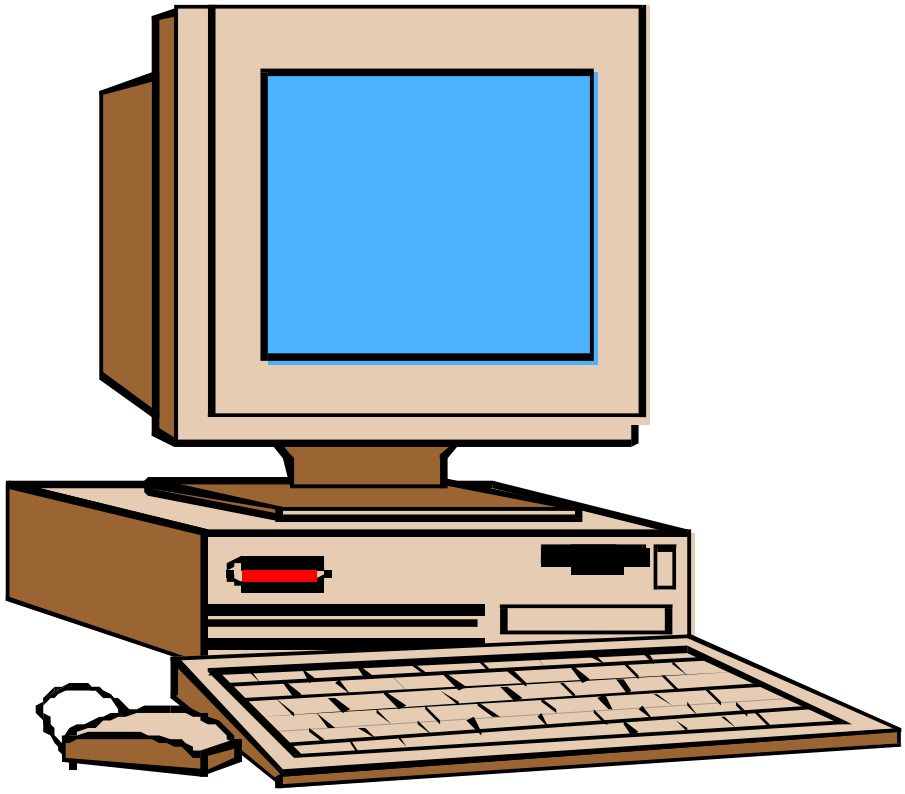
Department of Statistics

STA2601
Applied Statistics II



Study Guide

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

Introduction

Welcome to STA2601. If you are a student at the College of Science, Engineering and Technology, the four modules STA2601, STA2602, STA2603 and STA2604 form the second-year modules in statistics. The module is the followup on the module STA1502 (*Statistical Inference I*). The name *Applied Statistics* was chosen because of its double meaning: Data analysis is in effect *applied* statistical theory and you will learn how to *apply* the statistical software package JMP. This means that you **must have access to a suitable computer** for a component of practical work. (Please read carefully through the section "Role of computers and statistical calculators" following below.)

This module forms part of the new statistics curriculum and it will equip you with a proper basis in statistical knowledge, introduce you to a statistical package and highlight the value of thorough statistical know-how that the business and outside world require of students who major in Statistics! Knowledge of statistics will enable you to conduct quantitative research and statistical literacy will enable you to understand research reports you might encounter as a scientist in your everyday life or enable you to understand statistical reports you might encounter as a manager in your business.

There will be times when you feel frustrated and discouraged and then only your attitude will pull you through!

Learning outcomes

At the end of each study unit we will list the learning outcomes for that unit but there are also very specific overall outcomes for this module which we list below. Throughout your study of this module you must come back to this page, sit back and reflect upon these outcomes, think them through, digest them and feel confident in the end that you have mastered them.

- Describing various probability distributions and illustrating their applications as probabilities associated with critical values from tables.
- Describing desirable properties of estimators for population parameters and deriving these estimators through the methods of maximum likelihood and least squares.
- Evaluating the reliability of estimates of the population parameters by means of the sampling distributions of the corresponding sample statistics.
- Describing the behaviour of sample statistics (eg the sample mean, the sample variance et cetera) in repeated sampling focusing on various sampling distributions.
- Considering point and interval estimators for single or compound population parameters.

- Testing for normality by employing various tests (eg testing for skewness and for kurtosis, normal quantile plots et cetera).
- Statistical estimation and hypothesis testing involving population variances, means, correlation coefficients and regression coefficients.

The prescribed textbook(s)

You have to **buy** the following **prescribed textbook:** *Sall, J, Creighton, L and Lehman, A.* (2007 **fourth edition or any later edition**) **JMP**^{TM} **Start Statistics,** (ISBN 978-1-59994-572-9) Cary, NC: SAS Institute Inc.

This is the official handbook for JMP, the powerful statistical software developed by the **SAS Institute**. You will be instructed to study *specific sections from specific chapters,* and it is a guide book that you will use for *more than one module*, in other words for whatever statistical techniques you might encounter at different levels of your studies in statistics. (This includes modules such as STA2602, *Statistical Inference II,* STA2604, *Forecasting II,* and even for postgraduate modules such as STA4806, *Advanced Research Methods in Statistics.)*

You should also **buy** the following **prescribed book of tables:** *Stoker, DJ. (1977* 3rd *edition)* **Statistical tables,** Academica, Pretoria.

Feel free to use any other book of tables, but then it is up to you to find the correct table for a given problem.

The study guide, the textbook and the workbook

Your formal study material consists of a study guide, a textbook and a workbook **which are intertwined and together they cover the syllabus.** The study guide is more than what its name implies: it contains the major part of the theoretical contents of the course and it also serves as a guide through the textbook in a systematic way. There is a **separate workbook** which will provide you with an opportunity to apply your knowledge of the material that is covered in the guide and textbook. For each separate study unit you should first study the work in the study guide and/or textbook and then utilise the workbook to assess your progress, test your knowledge and prepare for the examination.

The *workbook* serves as an *interactive workbook,* where spaces are provided for your convenience. Should you so prefer, you are welcome to write and reference your solutions in your own book or file, if the space we supply is insufficient or not to your liking. The workbook will also serve as a kind of manual for beginners to help you with the computer exercises.

You will find the study of this module very unrewarding if you do not work actively through the workbook.

You should make sure that you receive both the study guide and the workbook. You cannot do your assignments without the workbook. Please feel free to give us feedback on any aspect of any study unit in the workbook. Negative feedback will motivate us to rectify what is wrong and positive feedback will give us inspiration to complete the workbook.

Study units and workload

We realise that you might feel overwhelmed by the volumes and volumes of printed matter that you have to absorb as a student! How do you eat an elephant? Bite by bite! Make very sure about the sections of the textbook in each study unit since some sections of the textbook are not included and we do not want you frustrated by working through unnecessary work. The study units vary in length but you should try to spend *on average 12 hours on each unit*. Practically everybody should be able to do statistics. It depends on the amount of TIME you spend on the subject. Regular contact with statistics will ensure that your study becomes personally rewarding.

Try to work through as many of the exercises and activities as possible

Doing exercises on your own will not only enhance your understanding of the work, but it will give you confidence as well. *Feedback* is given immediately after each activity in the workbook to help you check whether you understand the specific concept. The activities are designed (ie specific exercises are selected) so that you can reflect on a concept discussed in the study guide. You can only derive maximum benefit from this activity-feedback process if you discipline yourself *not to peep at the solution before you have attempted it on your own*! You should also not misuse it by merely glancing at sections needed for similar questions in the assignments.

Role of computers and statistical calculators

The emphasis in the study guide is well beyond the arithmetic of calculating statistics and the focus is on the identification of the correct technique, interpretation and decision making. This is achieved by a flexible design giving both manual calculations and computer steps. The statistical software package will give you the feeling that you are really practising statistics. I give the following quote from the textbook: "*If you give someone a large truck, they will find someone to drive it for them. But if you give them a sports car, they will learn to drive it themselves. Believe that statistics can be interesting and reachable so that people will want to drive that vehicle.*"

We try our best to illustrate every statistical technique that needs **computation** in a two-step approach:

Step 1 MANUALLY

Step 2 JMP

It is a good idea that you initially go through the laborious manual computations to enhance your understanding of the principles and mathematics. However, you must be able to manage the JMP

computations because using computers reflects the real world outside. The additional advantage of using a computer is that you can do calculations for larger and more realistic data sets.
It is impossible and impractical to do assessment of computer skills on computers in the examination but it does not preclude us from providing you with printed output which you have to interpret.
We will give you definite instructions on where and how to use a computer for your calculations in assignments. You must be able to use both a computer program and a statistical calculator as tool for your calculations. However, the emphasis in this module will always be on the interpretation and how to articulate the results.

Licence agreement

Unisa has a campus licence to supply one CD (a student version of JMP) free of charge to every student enrolled for STA2601. This is for your academic use only and you are not allowed to make copies of this product. Your licence will automatically expire after one year.

(This CD is included with your study material when you register at Unisa.)

You will be instructed in a tutorial letter on how to update your licence.

Access to a suitable computer

For the smooth running of JMP 8 you will need the following hardware:

CPU: At least a Pentium II or equivalent processor

RAM: 128 MB minimum, 256+ MB recommended

Drive space: 110 MB minimum

For your PC operating system JMP 8 requires:

Windows NT 4.X with service pack

or Windows 2000

or Windows XP

or Windows 7

Please note: JMP 8 will not run on Windows 98 or ME. **It is not compatible with Vista.**

Something about the author(s)

This study guide is a second revision by *Ms Suwisa Muchengetwa* after the major revision by *Dr Reina Nieuwoudt* of the previous STA203-N guide which was compiled by *Prof FE Steffens*, who has now retired.
As this is an applied statistics course, it needs continuous improvement since we are living in a dynamic world. A graduate of statistics needs to know about analysing data using statistical packages. It is a dream the authors shared to equip modern students interested in the world around them with the know-how to use a statistical package.

STUDY UNIT 1

Revision of statistical distributions

1.1 Introduction

The first study unit is designed to provide you with some background knowledge and to summarise what prior knowledge we assume you to have. These results are important building blocks and we often refer to them in the study units to follow. **You will not be examined explicitly on this section as the emphasis of this module is on applied statistics.**

A successful practical statistician must be courageous. Whereas a theoretical statistician can simply declare $X_1,\ X_2,\ ...,\ X_n$ to be independently normally distributed with mean μ and variance σ^2, the practical statistician has to worry about these assumptions: are they valid for my data, and what if they are not valid but approximately so?

Before the statistician can proceed with the analysis, he or she has to make a decision about this. In this respect one should avoid the two extremes: those who do not worry about the appropriateness of the analysis at all, who simply shove the data into the computer and believe what the computer says (also known as cookbook statisticians) and those who worry so much about the assumptions that they never get to analyse the data.

If a complex set of data is given to four highly skilled practical statisticians, then they could come up with four different analyses. Not that only one of them is correct and the others wrong! They will have analysed different aspects of the problem, and often such analyses are complementary. Combined it could lead to greater insight into the practical problem.

That is what makes Applied Statistics so exciting. It is not a rigid system, but allows an inventive person to use his or her originality to the full.

We trust that you will experience the thrill of practical statistics when you use JMP to enter data sets, perform the analyses and draw the final conclusions. Throughout this module, whenever a new technique is explained you should concentrate on the two aspects: what does it assume and what does it hope to achieve?

1.2 General concepts of continuous and discrete distributions

Many students get confused by statistics because different authors use different notation. In this study guide X will denote a random variable, and x a value assumed by X.

Definition 1.1

With every random variable X is associated a *distribution function* $F_X(x)$ which is defined as follows for all x:

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

The two main types of distribution functions are *discrete* and *continuous* distribution functions.

Discrete distributions

A discrete distribution function typically appears as in figure 1.1.

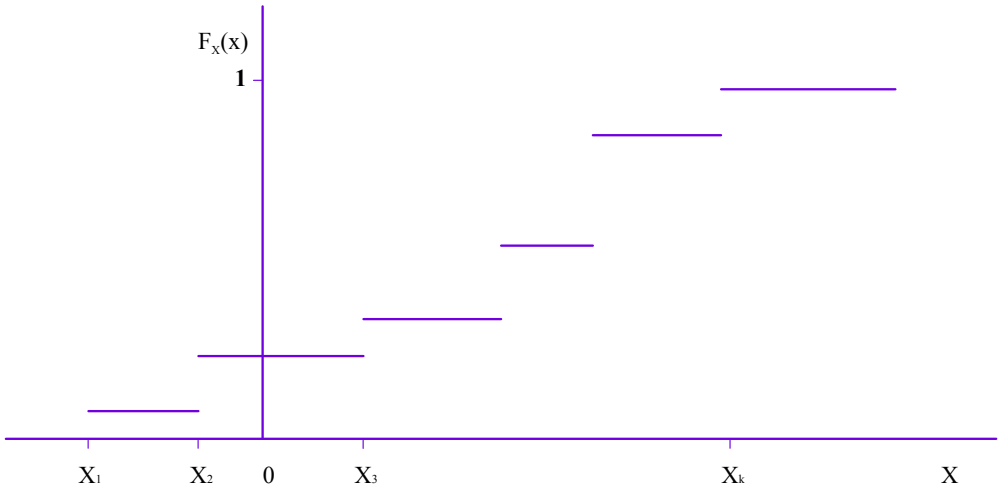


Figure 1.1: A discrete distribution function

The properties of a discrete distribution function are:

(a) $F_X(-\infty) = 0$ and $F_X(+\infty) = 1$.

(b) If $a > b$ then $F_X(a) \geq F_X(b)$, ie $F_X(x)$ is nondecreasing.

(c) $F_X(x)$ has jumps at a number of points called the discrete points of the distribution. A distribution can have at most countably many discrete points.

(d) $F_X(x)$ remains constant between the discrete points.

(e) The size of the jump at a discrete point x_i is the probability that the random variable X will assume the value x_i:

$$\begin{aligned} P(X = x_i) &= P(X \leq x_i) - P(X < x_i) \\ &= F_X(x_i) - F_X(x_i-) \end{aligned}$$

where $F_X(x_i-) = \lim_{\varepsilon \to 0} F_X(x_i - \varepsilon)$ (ε a small positive number).

Suppose now that $\mathcal{A}$ is the set of discrete points of X,

ie $P(X = x) > 0$ if $x \in \mathcal{A}$

$= 0$ if $x \notin \mathcal{A}$.

As was indicated before, $\mathcal{A}$ can have either finitely many or countably many elements.

Definition 1.2

The *probability function* of X is defined as

$$f_X(x) = P(X = x).$$

$f_X(x)$ has the properties:

(a) $f_X(x) \geq 0$ for all x.

(b) $\sum_{x \in \mathcal{A}} f_X(x) = 1$.

Moments and other special coefficients of a discrete variable X

The r-th central *moment* of X is computed as

$$\mu_r = E(X - \mu)^r$$

where μ is the *mean* or *expected value* of X.

$$\mu = E(X) = \sum_{x \in \mathcal{A}} x f_X(x)$$

μ_2 is called the variance of X and denoted as σ^2.

$$\sigma^2 = E(X - \mu)^2 = \sum_{x \in \mathcal{A}} (x - \mu)^2 f_X(x)$$

The *third central moment* of X is

$$\mu_3 = E(X - \mu)^3 = \sum_{x \in \mathcal{A}} (x - \mu)^3 f_X(x).$$

The *fourth central moment* of X is

$$\mu_4 = E\left(X-\mu\right)^4 = \sum_{x\in\mathcal{A}} \left(x-\mu\right)^4 f_X\left(x\right).$$

From the last two central moments we define the following two special coefficients:

The *coefficient of skewness* of X is

$$\beta_1 = \frac{\mu_3}{\sigma^3}.$$

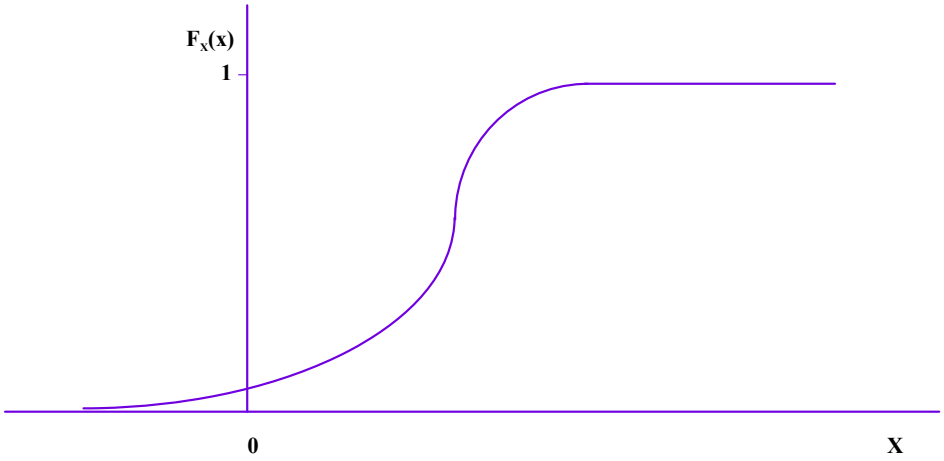
The *coefficient of kurtosis* of X is

$$\beta_2 = \frac{\mu_4}{\sigma^4}.$$

Continuous distributions

The distribution function $F_X\left(x\right) = P\left(X \leq x\right)$ of a continuous random variable X appears typically as follows:

(a)



(b)

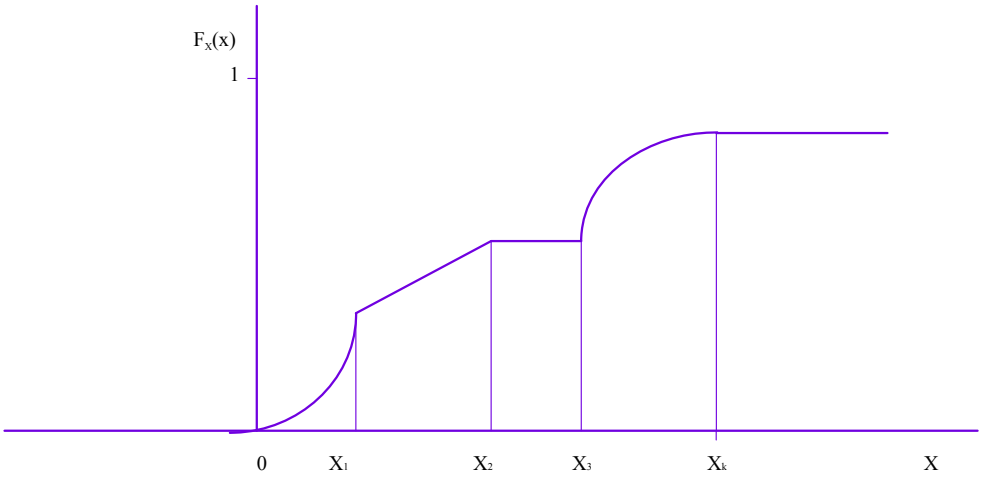


Figure 1.2: A continuous distribution function

A continuous distribution function has the following properties:

(a) $F_X(-\infty) = 0$ and $F_X(+\infty) = 1$.

(b) $F_X(x)$ is nondecreasing, ie if $a > b$ then $F_X(a) \geq F_X(b)$.

(c) $F_X(x)$ has no jumps, ie $F_X(x-) = F_X(x+)$ for all x, or $\lim_{\varepsilon \to 0} F_X(x - \varepsilon) = \lim_{\varepsilon \to 0} F_X(x + \varepsilon)$ for all x.

(d) $F_x(x)$ may have bend points (like $x_1,\ x_2,\ x_3$ and x_4 in figure 1.2(b) and it may remain constant in certain intervals, (eg between x_2 and x_3 in figure 1.2(b)). $F_X(x)$ can have at most countably many bend points.

Since $F_X(x)$ has no jumps, and at most countably many bend points, the derivative

$$F'_X(x) = \frac{dF_X(x)}{dx}$$

exists for all x except in the bend points.

Definition 1.3
The *probability density function* (pdf) $f_X(x)$ is defined as

$$f_X(x) = F'_X(x) = \frac{d}{dx} F_X(x).$$

$f_X(x)$ may have any arbitrary value if the derivative does not exist, eg

$$f_X(x) = 0$$

or

$$f_X(x) = \lim_{\varepsilon \to 0} \frac{F_X(x + \varepsilon) - F_X(x)}{\varepsilon} \qquad \text{(the derivative from the right)}$$

or

$$f_X(x) = \lim_{\varepsilon \to 0} \frac{F_X(x) - F_X(x - \varepsilon)}{\varepsilon} \qquad \text{(the derivative from the left)}$$

Probabilities concerning X are computed as follows:

If $a < b$ then $P(a < X \leq b) = F_X(b) - F_X(a)$

$$= \int_a^b f_X(x)\, dx.$$

This probability may be regarded as the area under the probability density function between $x = a$ and $x = b$ as in figure 1.3.

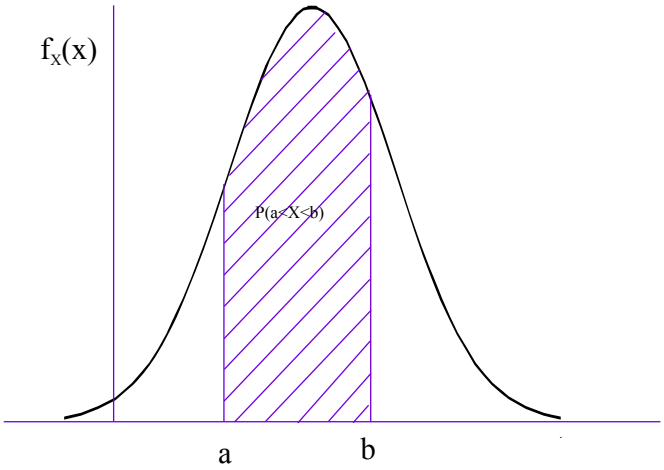


Figure 1.3: Probability equals area

Note that $P(X = a) = P(a \leq X \leq a)$

$$= \int_a^a f_X(x)\,dx$$

$$= 0$$

ie the probability that X assumes any specific value is 0. In this case (ie in the case of continuous random variables), we have

$$P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b).$$

Moments and other special coefficients of a continuous variable X

The r-th *central moment* of X is computed as $\mu_r = \int_{-\infty}^{\infty} (x-\mu)^r f_X(dx)$ where the *mean* or *expected value* of X is

$$\mu = E(X) = \int_{-\infty}^{\infty} x f_X(x)\,dx.$$

μ_2 is called the *variance* of X and is denoted as σ^2.

$$\sigma^2 = E(X-\mu)^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f_X(x)\,dx$$

The *third* and *fourth central moments* are

$$\mu_3 = E(X-\mu)^3 = \int_{-\infty}^{\infty} (x-\mu)^3 f_X(x)\,dx$$

$$\mu_4 = E(X-\mu)^4 = \int_{-\infty}^{\infty} (x-\mu)^4 f_X(x)\,dx.$$

From these two central moments we define the following two special coefficients:

The *coefficient of skewness* of X is

$$\beta_1 = \frac{\mu_3}{\sigma^3}$$

and the *coefficient of kurtosis* of X is

$$\beta_2 = \frac{\mu_4}{\sigma^4}.$$

If $\beta_1 = 0$ the distribution is called symmetric;
if $\beta_1 < 0$ the distribution is called negatively skew; and
if $\beta_1 > 0$ the distribution is called positively skew. A negatively skew distribution has a long tail to the left and a positively skew distribution has a long tail to the right:

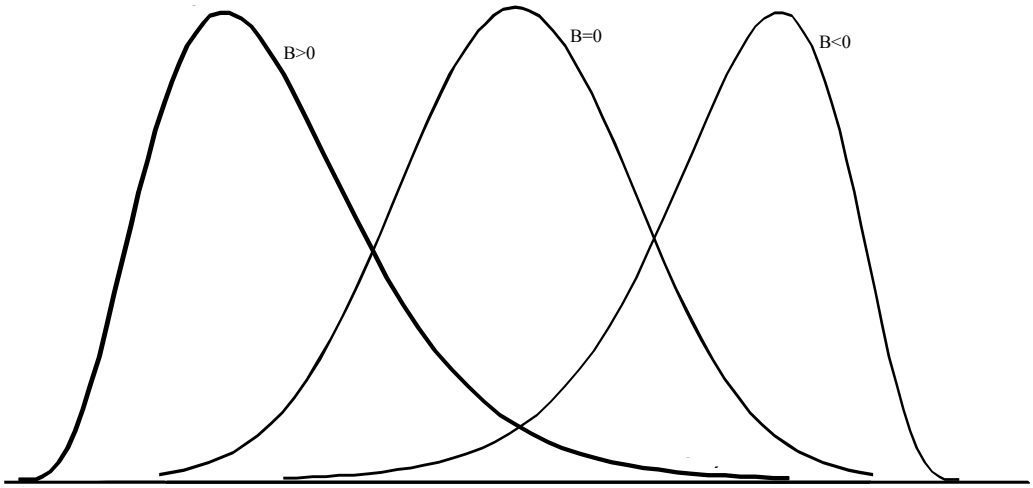


Figure 1.4: Types of skewness

Rare events

Given a random variable X with pdf $f_X(x)$, we have seen that, for given c,

$$P(X > c) = \int_c^{\infty} f_X(x)\, dx.$$

Suppose a very small value a has been chosen between 0 and 1, eg $\alpha = 0.05$ and the corresponding value of c calculated such that

$$P(X > c) = \alpha.$$

If a value of X is found which is larger than c, we say a *rare event* (or unlikely event) has occurred, ie an event with a small probability. Likewise we say a rare event has occurred if a value $X < d$ has been obtained where

$$P(X < d) = \int_{-\infty}^{d} f_X(x)\, dx = \alpha.$$

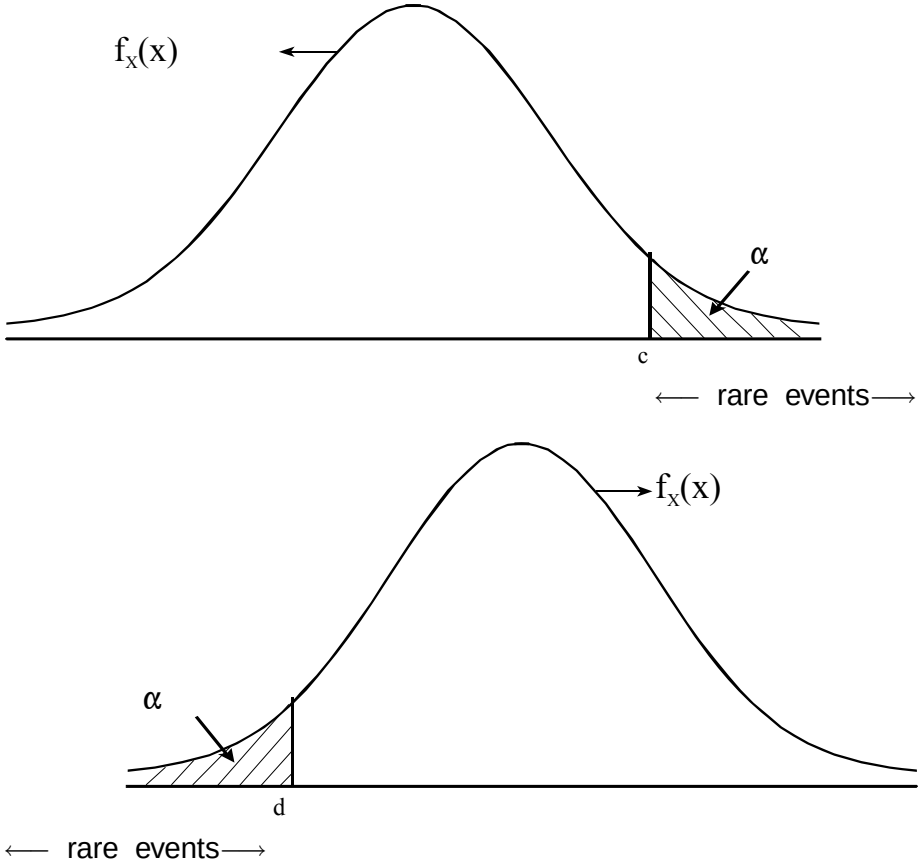


Figure 1.5: Rare events

Bivariate distributions

Sometimes one is interested in studying a number of variables jointly; their joint distribution may contain some information which is not available if they are studied separately. In this section some theory of bivariate distributions is given but this is treated in detail in STA2603. (In the next section we generalise it to multivariate distributions.)

Definition 1.4
Let X_1 and X_2 be two random variables. If a function $f_{X_1;X_2}(x_1;\ x_2)$ exists such that

$$P(X_1 \leq a_1;\ X_2 \leq a_2) = \int_{-\infty}^{a_2}\int_{-\infty}^{a_1} f_{X_1;X_2}(x_1;\ x_2)\,dx_1dx_2$$

for all a_1 and a_2, then $f_{X_1;X_2}(x_1;\ x_2)$ is called the *joint probability density function* of X_1 and X_2.

