the 10% level. We conclude that the data do not have kurtosis of a normal population.

Overall, we conclude that the sample was not from a normal distribution.

(b) The sample consists of 75 independent observations, which satisfy the assumption of independence.

The assumption of normality is, however, not met as indicated in the test using this data set.

QUESTION 29

The expected frequencies are $e_{ij} = \frac{N_{i.}N_{.j}}{N_{..}}$

		Sunday newspaper			Total
		A	B	C	
Income	Below R10 000	10	25	15	50
	R10 000-R20 000	20	50	30	100
	Above R20 000	10	25	15	50
Total		40	100	60	200

H_0 : There is no association between income and preferred Sunday paper.
H_1 : There is an association between income and preferred Sunday paper.

$$\begin{aligned} Y^2 &= \sum_{i=1}^{3}\sum_{j=1}^{3}\frac{(N_{ij}-e_{ij})^2}{e_{ij}} \\ &= \frac{(8-10)^2}{10}+\frac{(30-25)^2}{25}+\cdots+\frac{(21-15)^2}{15} \\ &= 0.4+1.0+\cdots 2.4 \\ &= 6.25 \end{aligned}$$

$\chi^2_{0.05;(h-1)(k-1)} = \chi^2_{0.05;2\times 2} = \chi^2_{0.05;4} = 9.488$. Reject H_0 if $Y^2 > 9.488$. Since $6.25 < 9.4888$, we cannot reject H_0. We conclude that there exists no association between income and preferred Sunday newspaper.

QUESTION 30

(a) Let $U = \frac{\Sigma (X_i - \mu)^2}{\sigma^2}$ then $U \sim \chi^2_n$ (known value of μ).

$$\begin{aligned} 1-\alpha &= P\left(\chi^2_{1-\frac{1}{2}\alpha;n} < U < \chi^2_{\frac{1}{2}\alpha;n}\right) \\ &= P\left[\chi^2_{1-\frac{1}{2}\alpha;n} < \frac{\Sigma (X_i - \mu)^2}{\sigma^2} < \chi^2_{\frac{1}{2}\alpha;n}\right] \\ &= P\left[\frac{1}{\chi^2_{1-\frac{1}{2}\alpha;n}} > \frac{\sigma^2}{\Sigma (X_i - \mu)^2} > \frac{1}{\chi^2_{\frac{1}{2}\alpha;n}}\right] \\ &= P\left[\frac{\Sigma (X_i - \mu)^2}{\chi^2_{\frac{1}{2}\alpha;n}} < \sigma^2 < \frac{\Sigma (X_i - \mu)^2}{\chi^2_{1-\frac{1}{2}\alpha;n}}\right] \end{aligned}$$

(b) (i)

$$\begin{aligned} \Sigma (X_i - \mu)^2 &= \Sigma x_i^2 - 2\mu\Sigma X_i + n\mu^2 \\ &= 380 - 2(4)(60) + 20(4)^2 \\ &= 380 - 480 + 320 \\ &= 220 \end{aligned}$$

The 90% confidence interval for σ^2 now becomes $\left(\frac{220}{\chi^2_{0.05;20;}}; \quad \frac{200}{\chi^2_{0.95;20}}\right)$, that is,

$\left(\frac{220}{31.4104}; \quad \frac{220}{10.8508}\right) = (7.004; \quad 20.275)$.

(ii) Since this interval includes 9, we cannot reject $H_0 : \sigma^2 = 9$ against $H_1 : \sigma^2 \neq 9$ at the 10% level of significance. (A two-sided confidence interval$\Longrightarrow$ two-sided hypothesis testing).

QUESTION 31

(a) It is reasonable to assume that the four samples are independent. The outcome of one group of mice cannot influence the outcome of another group because they are not related and they were randomly divided.

(b) To use JMP for this purpose we should have obtained normal quantile plots for each separate group or formal statistical goodness-of-fit tests.

(c) The assumptions can be formally tested.

$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$ against $H_1 : \sigma_p^2 \neq \sigma_q^2$ for at least one $p \neq q$

From the tests that the variances are equal (Output C) we see that all the p-values $>> 0.05$ which means that we cannot reject H_0. The assumption of equal variances is not violated.

(d) We have to test $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ against $H_1 : \mu_p \neq \mu_p$ for at least one pair $p \neq q$.

The test statistic is
$F = \dfrac{MSTr}{MSE} \approx F_{k-1;kn-k}.$

From **Output B: Computations for ANOVA** we see that $F = 18.7797$ which is highly significant with a p-value of < 0.0001 and we reject H_0. $\mu_p \neq \mu_p$ for at least one pair $p \neq q$.

(e) Manually, we should have computed for each pair of means, a test statistic $T = \frac{(\overline{X}_p - \overline{X}_q)}{S_{pooled}\sqrt{\frac{1}{n}+\frac{1}{n}}}$ where we have samples of equal sizes if we want to incorporate the principle of the Bonferroni inequality.

The Tukey-Kramer HSD that is shown in the JMP output performs individual comparisons that make adjustments for multiple tests.

Confidence intervals that do not include zero imply that the pairs of means differ significantly. Only the last two intervals include zero which means that **chemical B** and the **control group** do not differ and that **chemical C** and **chemical A** do not differ. From the diamonds graph as well as the table of means we may conclude that **chemicals C** and **A** are effective since they are significantly smaller means for the tumour masses.

QUESTION 32

x_i	y_i	$x_i - \overline{x}$	$(x_i - \overline{x})^2$	$y_i(x_i - \overline{x})$
3	65	-1	1	-65
4	90	0	0	0
2	22	-2	4	-44
3	81	-1	1	-81
5	132	1	1	132
7	180	3	9	482
			16	482

(a) $\widehat{\beta}_1 = \dfrac{\sum Y_i(X_i - \overline{X})}{\sum (X_i - \overline{X})^2} = \frac{482}{16} \approx 30.125$

$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X} = 95 - 30.125(4) = -25.5$

So, the estimated regression equation is

$$\widehat{Y} = -25.5 + 30.125X$$

(b) Since $0 \notin (22.18; 38.07)$, the hypothesis $H_0 : \rho = 0$ can be rejected at the 5% level. This implies that "week in the program" is a significant predictor for speed gain.The reading programme is successful.

(c) The gain in reading speed is

$$\begin{aligned} \widehat{Y} &= -25.5 + 30.125(6) \\ &= 155.25 \\ &\approx 155 \text{ words/min.} \end{aligned}$$

(d) $t = \dfrac{\widehat{\beta}_1}{s/d} \Longrightarrow \dfrac{ts}{d} = \widehat{\beta}_1 \therefore s = \dfrac{d\widehat{\beta}_1}{t} = \dfrac{16(30.125)}{10.53} \approx 45.774$. Thus, $\widehat{\sigma}^2 = s^2 = 2\,095.257$.

(e)

$$\left(\widehat{\beta}_0 + \widehat{\beta}_1 X\right) \pm t_{\alpha/2;n-2} \times S\sqrt{1 + \frac{1}{n} + \frac{\left(X - \overline{X}\right)^2}{\sum \left(X_i - \overline{X}\right)^2}}$$

$$155.25 \pm 2.228 \times (45.774)\sqrt{1 + \frac{1}{12} + \frac{4}{16}}$$

$$155.25 \pm 117.76$$

$$(37.49; 273.01)$$