$f_{X_1;X_2}(x_1;\ x_2)$ has the following characteristics:

(a) $f_{X_1;X_2}(x_1;\ x_2) \geq 0$ for all x_1 and x_2.

(b) $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X_1X_2}(x_1;\ x_2)\,dx_1dx_2 = P(X_1 \leq \infty;\ X_2 \leq \infty) = 1.$

Definition 1.5
The function

$$F_{X_1;X_2}(x_1;\ x_2) = \int_{-\infty}^{x_2}\int_{-\infty}^{x_1} f_{X_1;X_2}(u;\ v)\,dudv = P(X_1 \leq x_1;\ X_2 \leq x_2)$$

is called the *joint distribution function* of X_1 and X_2.

Note that

$$f_{X_1;X_2}(x_1;\ x_2) = \frac{\partial^2}{\partial x_1 \partial x_2} F_{X_1;X_2}(x_1;\ x_2)$$

ie $f_{X_1;X_2}(x_1;\ x_2)$ is the second order partial derivative of $F_{X_1;X_2}(x_1;\ x_2)$ with respect to x_1 and x_2.

Definition 1.6
The function

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} f_{X_1;X_2}(x_1;\ x_2)\,dx_2$$

is called the *marginal probability density function* of X_1.

Likewise,

$$f_{X_2}(x_2) = \int_{-\infty}^{\infty} f_{X_1;X_2}(x_1;\ x_2)\,dx_1$$

is called the marginal probability density function of X_2.

Definition 1.7
The function

$$f_{X_1|x_2}(x_1;\ x_2) = \frac{f_{X_1;X_2}(x_1;\ x_2)}{f_{X_2}(x_2)}$$

is called the *conditional probability density function* of X_1 given that $X_2 = x_2$.

The conditional pdf of X_2 given that $X_1 = x_1$ is defined in a similar manner.

Definition 1.8
The *conditional expectation* of the random variable X_1 given that $X_2 = x_2$, is defined as

$$E[X_1|\,X_2 = x_2] = \int_{-\infty}^{\infty} x_1 f_{X_1|x_2}(x_1;\ x_2)\,dx_1.$$

The conditional expectation of X_1 given that $X_2 = x_2$ is also called the *regression function* of X_1 on X_2.

The roles of X_1 and X_2 can of course be reversed in the above discussion.

Definition 1.9

$$E(X_1 - \mu_1)(X_2 - \mu_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x_1 - \mu_1)(x_2 - \mu_2) f_{X_1,X_2}(x_1;\ x_2)\, dx_1 dx_2$$

$$= E(X_1 X_2) - E(X_1) E(X_2)$$

is called the *covariance* of X_1 and X_2.

Definition 1.10

$$\rho = \frac{Cov(X_1, X_2)}{\sqrt{Var(X_1) Var(X_2)}}$$

is called the *correlation coefficient* between X_1 and X_2 if $Var(X_1) \neq 0$ and $Var(X_2) \neq 0$.

The correlation coefficient is a quantity which lies between -1 and 1.

Definition 1.11

If the correlation coefficient between two random variables is zero, the two variables are said to be *uncorrelated*.

NB From the definition ρ it follows that $\rho = 0$ if and only if the covariance is zero.

Definition 1.12

The random variables X_1 and X_2 are said to be *independent* if their joint pdf can be factorised:

$$f_{X_1;X_2}(x_1;\ x_2) = f_{X_1}(x_1) f_{X_2}(x_2).$$

The term independent has this special technical meaning when used in connection with random variables. Remember this! The following theorem is given without proof.

Theorem 1.1
If X_1 and X_2 are independent random variables then they are uncorrelated.

The converse is *not true*: two uncorrelated random variables need not be independent. This is very important.

Multivariate distributions

Generalisation of the bivariate distribution theory of the previous section is straightforward, and will not be done in detail here. The multivariate pdf of the n random variables $X_1,\ ...,\ X_n$ is the function

$$f_{X_1;X_2;...;X_n}(x_1;\ x_2;\ ...;\ x_n)$$

such that

$$P(X_1 \leq a_1;\ ...;\ X_n \leq a_n) = \int_{-\infty}^{a_n} ... \int_{-\infty}^{a_1} f_{X_1;...;X_n}(x_1;\ ...;\ x_n)\, dx_1 ... dx_n$$

for all $a_1,\ ...,\ a_n$.

The marginal pdf of X_1 is

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} f_{X_1;...;X_n}(x_1;\ ...;\ x_n)\, dx_2 dx_3 ... dx_n.$$

Similarly

$$f_{X_2}(x_2) = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} f_{X_1;...;X_n}(x_1;\ ...;\ x_n)\, dx_1 dx_3 ... dx_n,$$ et cetera

Definition 1.13
The random variables $X_1,\ ...,\ X_n$ are said to be *mutually independent* if

$$f_{X_1;...;X_n}(x_1;\ ...;\ x_n) = f_{X_1}(x_1) ... f_{X_n}(x_n)$$

for all values of $x_1;\ ...;\ x_n$.

An important special case is the following:

Let $X_1,\ ...,\ X_n$ be independent and identically distributed, ie

$$f_{X_1}(x) = f_{X_2}(x) = ... = f_{X_n}(x) = f_X(x),$$

say, for all x, and

$$f_{X_1;...;X_n}(x_1;\ ...;\ x_n) = f_X(x_1)...f_X(x_n)$$

then $X_1,\ ...,\ X_n$ are said to constitute a *random sample* of size n from the distribution with pdf $f_X(x)$.

1.3 Standard distributions

In this section a number of standard distributions are dealt with. These distributions are very important in statistical applications. The **binomial** and **Poisson** distributions are *discrete distributions*, while the **normal**, **chi-square**, $\mathbf{t}$- and F**-distributions** are *continuous distributions*. The bivariate normal distribution is an example of a continuous bivariate distribution.

You should remember the mathematical formulae for the binomial, Poisson, normal and bivariate normal distributions; it is not imperative that you memorise the probability density functions of the chi-square, t- and F-distributions. A random variable will sometimes be called a *variate* in this section.

The following book of tables is referred to in this section. This book is prescribed for this module.

DJ. Stoker: *Statistical tables*, Academia, Third Edition, 1977.

Bernoulli trials

Suppose the outcome of a random experiment is either a *success* or a *failure*. For example, if a patient is operated on, he or she may either recover (success) or die (failure). If we select a person at random and ask him or her whether he or she smokes, he or she may either say "Yes" (success) or "No" (failure). The labelling of one possible outcome as "success" and the other as "failure" is of course arbitrary, and may be switched according to the context of the problem.

The following mathematical model is used to describe such an experiment. Let the probability of a success be π where π is a constant with $0 < \pi < 1$. Define the random variable X as follows:

$X = 0$ if the outcome is a failure

$= 1$ if the outcome is a success.

$\therefore P(X = 1) = \pi$ and $P(X = 0) = 1 - \pi$.

Then $E(X) = \pi(1) + (1 - \pi)(0) = \pi$

and $E\left(X^2\right) = \pi(1)^2 + (1 - \pi)(0)^2 = \pi$

$\therefore Var(X) = E\left(X^2\right) - (E(X))^2 = \pi - \pi^2 = \pi(1 - \pi)$.

An experiment of this type is called a Bernoulli trial (named after the Swiss mathematician, Jacques Bernoulli (1664-1705)) and X is called a Bernoulli variate.

The binomial distribution

The binomial distribution was also derived by Jacques Bernoulli. Suppose a Bernoulli experiment is repeated n times, such that the outcomes $X_1,\ X_2,\ ...,\ X_n$ are *independent* Bernoulli variates with the *same* probability π of a success. The implications of these two assumptions, independence and constant probability of a success, are important. If a random sample is drawn from a finite population these conditions may hold if sampling is done with replacement. However, if the sample is drawn without replacement, the proportion of the population having the "success" property changes after each draw and the outcome of one draw depends on the outcomes of the previous draws. However, if the population is very large and the sample size n relatively small, the conditions of independence and constant probability of a success are approximately satisfied, and the model described here will be a good approximation to the true situation.

We are interested in the number of successes in the sample of size n. Let

$$Y = X_1 + X_2 + ... + X_n,$$

then Y represents the number of successes in the sample.

Definition 1.14

Y is a binomial variate, denoted by $Y \sim b(n;\ \pi)$ if

$$P(Y = y) = \binom{n}{y}\pi^y(1 - \pi)^{n-y};$$

where $y = 0;\ 1;\ 2;\ ...;\ n$ and $0 < \pi < 1$.

It can be shown that

$$E(Y) = n\pi \quad \text{and} \quad Var(Y) = n\pi(1-\pi).$$

Binomial variates have the following important *additive* property: If $Y_1 \sim b(n_1;\ \pi)$ and $Y_2 \sim b(n_2;\ \pi)$ and if Y_1 and Y_2 are independent, then $Y_1 + Y_2 \sim b(n_1 + n_2;\ \pi)$.

This property follows simply from the fact that, if Y_1 is the number of successes in n_1 independent Bernoulli trials and Y_2 is the number of successes in n_2 independent Bernoulli trials, if these $n_1 + n_2$ trials are mutually independent and the probability of a success is π throughout the $n_1 + n_2$ experiments, then $Y_1 + Y_2$ is the number of successes in $n_1 + n_2$ independent Bernoulli trials.

Table XI of Stoker gives the **cumulative binomial distribution**: $P(Y \leq y)$ for a given n and π. (In table XI r is used instead of our y and p instead of our π), ie

$$P(Y = r) = \binom{n}{r} p^r (1-p)^{n-r}$$

and

$$P(Y \leq r) = \sum_{k=0}^{r} \binom{n}{k} p^k (1-p)^{n-k}.$$

Individual probabilities are obtained by subtraction.

Example 1.1

Let $X \sim b(10;\ 0.4)$, then

$$P(X < 7) = P(X \leq 6) = 0.9452 \qquad \text{(table XI)}$$

$$P(X < 6) = P(X \leq 5) = 0.8338 \qquad \text{(table XI)}$$

$$P(X \geq 7) = P(X > 6) = 1 - P(X \leq 6) = 0.0548$$

$$P(X \geq 6) = P(X > 5) = 1 - P(X \leq 5) = 0.1662$$

$$P(X = 6) = P(X \leq 6) - P(X \leq 5) = 0.1114$$

$$P(3 \leq X \leq 8) = P(X \leq 8) - P(X \leq 2) = 0.9983 - 0.1673 = 0.8310 \qquad \text{(table XI).}$$

Note

Table XI gives the cumulative distribution of a $b(n;\ \pi)$ variate for $0 < \pi \leq \frac{1}{2}$. In the case $\pi > \frac{1}{2}$ we could interchange the successes and failures: if $X \sim b(n;\ \pi)$, ie if X is the number of successes in n trials with probability of success π, then $n - X$ is the number of failures; if we let $Y = n - X$ then Y may be regarded as the number of successes in n trials with probability of success $1 - \pi$, ie $Y \sim b(n;\ 1 - \pi)$.

Example 1.2

Let $X \sim b(12;\ 0.7)$. Find

(i) $P(X = 7)$

(ii) $P(3 < X < 7)$.

Solution

Let $Y = 12 - X$; then $Y \sim b(12;\ 0.3)$ and therefore probabilities concerning Y are catered for in table XI.

(i) $$\begin{aligned} P(X = 7) &= P(12 - X = 12 - 7) \\ &= P(Y = 5) \\ &= P(Y \leq 5) - P(Y \leq 4) \\ &= 0.8821 - 0.7237 \\ &= 0.1584 \end{aligned}$$

(ii) $$\begin{aligned} P(3 < X < 7) &= P(-7 < -X < -3) \\ &= P(12 - 7 < 12 - X < 12 - 3) \\ &= P(5 < Y < 9) \\ &= P(5 < Y \leq 8) \\ &= P(Y \leq 8) - P(Y \leq 5) \\ &= 0.9983 - 0.8821 \\ &= 0.1162 \end{aligned}$$

The Poisson distribution

Definition 1.15
Let X be a discrete random variable with probability function

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} \qquad x = 0;\ 1;\ 2;\ ...$$

where $\lambda > 0$ is a constant. Then X is said to be a Poisson variate with parameter λ which we denote by $X \sim Psn(\lambda)$.

The Poisson distribution (pronounced pwa-sòn) is named after its discoverer, the French mathematician Simeon Denis Poisson, who published the distribution in 1837. The Greek letter λ is pronounced "lambda".
The Poisson distribution is used extensively in practice for the number of occurrences of an event in a given time period. The number of telephone calls which arrive at an exchange in one minute, the number of customers who arrive at a supermarket in one hour and the number of vehicles which pass a certain point in five minutes are examples of random variables which have been found to be approximately distributed as Poisson variates under certain circumstances.

Important properties of the Poisson distribution

(i) If $X \sim Psn(\lambda)$ then $E(X) = \lambda$ and $Var(X) = \lambda$ (note that the mean and the variance are the same).

(ii) If X_1 and X_2 are independent $Psn(\lambda_1)$ and $Psn(\lambda_2)$ variates respectively, then $X_1 + X_2$ is a $psn(\lambda_1 + \lambda_2)$ variate (additive property).

Table XII gives the cumulative Poisson distribution.

Example 1.3
Let $X \sim Psn(2.5)$.
Then

$$P(X \leq 6) = 0.9858 \qquad \text{(table XII)}$$

$$P(X \leq 5) = 0.9580 \qquad \text{(table XII)}$$

$$P(X = 6) = P(X \leq 6) - P(X \leq 5) = 0.9858 - 0.9580 = 0.0278$$

$$P(X \geq 6) = P(X > 5) = 1 - P(X \leq 5) = 1 - 0.9580 = 0.0420$$

$$P(3 < X < 7) = P(3 < X \leq 6) = P(X \leq 6) - P(X \leq 3)$$
$$= 0.9858 - 0.7576 = 0.2282 \qquad \text{(table XII).}$$

In STA1503 (and also in STA1501) you learned that the normal distribution is used as a limiting distribution for the binomial distribution if n is large and π is close to 0.5 but the Poisson distribution is used as a limiting function for the binomial distribution if n is large and π is small.

Example 1.4

Let $X \sim b\,(20;\ 0.05)$, the largest n and smallest π in table XI. Then X is approximately $Psn\,(\lambda)$ with

$$\lambda = n\pi = 20\,(0.05) = 1.$$

We compare probabilities from tables XI and XII.

x	$P\,(X \leq x)$ table XII	$P\,(X \leq x)$ table XI
0	0.3679	0.3585
1	0.7358	0.7358
2	0.9197	0.9245
3	0.9810	0.9841

The approximations are fair, and for larger n the approximations become much better.

The normal distribution

Definition 1.16
If X is a continuous variate with pdf

$$f_X\,(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2/\sigma^2}; \qquad -\infty < x < \infty$$

then x is said to be a *normal variate* with mean μ and variance σ^2.
We write $X \sim n\left(\mu;\ \sigma^2\right)$.

Let $Z = \dfrac{X-\mu}{\sigma}$. Then Z is a *normal variate* with mean 0 and variance 1, ie $Z \sim n\,(0;\ 1)$.

The pdf of Z is obtained by setting $\mu = 0$ and $\sigma = 1$ in the pdf of X.

Definition 1.17

If Z is a continuous variate with pdf

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}; \qquad -\infty < x < \infty,$$

then Z is a *standard normal variate*.

This pdf is often denoted by $\phi(z)$ and the corresponding distribution function by $\Phi(z)$, ie

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

$$\Phi(z) = \int_{-\infty}^{z} \phi(u)\, du.$$

These two functions are depicted in the next two graphs.

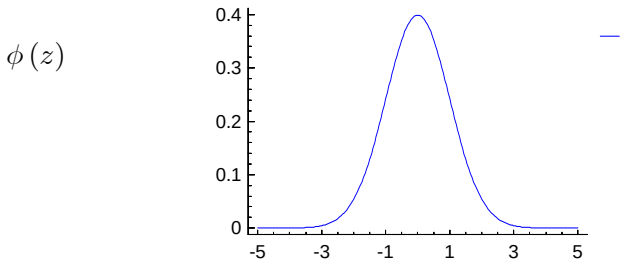


Figure 1.6:
The standardised normal probability density function

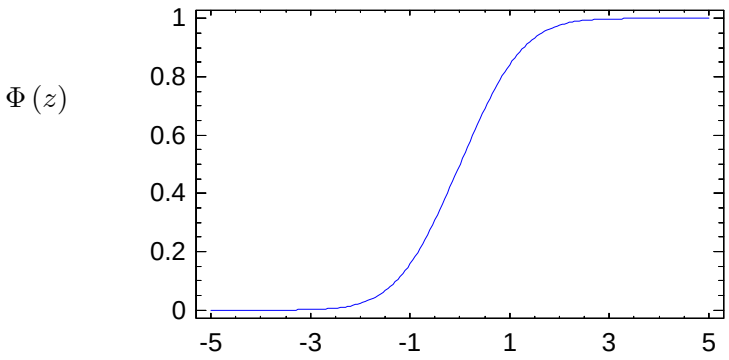
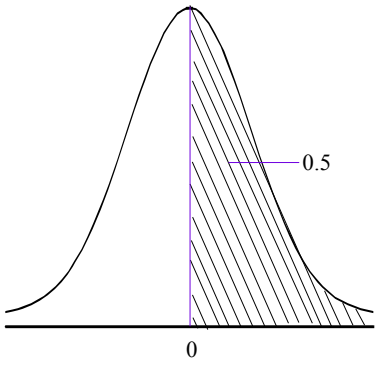


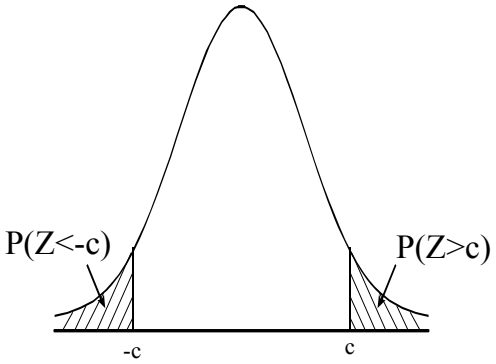
Figure 1.7:
The standardized normal distribution function

Notice that the standardised normal density function is symmetric about zero from which it follows that:

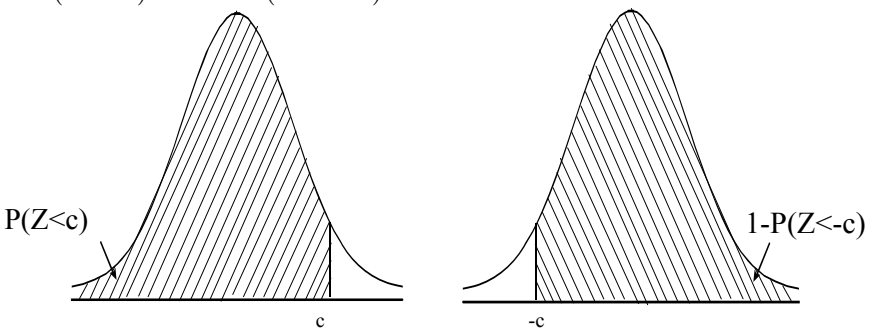
(i) $P(Z \leq 0) = P(Z \geq 0) = \frac{1}{2}$



(ii) $P(Z \geq c) = P(Z \leq -c)$



(iii) $P(Z < c) = 1 - P(Z < -c)$



In table I we find areas under the normal density function specifically $P(Z < c) = \int_{-\infty}^{c} \phi(u)\, du.$

Please note that some editions of normal tables tabulate $P(0 < Z < c) = \int_{0}^{c} \phi(u)\, du.$

These values enable us to compute other probabilities, as the following examples will show. **Always try to draw sketches which indicate the probabilities under consideration.**

Example 1.5

Let $Z \sim n\,(0;\ 1)$. Then

(a) $P\,(Z < 1.3) = 0.9032$ (table I)

(b) $P\,(Z > 1.3) = 1 - P\,(Z < 1.3) = 0.0968$

(c) $$\begin{aligned} P\,(-1.3 < Z < 1.3) &= P\,(Z < 1.3) - P\,(Z < -1.3) \\ &= P\,(Z < 1.3) - [1 - P\,(Z < 1.3)] \quad \text{from (b) above} \\ &= 2P\,(Z < 1.3) - 1 \\ &= 0.8064 \end{aligned}$$

(d) $P\,(|Z| < 1.3) = P\,(-1.3 < Z < 1.3) = 0.8064$ as above

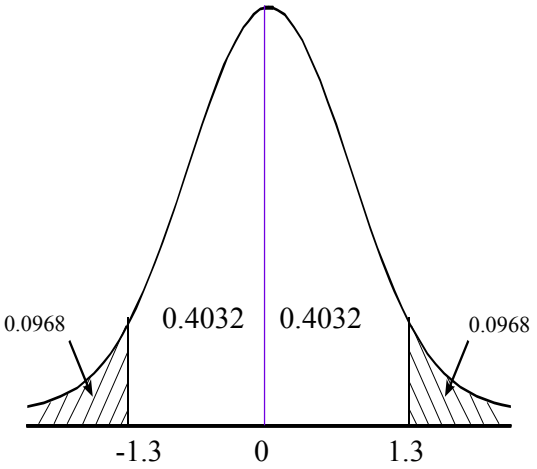
(*Please note*: $|Z|$ is the absolute value of Z, ignoring the sign of Z

$|Z| < c$ if and only if $-c < Z < c$ and

$|Z| > c$ if and only if $Z < -c$ or $Z > c$.)

(e) $P\,(Z < -1.3) = P\,(Z > 1.3) = 0.0968$ from (b)

(f) $P\,(Z > -1.3) = 1 - P\,(Z < -1.3) = 0.9032$



Furthermore

(g) $$\begin{aligned} P\,(-0.2 < Z < 1.8) &= P\,(Z < 1.8) - P\,(Z < -0.2) \\ &= P\,(Z < 1.8) - [1 - P\,(Z < 0.2)] \\ &= 0.9641 - 1 + 0.5793 \quad \text{(table I)} \\ &= 0.5434 \end{aligned}$$

(h) $P(0.55 < Z < 1.96) = P(Z < 1.96) - P(Z < 0.55)$
$= 0.9750 - 0.7088$ (table I)
$= 0.2662$

(i) $P(-1.32 < Z < -0.35) = P(Z < -0.35) - P(Z < -1.32)$
$= [1 - P(Z < 0.35)] - [1 - P(Z < 1.32)]$
$= (1 - 0.6368) - (1 - 0.9066)$
$= 0.9066 - 0.6368$
$= 0.2698$

You should acquaint yourself well with the use of table I - it is used more often by statisticians than any other table.

Example 1.6

Let $X \sim n(3;\ 16)$. Find $P(1 < X < 7)$.

Solution

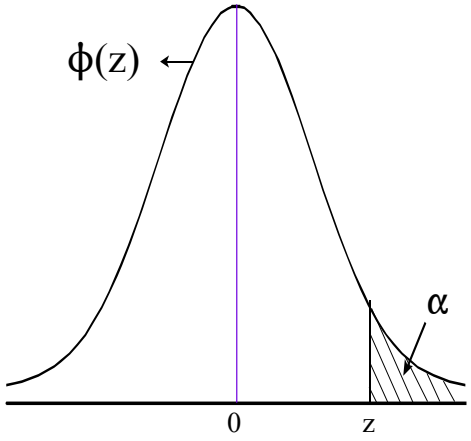
We use $Z = \dfrac{X - \mu}{\sigma}$ (where for this example $\mu = 3$ and $\sigma = \sqrt{16}$).

$$\therefore P(1 < X < 7) = P\left(\frac{1-3}{4} < \frac{X-3}{4} < \frac{7-3}{4}\right)$$
$$= P(-0.5 < Z < 1)$$
$$= P(Z < 1) - P(Z < -0.5)$$
$$= 0.8413 - (1 - 0.6915) \quad \text{(table I)}$$
$$= 0.5328$$

Table II works in an almost inverse manner, and we will use it mainly to obtain so-called critical values of the $n(0;\ 1)$ distribution. Given α, it gives z such that

$$P(Z > z) = \int_z^{\infty} \phi(u)\, du = \alpha$$

ie $P(Z < z) = \int_{-\infty}^{z} \phi(u)\,du = 1 - \alpha.$



We denote such values of Z by z_α ie it is our agreement that $P(Z > z_\alpha) = \alpha$.

We illustrate the use of table II by means of examples:

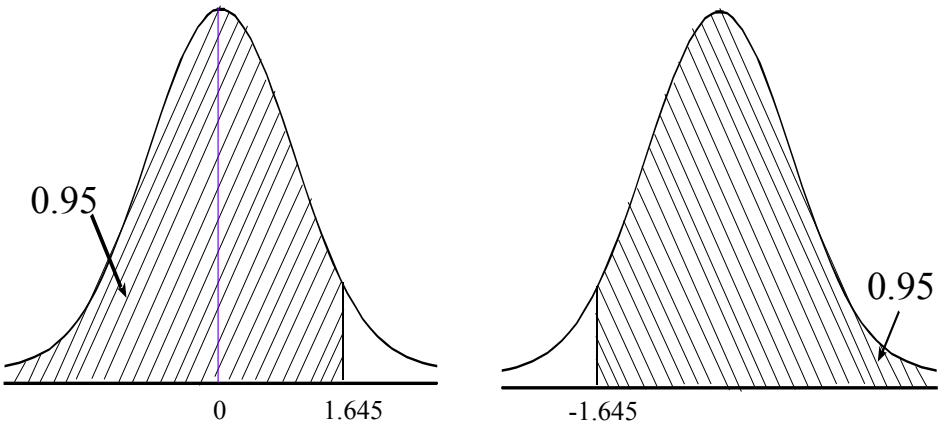
Example 1.7

(a) Find z such that $P(Z < z) = 0.95$.

(b) Find z such that $P(-z < Z < z) = 0.95$.

Solution

(a)

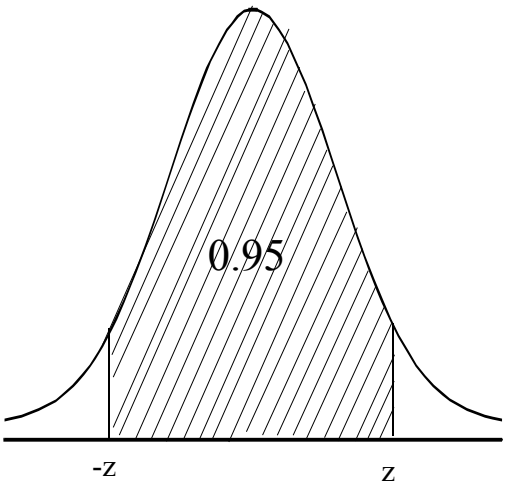


We can look up z in table II directly, because it is given that $P(Z \leq z) \equiv \Phi(z) = 0.95$, and we find $z = 1.645$.

$\therefore P(Z < 1,645) = 0.95.$

Therefore $P(Z > -1.645) = 0.95$ by symmetry.

(b) Now we cannot look up z in table II directly, because $P(-z < Z < z)$ is not $\Phi(z)$.



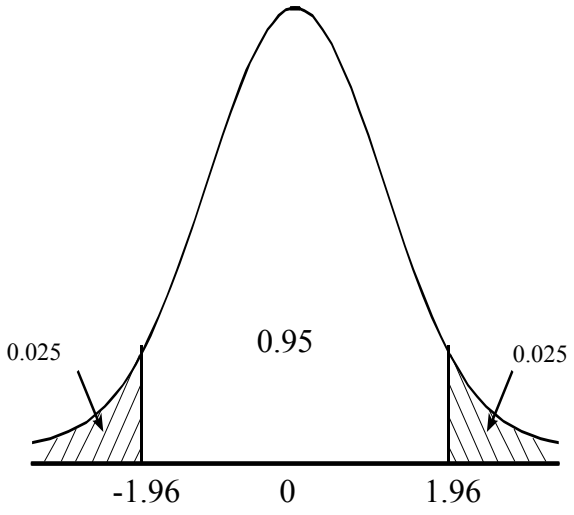
$$\begin{aligned} 0.95 &= P(-z < Z < z) \\ &= P(Z < z) - P(Z < -z) \\ &= P(Z < z) - [1 - P(Z > -z)] \\ &= P(Z < z) - 1 + P(Z < z) \\ &= 2P(Z < z) - 1 \end{aligned}$$

$\therefore 2P(Z < z) = 1.95$

$\therefore P(Z < z) = 0.975 = \Phi(z)$

$\therefore z = 1.960$ from table II

$P(-1.96 < Z < 1.96) = 0.95.$



Sums of independent normal variates

Let $X_1,\ X_2,\ ...,\ X_n$ be independent normal variates such that $X_i \sim n\left(\mu_i;\ \sigma_i^2\right)$. Let $Y = \sum_i^n c_i X_i$.

Then Y is a normal variate with mean and variance given by

$$E\left(Y\right) = \Sigma c_i \mu_i \quad \text{and} \quad Var\left(Y\right) = \Sigma c_i^2 \sigma_i^2.$$

In particular, if $\mu_1 = \mu_2 = ... = \mu_n = \mu$ and $\sigma_1^2 = \sigma_2^2 = ... = \sigma_n^2 = \sigma^2$ then

$$E\left(Y\right) = \mu \Sigma c_i \quad \text{and} \quad Var\left(Y\right) = \sigma^2 \Sigma c_i^2.$$

An important special case:

If $Y = \overline{X} = \frac{1}{n}\sum_1^n X_i$ then $E\left(\overline{X}\right) = \mu \sum_1^n \frac{1}{n} = \mu$ and $Var\left(\overline{X}\right) = \sigma^2 \sum_1^n \frac{1}{n^2} = \sigma^2/n.$

Theorem 1.2
If $X_1,\ ...,\ X_n$ are independent $n\left(\mu;\ \sigma^2\right)$ variates then

$$\overline{X} = \frac{1}{n}\sum_1^n X_i$$

is a $n\left(\mu;\ \sigma^2/n\right)$ variate.

The central limit theorem

We have seen above that, if $X_1,\ ...,\ X_n$ are independent $n\left(\mu;\ \sigma^2\right)$ variates, then $\overline{X} = \frac{1}{n}\sum_1^n X_i$ is a $n\left(\mu;\ \sigma^2/n\right)$ variate.

However, if $X_1,\ ...,\ X_n$ are independent variates from a general distribution having pdf $f_X\left(x\right)$, with mean μ and variance σ^2, then $\overline{X}$ still has mean and variance

$$E\left(\overline{X}\right) = \mu;\ Var\left(\overline{X}\right) = \sigma^2/n$$

whether the distribution of X_i is normal or not. However, in general the distribution of $\overline{X}$ is not normal. What is the distribution of $\overline{X}$ then? Unfortunately this depends on $f_X\left(x\right)$. However, the central limit theorem is of great use in this respect and reads as follows:

Theorem 1.3

Let $X_1,\ ...,\ X_n$ be independent variables from a general distribution with pdf $f_X(x)$ with mean μ and variance σ^2. Then $E\left(\overline{X}\right) = \mu$ and $Var\left(\overline{X}\right) = \sigma^2/n$ and

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \qquad \text{(provided } \sigma^2 \text{ is finite)}$$

is asymptotically normally distributed with mean 0 and variance 1.

In practical terms this means that the distribution of Z can, for large n, be approximated by the standardised normal distribution, and we may use tables I and II to obtain approximate probabilities and critical values with respect to $\overline{X}$. The condition that σ^2 must be finite is not a trivial one – there are distributions for which the variance does not even exist, for example the Cauchy distribution. The question of how large n should be before the approximation becomes satisfactory, is not easily answered. It depends on the specific $f_X(x)$.

Example 1.8

Let $X_1,\ X_2,\ ...,\ X_{36}$ be a random sample from a distribution with mean 10 and variance 25. Find an approximate value for $P\left(9 < \overline{X} < 11\right)$.

Solution

$Z = \dfrac{\overline{X} - 10}{\sqrt{25/36}} = 1.2\left(\overline{X} - 10\right)$ is approximately $n\left(0;\ 1\right)$.

$$\begin{aligned}
\therefore P\left(9 < \overline{X} < 11\right) &= P\left[1.2\left(9 - 10\right) < 1,2\left(\overline{X} - 10\right) < 1.2\left(11 - 10\right)\right] \\
&= P\left(-1.2 < Z < 1.2\right) \\
&\approx 0.7698 \qquad \text{(table I)}
\end{aligned}$$

The chi-square distribution

Definition 1.18
If Y is a random variable with pdf

$$f_Y(y) = \frac{y^{\frac{1}{2}d-1}e^{-\frac{1}{2}y}}{2^{\frac{1}{2}d}\Gamma\left(\frac{1}{2}d\right)}, \quad \text{for} \quad y > 0$$

$$= 0 \qquad\qquad \text{otherwise}$$

where d is a positive integer, then Y is said to have a **chi-square distribution** with d degrees of freedom. We write $Y \sim \chi^2_d$.

The pdf, which depends on the parameter d, is represented graphically for $d = 1; 4; 10$ and 20 in figure 1.8.

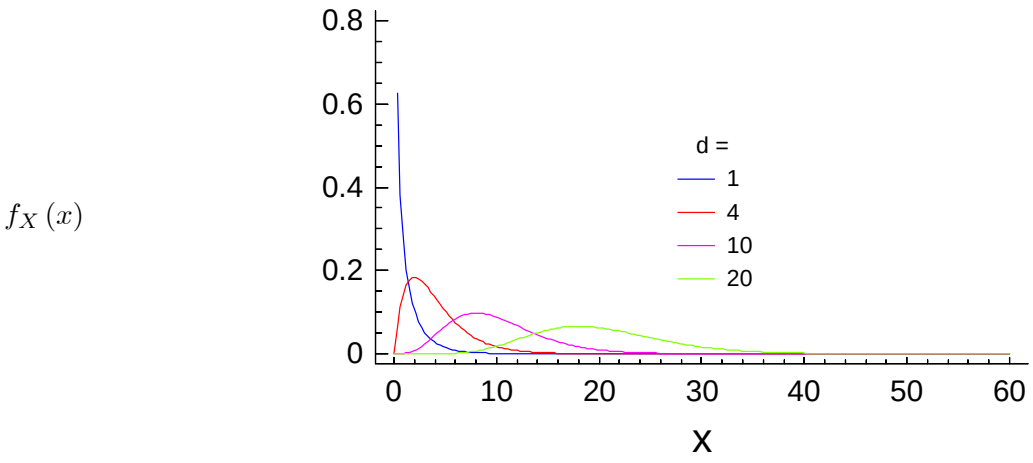


Figure 1.8:
The pdf of the chi-square distribution with d degrees of freedom

Table IV gives critical values of this distribution.

Example 1.9
Verify that, if $Y \sim \chi^2_6$, then

(a) $P(Y \geq 12.5916) = 0.05.$

(b) $P(Y \leq 10.6446) = 0.90.$

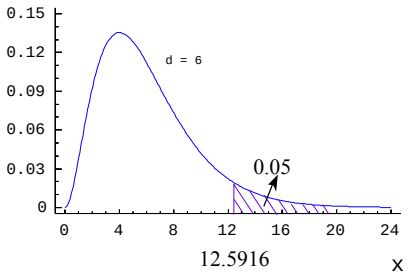
(c) $P(Y \geq 1.63539) = 0.95.$

(d) $P(Y \leq 0.872085) = 0.01.$

(e) $P(1.237347 \leq Y \leq 14.4494) = 0.95.$

Solution

(a) We can look up $P(Y \geq x) = 0.05$ directly in table IV because the one-sided exceedance probability P is given as 0.05. With $\nu = 6$ and $P = 0.05$ we therefore find $x = 12.5916$. We also write $\chi^2_{6;0.05} = 12.5916$.



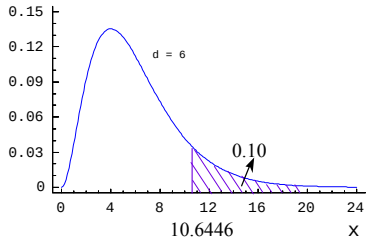
(b) $P(Y \leq x) = 0.90$

$\Rightarrow 1 - P(Y > x) = 0.90$

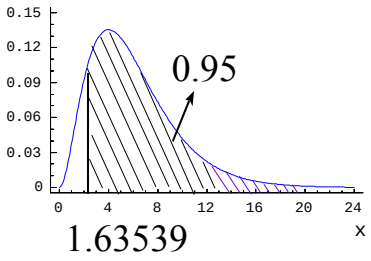
$\Rightarrow P(Y \geq x) = 0.10$

so that it follows from table IV that $x = 10.6446$.

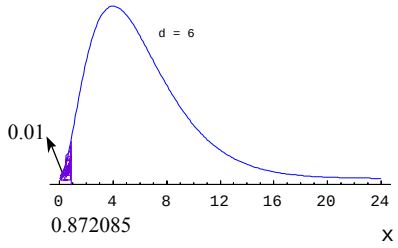
We also write $\chi^2_{6;0.10} = 10.6446$.



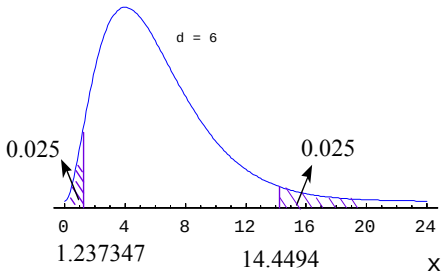
(c) We can also look up this value directly because now $P = 0.95$ and $P(Y \geq x) = 0.95$ so that it follows from table IV that $x = 1.63539$.



(d) $P(Y \leq x) = 1 - P(Y \geq x) \Rightarrow P(Y \geq x) = 0.99$ so that it follows from table IV that $x = 0.872085$.



(e)



It follows from table IV that $P(1.237347 \leq Y \leq 14.4494) = 0.975 - 0.025$

$= 0.95.$

Result 1.1
Properties of the chi-square distribution

(i) If $Y \sim \chi^2_d$, then $E(Y) = d$ and $Var(Y) = 2d$.

(ii) If Y_1 and Y_2 are independent and $Y_1 \sim \chi^2_{d_1}$ and $Y_2 \sim \chi^2_{d_2}$, then $Y_1 + Y_2 \sim \chi^2_{d_1+d_2}$ (additive property)

(iii) **Relation to normal sampling theory** (a very important result)
If $X_1, \ ..., \ X_n$ are independent $n(0; \ 1)$ variates (ie $X_1, \ ..., \ X_n$ is a random sample from a standardised normal distribution) and if

$$Y = X_1^2 + X_2^2 + ... + X_n^2$$

then $Y \sim \chi^2_n$.

(iv) **Special case** $n = 1$
If $X \sim n(0; \ 1)$ then $X^2 \sim \chi^2_1$.

Example 1.10

Let $X \sim n\,(0;\ 1)$ and $Y = X^2$. Then $Y \sim \chi^2_1$.

From table IV:

$$\begin{aligned} P\,(Y < 3.84146) &= 1 - P\,(Y \geq 3.84146) = 1 - 0.05 = 0.95 \\ \therefore P\left(X^2 < 3.84146\right) &= 0.95 \\ \therefore P\left(-\sqrt{3.84146} < X < \sqrt{3.84146}\right) &= 0.95 \\ \therefore P\,(-1.96 < X < 1.96) &= 0.95 \end{aligned}$$

which is in accordance with tables I and II.

From (iii) above we can deduce the following important result:

Result 1.2

If $X_1,\ ...,\ X_n$ are independent $n\left(\mu;\ \sigma^2\right)$ variates, then $\dfrac{X_1 - \mu}{\sigma},\ ...,$

$\dfrac{X_n - \mu}{\sigma}$ are independent $n\,(0;\ 1)$ variates and

$$Y = \sum_{i=1}^{n} \left[\frac{X_i - \mu}{\sigma}\right]^2 \sim \chi^2_n.$$

The following is also an important result. Note that we have "lost" one degree of freedom because μ was replaced by $\overline{X}$ and we already know that $\overline{X} \sim n\left(\mu;\ \sigma^2/n\right)$.

Result 1.3

Let $X_1,\ ...,\ X_n$ be independent $n\left(\mu;\ \sigma^2\right)$ variates, $\overline{X} = \dfrac{1}{n}\sum_{1}^{n} X_i$

and let $Y = \sum_{i=1}^{n} \left[\dfrac{X_i - \overline{X}}{\sigma}\right]^2$. Then $Y \sim \chi^2_{n-1}$ and $\overline{X}$ and Y are

independent variates, ie $f_{\overline{X};Y}\,(u;v) = f_{\overline{X}}\,(u)\,f_Y\,(v)$.

Note that Y can also be expressed as $Y = \dfrac{(n-1)\,S^2}{\sigma^2}$ with S^2 the sample variance.

Student's t distribution

This distribution was derived by WS Gossett who worked for a brewery and was not allowed to publish his results under his own name. He therefore used the pseudonym "Student".

Definition 1.19

If T is a random variable with pdf

$$f_T(t) = \frac{\Gamma\left[\frac{1}{2}(d+1)\right]}{\Gamma\left(\frac{1}{2}d\right)\Gamma\left(\frac{1}{2}\right)\sqrt{d}}\left(1+\frac{t^2}{d}\right)^{-\frac{1}{2}(d+1)} \quad ; \quad -\infty < t < \infty$$

where d is a positive integer, then T is called a *Student t-variate* with d degrees of freedom. We write $T \sim t_d$.

The pdf of this distribution is illustrated in the following graph:

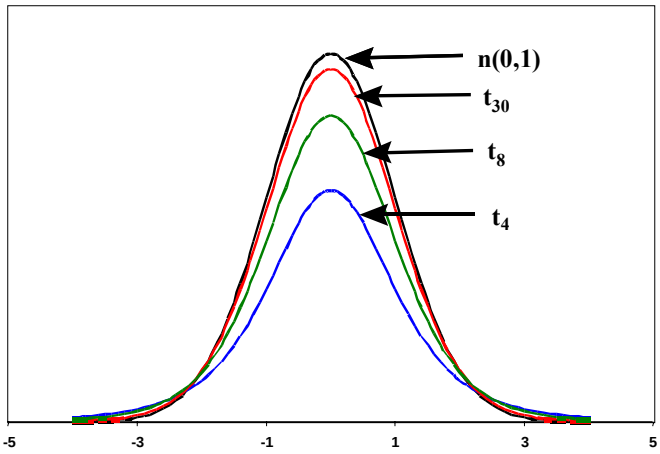


Figure 1.9:
The pdf of Student's t-distribution with $d = 4,\ d = 8$ and $d = 30$ degrees of freedom

When $d = \infty$ the pdf of T is identical to the standardised normal probability density function. (Compare the last line of table III with table II.)
Table III gives critical values of the t-distribution. Notice that the t-distribution is, like the standardised normal distribution, symmetric about zero.

Relation to normal sampling theory

Theorem 1.4
Let U and V be independent variates such that $U \sim n\,(0;\ 1)$ and $V \sim \chi^2_d$ and let

$$T = \frac{U}{\sqrt{V/d}}.$$ Then $T \sim t_d$.

Theorem 1.5
Let $X_1,\ ...,\ X_n$ be independent $n\left(\mu;\ \sigma^2\right)$ variates and let

$\overline{X} = \frac{1}{n}\sum_1^n X_i.$ Then $T = \dfrac{\sqrt{n}\left(\overline{X}-\mu\right)}{\sqrt{\dfrac{\Sigma\left(X_i-\overline{X}\right)^2}{n-1}}}$ is a t_{n-1} variate.

The F-distribution

Definition 1.20
Let X be a random variable with pdf

$$\begin{aligned} f_X(x) &= \frac{\Gamma\left[\frac{1}{2}(d_1+d_2)\right]}{\Gamma\left(\frac{1}{2}d_1\right)\Gamma\left(\frac{1}{2}d_2\right)} d_1^{\frac{1}{2}d_1} d_2^{\frac{1}{2}d_2} x^{\frac{1}{2}d_1-1}(d_2+d_1x)^{-\frac{1}{2}(d_1+d_2)} \quad x>0 \\ &= 0 \qquad\qquad \text{elsewhere,} \end{aligned}$$

where d_1 and d_2 are positive integers. Then X is said to have an F-distribution with d_1 and d_2 degrees of freedom. We write $X \sim F_{d_1;d_2}$.

This is a two-parameter family of distributions, and the pdf is illustrated for $d_1 = 10$ and $d_2 = 4; 10; 50$ and ∞ in the following graph:

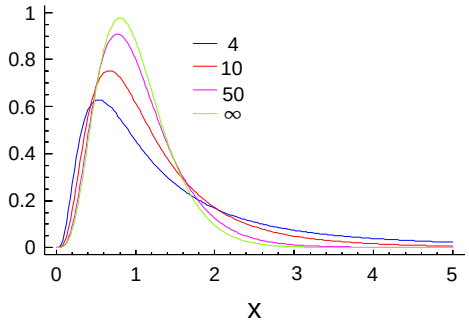


Figure 1.10:
The pdf of the F-distribution for $d_1 = 10$ and $d_2 = 4; 10; 50$ and ∞

Tables V, VI and VII list critical values of this distribution. Note that the first degrees of freedom d_1 is always listed at the top of the table and the second degrees of freedom d_2 on the left.
We will use the shorthand notation $F_{\alpha;d_1 d_2}$ for the upper-tail probability α.

Definition 1.21
Let X_1 and X_2 be *independent* random variables with

$X_1 \sim \chi^2_{d_1}$ and $X_2 \sim \chi^2_{d_2}$, and let $Y = \dfrac{X_1/d_1}{X_2/d_2}$.

Then $Y \sim F_{d_1;d_2}$.

Since the roles of X_1 and X_2 may be switched the following result is easily proved:

Result 1.4
If $X \sim F_{d_1;d_2}$ then $\dfrac{1}{X} \sim F_{d_2;d_1}$.

This result enables us to find a two-sided interval for an F-variate, as is shown in the next example.

Example 1.11

Find a 95% two-sided confidence interval for the F-variate X where $X \sim F_{8;20}$.

Solution

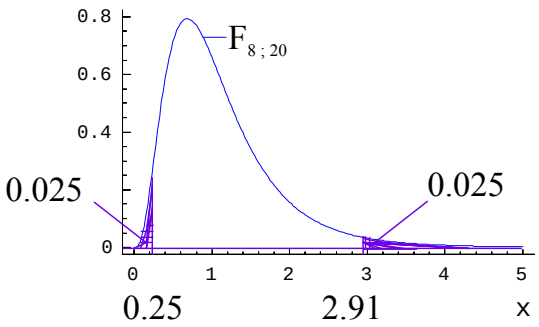
$P(X < 2.91) = 0.975$ (table VI). From the previous result $Y = \frac{1}{X} \sim F_{20;8}$.

$\therefore P(Y > 4) = 0.025$ (table VI) (ie $F_{0.025;20;8} = 4$)

$\therefore P\left(\frac{1}{X} > 4\right) = 0.025$

$\therefore P\left(X < \frac{1}{4}\right) = 0.025$ (ie $F_{0.975;8;20} = \frac{1}{F_{0.025;20;8}}$)

$\therefore P\left(\frac{1}{4} < X < 2.91\right) = P(X < 2.91) - P\left(X < \frac{1}{4}\right) = 0.975 - 0.025 = 0.95.$



The bivariate normal distribution

Definition 1.22

Let X_1 and X_2 be two random variables with joint pdf

$$f_{X_1;X_2}(x_1;x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{\left\{-\frac{1}{2}Q(x_1;x_2)\right\}};$$

for $-\infty < x_1 < \infty;\ -\infty < x_2 < \infty;\ \sigma_1 > 0;\ \sigma_2 > 0;\ -1 < \rho < 1$ where

$$Q(x_1;x_2) = \frac{1}{1-\rho^2}\left\{\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right)\right\}.$$

Then X_1 and X_2 are said to have a *bivariate normal distribution.*

The significance of the various constants is:

$$\mu_1 = E(X_1); \quad \mu_2 = E(X_2); \quad \sigma_1^2 = Var(X_1); \quad \sigma_2^2 = Var(X_2);$$

$\rho =$ correlation coefficient between X_1 and X_2.

Remember that it was pointed out previously that, if X_1 and X_2 are independent variates then they are uncorrelated; and also that the converse is not always true. The converse is true in the case of the bivariate normal distribution, however. If we set $\rho = 0$ in the joint pdf of X_1 and X_2, we obtain

$$\begin{aligned} f_{X_1;X_2}(x_1;x_2) &= \frac{1}{2\pi\sigma_1\sigma_2} e^{-\frac{1}{2}\left\{\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2+\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right\}} \\ &= \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{1}{2}\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2} \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{1}{2}\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2} \\ &= f_{X_1}(x_1) f_{X_2}(x_2). \end{aligned}$$

The following result is therefore true:

Result 1.5

Let X_1 and X_2 have a bivariate normal distribution. Then X_1 and X_2 are independent if and only if they are uncorrelated.

In the general case (ρ not necessarily equal to zero) it can be shown that, if X_1 and X_2 have a bivariate normal distribution, then the marginal distributions of X_1 and X_2 are normal distributions, ie

$$\int_{-\infty}^{\infty} f_{X_1;X_2}(x_1;x_2)\,dx_2 = f_{X_1}(x_1)$$

where $f_{X_1}(x_1)$ is the $n\left(\mu_1;\sigma_1^2\right)$ density function and likewise for X_2.

The exponential distribution

Definition 1.23

Let X be a random variable with pdf

$$f_X(x) = \frac{1}{\lambda} e^{-x/\lambda}, \qquad x \geq 0;\ \frac{1}{\lambda} > 0.$$

Then X is said to have an exponential distribution with parameter $\left(\frac{1}{\lambda}\right)$.

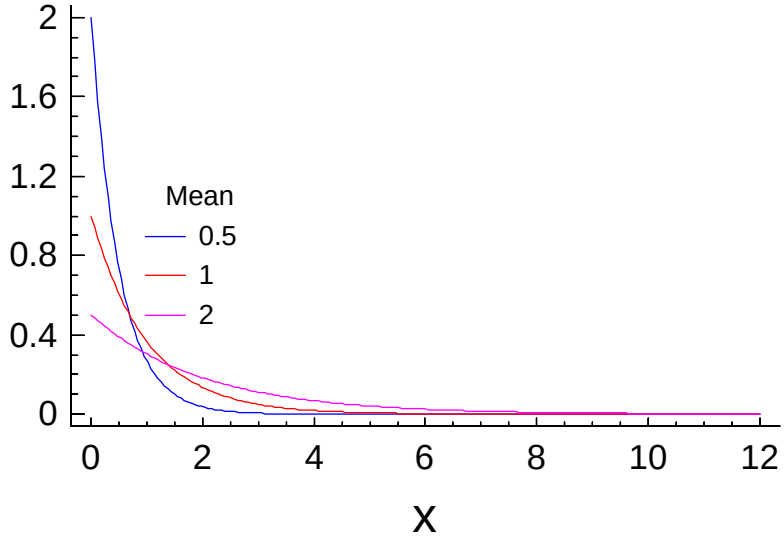


Figure 1.11:
The pdf of the exponential distribution

Result 1.6

For the exponential distribution:

(i) $E(X) = \lambda$

(ii) $Var(X) = \lambda^2$

(iii) $P(X \leq x) = 0$ for $x \leq 0$

$= 1 - e^{-x/\lambda}$ for $x > 0$

The parameter $\left(\frac{1}{\lambda}\right)$ is sometimes referred to as the failure rate.

Exercise 1.1

1. Verify that if $T \sim t_{10}$, then

$P(T \geq 2.764) = 0.01$ note that we write $t_{0.01;10} = 2.764$

$P(T < 2.764) = 0.99$ $\Longrightarrow P(T < -2.764) = 0.01$

$P(-2.764 < T < 2.764) = 0.98$

$P(-0.7 < T < 0.7) = 0.5.$

2. Verify that, if $X \sim F_{5;12}$ then

$P(X > 3.11) = 0.05$ (table V) $\therefore F_{0.05;5;12} = 3.11$

$P(X < 3.89) = 0.975$ (table VI) $\therefore F_{0.025;5;12} = 3.89$

$P(X < 5.06) = 0.99$ (table VIII) $\therefore F_{0.01;5;12} = 5.06.$

1.4 Learning outcomes

After studying unit 1 you should **know** the following concepts:

- *random variable*
- *probability density function* (pdf)
- the *mean* or *expected value* of a random variable
- the *variance* of a random variable
- the *covariance* of two random variables
- *uncorrelated* random variables
- the *probability functions* of the following two discrete random variables:

  $X \sim Psn(\lambda)$ (Poisson)
  $X \sim b(n;\ p)$ (binomial)
- the *probability density function* of $X \sim n\left(\mu;\ \sigma^2\right)$ (normal)
- the *central limit theorem*
- *properties* of the *chi-square* distribution
- the relation of the *t-distribution* to normal sampling theory
- the relation of the *F-distribution* to two independent χ^2-variables
- the pdf and properties of the *exponential* distribution

You should be able to look up a value in your prescribed book of tables that links an outcome of a variable with a given probability (or vice versa) for the following distributions:

- normal
- t
- F
- χ^2

STUDY UNIT 2

Concepts of estimation and inference

2.1 Introduction

It is always difficult to summarise a subject in a nutshell but we could say that Statistics as a science is focused on the following overall objective: *To collect, organise, analyse and interpret data for the purpose of making better decisions.*

In the previous study unit we stressed that the shape of the normal distribution is determined by the value of the mean μ and the variance σ^2, whilst the shape of the binomial distribution is determined by the sample size n and the probability of a success π. These critical values are called *parameters*. (If you might recall, *parameters are numerical measures that describe the characteristics of a population.*) We most often don't know what the values of the parameters are and thus we cannot "utilise" these distributions (ie use the mathematical formula to draw a probability density graph or compute specific probabilities) unless we somehow *estimate these unknown parameters*. In introductory courses it is usually simply stated that it makes perfect logical sense that to estimate the value of an unknown population parameter, we compute a corresponding or comparable characteristic of the sample. Is this always the best estimate? What does "best estimate" mean? In this study unit you will learn that there are mathematical techniques that will "lead" us to estimators of parameters!

In your first-year modules we dealt with probability and probability distributions, and emphasised that unless one has a proper understanding of the laws of probability, the mechanisms underlying statistical data analysis will not be understood properly. Probability theory is the tool that makes statistical inference possible. In dictionary terms, *inference* is the act or process of inferring and to *infer* means *to conclude or judge from premises or evidence* which means to derive by reasoning. In general the term implies a conclusion based on experience or knowledge. More specifically in statistics, 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. The set of principles, procedures and methods that we use to study populations by making use of information obtained from samples is called *statistical inference.* Thus our objective will be to *draw inference* about a population (a complete set of data) based on the limited information contained in a sample.

How will we link the information from a sample to a population? You have already learned from first-year modules that the **sampling distribution of a statistic** is the vehicle to move between the sample and the population. For example, we showed you how to *derive* the sampling distribution of the sample mean, $\overline{X}$, and how to *apply* this sampling distribution in developing an interval estimate

for a population mean and how to perform a hypothesis test. In this study unit we will return to concepts of hypothesis testing and confidence intervals in general.

2.2 Defining a random sample and a statistic

At first-year level, we were very specific with our examples and explanations of the **sampling distribution of a statistic** in developing an interval estimate for a population parameter or to perform a hypothesis test for a population parameter. For example, we explained how a confidence interval is derived for μ using the sampling distribution of $\overline{X}$, how a confidence interval is derived for p using the sampling distribution of $\widehat{p}$ and how a confidence interval is derived for $\mu_1 - \mu_2$ using the sampling distribution of $\overline{X}_1 - \overline{X}_2$.

How can we generalise these principles?

In general, we are interested in a random variable X with probability density function (pdf) $f_X(x)$ which depends on a parameter θ, which is (usually) unknown. We sometimes write $f_X(x;\theta)$ to emphasise that the pdf depends on θ. We are interested specifically in obtaining information about the parameter θ, for example that $\theta = E(X)$; $\theta = Var(X)$ or $\theta = P(X \geq c)$ for a specific c.

Consider for example the random variable X which represents the life (in thousands of kilometres) of a tyre of given size and manufacture. We may be interested in the expected life, in which case $\theta = E(X)$, or in the probability that the tyre will last for more than 50 000 km, in which case $\theta = P(X > 50)$.

In order to obtain information about the unknown parameter θ we usually make use of a random sample. Suppose in the above example we select five tyres at random and determine the life of each tyre, say $X_1, \ ..., \ X_5$. In order for $X_1, \ ..., \ X_5$ to be regarded as a *random sample* for a given distribution, we require that $X_1, \ ..., \ X_5$ be *independent* and that each of them has the prescribed distribution.

Definition 2.1

The random variables $X_1, \ ..., \ X_n$ constitute a *random sample* from the distribution with pdf $f_X(x)$ if $X_1, \ ..., \ X_n$ are independent random variables, each with pdf $f_X(x)$.

It follows that the joint pdf of $X_1, \ ..., \ X_n$ is given by

$$f_{X_1,...,X_n}(x_1, \ ..., \ x_n) = f_X(x_1)...f_X(x_n).$$

After such a random sample has been obtained, it must be analysed. This may be done in many ways, depending on the objective. Firstly the data are represented graphically in different ways in

order to try to find out what information about the population may be obtained from the sample. This graphical analysis is followed by statistical computations. These computations lead to quantities which we shall call *statistics*.

Definition 2.2

Any function $T \equiv T(X_1,\ ...,\ X_n)$ of the random sample $X_1,\ ...,\ X_n$ is called a *statistic* if it can be computed without using unknown parameters.

NB: Since a statistic is a function of random variables, it is itself a *random variable*.

Example 2.1

If $X_1,\ ...,\ X_n$ is a random sample from a distribution with mean μ and variance σ^2, then the following functions are examples of statistics:

(a) $\frac{1}{n}\sum_{1}^{n} X_i$

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

(c) $\max(X_1,\ ...,\ X_n)$

(d) X_3.

The following is not a statistic:

$$\sum_{1}^{n}(X_i - \mu)^2$$

(unless μ is known).

The remainder of this study unit is devoted to the general introduction to the three main subjects falling under statistical inference:

- **Point estimation**

  We want to find a statistic T which may be used as an estimator for the unknown parameter.

- **Hypothesis testing**

  We are looking for a decision rule by means of which we may choose between the two hypotheses H_0 (the null hypothesis) and H_1 (the alternative hypothesis). Such a hypothesis is some or other statement about the unknown parameter θ, for example $\theta = 0; \theta > 10; \theta \neq 6$.

- **Interval estimation**

  We are trying to find two statistics $T_1 \equiv T_1(X_1, ..., X_n)$ and $T_2 \equiv T_2(X_1, ..., X_n)$ such that $P(T_1 < \theta < T_2) = 1 - \alpha$ where α is a small number between 0 and 1, for example $\alpha = 0.05$.

2.3 Point estimation

Given a random sample $X_1, ..., X_n$ from a distribution with pdf $f_X(x; \theta)$ which depends on the unknown parameter θ, we wish to find a statistic $T \equiv T(X_1, ..., X_n)$ which may serve as an *estimator* for θ. An estimator for θ is sometimes denoted by $\hat{\theta}$ ("theta-hat"). The sample $X_1, ..., X_n$ consists of n random variables and the estimator T is also a random variable. The values which $X_1, ..., X_n$ assume in a specific example, $x_1, ..., x_n$ say, are constants and the corresponding value of T, namely $T(x_1, ..., x_n)$ which is a realisation of the estimator, is called an *estimate* of θ.

An *estimator* may be regarded as a formula by means of which an *estimate* is obtained from a given set of data.

Thus, for example, we shall show that $\overline{X} = \frac{1}{n}\sum_{1}^{n} X_i$ is a possible estimator for the population mean μ; if we obtain a sample $x_1 = 10; x_2 = 20; x_3 = 15$ then $\overline{x} = (10 + 20 + 15)/3 = 15$ is an estimate of μ.

In order to ensure that there is some connection between the estimator and the parameter, in other words to prevent the possibility that just any old statistic be used as an estimator of θ, certain restrictions are imposed on the estimator. Such restrictions are treated more fully in advanced courses, but we mention briefly the property of *unbiasedness*. This is a logical property for an estimator to have, but the requirement of unbiasedness is sometimes replaced by other requirements which may lead to better estimators.

Definition 2.3

The statistic T is called an *unbiased estimator* for the parameter θ if $E(T) = \theta$.

Example 2.2

Let $X_1, \ ..., \ X_n$ be a random sample from a distribution with expected value θ. Prove that $\overline{X} = \frac{1}{n}\sum_{1}^{n} X_i$ is an unbiased estimator for θ.

Solution

$$\begin{aligned} E\left(\overline{X}\right) &= E\left(\frac{1}{n}X_1 + ... + \frac{1}{n}X_n\right) \\ &= \frac{1}{n}E\left(X_1\right) + ... + \frac{1}{n}E\left(X_n\right) \\ &= \frac{1}{n}\theta + ... + \frac{1}{n}\theta \\ &= \theta \end{aligned}$$

Example 2.3

Let X_1 and X_2 be random variables from a $n\left(\mu; \ \sigma^2\right)$ distribution. Show that both

$$\hat{\mu}_1 = \frac{1}{3}X_1 + \frac{2}{3}X_2$$

and

$$\hat{\mu}_2 = \frac{1}{2}X_1 + \frac{1}{2}X_2 \left(= \overline{X}\right)$$

are unbiased estimators of the mean.

Solution

$$\begin{aligned} E\left(\hat{\mu}_1\right) &= E\left(\frac{1}{3}X_1 + \frac{2}{3}X_2\right) \\ &= \frac{1}{3}E\left(X_1\right) + \frac{2}{3}E\left(X_2\right) \\ &= \frac{1}{3}\mu + \frac{2}{3}\mu \\ &= \mu \end{aligned}$$

$$\begin{aligned} E(\hat{\mu}_2) &= E\left(\frac{1}{2}X_1 + \frac{1}{2}X_2\right) \\ &= \frac{1}{2}\mu + \frac{1}{2}\mu \\ &= \mu \end{aligned}$$

Thus both $\hat{\mu}_1$ and $\hat{\mu}_2$ are unbiased estimators of μ.

Definition 2.4

If we have two or more unbiased estimators for the parameter θ, then we select the estimator with the smallest variance. Such an estimator is called the most *efficient* of the estimators.

Example 2.4

For example 2.3 we have

$$\begin{aligned} Var(\hat{\mu}_1) &= Var\left(\frac{1}{3}X_1 + \frac{2}{3}X_2\right) \\ &= \frac{1}{9}Var(X_1) + \frac{4}{9}Var(X_2) \\ &= \frac{1}{9}\sigma^2 + \frac{4}{9}\sigma^2 \\ &= \frac{5}{9}\sigma^2. \end{aligned}$$

$$\begin{aligned} Var(\hat{\mu}_2) &= Var\left(\frac{1}{2}X_1 + \frac{1}{2}X_2\right) \\ &= \frac{1}{4}\sigma^2 + \frac{1}{4}\sigma^2 \\ &= \frac{1}{2}\sigma^2. \end{aligned}$$

Therefore $\hat{\mu}_2$ is a more efficient estimator of μ than $\hat{\mu}_1$ is.

2.4 Methods of finding estimators

Some highly sophisticated methods exist for finding estimators. Some of these methods involve complicated theories and are treated in more advanced courses. We discuss two methods here:

(A) least squares

(B) maximum likelihood.

(A) Least squares estimation

The method of least squares is used especially in problems where the unknown parameters are linear functions of known constants.

Theorem 2.1

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

$$E(X_i) = c_{i1}\theta_1 + c_{i2}\theta_2 + ... + c_{ik}\theta_k;$$

$$Var(X_i) = \sigma^2; \quad i = 1,\ ...,\ n; \quad \text{where}$$

$c_{ij}, \quad j = 1,\ ...,\ k$ and $i = 1,\ ...,\ n$ are known constants.

The least squares estimators of $\theta_1,\ ...,\ \theta_k$ are found by minimising

$$\begin{aligned} Q(\theta_1,\ ...,\ \theta_k) &= \sum_{i=1}^{n} (X_i - E(X_i))^2 \\ &= \sum_{i=1}^{n} (X_i - c_{i1}\theta_1 - c_{i2}\theta_2 - ... - c_{ik}\theta_k)^2 . \end{aligned}$$

This is achieved by setting

$$\frac{\partial Q}{\partial \theta_j} = 0; \quad j = 1,\ ...,\ k$$

thus obtaining k equations with k unknowns, which are solved to obtain $\hat{\theta}_1,\ ...,\ \hat{\theta}_k$.

Example 2.5

Let $X_1,\ X_2,\ \ldots,\ X_n$ be independent random variables from a distribution with expected value θ. Show that $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the least squares estimator for θ.

Solution

$E(X_i) = \theta \qquad i = 1,\ 2,\ \ldots,\ n$

$$Q(\theta) = \sum_{i=1}^{n} [X_i - E(X_i)]^2$$

$$= \sum_{i=1}^{n} (X_i - \theta)^2$$

$$\frac{\partial Q}{\partial \theta} = \sum_{i=1}^{n} 2(X_i - \theta)(-1)$$

$$= -2\sum_{i=1}^{n} X_i + 2n\theta \qquad \text{(Do you recall that } \sum_{i=1}^{n} k = nk?)$$

Set $\frac{\partial Q}{\partial \theta} = 0$

$$\Rightarrow 2n\theta = 2\sum_{i=1}^{n} X_i$$

$$\hat{\theta} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Example 2.6

Let $X_1,\ ...,\ X_n$ be independent random variables such that $E(X_i) = c_i\theta,\ i = 1,\ ...,\ n$ where $c_1,\ ...,\ c_n$ are known constants. Find the least squares estimator for θ.

Solution

We estimate θ by minimising $Q(\theta)$ where

$$Q(\theta) = \sum_{i=1}^{n} (X_i - c_i\theta)^2 = \sum_{i=1}^{n} \left(X_i^2 - 2c_iX_i\theta + c_i^2\theta^2\right)$$

$$\begin{aligned}\frac{\partial Q(\theta)}{\partial \theta} &= -2\sum_{i=1}^{n}\left(c_i X_i - c_i^2\theta\right) \\ &= -2\left(\Sigma c_i X_i - \theta \Sigma c_i^2\right) \\ &= 0 \quad \text{if} \quad \theta = \frac{\Sigma c_i X_i}{\Sigma c_i^2}.\end{aligned}$$

Thus the least squares estimator of θ is $\hat{\theta} = \dfrac{\sum_{i=1}^{n} c_i X_i}{\sum_{i=1}^{n} c_i^2}$.

(B) Maximum likelihood estimation (mle)

Before we formally define this method, consider the following concrete example:

Example 2.7

Suppose the number of visits a child pays the dentist per year, has a Poisson distribution with unknown parameter θ. A random sample of 4 children paid the following observed number of visits to the dentist:

$$X_1 = 0; \qquad X_2 = 2; \qquad X_3 = 1 \qquad \text{and} \qquad X_4 = 3.$$

This means we have a random sample of size $n = 4$ from a distribution with pdf

$$f_X(x;\ \theta) = \frac{e^{-\theta}\theta^x}{x!} \qquad \text{for } x = 0;\ 1;\ 2;\ \ldots \text{ (see definition 1.15)}$$

where $E(X) = \theta$ and $Var(X) = \theta$.

The probability of any outcome X therefore depends on θ only so that we can write

$$\begin{aligned}P(X_1 = 0) &= f_X(X_1;\ \theta) = \frac{e^{-\theta}\theta^0}{0!}; \\ P(X_2 = 2) &= f_X(X_2;\ \theta) = \frac{e^{-\theta}\theta^2}{2!}; \\ P(X_3 = 1) &= f_X(X_3;\ \theta) = \frac{e^{-\theta}\theta^1}{1!} \text{ and} \\ P(X_4 = 3) &= f_X(X_4;\ \theta) = \frac{e^{-\theta}\theta^3}{3!}.\end{aligned}$$

Assuming independence of $X_1,\ X_2,\ \ldots,\ X_4$ the joint probability function of the sample can therefore be written as

$$\begin{aligned} P\left(X_1;\ X_2;\ X_3;\ X_4;\ \theta\right) &= \frac{e^{-\theta}\theta^0}{0!} \cdot \frac{e^{-\theta}\theta^2}{2!} \cdot \frac{e^{-\theta}\theta^1}{1!} \cdot \frac{e^{-\theta}\theta^3}{3!} \\ &= \frac{e^{-4\theta}\theta^{0+2+1+3}}{1 \times 2 \times 1 \times 1 \times 3 \times 2 \times 1} \\ &= \frac{e^{-4\theta}\theta^6}{12}. \end{aligned}$$

Please note that this specific probability expression is not a "general case" but specifically derived for a sample of four, with very specific outcomes, ie it is $P\left(X_1 = 0,\ X_2 = 2,\ X_3 = 1,\ X_4 = 3\right)$.

Since it is a function of θ only, we denote it by $L(\theta)$ and call it the likelihood function (from there the $\underline{L}$).

For each different value of θ, we can compute a different value for $L(\theta)$. This means we could use the "connect-the-dots" method to draw a graph of $L(\theta)$. (See figure 2.1.)

If $\theta = 1$ we have $L\left(1\right) = \frac{e^{-4}1^6}{12} = \frac{0.0183156}{12} = 0.0015263$ (and we interpret it as the joint probability of the specific sample for the case where $\theta = 1$).

If $\theta = 2$ we have $L\left(2\right) = \frac{e^{-8}2^6}{12} = \frac{0.0003355 \times 64}{12} = 0.001789.$

In the following table $L(\theta)$ has been computed for seven different values of θ.

θ	$L(\theta)$
0	0
0.5	0.000176
1.0	0.001526
1.5	0.002353
2.0	0.001789
2.5	0.000924
3.0	0.000373

These likelihood values are plotted in figure 2.1 and it is obvious from the graph that $L(\theta)$ reaches a maximum at $\theta = 1.5$. We therefore say that $\theta = 1.5$ is our maximum likelihood estimator for this specific sample.

But, what is now very interesting is to note that $\overline{X} = \dfrac{\sum_{i=1}^{4} X_i}{4} = \dfrac{0+2+1+3}{4} = 1.5.$

Instead of this "trial and error" or graphical method we will mostly use analytical methods to determine θ – although it is instructive to look at the problem this way.

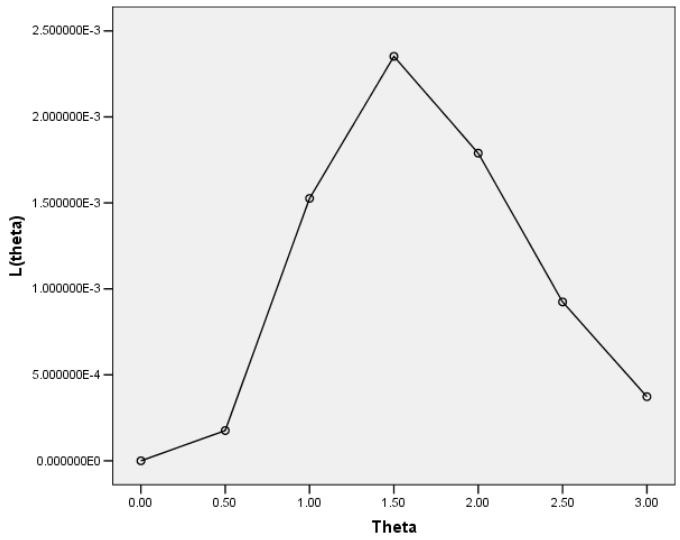


Figure 2.1

The method of maximum likelihood is in effect that one has to find that value of θ that will maximise $L(\theta)$ for the observed sample.

Definition 2.5

The method of *maximum likelihood* for estimating a parameter θ, selects that value of θ as a point estimator that maximises the *likelihood function*

$$L(\theta) = f_X(X_1;\ \theta)\, f_X(X_2;\ \theta) \ldots f_X(X_n;\ \theta) = \prod_{i=1}^{n} f_X(X_i;\ \theta).$$

In the same way that the symbol $\sum_{i=1}^{n}$ denotes the sum of n terms, the symbol $\prod_{i=1}^{n}$ can be used to denote the **product** of n terms.

Theorem 2.2

In many problems it is easier to maximise $\log L(\theta)$ than $L(\theta)$. The value of θ that maximises $\log L(\theta)$ will also maximise $L(\theta)$ since $\log L(\theta)$ is a strictly increasing function of $L(\theta)$.

Example 2.7 (continued)

Suppose that we now want to keep it abstract and use $X_1,\ X_2,\ X_3$ and X_4 without replacing them with the observed values.

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{4} f_X(X_i;\ \theta) \\ &= \frac{e^{-\theta}\theta^{X_1}}{X_1!} \cdot \frac{e^{-\theta}\theta^{X_2}}{X_2!} \cdot \frac{e^{-\theta}\theta^{X_3}}{X_3!} \cdot \frac{e^{-\theta}\theta^{X_4}}{X_4!} \\ &= \frac{e^{-4\theta}\theta^{\sum_{i=1}^{4} X_i}}{\prod_{i=1}^{4} X_i!} \end{aligned}$$

Now $\log L(\theta) = -4\theta + \sum_{i=1}^{4} X_i \log\theta - \sum_{i=1}^{4} \log(X_i!)$

$$\frac{\partial \log L(\theta)}{\partial\theta} = -4 + \sum_{i=1}^{4} X_i \cdot \frac{1}{\theta} + 0.$$

Let $\frac{\partial \log L(\theta)}{\partial\theta} = 0$ then $\sum_{i=1}^{4} X_i = 4\theta.$

Therefore $\widehat{\theta} = \frac{\sum_{i=1}^{4} X_i}{4} = \overline{X}$ is the maximum likelihood estimator of θ.

[Strictly speaking (mathematically) we should also inspect the **second-order derivative** to ascertain whether we in fact have a maximum value and not a minimum!]

We already know that (see exercise 2.1, question 1) $\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$ is an unbiased estimator for the variance of a distribution and also that $\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$ is an unbiased estimator for the mean.

We began example 2.7 with the statement that the *number of visits* has a Poisson distribution. Paging back to the properties of this discrete distribution (see the heading after definition 1.15) we may write down that

$$\begin{aligned} E(X) &= \theta \qquad \text{and} \\ Var(X) &= \theta \end{aligned}$$

Why don't we choose $S^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$ as our unbiased estimator for θ?

It is because the method of maximum likelihood has "guided" us to $\overline{X}$ (and not S^2) and usually (under general conditions) the maximum likelihood estimators are more efficient than other estimators, **but they are not necessarily always unbiased**.

Example 2.8

Let $X_1,\ \ldots,\ X_n$ be a random sample from a distribution with pdf

$$f(x)=cx^{-c-1} \qquad x>1.$$

Find the MLE for c.

Solution

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

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

$$\frac{\partial \log L}{\partial c}=\frac{n}{c}-\sum_{i=1}^{n}\log X_i=0 \text{ if}$$

$$c=\frac{n}{\sum_{i=1}^{n}\log X_i}.$$

Therefore the MLE of c is

$$\hat{c}=\frac{n}{\sum_{i=1}^{n}\log X_i}.$$

In cases where more than one, say k, unknown parameters are to be estimated, the partial derivative of L (or $\log L$) with respect to each parameter is equated to zero to obtain k equations with k unknowns. These equations cannot always be solved very easily if they are nonlinear; sometimes it is necessary to employ an iterative method to obtain a numerical solution.

2.5 Hypothesis testing

One of the most commonly used techniques in the analysis of data is hypothesis testing. The basis for it was laid in the thirties of the previous century by two statisticians: Jerzy Neyman, originally from Poland, and Egon S Pearson of the UK, son of the famous statistician Karl Pearson.

The technique of hypothesis testing is discussed here generally, but using an example to illustrate the concepts. The details of how to find the decision rule in specific types of problems will be discussed in later study units. The process is described here in a number of steps. The order of these steps represents more or less the ideal order. In practice one may sometimes have to change the order due to practical necessity. Such changes in the order may, however, change the characteristics of the test.

Step 1. The brainwave

A researcher develops a theory about a natural phenomenon, economic law, production process, et cetera which he or she is in the process of investigating. The researcher decides that the theory is of sufficient importance to try to verify or discard by means of an experiment.

Example 2.9

A farmer wants to find a better feed which will make his piglets grow faster. He knows from past experience that his piglets seldom reach a mass of 40 kg or more after four months. A salesman assures him that his piglets will on average weigh more that 40 kg after four months if they are fed on Yumyum Balanced Pig Feed. He decides to try Yumyum on a few piglets for a trial period.

Step 2. Choice of a model

At this stage it is desirable that a statistical model be formulated as carefully as possible for the proposed experiment. Sometimes the model can be formulated only partially at this stage, since one may want to gain information about the model from the data after experimentation. In such a case one would formulate a tentative model with the idea that it may be altered later.

Example 2.9(a) (example 2.9 continued)

For the piglets we could formulate the following model: Let X denote the mass after four months of a piglet selected at random and fed on Yumyum. Let μ be the mean and σ^2 the variance of the distribution of X. We assume for the moment that

$$X \sim n(\mu; \sigma^2).$$

In some applications the researcher may know from past experience that data of the type which he or she is going to collect, usually follow a certain distribution. Sometimes, however, the distribution may have to be investigated after the data become available and the model adjusted accordingly.

Step 3. Specification of the hypothesis and significance level

At this stage the null and alternative hypotheses must be specified. These hypotheses consist of specifications for one or more parameters. The null hypothesis usually specifies a single value for each parameter being tested; the alternative is usually less specific.

Example 2.9 (b) (example 2.9 continued)

In the piglet example the obvious null hypothesis is

$$H_0 : \mu = 40$$

and the alternative

$$H_1 : \mu > 40.$$

This is a *one-sided* alternative: the farmer only wants to know whether the expected mass is more than 40 kg; he will not be interested in Yumyum if the expected mass is less than 40 kg. Actually one could say that the null hypothesis is $H_0 : \mu \leq 40$, but usually only the extreme value (closest to H_1) is specified.

In many problems the alternative would be *two-sided*.

Suppose, for example, a dealer orders ball bearings with the specification that the mean diameter must be $\mu = 10\ mm$. Ball bearings which are too large or too small are unacceptable. Thus $H_0 : \mu = 10$ is regarded as false if either $\mu < 10$ or $\mu > 10$ and the alternative is $H_1 : \mu \neq 10$. This is an example of a two-sided alternative.

Note

The research worker must know before the experiment is conducted what the null and alternative hypotheses are. If he or she does not know which specific hypotheses will be tested, he or she must specify the hypotheses as generally as is necessary in order to provide for all possibilities. The practice of generating hypotheses by first studying the data is not to be recommended. It may promote the drawing of false conclusions. If one searches carefully enough, one could find false hypotheses in any set of data. It may be necessary to collect further data to confirm hypotheses generated from the original data. The statistician who does consultation work may have to question his or her client carefully in order to establish whether the latter had good reason to expect the hypotheses **before seeing the data**.

As was said earlier, an experiment will be conducted in order to gain information which will enable the investigator to choose between H_0 and H_1. In the final decision two types of error can be committed:

A **type I** error is committed if we reject H_0 when H_0 is in fact true.
A **type II** error is committed if we do not reject H_0 when H_0 is false.

Note: We never say *"we accept H_0"*, we say *"we do not reject H_0"* or *"we fail to reject H_0"*.

This is represented in the following table:

		Decision based on the data	
		Do not reject H_0	Reject H_0
The true state of nature	H_0 is true	Good decision	Type I error
	H_1 is true	Type II error	Good decision

The decisions to "fail to reject" or "reject" H_0 must be interpreted as follows: If H_0 is rejected (and H_1 is not rejected) it means either that H_0 is true and a rare event has occurred, or that H_1 is true. Since a rare event occurs only rarely, however, we are inclined to lean towards the belief that H_1 is true. If H_0 is not rejected (and H_1 is rejected) it does not mean that we have **proved** that H_0 is true; we could have made a type II error. It means only that there is not sufficient evidence in the data to reject H_0.

In every hypothesis testing procedure, there are probabilities associated with the two types of error:

P(type I error) = α
P(type II error) = β.

We consider the two cases: H_0 true and H_1 true.

(a) H_0 is true.

Example 2.9 (c) (example 2.9 continued)

Assume for illustration purposes that we know that $\sigma^2 = 4$, so that $X \sim n(\mu;\ 4)$.

Now "H_0 is true" means that $\mu = 40$ (ie $X \sim n(40;\ 4)$) and we graphically represent the mass distribution of the piglets by drawing a normal curve.

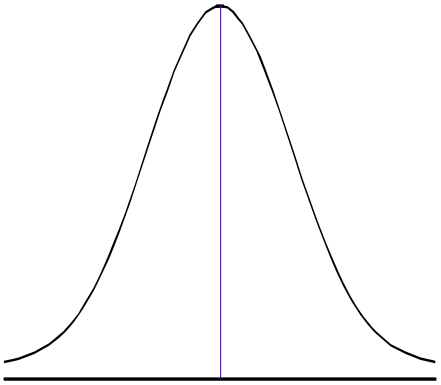


Figure 2.2: Curve of a $n(40;\ 4)$ distribution

With some manipulation and using table I (Stoker) we find that $P(X \geq 43.29) = 0.05$.

To graphically display a type I error, we shade the area where H_0 is rejected. In this example, if $\alpha = 0.05$ then $P(X \geq 43.29) = \alpha$ (assuming $x \sim n(40;\ 4)$).

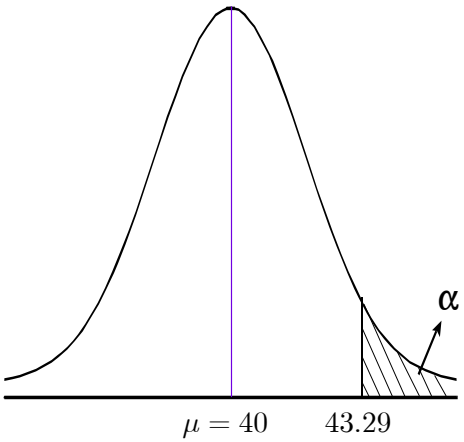


Figure 2.3: $P(H_0$ is rejected$|\ H_0$ is true$) = \alpha$

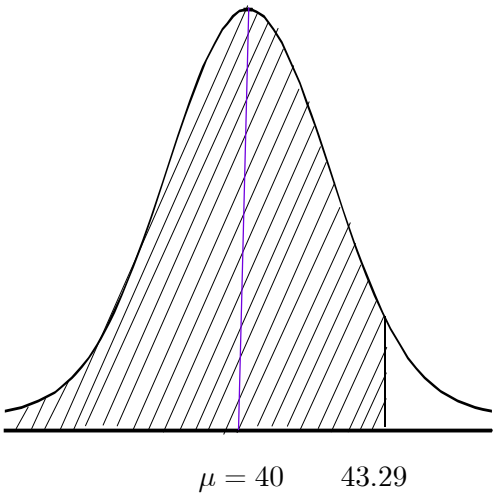


Figure 2.4: $P(H_0$ is not rejected$|\ H_0$ is true$) = 1 - \alpha$

There are two possibilities: H_0 may not be rejected or rejected. The probabilities are

$P(H_0$ is not rejected$|\ H_0$ is true$) = 1 - \alpha;$

$P(H_0$ is rejected$|\ H_0$ is true$) = \alpha.$

Definition 2.6

α is called the significance level of the test, if $P(H_0$ is rejected$|\ H_0$ is true$) = \alpha.$

The significance level is selected in advance, depending on the seriousness of a type I error. If a type I error means that the farmer will use a somewhat poorer feed for his pigs, he may use $\alpha = 0.05$ or even $\alpha = 0.10$. However, if a type I error means that a patient will die, a much smaller α (like $\alpha = 0.001$) will have to be used. The most generally used choices of α are 0.05 and 0.01. To a certain extent the choice of α is restricted by the availability of statistical tables, when we perform hypothesis tests manually. However, when you perform a hypothesis test using a statistical package, the p-value will be used more often to draw a conclusion. (The definition and interpretation of a p-value is discussed at the end of this section.)

Although α is selected in advance, the eventual significance level may differ from α. The assumptions in the model, like normality and independence, are not always satisfied exactly. There is probably no such thing as a normal population in real life. The model being used will only be an approximation to the true situation. Certain types of deviations from the model may cause the true significance level to be larger than the chosen α; other deviations may cause it to be smaller.

(b) **H_1 is true**

There are again two possibilities: we may reject H_0 and not reject H_1 or do not reject H_0 and reject H_1. The probabilities are

$P($not rejecting $H_0 \mid H_1$ is true$) = \beta$

$P($not rejecting $H_1 \mid H_1$ is true$) = 1 - \beta.$

Example 2.9 (d) (example 2.9 continued)

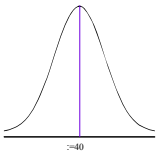
If we continue with the example of the piglets and assume that $\sigma^2 = 4$ (once again simply for illustration purposes) then X is still $\sim n(\mu;\ 4)$. Now "H_1 is true" means that $\mu > 40$. There is not simply a single graph which captures this scenario but trillions of possible graphs! *How can you draw a graph where* "$\mu > 40$"? *What value will you choose*?

If we want to try to represent this graphically, we have to assign specific values to μ (where of course $\mu > 40$).

For example, let us consider where $\mu = 40.5;\ \mu = 43.29$ and $\mu = 45.362$. This means we draw the following three normal probability distributions:

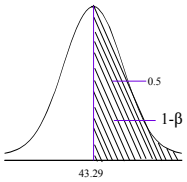
$$n(40.5;\ 4), \qquad n(43.29;\ 4) \qquad \text{and} \qquad n(45.362;\ 4).$$

$(1)\ \mu = 40.5$



μ=40.5 43.29

$(2)\ \mu = 43.29$



$(2)\ \mu = 45.362$

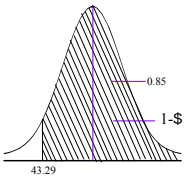


Figure 2.5: $P\,(\text{not rejecting } H_1 | H_1 \text{is true}) = 1 - \beta$

Definition 2.7

The probability, $1 - \beta$, is called the **power** of the test, where $P(\text{not rejecting } H_0 \mid H_1 \text{ is true}) = \beta$.

The power of the test depends on the following factors:

(i) **The significance level** α:

The larger α is, the smaller is β and thus the larger the power. In the choice of α we have a trade-off between α and β. If α is small the test is called conservative and the result is that the power is small. Similarly, if α is large then the power is large.

(ii) **The correctness of the model:**

Just as deviations from the model influence the significance level, they may cause the power to decrease or increase.

(iii) **The value of** θ:

Suppose, as before, the null hypothesis and the alternative are $H_0 : \theta = \theta_0$ and $H_1 : \theta > \theta_0$ where θ_0 is a specified constant. The power of the test will depend on the deviation of the true value of θ from the hypothesised value θ_0. In general $\beta \to 0$ and $(1 - \beta) \to 1$ as $\theta \to \infty$ (for the alternative $\theta > \theta_0$). A graph of $1 - \beta$ versus θ will have the following general form:

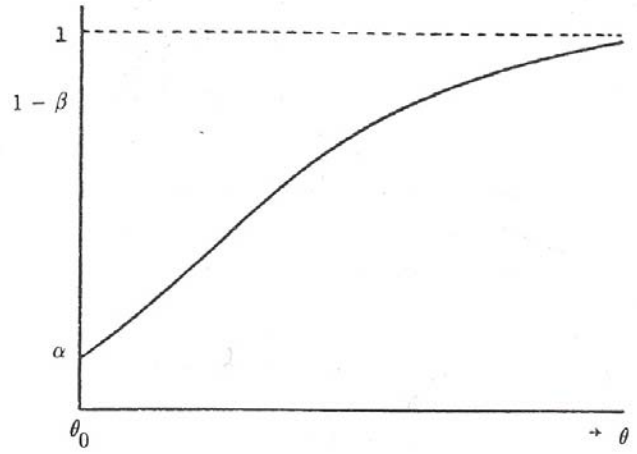


Figure 2.6

On the other hand if $H_0 : \theta = \theta_0$ is tested against $H_1 : \theta \neq \theta_0$, the power curve will appear as follows:

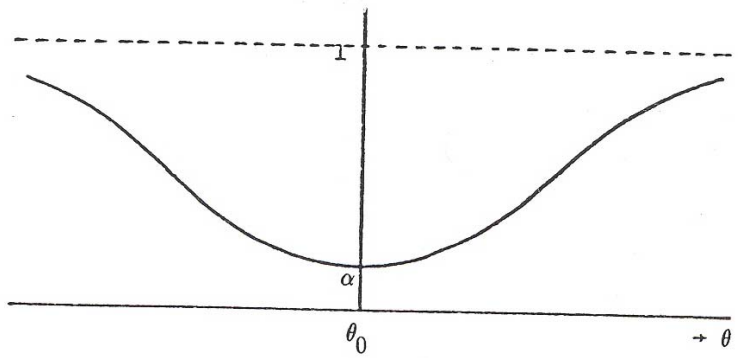


Figure 2.7

If the curve is symmetrical about θ_0, one would obtain a graph like the graph for the one-sided test by plotting $|\theta - \theta_0|$ on the horizontal axis.

(iv) **The planning of the experiment including the choice of the sample size** n:

The larger the sample size, the larger the power. If all other factors remain constant and H_1 is true, it will be true in general that $\beta \to 0$ as $n \to \infty$ and thus the power increases to 1 as the sample size increases. Thus it would seem as if the ideal situation can be approached by simply collecting a very large sample. If even the smallest little deviation from H_0 is of practical importance, this would be a good strategy. However, one must remember that, if $H_0 : \theta = \theta_0$ is not true, θ could still be equal to $\theta_0 + \delta$ where δ is a very small number; in fact δ could be so small as to be of no practical importance. Yet if a very large sample is taken, the power could be close to 1 even if $\theta = \theta_0 + \delta$. A very large sample may often be analysed more informatively by constructing confidence intervals rather than by testing hypotheses. The power curves look something like the following for different sample sizes:

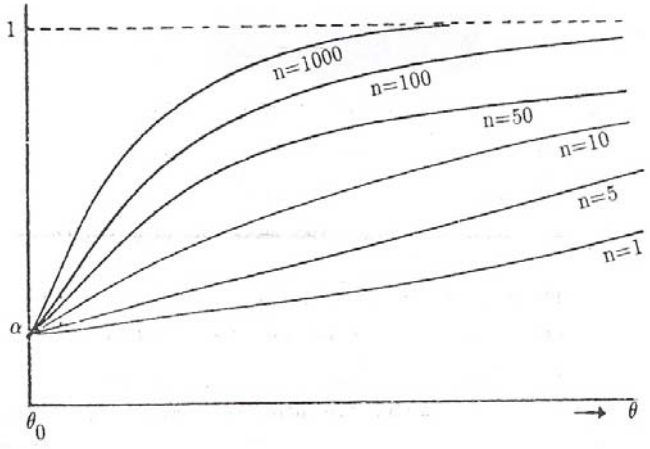


Figure 2.8

Such power curves may be used to select the sample size which would ensure that the test will have a selected power for a given value of θ.

(v) **Other factors:**

The amount of variation in the population may, for example, also play a role in determining the power of the test, depending on the parameter being tested. In general, if the parameter being tested is a mean of a population, then $\beta \to 0$ as $\sigma^2 \to 0$ where σ^2 is the population variance, so that the power $\to 1$. Such a parameter, like σ^2 above, which is of no importance in itself, but which has a profound influence on the test, is sometimes called a "nuisance parameter".

Step 4. Planning of the experiment

We shall not say much about this aspect here except to stress the importance of designing a well-planned experiment. The module STA2602 treats the subject more fully. For some types of experiment tables and graphs are available to enable one to select a sample size – more about this later.

Example 2.9 (e) (example 2.9 continued)

Suppose we assume that a significance level of 0.05 has been selected and that a sample of size $n = 10$ was decided upon. For the piglet example, this means the farmer must try to obtain 10 *independent* observations of four-month-old piglets which have been raised on Yumyum. It is preferable that 10 piglets out of *different litters* be selected rather than ten of the same litter. Ten piglets out of the same litter cannot be considered as a random sample of independent observations and will be much more similar than 10 piglets from different litters. The same litter possesses certain common factors. In the final analysis the farmer wants to say something about all his pigs, not just about the one litter. The crucial question will always be which population does the sample represent?

Step 5. Choice of a test

At this stage (and remember the experiment has not even been started) the researcher must already know how the data is going to be analysed once they have been obtained. If difficulties arise at this stage, the plan could still be altered. Once the experiment is started, it would probably be too late to change the plan. Planning the analysis of the data at this stage helps to ensure objectivity.

A *decision rule* is formulated, and very often this can be done in terms of an estimator of θ. If θ is the population mean, for example, one may decide that H_0 will be rejected if $\overline{X}$, the sample mean, lies in a certain region, called the *critical region*.

[Note that in figure 2.3, for example, the "critical region" was taken as $\{X : X > 43.29\}$ for $\alpha = 0.05$. Here the distribution of $\overline{X}$ was not yet taken into account. We do however know that if $X_1,\ X_2,\ ...,\ X_n$ constitute a random sample from a $n\left(\mu;\ \sigma^2\right)$ distribution, then $\overline{X} \sim n\left(\mu;\ \sigma^2/n\right)$. Only for a sample size of $n = 1$ will 43.29 be the critical value. If $n = 4$ then $\overline{X} \sim n\left(\mu;\ 4/4\right) = n\left(\mu;\ 1\right)$ and if $H_0 = \mu = 40$ is true, then $P\left(\overline{X} > 41.645\right) = 0.05$ so that the critical value becomes $c = 41.645$.]

In the case of a population mean, we say $\overline{X}$ is significantly different from θ_0, the hypothesised value at the α-level (or $100\alpha\%$ level). In this case (θ the population mean) the critical region is usually of the form:

$$\left\{\overline{X} : \overline{X} > c\right\} \text{ if the alternative is } H_1 : \theta > \theta_0;$$

$$\left\{\overline{X} : \overline{X} < c\right\} \text{ if the alternative is } H_1 : \theta < \theta_0 \text{ or}$$

$$\left\{\overline{X} : \left|\overline{X} - \theta_0\right| > c\right\} \text{ if the alternative is } H_1 : \theta \neq \theta_0.$$

(We are going to devote a whole study unit to the testing of means, where the specific details of how the critical region must be obtained, will be discussed in detail. This is simply an overview to refresh your lnowledge of the statistical jargon of hypothesis testing.)

The constant c is determined by the sample size, the significance level and the variance (the population variance if known; otherwise the sample variance).

If $\overline{X}$ does not lie in the critical region, in other words if $\overline{X}$ lies in the complement of the critical region, called the *acceptance region*, then H_0 is not rejected and we say that $\overline{X}$ is not significantly different from θ_0 at the α-level. It does not make sense to say that $\overline{X}$ is (or is not) significantly different from θ_0 without specifying the level of significance. The words *significant* and *significantly different* imply that a statistical test has been performed at a certain level.

Step 6. The experiment

We shall not elaborate on this step except to say that the statistician should, if possible, observe the experimentation. In this way he or she may prevent unwanted factors from confusing the experiment without his or her being aware of it, like operator fatigue which could have the effect that some observations are made less carefully than others, or a breakdown of the machine with the result that the machine setting is changed during the experiment. Even the statistician may not always be able to prevent these occurrences,he or she may be able to take them into account when analysing the data.

Step 7. Analysis of the data

Once the data have been received, the statistician will start analysing them. The first step is to draw graphs and represent the data in various ways in order to decide whether the chosen model is a reasonable approximation or not. With some types of experiment one may know from past experience that the chosen model usually holds in similar situations, but sometimes one may have to rely almost entirely on the data. A word of warning, however. The fact that the model may possibly be changed after the experiment will certainly have an effect on the ultimate significance level, but the size of this effect is unknown. However, this is not a good reason to be blind to obvious and gross deviations from the model. If the model, and possibly the hypotheses, are changed drastically after the data have been studied, one may have to confirm the conclusions by means of a further experiment. Remember that no two samples from the same population are the same, and the danger always exists that a phenomenon in the sample which is due to sampling variation, will be interpreted as a phenomenon in the population.

Finally a choice between H_0 and H_1 is made. In a research environment this usually leads to further theories which are investigated in turn.

This concludes the description of the steps in hypothesis testing.

The $\mathbf{p}$-value

One of the criticisms against hypothesis testing is that it is too much of an all-or-nothing procedure: the final decision is either that H_0 is true or that H_1 is true without specifying how close to the truth H_0 is. The procedure makes no distinction between $H_1 : \theta = \theta_0 + \delta$ and $H_1 : \theta = \theta_0 + \omega$ where δ is very small and ω is very large. Thus if the decision rule is to reject H_0 if $\overline{X} > 45$, we shall reject H_0

if $\overline{X} = 45.1$ and if $\overline{X} = 1\,045.1$ where the values $\overline{X} = 45.1$ and $\overline{X} = 1\,045.1$ are treated as completely equivalent results with regard to the procedure of hypothesis testing. One way of overcoming this criticism at least partially, is by quoting the so-called p-*value* or *exceedance probability*.

Definition 2.8
The p-*value* is the **probability** that a value of the statistic, which is equal to or more than the observed value, will be obtained if H_0 is true.

For example if $\overline{X}$ is the statistic and $\overline{x}$ the observed value and we have the case where the alternative is $H_1 : \theta > \theta_0$ we will compute the p-value as

$$p\text{-value} = P\left(\overline{X} \geq \overline{x} \mid H_0 \text{ is true}\right).$$

If we have the two-sided alternative where $H_1 : \theta \neq \theta_0$, we will compute the p-value as

$$p\text{-value} = P\left(\left|\overline{X} - \theta_0\right| > |\overline{x} - \theta_0| \mid H_0 \text{ is true}\right)$$

- If this p-value is very small, $\overline{x}$ is said to be highly significant (usually if $p \ll \alpha$).
- If the p-value is fairly small, $\overline{x}$ is said to be significant (usually if $p < \alpha$).
- If the p-value is large, $\overline{x}$ is said to be not significant (usually if $p > \alpha$).

[NB We read the symbol "$\ll$" as "is much smaller than" whereas we read "$<$" only as "is smaller than".]

2.6 Confidence intervals

It was said in the previous paragraph that a criticism against hypothesis testing is that it is too much of an all-or-nothing procedure. If we decide that $\theta \neq \theta_0$ we still do not know by how much θ differs from θ_0. Unfortunately it is not possible to say what the exact value of θ is, but we may be able to construct an interval such that we can say with a given certainty that θ lies within the interval. This interval is called a confidence interval. Hypothesis testing and the construction of a confidence interval are not mutually exclusive or opposing procedures. They are based on the same statistical theory. In fact one may test the hypothesis $H_0 : \theta = \theta_0$ by first constructing a confidence interval for θ and rejecting H_0 if θ_0 is outside the interval.

A confidence interval may be two-sided, ie of the form $(T_1;\ T_2)$ where T_1 and T_2 are statistics, or one-sided, ie of the form $(-\infty;\ T)$ or $(T;\ \infty)$ according to the needs of the experimenter. Basically the construction of a two-sided confidence interval implies that one finds the smallest and largest values of θ such that the sample is not a rare event. For example, if θ is the mean of the distribution and $\overline{X}$ the sample mean, we may represent T_1 and T_2 as follows:

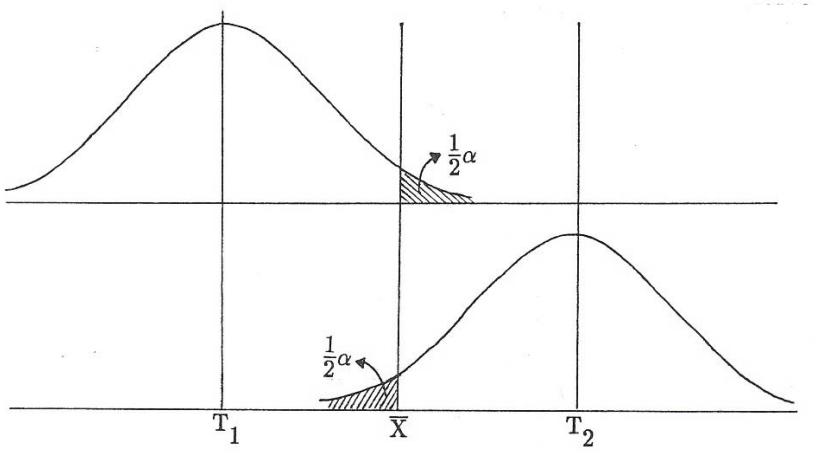


Figure 2.9

If $\theta < T_1$ then $\overline{X}$ is a rare event; similarly if $\theta > T_2$.

The method of construction is usually based on a function of θ, say $U(\theta)$, which would be a statistic if θ were known. $U(\theta)$ must be a function of θ but not of any other unknown parameter. The distribution of $U(\theta)$ must be known and independent of θ, so that a probability statement of the form

$$P\left(a \leq U(\theta) \leq b\right) = 1 - \alpha$$

may be made, where a and b are found from tables of the distribution of $U(\theta)$. For a one-sided interval we select $a = -\infty$ or $b = +\infty$ as the case may be. In the above equation α is again a small number between 0 and 1, such as $\alpha = 0.05$ or $\alpha = 0.01$.

The number $1 - \alpha$ is called the *confidence level* of the interval (compared to the term "significance level" for α in hypothesis testing). The inequality $a \leq U(\theta) \leq b$ is then manipulated algebraically to obtain an inequality of the form $T_1 \leq \theta \leq T_2$, so that we may say that

$$P\left(T_1 \leq \theta \leq T_2\right) = 1 - \alpha.$$

Note that we now have the unknown parameter inside the interval $(T_1;\ T_2)$. The end points of the interval are *statistics* and therefore *random variables*. Technically, this means we cannot say that the *probability* that θ lies between T_1 and T_2, is $(1 - \alpha)$. We therefore call it a *confidence interval* and not a probability interval. An interpretation of the confidence interval is the following:

If we draw repeated samples from the same population and compute the confidence interval every time, the true value of θ will lie inside the interval $100\,(1-\alpha)\,\%$ of the time and outside the interval $100\alpha\%$ of the time.

When we deal with a specific sample, θ lies either inside or outside the interval; it would seem strange to write

$$P\,(16 \leq \theta \leq 20) = 0.95$$

if we obtain $T_1 = 16$ and $T_2 = 20$ in a specific sample, because it would appear that θ is regarded as a random variable. Rather, we regard $(16;\ 20)$ as an interval chosen at random from a **population** of **intervals**, 95% of which contain θ and 5% of which do not contain θ.

Example 2.10

Let $X_1,\ ...,\ X_n$ be a random sample from a $n\left(\theta;\ \sigma^2\right)$ distribution with σ^2 unknown. Let

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

From theorem 1.2, study unit 1, we know that $\overline{X} \sim n\left(\theta;\ \sigma^2/n\right)$ and therefore

$$Z = \frac{\overline{X} - \theta}{\sigma/\sqrt{n}} = \sqrt{n}\left(\overline{X} - \theta\right)/\sigma$$

is a $n\,(0;\ 1)$ variate which is independent of $(n-1)\,S^2/\sigma^2$, which in turn is a χ^2_{n-1} variate (see result 1.3). Employing theorem 1.4 we obtain the following Student's t-variate:

$$\begin{aligned} U(\theta) &= \frac{\sqrt{n}\left(\overline{X} - \theta\right)/\sigma}{\sqrt{\left((n-1)\,S^2/\sigma^2\right)/\,(n-1)}} \\ &= \sqrt{n}\left(\overline{X} - \theta\right)/S, \end{aligned}$$

which is a Student's t-variate with $n-1$ degrees of freedom. From tables of the t-distribution we obtain $t = t_{\frac{1}{2}\alpha;\, n-1}$ such that

$$\begin{aligned}
1-\alpha &= P\left[-t \leq U(\theta) \leq t\right] \\
&= P\left[-t \leq \frac{\sqrt{n}\left(\overline{X}-\theta\right)}{S} \leq t\right] \\
&= P\left[\frac{-tS}{\sqrt{n}} \leq \overline{X}-\theta \leq \frac{tS}{\sqrt{n}}\right] \\
&= P\left[-\overline{X}-\frac{tS}{\sqrt{n}} \leq -\theta \leq -\overline{X}+\frac{tS}{\sqrt{n}}\right] \\
&= P\left[\overline{X}-\frac{tS}{\sqrt{n}} \leq \theta \leq \overline{X}+\frac{tS}{\sqrt{n}}\right]
\end{aligned}$$

therefore the interval $\left[\overline{X}-\frac{tS}{\sqrt{n}};\ \ \overline{X}+\frac{tS}{\sqrt{n}}\right]$ is a $100\,(1-\alpha)\,\%$ confidence interval for θ.

2.7 Simultaneous inference

In analysing the results of complex experiments, one may sometimes want to test a number of hypotheses or construct a number of confidence intervals.

Social scientists, for example, who carry out surveys may sometimes want to test several hundred hypotheses on the results of one survey. The problem is that the probability of a type I error increases as the number of tests or confidence intervals increases. For example, if 100 significance tests are performed, each at a 5% level of significance, then the probability of one or more type I errors could be very close to 1.

Definition 2.9

If k hypotheses $H_{01},\ ...,\ H_{0k}$ are tested **simultaneously**, then
the *overall significance level* is defined as

$P\,(\text{at least one } H_{0j} \text{ is rejected}|\ \text{all } H_{0j} \text{ are true}) = 1 - P\,(\text{no } H_{0j} \text{ is rejected}|\ \text{all } H_{0j} \text{ are true})\,.$

Definition 2.10

If k confidence intervals $I_1,\ ...,\ I_k$ are constructed for parameters $\theta_1,\ ...,\ \theta_k$, then the *overall significance level* is defined as

$$P\left(\theta_j \in I_j;\quad j = 1,\ ...,\ k\right)$$

ie the probability that all the intervals will contain the true values of the respective parameters.

The problem is: how to perform the k significance tests so that the overall significance level is α, or how to construct the k confidence intervals so that the overall confidence level is $1-\alpha$. This problem has been studied in great detail in the literature, and the best solution depends on the specific type of problem. One very general solution that can be applied to any simultaneous inference problem is based on the *Bonferroni inequality*.

Theorem 2.3

Let $E_1,\ E_2,\ ...,\ E_k$ be any k events in a sample space S. Then

$$P\left(E_1 \cup E_2 \cup ... \cup E_k\right) \leq P\left(E_1\right) + P\left(E_2\right) + ... + P\left(E_k\right).$$

To apply this theorem to a simultaneous testing problem, assume $H_{01},\ ...,\ H_{0k}$ are true and let

$$E_j = P\left(\text{reject } H_{0j} | \, H_{0j} \text{ is true}\right) = \alpha_j$$

say, the significance level of the j-th test. Then the overall significance level

$$= P\left(\text{at least one } H_{0j} \text{ rejected} | \text{ all } H_{0j} \text{ true}\right)$$

$$= P\left(E_1 \cup E_2 \cup ... \cup E_k | \text{ all } H_{0j} \text{ true}\right)$$

$$\leq \alpha_1 + \alpha_2 + ... + \alpha_k.$$

Thus if we choose $\alpha_j = \dfrac{\alpha}{k},\ j = 1,\ ...,\ k$, then the overall significance level is $\dfrac{\alpha}{k} + \dfrac{\alpha}{k} + ... + \dfrac{\alpha}{k} = \alpha$.

Thus if each test is performed at level $\dfrac{\alpha}{k}$ then the overall significance level is *at most* equal to α. Similarly, if each of k confidence intervals has confidence level $1 - \dfrac{\alpha}{k}$ then the overall confidence level is *at least* $1-\alpha$.

One point of criticism against using the Bonferroni inequality for this purpose is that the resulting procedure may be very conservative if k is large: the individual tests may have low power or the individual confidence intervals may be very wide because $\dfrac{\alpha}{k}$ is so very small.

One solution may be that the investigator (the biologist, engineer, social scientist, et cetera) should formulate the research problem better before starting, thus eliminating any fancy hypotheses that may have very little meaning. A practice that should be guarded against very carefully, and that is all too prevalent in certain disciplines, unfortunately, is to test many hypotheses on the same data and then to report only the significant ones as if they were the only ones tested. While this is downright dishonest, many scientists without a statistical training fail to see it that way. If you have trouble convincing a client, ask him or her whether he or she would be willing to play the following game:
We toss a coin: "Heads" I win, "Tails" we toss again. "Heads" I win, "Tails" we toss again. "Heads" I win, "Tails" we toss again ...

2.8 Bayesian inference

In the classical inference theory, as described in sections 2.4 and 2.5, we test a hypothesis about a parameter θ or construct a confidence interval for θ where θ is regarded as a fixed (but unknown) constant for a specified population. An alternative view is that θ is a random variable (or may be treated as if it were a random variable), and this leads to Bayesian inference. $X_1,\ X_2,\ ...,\ X_n$ is a random sample from a distribution with pdf $f_X\left(x|\,\theta\right)$ that depends on the parameter θ; θ is regarded as a random variable with *prior* distribution with pdf $g\left(\theta\right)$. Using Bayes' theorem, the *posterior* pdf of θ, given $X_1,\ X_2,\ ...,\ X_n$ is found, say $h\left(\theta|\,X_1,\ ...,\ X_n\right)$ and then the Bayes estimator of θ is the expected value of the posterior distribution. Significance tests and confidence limits are likewise based on the posterior distribution, but the subject is not pursued further in this module.

Exercise 2.1

1. Let $X_1,\ ...,\ X_n$ be a random sample from a distribution with expected value μ and variance θ.

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

   (Hint: Remember that $E\left(X_i^2\right)=\theta+\mu^2$ and $E\left(\overline{X}^2\right)=\frac{\theta}{n}+\mu^2$.)

2. Let $X_1,\ ...,\ X_n$ be a random sample from a $n\left(\mu;\ \theta\right)$ distribution with μ known. Prove that the two statistics

   $$T_1=\frac{1}{n}\sum_{1}^{n}\left(X_i-\mu\right)^2 \text{ and } T_2=\frac{1}{n-1}\sum_{1}^{n}\left(X_i-\overline{X}\right)^2$$

   are both unbiased estimators for θ and that T_1 has a smaller variance than T_2.

   (Hint: from study unit 1 it is known that multiples of T_1 and T_2 are χ^2 variates.)

3. Let $X_1,\ X_2,\ \ldots,\ X_n$ be independent random variables from a distribution such that

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

   where θ_1 and θ_2 are known parameters while $c_1,\ c_2,\ \ldots,\ c_n$ are known constants. Find the least squares estimators for θ_1 and θ_2.

4. Let $X_1,\ X_2,\ \ldots,\ X_n$ be independent random variables such that

   $$E(X_i)=\theta_1, \qquad i=1,\ \ldots,\ (n-1)$$

   $$E(X_n)=\theta_1+\theta_2.$$

   Find the least squares estimators for θ_1 and θ_2.

5. Let $X_1,\ \ldots,\ X_n$ be a random sample from a $n(\mu;\ \theta)$ distribution with μ known.

   Show that the MLE of θ is $\frac{1}{n}\Sigma\left(X_i-\mu\right)^2$.

6. Let $X_1, \ldots, X_n$ be a random sample from a distribution with pdf

$$f_X(x;\ \theta) = \theta(1-\theta)^{x-1}; \qquad x > 1.$$

Find the maximum likelihood estimator of θ.

7. Let $X_1, \ldots, X_n$ be a random sample from an exponential distribution with pdf

$$f_X(x;\ \theta) = \frac{1}{\theta} e^{-\frac{x}{\theta}} \qquad x > 0.$$

Find the M.L.E. for θ.

8. Let $X_1, \ ..., \ X_n$ be a random sample from a $n\left(\theta;\ \sigma^2\right)$ distribution with $\boldsymbol{\sigma^2}$ **known**. Use the distribution of $U(\theta) = \sqrt{n}\left(\overline{X} - \theta\right)/\sigma$ to 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. (Hint: $U(\theta)$ does not have a t-distribution.)

2.9 Learning outcomes

Use the following learning outcomes as a checklist after you have completed this study unit to evaluate the knowledge you have acquired.

After studying study unit 2, you should **know** (and understand!) the following definitions:

- a *random sample*
- a *statistic*
- an *unbiased estimator*
- the *most efficient* estimator
- the method of *least squares* estimation
- the *likelihood function* of a random sample
- the method of *maximum likelihood* estimation
- a *type I error* for hypothesis testing
- a *type II error* for hypothesis testing
- the *significance level* of a hypothesis test
- the *power* of a test
- the *exceedance probability* for hypothesis testing
- a *confidence interval*
- the *overall significance level* for k simultaneous hypothesis tests

STUDY UNIT 3

Introduction to statistical software: JMP

3.1 Introduction

Before we continue with any new statistical concepts in our study guide it would be a good idea nice to make the contents of the previous two study units more alive and applied – which calls for the use of a statistical package.

In the preface of your textbook, you will read that JMP is "statistical discovery software" created by the SAS Institute whose principal commercial product is the *SAS System*. Whereas the SAS System is used by large institutions such as STATSA or large banks to perform large-scale statistical data processing, JMP is used to perform smaller, personal data analyses. You will also read that the textbook is a mix of software manual and statistics text. This study unit will reflect that same mix – slanting a bit more towards the statistics text whereas the workbook will slant a bit more towards the software manual. You should also always keep in mind that the statistical software includes many advanced methods that will only be dealt with at honours level. Even the textbook deals with and include methods that are not in the syllabus of STA2601. **Hence, it is very important for this specific study unit that you only go to your workbook when I instruct you to do so and that you do not study sections at random.**

If you are using your computer for the first time I advise you to do activity 3.1 before you continue with the next section.

3.2 Familiarise yourself with JMP

I do hope that your brain tricked you into reading the three letters JMP as jump? That is correct! It is exactly why the textbook is called **JMP Start Statistics!** In this section we are even more bold to jump right into the software! It means that the time has come to get practical and to open your prescribed textbook: *Sall, J, Creighton, L and Lehman, A. (2007 4^{th} edition)* **JMP^{TM} Start Statistics.** The only way to familiarise yourself with JMP and to get to know the program is to work with the program! Of course the first step will be to install the software on your computer. You will notice as we proceed through the study guide that whenever you have to perform an action, the workbook will guide you step by step. Hence it seems logical that the workbook on study unit 3 will

be rather lengthy! Especially section 3.2 of the workbook will guide you in detail through the first sessions on the computer and I do hope that you enjoy your introduction to statistical software!

Data analysis starts with a data set. In this module our focus is not the methods of obtaining data but rather on the methods of analysing data. Don't get confused by the action of "obtaining data" in the context of a computer program – it will mostly mean **the capturing of values** such that you (and the computer) will have them displayed on the screen. We will guide you in the workbook to create new data tables and to open existing data tables.

Please note that I deviate from the chronological order of the textbook in a systematic and logical way to synchronise with the syllabus for STA2601. This different manner (which now differs from the authors' order) will seem haphazard if you do not follow my guidance. Thus I urge you to do all the activities in the workbook and also to try to stick to the order in which they are given.

Please work through section 3.2 of the workbook and do activities 3.2-3.4 before you continue with the next section.

3.3 Generating random data

The heading of this section is in itself an important concept to grasp. To "generate data" will imply that the computer goes through a process whereby random sampling from a specific population is simulated. (This seems like the marriage of the different nuances of the concept of "obtaining data" as explained in the context of a computer program as the capturing of values – both happening at the same time!)

The end result is that you will have a set of observations (data) that was drawn from a familiar distribution. "Familiar" means that we know the parameters which underlie the theoretical model. This is hardly the scenario when you are a real-life botanist or market researcher or whatever you do when you are busy with statistics in the outside world! However, analysing simulated data is useful because you more or less know what to expect of the data and thus it enhances your understanding of statistical theory. It also helps you to learn in an almost relaxed manner how to work with the powerful analysing and graphing techniques of JMP.

As we have stressed in the previous section, your first step with any statistical software application will be to have a data file in front of you – whether you play around with simulated data or have the task of analysing proper real-world data. In this specific section we will only work with **generated data, which are synonymous to simulated data.**

READ THROUGH

Sall, Creighton and Lehman, **Chapter 7** **Univariate distributions: one variable, one sample**

Start reading on page 122 "*Probability distributions*" **and read up to** ..."*Generating random data*".

Now it is time to go back to your workbook!

Work through section 3.3 and do activities 3.5–3.7.

It is important that you understand how to use "**Randdist.Jmp"** to create a "Randdist Data Table" on your computer screen and to make use of the "**Random Number Functions**" to generate a random sample of any specified size from a normal distribution with specified parameters.

From this point onwards I assume that you have worked through **activities 3.5–3.7** of the workbook and that the table of values shown below, makes sense to you. (I have used the "Random Number Functions" to generate a random sample of size $n = 200$ from a normal distribution with $\mu = 100$ and $\sigma = 15$ and copied table 3.1 of the workbook.**)**

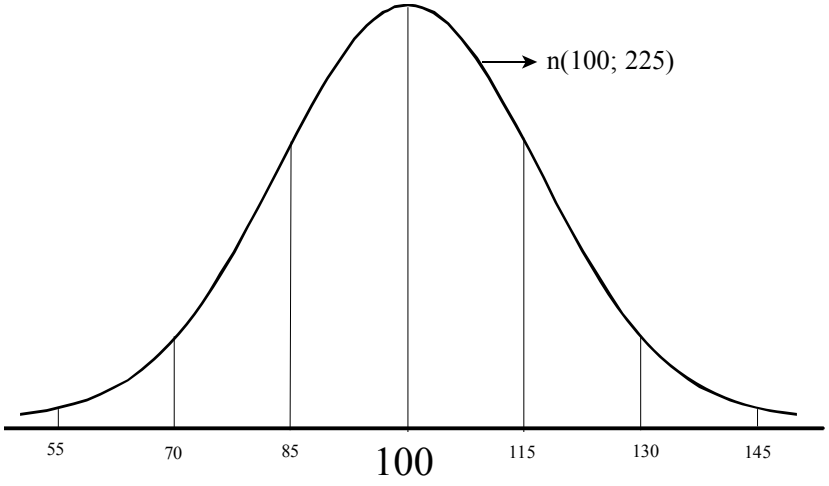
Table 3.1:
Random sample of size $n = 200$ from a $n(100; 225)$ distribution

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

We started this study unit with the wish to make the contents of the previous two study units more alive and applied and then detoured to introduce you to the very basics of JMP. So, you might well ask now, what do we make of the generated data and what they have to do with the previous two study units?

For starters, they cement your understanding of **section 1.3 "Standard distributions**" and more specifically it enhances your understanding of the normal distribution. (This is not trivial since the normal distribution is the workhorse of statistics!)

In section 1.3 you learned that if the theoretical model of a variable X is a normal distribution with mean μ and variance σ^2 , we write it as $X \sim n(\mu; \sigma^2)$. Thus, if we know that $\mu = 100$ and $\sigma^2 = 225$ we write it as $X \sim n(100; 225)$. Since the two parameters are known, it means that we have a workable probability distribution for which we may draw the following bell-shaped normal probability graph:



Note that $\sqrt{225} = 15 = \sigma$ (the SD) and that there are vertical lines at respectively one, two and three standard deviations above and below the mean.

From what we have learned in study unit 1 and employing the table of normal probabilities, we are 99% sure that the theoretical X-values will vary between 55 and 145.

Suppose we plan to draw a random sample from this specific normal population, what could we expect?

- We would expect that the smallest observed value will be > 55 and that the largest observed value will be < 145.
- We would expect 50% of the sample values to be above the mean $\mu = 100$ and 50% of the sample values to be below the mean $\mu = 100$. (This follows from the symmetrical property of the normal distribution.)
- Furthermore, we would not expect the "tail values" to dominate the sample as we would expect most ($\pm 68\%$) of the sample values to lie within one standard deviation below and above the mean, ie between 85 and 115.

Let us return to the generated sample of size $n = 200$ given in table 3.1: Keep in mind that what we expect is based on a theoretical model and always **remember that anything is possible in sampling and that randomness makes the world interesting.** This means that we can never be certain how a sample is going to turn out! This is of course also true for a generated sample. The authors of the textbook talk about the two sides of statistics that are "forever interacting, catalyzed by Random, the agent of uncertainty".

Did what-we-may-expect happen with the generated sample?

- The smallest observed value was 62.42448 and the largest observed value was 141.1041 (within our expectancy of the theoretical X-values varying between 55 and 145).
- There were $97 = 48.5\%$ of the sample values above the mean and $103 = 51.5\,\%$ of the sample values below the mean (again within our expectancy of the $50/50$ split).
- Did the "tail values" dominate the sample? We observe that there were 37 values below 85 and 31 values above 115, hence there were $200 - 37 - 31 = 132$ values between 85 and 115, in other words, $\frac{132}{200} = 66\%$ (again within our expectancy that $\pm 68\%$ of the sample values will lie within one standard deviation below and above the mean).

Big deal! The sample behaved as we would expect of a sample from a normal population because it came from a normal population! So what did you learn from this? Somehow we would like to assess if a sample really "passes a test" as coming from a normal population. In real life this whole process will be in reversed order! We will not know from what kind of distribution our sample comes. Remember that statistics, as seen as a discovery tool, would like to find patterns in the data and to fit models. What we did above was merely an intuitive test. In the next study unit you will formally learn about "**Testing for normality**".

What does this sample have to do with study unit 2?

Firstly, it illustrates in a practical way the concepts *random sample* and *statistic* which we defined in section 2.2. (In activities 3.8 and 3.9 of the workbook you will learn how to compute various statistics for this sample.)

According to the definition of a statistic, we may say that the following statistics have been computed for the generated sample of table 3.1 above:

$\sum_{i=1}^{200} X_i = 19\,980.7108$; $\sum_{i=1}^{200} X_i^2 = 2\,045\,288.17$; $\overline{X} = 99.90$; $var(X) = 246.956$; $median = 99.8152$

$Q_1 = 89.48$ and $Q_3 = 108.98$

Secondly, we could use this sample to illustrate concepts of *estimation*:
We could say that the sample mean, $\overline{X} = 99.90$, and the sample median, $me = 99.8152$, are both unbiased *point estimates* (section 2.3) of the population mean μ. With the knowledge of section 2.4 we could even go a step further and say that $\overline{X}$ is a maximum likelihood estimator for μ.

We could also say that $\overline{X}$ is a more efficient estimator than the median. This is the kind of theoretical information you will hardly ever see as output from a computer!

Thirdly, we could use this sample to illustrate sections 2.5 and 2.6 regarding *hypothesis testing* and *confidence intervals*.
From activities 3.8 and 3.9 we may state that a 95% confidence interval for the population mean μ is given by $(97.71; 102.09)$.

This confidence interval was computed in the blink of an eye by the computer. Are you able to do it manually? How do you interpret the interval? What about the hypothesis test? Questions like these will be discussed in detail when we deal with "**Tests for means**" in study unit 7.

To summarise:

In this study unit you were introduced to JMP which will not only be used to perform smaller, personal data analyses but which must be seen as "statistical discovery software". This powerful tool enhances understanding of the terminology of statistics and statistical thinking. One such an application was to create simulated data or generated data which implies that the computer goes through a process whereby random sampling from a specific population is simulated. The end result is that you will have a set of observations (data) that was drawn from a familiar distribution of whom you know the parameters which underlie the theoretical model. In this study unit we have only illustrated generated data for the normal distribution. However, if you are enrolled for the module *STA2603: Distribution Theory II*, you will use JMP again to generate random samples from other important theoretical distributions.

3.4 Learning outcomes

After studying study unit 3, you should **be able to**

- create a new data table in JMP
- open an existing data table in JMP
- *generate* data (ie simulate a random sample) from a specified $n(\mu; \sigma^2)$ distribution using JMP
- draw a *histogram* for a given sample using JMP
- draw an *Outlier and Quantile Box Plot* for a given sample using JMP
- compute basic *sample statistics* for a given sample using JMP

STUDY UNIT 4

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

4.1 Introduction

The normal distribution is probably the distribution which is used most often as model for statistical experiments. To some extent the use of the normal distribution can be justified because of the *central limit theorem*. If an observation X can be regarded as the sum of a large number of random components, for example

$$X = \mu + Y_1 + Y_2 + ... + Y_p$$

where $\mu = E(X)$, then X will, under fairly general conditions, be approximately normally distributed. If X is the size of a product manufactured in a factory, the deviation of X from its expected value may be the result of such factors as variation in the electrical current, machine setting, variations in the raw materials and the fact that the operator does not repeat his or her actions identically each time.

The analysis of the observations is therefore often based on the assumption that they come from a normal distribution. Sometimes this assumption is **not very crucial**, especially when the sample is large and the parameters which are being investigated are *expected values*. As a result of the central limit theorem it may be shown that Student's t-distribution is, for large samples, a good approximation to the distribution of

$$T = \sqrt{n}\left(\overline{X} - \mu\right)/S$$

even if $X_1,\ ...,\ X_n$ are not normally distributed. However, if the parameters under investigation are *variances* or *correlation coefficients*, the assumption of normality becomes more crucial. There is no "central limit theorem" which states, for example, that $(n-1)S^2/\sigma^2$ (where $S^2 = \Sigma\left(X_i - \overline{X}\right)^2/(n-1)$) is asymptotically distributed as a χ^2_{n-1} variate.

How do we know whether a sample comes from a normal distribution? How can we test whether a sample comes from a normal distribution?

4.2 Graphical techniques

Suppose we have a random sample $X_1,\ X_2,\ ...,\ X_n$ from a distribution. How will we investigate the possibility that this is a sample from a normal distribution?

A. Drawing a histogram

If the sample is **large enough**, we could construct a histogram in order to see whether it resembles the typical bell-shaped pdf of the normal distribution. If we draw a histogram with JMP, there is the option to superimpose the normal density curve corresponding with $\mu = \overline{X}$ and $\sigma = S$ over the histogram. If this superimposed pdf fits snugly over the histogram, and the intervals of the histogram are not too wide, we may subjectively conclude "a good fit". The problem is, when will you decide the "fit is not good"? Secondly, a histogram based on a small sample will not be very informative.

B. Using normal probability paper

We know from section 1.3 that the (cumulative) standardised normal distribution function is given by

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{1}{2}u^2} du \qquad \text{(see definition 1.17).}$$

Let for example $X \sim n(10;\ 4)$, ie $Z = \dfrac{X-10}{2} \sim n(0;\ 1)$.

$$\therefore P(X < x) = P\left(\frac{X-10}{2} < \frac{x-10}{2}\right) = P\left(Z < \frac{x-10}{2}\right) = \Phi\left(\frac{x-10}{2}\right).$$

In pre-computer days, special graph paper, called normal probability paper, was constructed such that, if $\Phi(z)$ was plotted against z, the result was a straight line. (See figure 4.1.) Note that $100\Phi(z)$ is marked on the vertical axis rather than $\Phi(z)$.

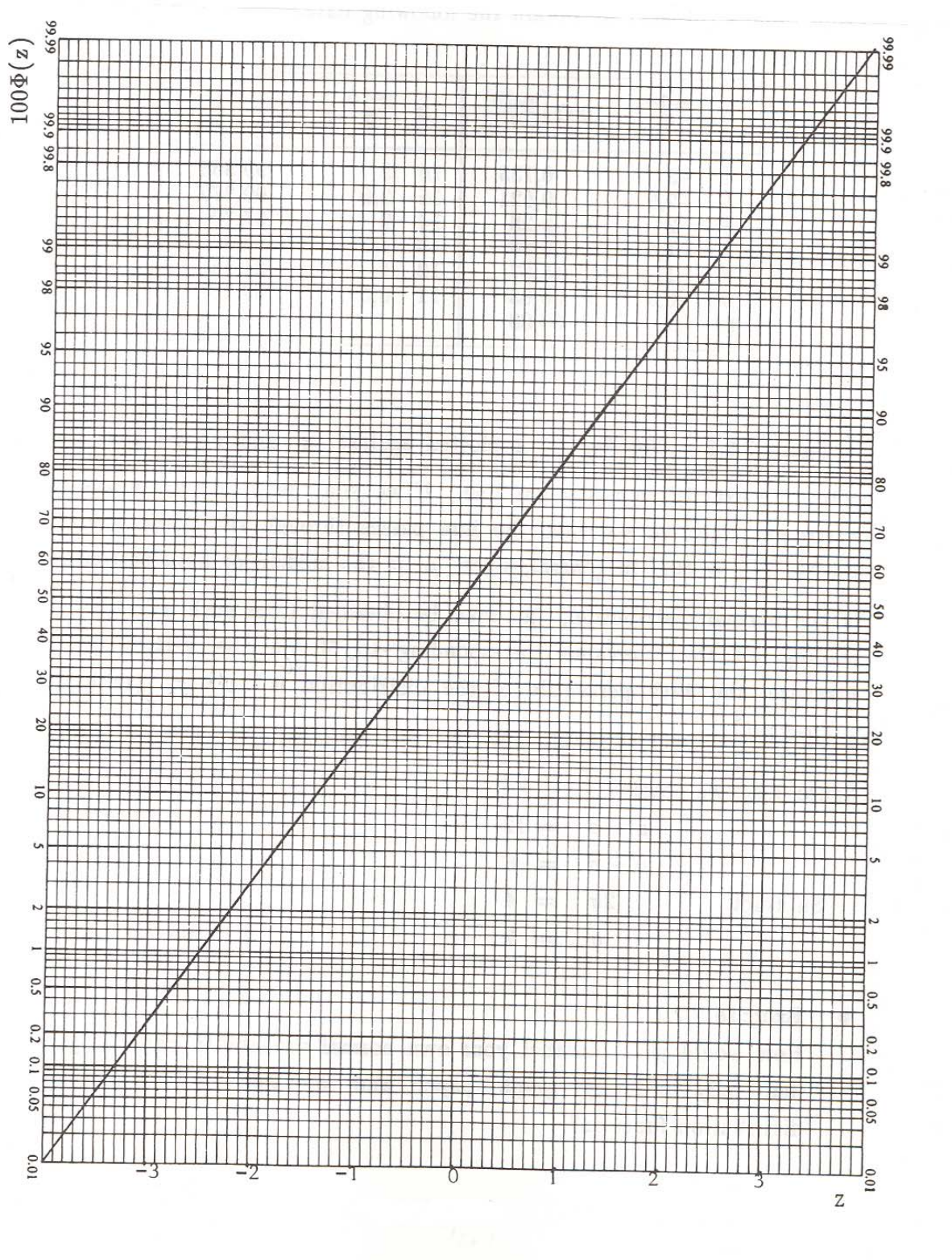

100Φ(z)
99.99
99.9
99.8
99
98
95
90
80
70
60
50
40
30
20
10
5
2
1
0.5
0.2
0.1
0.05
0.01
−3
−2
−1
0
1
2
3
z

Figure 4.1

The idea behind the graphical technique is to compare computed percentiles of the observed sample with theoretical percentiles of a normal distribution. In other words, you will have to draw the set of "paired" points (observed; expected) on the graph and hope that they fall more or less on the straight line. The interpretation remains subjective (whether by hand or by computer). Keep in mind that because points constitute a random sample, they will not lie exactly on a straight line and we will only conclude non-normality if there appears to be a **systematic deviation** from the line. Doing this by hand is rather outdated and the special probability paper is difficult to obtain because computers have taken over all the tedious tasks!

Although we will not draw such graphs by hand, you need to understand the principle behind the technique.

Theoretical percentiles of normal distribution

For a normal distribution, the mean μ is also the median or the 50th percentile. Note that in figure 4.1 the mean of Z is zero and hence the value $z = 0$ corresponds to $100\Phi(z) = 50$.

For any normal distribution, the value $\mu + \sigma$ will represent the 84th percentile. For the $n(0;\ 1)$ distribution $\mu + \sigma = 0 + 1$ and hence the value $z = 1$ corresponds to $100\Phi(z) = 84$ in figure 4.1.

Why is this the case? (Please see activity 4.4 of the workbook.)

Similarly the value $\mu - \sigma$ will represent the 16th percentile. Using table I of Stoker, we could compile the following table:

Table 4.1

z	$100\Phi(z)$
-3	0.135
-2.5	0.621
-2	2.28
-1.5	6.68
-1	15.87
-0.5	30.85
0	50.00
0.5	69.15
1.0	84.13
1.5	93.32
2	97.72
2.5	99.379
3	99.865

If we plot these points on *ordinary graph paper* we will get the curved standardised normal distribution function looking like the one in figure 1.7 of section 1.3, but if we plot them on the paper of figure 4.1 they will all fall on the straight line.

The special probability paper makes it easier to detect deviations from the cumulative distribution function because our eyes are trained to detect deviations from a straight line.

C. Normal quantile plots

The discussion in Sall, Creighton and Lehman is a little confusing at first glance because their histograms and accompanying normal quantile plots look "tilted by 90°". Let us first understand the principle in terms of an ordinary XY-graph before you work through this section in the textbook. The horizontal axis will represent the observed values and the vertical axis will represent the expected value under the normal distribution associated with a specific probability p. One of the problems will be to decide on the value of this probability p.

Let $X_1,\ ...,\ X_n$ be a random sample from a $n\left(\mu;\ \sigma^2\right)$ distribution (with μ and σ^2 unknown). **Arrange the observations in order of magnitude**, and call the result $X_1^*,\ ...,\ X_n^*$, so that $X_1^* < X_2^* < ... < X_n^*$. Then $X_1^*,\ ...,\ X_n^*$ are the *order statistics* of a sample of size n from a normal distribution. On the probability paper X_i^* will be plotted on the horizontal axis. What is the vertical coordinate which corresponds to X_i^*?

You will see that Sall, Creighton and Lehman state that the normal quantile values are $\Phi^{-1}\left(\frac{r_i}{n+1}\right)$ where r_i is the rank of the observation being scored.

How will you do this manually?

Example 4.1

Consider the following sample $(X_1,\ X_2,\ ...,\ X_{19})$ of 19 observations:

2.75	6.80	4.51	7.45	6.49	4.99	8.72	6.28	6.12	3.40	7.30
7.00	7.66	5.34	4.88	4.20	9.47	5.81	8.30			

If we rank the values from small to large (to obtain X_i^*) and compute $\frac{ri}{n+1}$ for each ranked value, we get:

Table 4.2

r_i	1	2	3	4	5	6	7	8	9	10
X_i^*	2.75	3.40	4.20	4.51	4.88	4.99	5.34	5.81	6.12	6.28
$r_i/(n+1)$	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50

r_i	11	12	13	14	15	16	17	18	19
X_i^*	6.49	6.80	7.00	7.30	7.45	7.66	8.30	8.72	9.47
$r_i/(n+1)$	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95

Please note that in pre-computer days you would plot $\left(X_i^*;\ 100\left(\frac{r_i}{(n+1)}\right)\right)$ on the special probability paper and if the coordinates fell more or less on the straight line, without any systematic deviation, you could conclude that the sample comes from a normal distribution.

But, the computer does not use the special probability paper because it CONVERTS $\left(\frac{r_i}{(n+1)}\right)$ to an expected normal score. What does this mean?

The formula $\frac{r_i}{(n+1)}$ is called Van der Waerden's formula. Van der Waerden argued that we may associate a probability of 0.05 with the smallest observation in a sample; we may associate a cumulative probability of 0.10 with the second smallest observation, et cetera, up to a cumulative probability of 0.95 with the largest observation. [This is when $n = 19$. For a sample of size $n = 99$ we will assume a probability of 0.01 with the smallest observation and a cumulative probability of 0.99 with the largest observation.] Other statisticians have proposed different formulae to compute the corresponding cumulative probability associated with the rank r_i.

For example Tukey's formula is $\frac{(3r_i - 1)}{(3n + 1)}$ and

Blom's formula is $\frac{(8r_i - 3)}{(8n + 2)}$.

We will only consider Van der Waerden's method seeing that this is the one the authors of the textbook also prefer.

So far so good! Now, how will you compute $\Phi^{-1}\left(\frac{r_i}{n+1}\right)$?

For example, $\Phi^{-1}(0.05)$ translated into ordinary English means "find a z-value such that $P(Z \leq z) = 0.05$". This means that we have to use the inverse normal table. From first-year applications of the normal distribution we know that we have to manipulate table II if $p < 0.50$. Are you able to show that $\Phi^{-1}(0.05) = -1.645$: (See activity 4.5 of the workbook.)

This means that $P(Z \leq -1.645) = 0.05$

In a similar fashion,
$$\begin{aligned} \Phi^{-1}(0.10) &= -1.282 \\ \Phi^{-1}(0.15) &= -1.036 \\ &\text{et cetera} \\ &\vdots \\ \Phi^{-1}(0.95) &= 1.645 \end{aligned}$$

Keep in mind that these $\Phi^{-1}\left(\frac{r_i}{n+1}\right)$ values are the standardised z-values and we are interested in the values corresponding to the X_i^*-scale.

Thus, the final step is to transform the variable Z to X^* and for this we need μ and σ. We do not have μ and σ^2 for the population but we use the estimates from the sample.

$$\hat{\mu} = \overline{X} = 6.1826$$

$$\hat{\sigma} = S = 1.7973$$

Hence, the expected X_i^*-value for a $n\left(6.1826;\ (1.7973)^2\right)$ distribution associated with a probability of 0.05 is $(-1.645)(1.7973) + 6.1826 = 3.23$. Similarly, the expected X_i^*-value for $\Phi^{-1}(0.10)$ is 3.88.

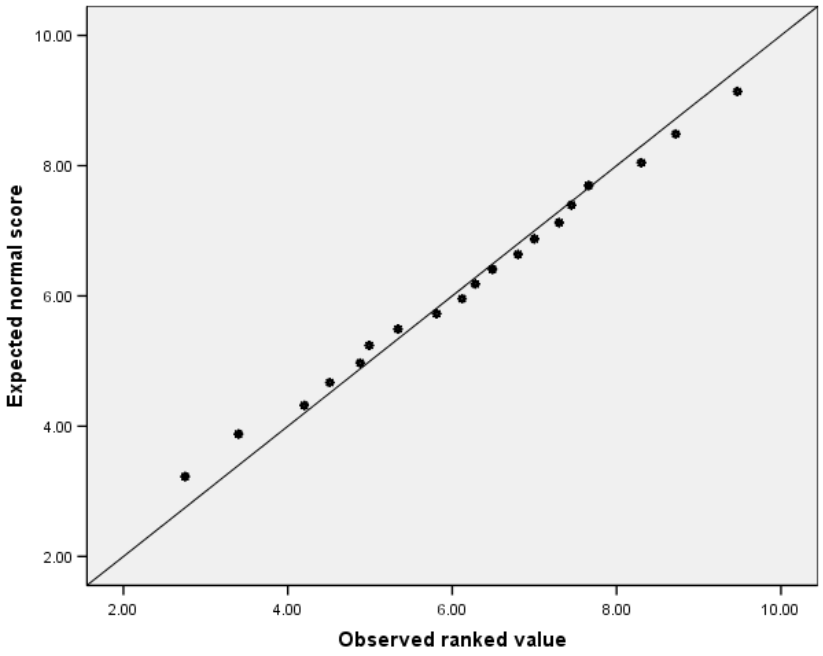
Do you agree that it is very laborious to do this for all 19 observations by hand? This is why we rely on JMP to draw the **normal quantile plot**.

We can summarize all the calculations in the following table:

Table 4.3

r_i	X_i^*	$\frac{r_i}{(n+1)}$	$\Phi^{-1}\left(\frac{r_i}{n+1}\right)$	Expected X_i^*-score
1	2.75	0.05	−1.645	3.23
2	3.40	0.10	−1.282	3.88
3	4.20	0.15	−1.036	4.32
4	4.51	0.20	−0.842	4.67
5	4.88	0.25	−0.674	4.97
6	4.99	0.30	−0.524	5.24
7	5.34	0.35	−0.385	5.49
8	5.81	0.40	−0.253	5.73
9	6.12	0.45	−0.126	5.96
10	6.28	0.50	0.000	6.18
11	6.49	0.55	0.126	6.41
12	6.80	0.60	0.253	6.64
13	7.00	0.65	0.385	6.87
14	7.30	0.70	0.524	7.12
15	7.45	0.75	0.674	7.39
16	7.66	0.80	0.842	7.70
17	8.30	0.85	1.036	8.04
18	8.72	0.90	1.282	8.49
19	9.47	0.95	1.645	9.14

To draw a normal quantile plot similar to the one produced by JMP, you will have to draw a scatter plot of the data pairs $(X_i^*;$ expected X_i^*-score$)$ on ordinary graph paper.



READ THROUGH

Sall, Creighton and Lehman, **Chapter 7** **Univariate distributions: one variable, one sample**

Start reading on page 127 "*Histograms*" **and read up to** ....."*Outlier and quantile box plots*".
Then read page 152 "*Examining for normality- normal quantile plots*".

It will not be expected of you to draw a normal quantile plot manually, but you must be able to do it with JMP. (See activity 4.6.)

Please note that the data for example 4.1 were in fact generated from a $n(6;4)$ distribution for illustrative purposes. In order to show what a sample from a non-normal distribution may look like when plotted on probability paper and converted to a normal quantile plot, consider the following 19 observations:

Example 4.2

The following are the order statistics of a random sample from a non-normal distribution:

6.06	7.32	8.64	9.06	9.48	9.66
10.08	10.50	10.86	11.04	11.22	11.52
11.70	11.94	12.06	12.24	12.72	13.02
13.50					

In order to perform a manual normal quantile plot we have to go through the same laborious process as in example 4.1.
For this sample $\overline{X} = 10.66421$ and $S = 1.94369$.

Table 4.4

r_i (rank)	Observed value	$\frac{r_i}{20}$	$\Phi^{-1}\left(\frac{r_i}{n+1}\right)$	Expected normal quantile
1	6.06	0.05	−1.645	7.47
2	7.32	0.10	−1.282	8.17
3	8.64	0.15	−1.036	8.65
4	9.06	0.20	−0.842	9.03
5	9.48	0.25	−0.674	9.35
6	9.66	0.30	−0.524	9.64
7	10.08	0.35	−0.385	9.92
8	10.50	0.40	−0.253	10.17
9	10.86	0.45	−0.126	10.42
10	11.04	0.50	0.000	10.66
11	11.22	0.55	0.126	10.91
12	11.52	0.60	0.253	11.16
13	11.70	0.65	0.385	11.41
14	11.94	0.70	0.524	11.68
15	12.06	0.75	0.674	11.97
16	12.24	0.80	0.842	12.30
17	12.72	0.85	1.036	12.68
18	13.02	0.90	1.282	13.16
19	13.50	0.95	1.645	13.86

If we plot the observed values versus the expected normal values on ordinary graph paper we get figure 4.3.

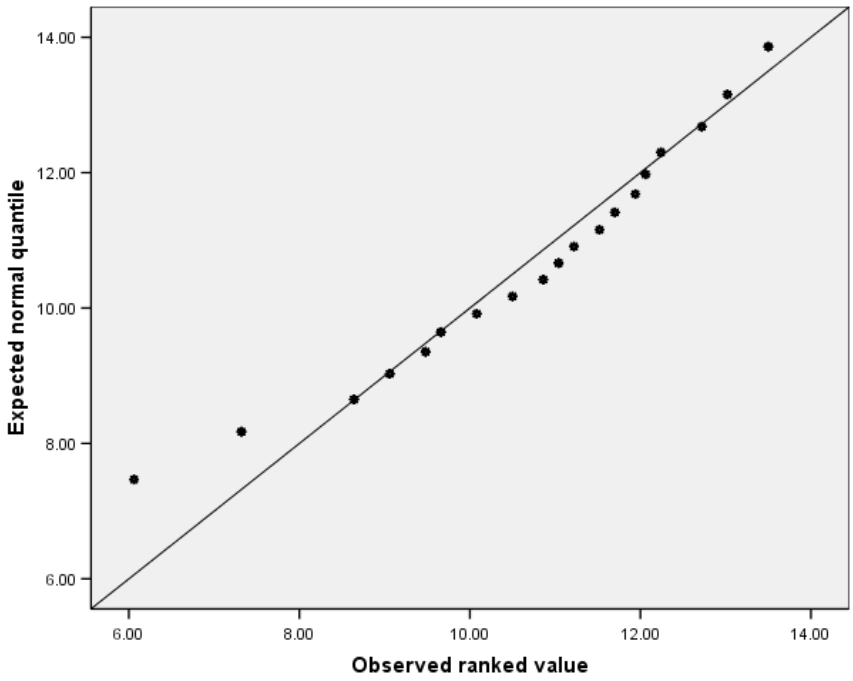


Figure 4.3

If we plot the observed values versus $100\left(\frac{i}{n+1}\right)$ on the special probability paper we get figure 4.4.

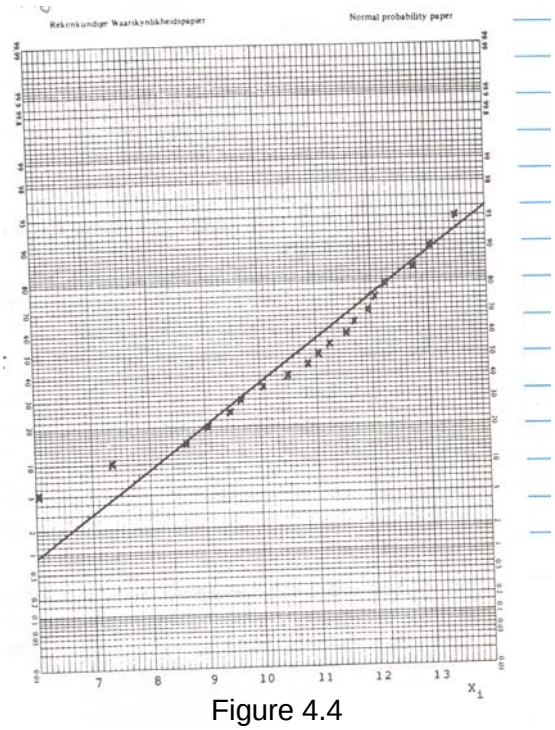


Figure 4.4

We get exactly the same picture for both methods and we see that there is a **systematic deviation** from a straight line. The first few points on the left and the last few points on the right are above the line, while the points in the middle are below the line. We conclude that the sample is probably not from a normal distribution.

This graphical method is rather subjective, but is often sufficient to enable us to make a decision. Usually we only want to know whether the normal distribution is a fair approximation to the true (but unknown) distribution from which the sample came.

If a subjective graphic investigation of the data is not sufficient, one may decide to perform a test for normality. A number of tests for normality exist, for example one based on the correlation coefficient of the points on the probability plot – if the points fall close to a straight line, one would expect the correlation coefficient to be close to 1. Special tables are needed for this test, **and we shall not consider it further**.

We will consider two other possible tests for normality, namely the goodness-of-fit test and the method-of-moments test.

4.3 Goodness-of-fit test for normality

We started our discussion on graphical techniques with the possibility of drawing a *histogram* which we had to judge subjectively to decide whether it deviates from the form of a normal distribution. In a sense we are now going to continue with a *histogram* but we are going to try and "measure" how far it deviates from a histogram of a normal distribution and we do it by way of a proper hypothesis test.

In general a test for *goodness of fit* checks the agreement (consistency) between a set of *observed data* and a proposed model. In other words, it is a test that can be used to test a *distribution type*. (In this section we specify the type as *normal* but any other known statistical distribution can be specified.)

The **null hypothesis** must always **specify the distribution** that is being tested, and the distribution must be fully specified (no unknown parameters) in order to compute the theoretical or expected values for the given intervals. Suppose there are k *intervals* into which the data are classified. For the time being, please accept the following result which is an application of theorem 4.1 that follows in the next section of this study unit. (In section 4.4 you will also see why it makes sense to denote the test statistic as a squared value.)

The appropriate test statistic is the chi-square statistic

$$Y^2 = \sum_{i=1}^{k} \frac{(\text{observed frequency } - \text{ expected frequency})^2}{\text{expected frequency}}$$

which is approximately distributed as χ^2_{k-1}. This is only true if the theoretical distribution is completely specified (for example $H_0 : X$ has a $n\left(2.5;\ 46\right)$ distribution).

If the distribution is not completely specified (for example $H_0 : X$ has a $n\left(\mu;\ \sigma^2\right)$ distribution) the test statistic will be approximately a χ^2_{k-1-r} variable where $r =$ number of unknown parameters that are estimated).

Example 4.3

Suppose we have the following random sample of 100 observations and we wish to test the null hypothesis that the sample comes from a $n\left(50;\ 100\right)$ distribution. Use **ten class intervals of equal expected frequencies** to perform the test.

(Please note that the sample values have been ordered from small to large to ease the classification into intervals.)

32.0	32.5	33.3	33.4	33.8	34.0	34.4	34.6	35.0	35.4
36.0	36.4	36.8	37.0	37.4	37.5	37.7	38.1	38.6	38.7
39.1	39.4	39.7	40.2	40.3	40.5	40.8	41.0	41.1	41.5
41.6	42.3	42.8	43.5	43.7	44.1	44.4	44.7	44.9	45.4
45.7	46.3	46.8	47.4	47.5	47.5	47.7	47.8	48.1	48.3
48.4	48.8	49.3	49.7	49.9	50.1	50.3	50.6	51.4	51.7
51.9	52.4	52.6	53.7	54.1	54.8	55.2	55.3	56.4	56.8
57.3	57.6	58.2	58.8	59.0	59.1	59.3	59.8	60.2	60.6
61.0	61.3	61.9	62.4	62.6	62.7	62.9	63.2	63.5	63.8
64.1	64.3	65.0	65.4	65.7	66.5	66.8	67.2	67.7	68.0

Solution

If we have to use **10** class intervals of equal expected frequencies, it means the theoretical model (which is the normal distribution) will have **five** classes below the mean and **five** classes above the mean. (This is because the normal distribution is symmetrical.) The most difficult part of this problem is to find the limits of the intervals in order to classify the observed values.

Since we know that the probability of each interval must be $\frac{1}{10}$, we will use table II of Stoker which gives a z-value for a known area. Thus, the first step will be to find the limits in terms of the Z-scale and then to transform back to the X-scale where

$$Z = \frac{X - \mu}{\sigma} = \frac{X - 50}{\sqrt{100}}.$$

(We can even take a "shortcut" for our use of table II and use every second line of table 4.3 where $\Phi^{-1}\left(\frac{r_i}{n+1}\right)$ is actually the z-value associated with a given probability!)

Hence we know that
$$\begin{aligned} P(Z < -1.282) &= 0.10 \\ P(Z \leq -0.842) &= 0.20 \\ P(Z \leq -0.524) &= 0.30 \\ \vdots \text{ et cetera} &\quad \vdots \\ P(Z \leq 0.842) &= 0.80 \\ P(Z \leq 1.282) &= 0.90 \end{aligned}$$

From this it follows that the 10 intervals are

Z-scale	X-scale
$Z \leq -1.282$	$X \leq 37.18$
$-1.282 \leq Z \leq -0.842$	$37.18 \leq X \leq 41.58$
$-0.842 \leq Z \leq -0.524$	$41.58 \leq X \leq 44.76$
$-0.524 \leq Z \leq -0.253$	$44.76 \leq X \leq 47.47$
$-0.253 \leq Z \leq 0$	$47.47 \leq X \leq 50.00$
$0 \leq Z \leq 0.253$	$50.00 \leq X \leq 52.53$
$0.253 \leq Z \leq 0.524$	$52.53 \leq X \leq 55.24$
$0.524 \leq Z \leq 0.842$	$55.24 \leq X \leq 58.42$
$0.842 \leq Z \leq 1.282$	$58.42 \leq X \leq 62.82$
$Z \geq 1.282$	$X \geq 62.82$

This conversion can be represented in the following figure:

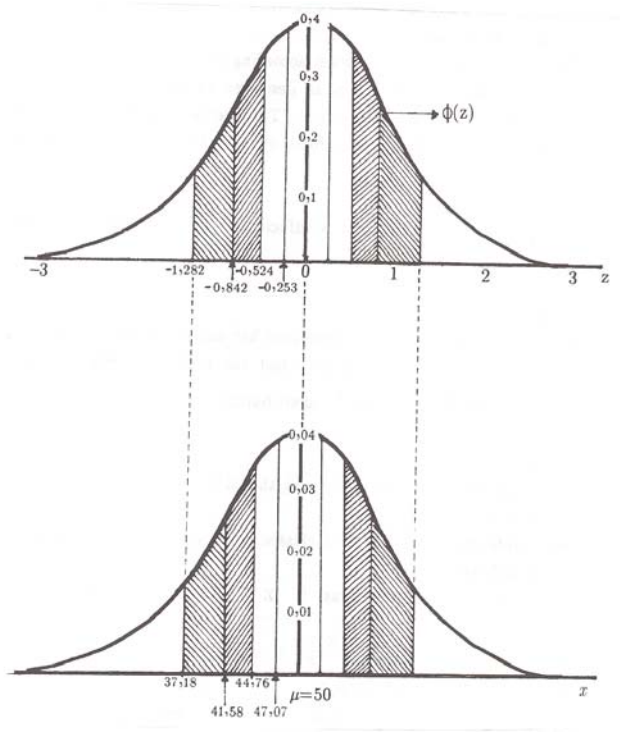


Figure 4.5

Using the 10 intervals, we may now classify the data to obtain the following table:

(I have added a second column called "Tally marks" which is what one would normally have to do if your data are *not arranged* from small to large and you have to classify them by hand. It is a simple way of counting where 𝍸 represents five observations.)

Table 4.5

Interval	Tally marks	Observed frequency, N_i	Expected frequency, $\hat{e}_i = n\hat{\pi}_i$	$(N_i - \hat{e}_i)$
$X < 37.18$	𝍸 𝍸 ||||	14	10	$+4$
$37.18 \leq X < 41.58$	𝍸 𝍸 𝍸 |	16	10	$+6$
$41.58 \leq X < 44.76$	𝍸 |||	8	10	-2
$44.76 \leq X < 47.47$	𝍸 |	6	10	-4
$47.47 \leq X < 50.00$	𝍸 𝍸 |	11	10	1
$50.00 \leq X < 52.53$	𝍸 ||	7	10	-3
$52.53 \leq X < 55.24$	𝍸	5	10	-5
$55.24 \leq X < 58.42$	𝍸 |	6	10	-4
$58.42 \leq X < 62.82$	𝍸 𝍸 |||	13	10	$+3$
$X \geq 62.82$	𝍸 𝍸 ||||	14	10	$+4$
Totals		100	100	

We have to test the following hypotheses:

H_0 : The sample comes from a $n\left(50;\ 100\right)$ distribution.
H_1 : The sample does not come from a $n\left(50;\ 100\right)$ distribution.

We compute the test statistic as

$$\begin{aligned} Y^2 &= \sum_{i=1}^{10} \left(N_i - \hat{e}_i\right)^2 / \hat{e}_i \\ &= \frac{16}{10} + \frac{36}{10} + ... + \frac{16}{10} \\ &= 14.8. \end{aligned}$$

We will reject the null hypothesis at the 5% level of significance if $Y^2 \geq \chi^2_{0.05;\ 10-1} = \chi^2_{0.05;9} = 16.919$.

Since $14.8 < 16.919$ we cannot reject the null hypothesis and conclude that the sample could have come from a $n\left(50;\ 100\right)$ distribution. Suppose, however, we had chosen $\alpha = 0.10$. Now $\chi^2_{0.10;9} = 14.6837$. Since $Y^2 > 14.6837$ we reject H_0 at the 10% level of significance and conclude that the underlying distribution is not normal.

It is informative in this case to look at the discrepancies $N_i - \hat{e}_i$. We see that these are mostly positive in the tails and negative in the middle. This suggests that the distribution is rather leptokurtic compared to the normal distribution. (This will be discussed in detail in section 4.5.)

In a more realistic or real-life situation, we will most often not know what the parameters of the distribution are, and the instruction for the hypothesis test will change to: "Use ten class intervals of equal expected frequencies and perform a hypothesis test to test for normality".

How will this change the solution to example 4.3?

Example 4.4

Refer to the data of example 4.3. Use ten class intervals of equal expected frequencies and test whether the data come from a $n\left(\mu;\ \sigma^2\right)$ distribution.

Solution

The first part of the solution will be the same as the first part for example 4.3 (in other words where we find the z-values corresponding to probabilities of $\frac{1}{10}$.)

The difference is that μ and σ are unknown and have to be estimated from the sample. We have to use the **maximum likelihood estimators** of μ and σ^2. (This is an application of theorem 4.3 which follows towards the end of the following section.)

For this sample $\hat{\mu} = \overline{X} = \frac{4\,945.7}{100} = 49.457$

and the M.L.E. $\hat{\sigma} = \sqrt{\frac{\Sigma\left(X_i - \overline{X}\right)^2}{n}}$ (note that we divide by n and not by $(n-1)$)

hence $\hat{\sigma} = \sqrt{\frac{10\,812.41}{100}} = 10.398.$

If we now use $Z = \frac{X - \hat{\mu}}{\hat{\sigma}} = \frac{X - 49.457}{10.398}$ we will get the following 10 intervals:

Z-scale	X-scale
$Z \leq -1.282$	$X \leq 36.13$
$-1.282 \leq Z \leq -0.842$	$36.13 \leq X \leq 40.70$
$-0.842 \leq Z \leq -0.524$	$40.70 \leq X \leq 44.01$
$-0.524 \leq Z \leq -0.253$	$44.01 \leq X \leq 46.83$
$-0.253 \leq Z \leq 0$	$46.83 \leq X \leq 49.46$
$0 \leq Z \leq 0.253$	$49.46 \leq X \leq 52.09$
$0.253 \leq Z \leq 0.524$	$52.09 \leq X \leq 54.91$
$0.524 \leq Z \leq 0.842$	$54.91 \leq X \leq 58.21$
$0.842 \leq Z \leq 1.282$	$58.21 \leq X \leq 62.79$
$Z \geq 1.282$	$X \geq 62.79$

Classifying the data into these classes leads to the following table:

Table 4.6

Interval	Observed frequency, N_i	Expected frequency, $\hat{e}_i = n\hat{\pi}_i$	$(N_i - \hat{e}_i)^2$
$X \leq 36.13$	11	10	1
$36.13 \leq X \leq 40.70$	15	10	25
$40.70 \leq X \leq 44.01$	9	10	1
$44.01 \leq X \leq 46.83$	8	10	4
$46.83 \leq X \leq 49.46$	10	10	0
$49.46 \leq X \leq 52.09$	8	10	4
$52.09 \leq X \leq 54.91$	5	10	25
$54.91 \leq X \leq 58.21$	7	10	9
$58.21 \leq X \leq 62.79$	13	10	9
$X \geq 62.79$	14	10	16
Totals	100	100	

We have to test:

H_0 : The sample comes from a normal distribution.
H_1 : The sample does not come from a normal distribution.

$$\begin{aligned} Y^2 &= \sum_{i=1}^{10} \frac{(N_i - \hat{e}_i)^2}{\hat{e}_i} \\ &= \frac{1}{10} + \frac{25}{10} + \frac{1}{10} + \frac{4}{10} + 0 + \frac{4}{10} + \frac{25}{10} + \frac{9}{10} + \frac{9}{10} + \frac{16}{10} \\ &= 9.40 \end{aligned}$$

We have $k - 1 = 9$ and $k - r - 1 = 7$; $\chi^2_{0.05;7} = 14.0671$.

Since $9.40 < 14.0671$ we cannot reject H_0. We may conclude that the sample comes from a normal distribution.

A variation on the theme of goodness of fit for a normal distribution, is that a specific set of intervals with observed data is given and then one has to test for normality. In other words you are given the tabular equivalent of a histogram (which most often consists of a number of intervals with the same length). This means that you need not compute the limits because you are given a set of intervals (all with the same lengths) as well as the observed frequencies. The problem will be to *find the expected frequencies* under the assumption that a normal curve will be superimposed over these intervals. So, here we have a proper statistical test appropriate for the first graphical technique of the previous section.

Example 4.5

Refer to the data of example 4.3. These 100 values can be classified into the following frequency table:

Table 4.7

Interval	Observed frequency
$29.95 - 33.95$	5
$33.95 - 37.95$	12
$37.95 - 41.95$	14
$41.95 - 45.49$	10
$45.95 - 49.95$	14
$49.95 - 53.95$	9
$53.95 - 57.95$	8
$57.95 - 61.95$	11
$61.95 - 65.95$	12
$65.95 - 69.95$	5
Total	100

Suppose the instruction is similar to that of example 4.3: "**Test the null hypothesis that the sample comes from a** $n\,(50;\ 100)$ **distribution**".

Solution

The trap is to assume that the expected frequencies are 10 for each interval (as we had in the previous two examples). Please note that this is not the case. We now have a different scenario where the expected probability for each interval has to be computed by making use of table I (Stoker).

The first step is to standardise the interval limits of the X-scale to the corresponding interval limits of the Z-scale. Since it was given as part of the null hypothesis that $\mu = 50$ and $\sigma = 10$, we use $Z = \dfrac{X - 50}{10}$.

The second step is to compute the corresponding probabilities $P\,(a \leq Z \leq b)$ for each interval by making use of table I (Stoker). This is laborious work!

Both these steps are summarised in the following table:

Table 4.8

Intervals		Expected
X-scale	Z-scale	probability (π_i)
$29.95 - 33.95$	$Z \leq -1.61$	0.0537
$33.95 - 37.95$	$-1.61 \leq Z \leq -1.21$	0.0594
$37.95 - 41.95$	$-1.21 \leq Z \leq -0.81$	0.0959
$41.95 - 45.49$	$-0.81 \leq Z \leq -0.41$	0.1319
$45.95 - 49.95$	$-0.41 \leq Z \leq -0.01$	0.1551
$49.95 - 53.95$	$-0.01 \leq Z \leq 0.40$	0.1594
$53.95 - 57.95$	$0.40 \leq Z \leq 0.80$	0.1327
$57.95 - 61.95$	$0.80 \leq Z \leq 1.20$	0.0968
$61.95 - 65.95$	$1.20 \leq Z \leq 1.60$	0.0603
$65.95 - 69.95$	$Z \geq 1.60$	0.0548
		1.0033

The expected frequencies for the intervals are found by multiplying the expected probability by the sample size.

Do you notice that the first and the last interval for the Z-scale are *open-ended*? This is necessary to ensure that $\sum_{i=1}^{10} \pi_i = 1$. However, if we add the values in the last column we get 1.0033. This is due to rounding in table 1 which results in a cumulative rounding error.

We use the same goodness-of-fit test statistic:

$$\begin{aligned} Y^2 &= \sum_{i=1}^{10} \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \\ &= \frac{(5-5.37)^2}{5.37} + \frac{(12-5.94)^2}{5.94} + \frac{(14-9.59)^2}{9.59} + ... + \frac{(5-5.48)^2}{5.48} \\ &= 0.086 + 6.182 + 2.028 + ... + 0.042 \\ &= 20.462 \end{aligned}$$

Since the number of classes did not change, we use the same critical value χ^2 as for example 4.3.

$$\chi^2_{0.05;10-1} = \chi^2_{0.05;9} = 16.919$$

We notice that $20.462 > 16.919$ and hence we **reject the null hypothesis**.

Table 4.7 can be displayed graphically as the following histogram:

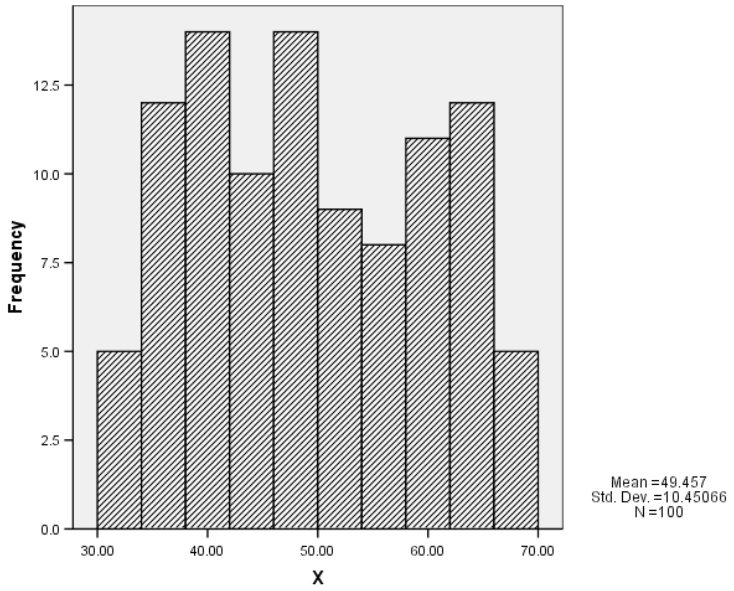


Figure 4.6: Histogram of sample data

Looking at this graph, would you say this is a sample from a normal distribution?

Superimposing a normal curve over the histogram makes our decision easier and it seems as if the sample does not come from a normal distribution. Is this what you conclude from the next figure?

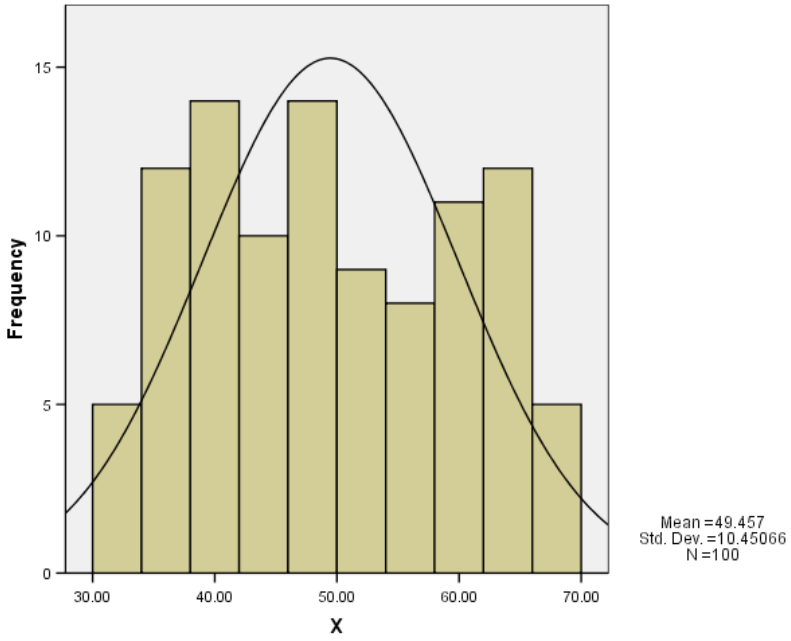


Figure 4.7: Histogram and normal curve

Our subjective conclusion based on the graphical method is confirmed by the formal hypothesis test. We conclude that the sample is most probably not from a normal distribution.

So, why is there a discrepancy between the results of the χ^2-test of example 4.3 and this example? Please see activity 4.8 of the workbook.

The χ^2 goodness-of-fit test can be used to test for any distribution type where the null hypothesis always specifies the type of distribution.

4.4 Goodness-of-fit tests in general

A. The multinomial distribution

The multinomial distribution is a generalisation of the binomial distribution in the sense that the latter is a special case of the former.

Consider an infinite population of items, each of which belongs to one of k categories. Let the proportion belonging to category i be π_i, thus

$$\pi_1 + \pi_2 + ... + \pi_k = 1.$$

If we select an element of the population at random, the probability that it will belong to category i is π_i. Suppose we draw a random sample of size n from a population. Let N_i be the number of elements of the sample which belong to category i. Thus $N_1+N_2+...+N_k=n$. The joint distribution of $N_1,\ ...,\ N_k$ is called the **multinomial distribution. (In the special case $k=2$, N_1 is a binomial variate.)**

Suppose now we have a number of random variables $X_1,\ ...,\ X_n$ and suppose we select class intervals $(a_0;\ a_1)\,;\ (a_1;\ a_2)\,;\ ...;\ (a_{k-1};\ a_k)$ which cover the whole range of variation of these variables. (We could choose $a_0=-\infty$ and $a_k=+\infty$ if necessary.) If $X_1,\ ...,\ X_n$ is a random sample from a continuous distribution with pdf $f_X(x)$, let

$$\pi_i = P(a_{i-1} < X \leq a_i) = \int_{a_{i-1}}^{a_i} f_X(x)\,dx.$$

On the other hand, if $X_1,\ ...,\ X_n$ is a random sample from a discrete distribution, $\pi_i = P(a_{i-1} < X \leq a_i)$ is found *by summation* rather than by integration.

If we now let N_i be the number of Xs which fall in the i-th class interval, then $N_1,\ ...,\ N_k$ will have a multinomial distribution with parameters $\pi_1,\ ...,\ \pi_k$. We use this fact to test whether a sample comes from a given distribution.

We distinguish between two types of problems:

(i) The distribution is completely specified by the null hypothesis, including all parameters, for example $H_0 : X$ is $n(2.5;\ 46)$.

(ii) The type of distribution is specified but not all the parameters, for example $H_0 : X$ is $n(\mu;\ 5)$ with μ not specified; or $H_0 : X$ is $n\left(\mu;\ \sigma^2\right)$ with μ and σ^2 not specified.

B. Distribution completely specified

We make use of the following theorem which we shall prove for a special case only.

Theorem 4.1

Let $N_1,\ ...,\ N_k$ be observed frequencies in a random sample of size n from a multinomial distribution with probabilities $\pi_1,\ ...,\ \pi_k$ where $N_1+...+N_k=n$ and $\pi_1+...--+\pi_k=1$. Then

$$Y^2 = \sum_{i=1}^{k} (N_i - n\pi_i)^2 / n\pi_i$$

is approximately distributed as χ^2_{k-1}.

The proof of this theorem falls beyond the scope of this module. However, it is interesting to look at the case where $k = 2,$ please see activity 4.9 of the workbook.

The reason why we have only $k-1$ degrees of freedom is the linear restriction $N_1 + ... + N_k = n$, in other words we have freedom to vary $k-1$ of the frequencies, but after $k-1$ frequencies have been chosen the k-th frequency is fixed.

An interesting fact to prove is to show that $E\left(Y^2\right) = k - 1.$ This will strengthen our belief in theorem 4.1 since we know that the expected value of a chi-squared variate is equal to its degrees of freedom.

Theorem 4.2

$$E\left(Y^2\right) = k - 1.$$

Proof

Every observation can fall in category i with probability π_i and not in category i with probability $1 - \pi_i$. Therefore N_i, the number of observations falling in category i, is a binomial variate with expectation $n\pi_i$ and variance $n\pi_i\left(1 - \pi_i\right).$
Therefore

$$\begin{aligned}
E\left(N_i - n\pi_i\right)^2 &= n\pi_i\left(1 - \pi_i\right) \\
\therefore E\left(N_i - n\pi_i\right)^2 / n\pi_i &= 1 - \pi_i \\
\therefore E\left(Y^2\right) &= E\sum_{i=1}^{k}\left(N_i - n\pi_i\right)^2 / n\pi_i \\
&= (1 - \pi_1) + (1 - \pi_2) + ... + (1 - \pi_k) \\
&= \underbrace{1 + 1 + 1 + \cdots + 1}_{k\ times} - \underbrace{(\pi_1 + \pi_2 + ... + \pi_k)}_{=1} \\
&= k - 1.
\end{aligned}$$

The quantities $n\pi_i$ are usually called *expected frequencies* (they need not be integers). Y^2 is sometimes written as

$$Y^2 = \sum_{i=1}^{k}\left(N_i - e_i\right)^2 / e_i \quad \text{where } e_i = n\pi_i.$$

[It is easier to remember this formula as: $\sum_{i=1}^{k}\frac{[\text{observed } - \text{ expected}]^2}{\text{expected}}$.]

How do we use theorem 4.1 to test goodness of fit?

We divide the data into categories (if the distribution is discrete then the data will already form categories; otherwise we group the data into intervals). We compute the probabilities that an observation will fall into each class according to the distribution specified by the null hypothesis, and compute Y^2. The value we obtain is compared with a critical value of the appropriate χ^2-distribution. We illustrate applications other than the normal distribution by means of examples.

Example 4.6

According to genetic theory the offspring of parents of genetic types AA and aa will be the following:

type AA with probability $\frac{1}{4}$;

type aa with probability $\frac{1}{4}$ and

type Aa with probability $\frac{1}{2}$.

In an experiment with pea plants a geneticist crossed plants of type AA with plants of type aa and from 132 seeds he reported the following counts:

$$AA = 35; \quad aa = 30 \quad \text{and} \quad Aa = 67.$$

Test this genetic theory at the 10% level.

Solution

We want to test $H_0 : \pi_1 = \frac{1}{4}; \quad \pi_2 = \frac{1}{4}; \quad \pi_3 = \frac{1}{2}.$

We have $N_1 = 35; \quad N_2 = 30; \quad N_3 = 67; \quad n = 132;$ so that $n\pi_1 = 33; \quad n\pi_2 = 33; \quad n\pi_3 = 66.$

We use the test statistic $Y^2 = \sum_{i=1}^{4} (N_i - n\pi_i)^2 / n\pi_i$

$$= \frac{(35-33)^2}{33} + \frac{(30-33)^2}{33} + \frac{(67-66)^2}{66}$$

$$= 0.4091.$$

From table IV we see that $\chi^2_{0.10;2} = 4.60517$. This implies that we will reject H_0 if $Y^2 \geq 4.60517$.

Since $Y^2 < 4.60571,$ we cannot reject H_0 at the 10% level.

Three points to remember

1. Note that large values of Y^2 are obtained when the differences between the theoretical and observed frequencies are large. Small values of Y^2 are obtained when the observed and theoretical frequencies are close. Therefore we reject H_0 if Y^2 is large, in other words we do a *one-sided test*.
2. Large values of Y^2 may also be obtained by having small values of $n\pi_i$ (because we divide by $n\pi_i$), and large values obtained in this way do not necessarily imply that H_0 is not true. We should therefore not have small frequencies $n\pi_i$. **If we have small expected frequencies, we pool two or more cells** by adding both their observed and expected frequencies. As a general rule **we should not have expected frequencies of less than five**, otherwise the approximation of the distribution of Y^2 by χ^2_{k-1} may not be adequate.
3. Large values of Y^2 can also arise from **very large samples**.

Example 4.7

The times to failure of 50 electronic components were recorded in minutes and are given below:

Using an Excel spreadsheet, the observations have been arranged from small to large:

10.6	11.3	15.7	18.9	19.2
21.3	22.4	23.6	27.1	28.2
30.9	34.6	36.0	39.5	40.6
45.9	47.8	49.2	50.8	62.1
67.3	71.8	74.2	83.7	85.1
89.2	90.4	96.7	107.1	122.2
127.8	135.1	136.8	139.1	142.6
147.4	150.6	153.4	157.3	162.9
169.3	171.2	178.3	185.8	190.2
193.5	199.4	203.8	211.6	219.4

Test the null hypothesis that the data are from an *exponential distribution* with pdf

$$f_X(x) = \frac{1}{100} e^{-\frac{x}{100}} \quad \text{for } x \geq 0.$$

Perform a goodness-of-fit test by making use of five classes with equal expected frequencies.

Solution

If X is the time to failure of an electronic component, we have to test whether

$H_0 : X$ has an exponential distribution with $\theta = 100.$

We have to use five class intervals of equal expected frequencies. In other words, we have to find the unknown class limits such that if we divide the observations into these classes we will know that the expected frequency for each class is $\frac{50}{5} = 10.$

In other words $\pi_i = 0.2 \quad$ for $i = 1,\ 2,\ ...,\ 5.$

Unlike example 4.3 we do not have tables for the exponential distribution and thus we have to follow the theoretical route!

For any continuous distribution, we know from calculus that

$$P\left(a \leq X \leq b\right) = \int_a^b \left(p.d.f\right) dx.$$

So, if we assume that $\theta = 100,$ for this specific exponential distribution, we may write that

$$\begin{aligned} P\left(a \leq X \leq b\right) &= \frac{1}{100}\int_a^b e^{-\frac{x}{100}} dx \\ &= e^{-\frac{a}{100}} - e^{-\frac{b}{100}} \qquad \text{(which is a result from calculus).} \end{aligned}$$

For the first interval we know that $a = 0$ and we also know that $e^0 = 1.$ If we set $P\left(0 \leq X \leq b\right) = 0.2$ we obtain $P\left(0 \leq X \leq b\right) = e^{-\frac{0}{100}} - e^{-\frac{b}{100}} = 0.2.$

In other words $\begin{aligned} 1 - e^{-\frac{b}{100}} &= 0.2 \\ \therefore e^{-\frac{b}{100}} &= 0.8 \\ \therefore -\frac{b}{100} &= \ln\left(0.8\right) \\ -\frac{b}{100} &= -0.2231 \\ \Rightarrow b &= 22.31. \end{aligned}$

For the second interval we replace a by 22.31 and hence

$$\begin{aligned} P(22.31 \leq X \leq b) &= e^{-\frac{22.31}{100}} - e^{-\frac{b}{100}} = 0.2 \\ \therefore e^{-\frac{b}{100}} &= e^{-\frac{22.31}{100}} - 0.2 \\ &= 0.8 - 0.2 \\ &= 0.6 \\ \therefore b &= (-100)(\ln 0.6) = 51.08. \end{aligned}$$

In a similar fashion we derive $\alpha_3 = 91.63$ and $a_4 = 160.94$. Thus $a_0 = 0; \quad a_1 = 22.31; \quad a_2 = 51.08; \quad a_3 = 91.63$ and $a_4 = 160.94$ in figure 4.8 showing the pdf of an exponential distribution with $\theta = 100$.

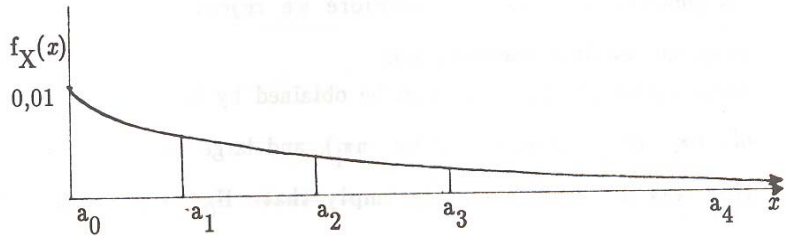


Figure 4.8: The pdf of an exponential distribution with $\theta = 100$

If we classify the 50 observations into these intervals we get the following:

Time to failure in minutes	Observed frequencies	Expected frequencies
$0 \leq x < 22.31$	6	10
$22.31 \leq x < 51.08$	13	10
$51.08 \leq x < 91.63$	8	10
$91.63 \leq x < 160.94$	12	10
$160.94 \leq x < \infty$	11	10
Total	50	50

Thus,

$$Y^2 = \frac{(6-10)^2}{10} + \frac{(13-10)^2}{10} + \frac{(8-10)^2}{10} + \frac{(12-10)^2}{10} + \frac{(11-10)^2}{10} = 3.4$$

Since $\chi^2_{0.10;4} = 7.77944$ we do not reject the null hypothesis that the sample is from an exponential distribution with $\theta = 100$ at the 10% level of significance.

Note

In the above example we chose to divide the range of the observations into classes with equal expected frequencies, since that makes the computations easier. The problem is dealt with differently, namely by choosing intervals of equal length (eg $0 \leq x < 30; \quad 30 \leq x < 60; \quad 60 \leq x < 90$ et cetera) and the corresponding expected frequencies are computed by integration. This is a valid method, but the computations are more messy because the expected frequencies are usually not integers.

C. Distribution not completely specified

We use the following theorem which we shall also not prove.

Theorem 4.3

Let $N_1, \ ..., \ N_k$ be observed frequencies with $N_1 + ... + N_k = n$ and let $\pi_1, \ ..., \ \pi_k$ be the corresponding cell probabilities, with $\pi_1 + ... + \pi_k = 1$, such that $\pi_1, \ ..., \ \pi_k$ depend on r unknown parameters $\theta_1, \ ..., \ \theta_r$.

Then $Y^2 = \sum_{i=1}^{k} \left(N_i - n\hat{\pi}_i\right)^2 / n\hat{\pi}_i$ is approximately a χ^2_{k-r-1} variate provided the $\hat{\pi}_i$ are computed by substituting the **maximum likelihood estimators** of $\theta_1, \ ..., \ \theta_r$.

Example 4.8

A sociologist is studying the distribution of TV sets per household in a certain area. According to a theory developed by him, the ratio of the number of TV sets in a household will be $\theta : 5\theta : 1-6\theta$ where the first group represents households with no TVs; the second group represents households with 1 TV and the last group represents households with 2 or more TVs. (In other words, if X represents the number of TV sets in a household chosen at random from this specific area, the probabilities should be related as follows:

$$P(X=0)=\theta; \quad P(X=1)=5\theta; \quad P(X\geq 2)=1-6\theta$$

where θ is an *unknown constant*.

In a random sample of 50 households the sociologist observed the following distribution:

Number of TV sets	Observed frequency
0	12
1	33
≥ 2	5
Total	50

Is this distribution in accordance with the theory?

Solution

We have to test H_0 : The probabilities for the three classes will be in the ratio $\theta : 5\theta : 1-6\theta$.
We first have to estimate θ according to the maximum likelihood method.
Let $N_0,\ N_1$ and N_2, respectively, denote the number of households with 0, 1 and more than 1 TV set where $N=N_0+N_1+N_2$. The likelihood function is the product of the probabilities for the observed sample. (Revise this in section 2.4 of the study guide.)

$$\begin{aligned}
L(\theta) &= \prod_{i=1}^{N} P(X_i=r_i) \\
&= \underbrace{\theta.\theta.\theta.\ ...\theta.}_{N_0 \text{ times}}\ \underbrace{5\theta.5\theta.\ ...\ 5\theta.}_{N_1 \text{ times}}\ \underbrace{(1-6\theta)\,.\ (1-6\theta)\,.\ ...\ (1-6\theta)\,.}_{N_2 \text{ times}} \\
&= \theta^{N_0}(5\theta)^{N_1}(1-6\theta)^{N_2} \\
\therefore \ln L(\theta) &= N_0\ln(\theta)+N_1\ln(5)+N_1\ln(\theta)+N_2\ln(1-6\theta) \\
\therefore \frac{\partial \ln L(\theta)}{\partial\theta} &= \frac{N_0}{\theta}+\frac{N_1}{\theta}+\frac{-6N_2}{1-6\theta}
\end{aligned}$$

Setting $\frac{\partial \ln L(\theta)}{\partial\theta}=0$ (to obtain the maximum value) we get $\frac{N_0}{\theta}+\frac{N_1}{\theta}=\frac{6N_2}{1-6\theta}$.

$$\begin{aligned}
\therefore \frac{N_0+N_1}{\theta} &= \frac{6N_2}{1-6\theta} \\
\therefore 6N_2\theta &= (N_0+N_1)(1-6\theta) \\
\therefore 6(N_0+N_1+N_2)\theta &= N_0+N_1 \\
\therefore \hat{\theta} &= \frac{N_0+N_1}{6(N_0+N_1+N_2)}
\end{aligned}$$

In the present example

$$\hat{\theta} = \frac{12+33}{6\,(50)} = \frac{45}{300} = 0.15.$$

The estimated probabilities are therefore

$$\hat{\theta} = 0.15; \quad 5\hat{\theta} = 0.75; \quad 1 - 6\hat{\theta} = 0.10.$$

Multiplying by 50 we obtain the expected frequencies:

Class	Observed frequencies	Expected frequencies
0	12	7.5
1	33	37.5
≥ 2	5	5.0

Therefore

$$\begin{aligned} Y^2 &= \frac{(12-7.5)^2}{7.5} + \frac{(33-37.5)^2}{37.5} + \frac{(5-5)^2}{5} \\ &= 2.7 + 0.54 + 0 \\ &= 3.24. \end{aligned}$$

We have $3 - 1 - 1 = 1$ degree of freedom (one parameter estimated) and $\chi^2_{0.05;1} = 3.84146$. Since $3.24 < 3.84146$ the theory cannot be rejected at the 5% level of significance.

We conclude this section where *the distribution is not completely specified* by returning to the goodness-of-fit test for a normal distribution (which was illustrated in section 4.3).

The MLEs based on the *ungrouped* data are as follows:

μ known:	$\hat{\sigma}^2 = \frac{1}{n}\Sigma\,(X_i - \mu)^2$
σ^2 known:	$\hat{\mu} = \frac{1}{n}\Sigma X_i = \overline{X}$
μ and σ^2 unknown:	$\hat{\mu} = \frac{1}{n}\Sigma X_i = \overline{X}$
	$\hat{\sigma}^2 = \frac{1}{n}\Sigma\left(X_i - \overline{X}\right)^2$

[NB You should be able to derive the above yourself!]

The MLEs based on the *grouped* data present some computational difficulties. If N_i is the number of observations lying in the interval $(a_{i-1},\ a_i)$ for $i = 1,\ ..,\ k$ and $\hat{\mu}$ and $\hat{\sigma}^2$ are the MLEs to be computed, then the likelihood function is

$$\begin{aligned} L\left(\mu,\ \sigma^2\right) &= \prod_{i=1}^{k} \left[P\left(a_{i-1} < X \leq a_i\right)\right] \\ &= \prod_{i=1}^{k} \left[\int_{a_{i-1}}^{a_i} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\alpha}\right)^2} dx\right]^{N_i}. \end{aligned}$$

Maximising this likelihood function with respect to μ and σ^2 will not be done easily without a computer. Consequently we are faced with a dilemma:

- If we use the MLEs based on the *grouped* data, then the distribution of Y^2 is asymptotically χ^2_{k-r-1} where r is the number of parameters so estimated; the problem is that the MLEs are not easily computed.
- The MLEs based on the *ungrouped* data are easily computed, but now the distribution of Y^2 is not easily computed. It has been found that the distribution of Y^2 lies between χ^2_{k-1} and χ^2_{k-r-1} in this case.

  A pragmatic solution would be as follows:

  Compute the MLEs based on the ungrouped data. Compute Y^2 as before.

  If $Y^2 < \chi^2_{\alpha;k-r-1}$: do not reject H_0

  If $Y^2 > \chi^2_{\alpha;k-1}$: reject H_0

  If $\chi^2_{\alpha;k-r-1} < Y^2 < \chi^2_{\alpha;k-1}$: decision uncertain

  In the latter case there are two possibilities:

  (a) Obtain a larger sample.

  (b) Choose another significance level according to the circumstances.

For the purpose of this module it is sufficient simply to state: "Decision uncertain".

D. The Kolmogorov-Smirnov test

We briefly mention an alternative test which can be applied to test whether a random sample comes from a specified distribution (with all the parameters specified). For any x we have

$$\begin{aligned} F\left(x\right) &= F_X\left(x\right) = P\left(X \leq x\right) \text{ which is completely specified} \\ F_n\left(x\right) &= \hat{F}_X\left(x\right) = \frac{\text{number of observations in the sample } \leq x}{\text{total number of observations}} \\ &= \text{cumulative relative frequency} \end{aligned}$$

A **one-sided test** is based on

$$D_n^+ = \text{supremum over all } x \text{ of } \{F_n(x) - F(x)\}.$$

The value thus computed is compared to a critical value read from table XIX (Stoker). If D_n^+ is larger than the critical value, reject H_0.

For a **two-sided test**, compute

$$D_n = \text{supremum over all } x \text{ of } |F_n(x) - F(x)|.$$

The critical value for a two-sided α-level test is approximately the same as the critical value for the one-sided $\frac{1}{2}\alpha$-level test. Reject H_0 if D_n is larger than this critical value.

This is the test JMP employs for a goodness-of-fit test. (See activity 4.10 of the workbook.) A computer, however, does not use critical values but only computes the p-value which has to be interpreted.

You will not be required to know this test for examination purposes.

4.5 Using the method of moments to test for normality

Another test procedure is based on *skewness* and *kurtosis*. For any distribution with mean μ and pdf $f_X(x)$ the r-th central moment is defined as

$$\mu_r = \int_{-\infty}^{\infty} (x-\mu)^r f_X(x)\, dx.$$

The third moment is zero if $f_X(x)$ is symmetric.

The *third standardised moment*

$$\beta_1 = \frac{\mu_3}{\sigma^3} = \frac{\mu_3}{(\mu_2)^{\frac{3}{2}}}$$

is a *measure* of the *skewness* of the distribution. For the normal distribution, as for any symmetric distribution, $\beta_1 = 0$.

The *fourth standardised moment*

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{\mu_4}{\sigma^4}$$

is a measure of the kurtosis of the distribution. For the normal distribution $\beta_2 = 3$.

If $X_1, \ ..., \ X_n$ is a random sample from a normal distribution we can estimate μ_r by

$$\hat{\mu}_r = \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^r .$$

A. Test for skewness

We can test the null hypothesis that the distribution is symmetrical,that is $H_0 : \beta_1 = 0$, against a two-sided or one-sided alternative. The critical values are tabulated in table A for different sample sizes but not for different levels of significance.

The null hypothesis is that the distribution is normal, namely $H_0 : \beta_1 = 0.$

If the alternative hypothesis is positive skewness (one-sided testing), namely $H_1 : \beta_1 > 0,$
we reject H_0 at the 5% level if $B_1 >$ tabulated percentage point.

If the alternative is negative skewness (one-sided testing), that is $H_1 : \beta_1 < 0,$
we reject H_0 at the 5% level if $B_1 < -$ (tabulated percentage point) .

If the alternative is skewness (two-sided testing), namely $H_1 : \beta_1 \neq 0,$
we reject H_0 at the 10% level if $|B_1| >$ tabulated percentage point.

We use the test statistic

$$B_1 = \frac{\hat{\mu}_3}{(\hat{\mu}_2)^{\frac{3}{2}}} = \frac{\frac{1}{n} \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^3}{\left(\frac{1}{n} \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2\right)^{\frac{3}{2}}}.$$

For a symmetrical distribution (as the normal distribution) we would expect this test statistic to vary in the region of 0. We do not expect you to know the distribution of B_1, but critical values have been computed for this distribution and are summarised in the table below. We are restricted to either test 5% one-sided or 10% two-sided. In other words, we cannot choose freely what the significance level is going to be.

Table A: Percentage points for the distribution of B_1
(Lower percentage point $= -$ (tabulated upper percentage point)

Size of sample	Percentage points	Size of sample	Percentage points
n	5%	n	5%
25	0.711	200	0.280
30	0.662	250	0.251
35	0.621	300	0.230
40	0.587	350	0.213
45	0.558	400	0.200
50	0.534	450	0.188
		500	0.179
60	0.492	550	0.171
70	0.459	600	0.163
80	0.432	650	0.157
90	0.409	700	0.151
100	0.389	750	0.146
		800	0.142
125	0.350	850	0.138
150	0.321	900	0.134
175	0.298	950	0.130
200	0.280	1000	0.127

Please note:
Because the sampling distribution of B_1 is symmetrical about zero, the same values, with negative sign, correspond to the lower limits.

B. Test for kurtosis

To test for kurtosis the null hypothesis is that the distribution is normal, namely. $H_0 : \beta_2 = 3.$

A distribution with $\beta_2 > 3$ is called *leptokurtic*: the pdf has a sharper peak than the normal distribution and has longer tails.

A distribution with $\beta_2 < 3$ is said to be *platykurtic*: the pdf is flat and has shorter tails than the normal distribution.

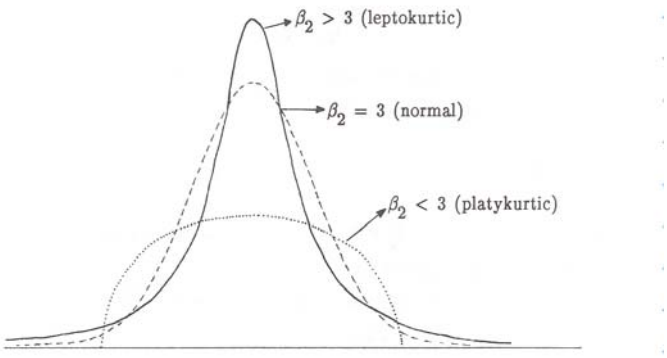


Figure 4.9: Degrees of kurtosis

To test for kurtosis, we could use two different test statistics.

We may use the test statistic

$$B_2 = \frac{\hat{\mu}_4}{(\hat{\mu}_2)^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^4}{\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2\right)^2}.$$

For the normal distribution we would expect this test statistic to vary in the region of 3. We do not expect you to know the distribution of B_2, but only to realise that critical values (associated with a significance level of 5%) have been computed and are tabulated in table B. Again (as with B_1) we are restricted to test 5% one-sided or 10% two-sided, and we cannot freely choose the significance level.

Table B:. Percentage points of the distribution of B_2

Size of sample n	Percentage points	
	Upper 5%	Lower 5%
50	3.99	2.15
75	3.87	2.27
100	3.77	2.35
125	3.71	2.40
150	3.65	2.45
200	3.57	2.51
250	3.52	2.55
300	3.47	2.59
350	3.44	2.62
400	3.41	2.64
450	3.39	2.66
500	3.37	2.67
550	3.35	2.69
600	3.34	2.70
650	3.33	2.71
700	3.31	2.72
800	3.29	2.74
900	3.28	2.75
1000	3.26	2.76

Test based on $\mathbf{B_2}$

If the alternative is $\beta_2 < 3$, reject H_0 at the 5% level if $B_2 <$ lower 5% point in table B.

If the alternative is $\beta_2 > 3$, reject H_0 at the 5% level if $B_2 >$ upper 5% point in table B.

If the alternative is $\beta_2 \neq 3$, reject H_0 at the 10% level if $B_2 <$ lower 5% point or if $B_2 >$ upper 5% point in table B.

Example 4.9

From a random sample of size $n = 100$ the following were computed:

$$\Sigma X_i = 200; \quad \Sigma X_i^2 = 416; \quad \Sigma \left(X_i - \overline{X}\right)^3 = 12.8; \quad \Sigma \left(X_i - \overline{X}\right)^4 = 10.24$$

We wish to test the sample for normality. We shall test

(a) for skewness (two-sided) at the 10% level;

(b) for kurtosis (two-sided) at the 10% level.

A sample from a normal distribution should pass both tests with a high probability.

Solution

(a) **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 $|B_1| > 0.389$ (in other words if $B_1 < -0.389$ or if $B_1 > 0.389$ (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.

$$\begin{aligned} \Sigma \left(X_i - \overline{X}\right)^2 = \Sigma X_i^2 - n\overline{X}^2 &= 416 - 100 \left(\frac{200}{100}\right)^2 \\ &= 416 - 400 \\ &= 16 \end{aligned}$$

$$\therefore B_1 = \frac{\frac{12.8}{100}}{\left(\sqrt{\frac{1}{100}(16)}\right)^3} = \frac{0.128}{(0.4)^3} = 2$$

Since $2 > 0.389$ we reject H_0 at the 10% level.

(b) **Test for kurtosis**

We have to test $H_0 : \beta_2 = 3$ against
$H_1 : \beta_2 \neq 3.$

We will reject H_0 at the 10% level of significance (two-sided) if $B_2 > 3.77$ or if $B_2 < 2.35$ (from table B).

The value of the test statistic is $B_2 = \dfrac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^4}{\left[\frac{1}{n}\Sigma\left(X_i - \overline{X}\right)^2\right]^2} = \dfrac{\frac{10.24}{100}}{[0.16]^2} = \dfrac{0.1024}{0.0256} = 4.$

Since $4 > 3.77$ we reject H_0 at the 10% level.

The sample **failed both tests** and hence we conclude that the sample is not from a normal population.

Another statistic which is a measure of kurtosis is the **standardised mean deviation,**

$$A = \frac{\frac{1}{n}\Sigma\left|X_i - \overline{X}\right|}{\sqrt{\frac{1}{n}\Sigma\left(X_i - \overline{X}\right)^2}} = \frac{\text{mean deviation}}{\text{standard deviation}}.$$

($\left|X_i - \overline{X}\right|$ is read as "the absolute value of $X_i - \overline{X}$" and it means you take the *positive value* of the difference.)

The test statistic you choose depends on the sample size: for small samples $(n < 50)$ we usually use A; for larger samples $(N \geq 50)$ use B_2.

Table C: Percentage points for the distribution of $A = \dfrac{\text{mean deviation}}{\text{standard deviation}}$

Size of sample n	$n-1$	Percentage points			
		Upper 5%	Upper 10%	Lower 10%	Lower 5%
11	10	0.9073	0.8899	0.7409	0.7153
16	15	0.8884	0.8733	0.7452	0.7236
21	20	0.8768	0.8631	0.7495	0.7304
26	25	0.8686	0.8570	0.7530	0.7360
31	30	0.8625	0.8511	0.7559	0.7404
36	35	0.8578	0.8468	0.7583	0.7440
41	40	0.8540	0.8436	0.7604	0.7470
46	45	0.8508	0.8409	0.7621	0.7496
51	50	0.8481	0.8385	0.7636	0.7518
61	60	0.8434	0.8349	0.7662	0.7554
71	70	0.8403	0.8321	0.7683	0.7583
81	80	0.8376	0.8298	0.7700	0.7607
91	90	0.8353	0.8279	0.7714	0.7626
101	100	0.8344	0.8264	0.7726	0.7644

Test based on A

If the alternative is that the distribution is leptokurtic, namely $H_1 : \beta_2 > 3,$
we reject H_0 at the 5% level of significance if $A >$ upper 5% point in table C (or at the 10% level if $A >$ upper 10% point in table C).

If the alternative is that the distribution is platykurtic, namely $H_1 : \beta_2 < 3,$
we reject H_0 at the 5% level of significance if $A <$ lower 5% point in table C (or at the 10% level if $A <$ lower 10% point in table C).

If the alternative is two-sided, namely $H_1 : \beta_2 \neq 3,$
we reject H_0 at the 10% significance level if $A <$ lower 5% point or if $A >$ upper 5% point in table C.

Example 4.10

We wish to test the kurtosis of the following sample:

$$18 \quad 26 \quad 21 \quad 25 \quad 20 \quad 16 \quad 12 \quad 24 \quad 17 \quad 19 \quad 22$$

Test two-sided at the 10% level of significance.

Solution

We have to test $H_0 : \beta_2 = 3$ against
$H_1 : \beta_2 \neq 3.$

Since $n = 11 < 50,$ we will use the test statistic A.

$$A = \frac{\frac{1}{n}\sum_{i=1}^{n}\left|X_i - \overline{X}\right|}{\sqrt{\frac{1}{n}\Sigma\left(X_i - \overline{X}\right)^2}}$$

where $\overline{X} = \frac{220}{11} = 20; \quad \Sigma\left|X_i - \overline{X}\right| = 36$ and $\Sigma\left(X_i - \overline{X}\right)^2 = 176.$

Thus $A = \frac{\frac{36}{11}}{\sqrt{\frac{176}{11}}} = 0.8182.$

We will reject H_0 two-sided if $A < 0.7153$ or if $A > 0.9073$ (from table C).

Since $0.7153 < 0.8182 < 0.9073$ we conclude that the kurtosis of the sample is not significantly different from the kurtosis of the normal distribution, at the 10% level (two-sided).

What about statistical packages and moments?

The statistical package SPSS computes the third and fourth moments as standard output under "descriptive statistics" for any data set. It is, however, not part of the standard output of JMP. We need to manipulate our output if we want to compute B_1 and B_2. (Please see activity 4.15 of the workbook.)

Exercise 4.1

1. The blood of a random sample of 1 000 people from a certain population was classified into 4 blood groups, and the results are as follows:

$$N_1 = 125; \quad N_2 = 185; \quad N_3 = 230; \quad N_4 = 460$$

It is postulated that the population is divided into the four blood groups in the following proportions:

$$\pi_1 = 0.10; \quad \pi_2 = 0.20; \quad \pi_3 = 0.20; \quad \pi_4 = 0.50$$

Test this hypothesis at the 1% level.

2. According to a seed man's claim, of the plants that germinate from a packet of "Colorglo" Namaqualand daisy seeds, there will be twice as many plants bearing yellow flowers as white flowers, and twice as many bearing orange flowers as yellow flowers. It is admitted implicitly that a certain proportion will not germinate at all. The theory can be written as a model as follows:

$$P(\text{White}) = \theta; \quad P(\text{Yellow}) = 2\theta; \quad P(\text{Orange}) = 4\theta; \quad P(\text{Fail to germinate}) = 1 - 7\theta$$

I sow 100 of these seeds (presumably a random sample) and 84 germinate. Of these 84 plants, 16 bear white flowers, 28 bear yellow flowers and 40 bear orange flowers. Can the seed man's claim be rejected at the 5% level of significance?

3. A sample of size 40 from a distribution with known variance $\sigma^2 = 100$ has mean $\overline{X} = 10$. The following classification was obtained:

X	Frequency
$X < 3.255$	7
$3.255 \leq X < 10$	6
$10 \leq X < 16.745$	15
$X \geq 16.745$	12

Compute the goodness-of-fit statistic to test whether the distribution is normal. Determine whether the sample is significantly different from normal

(a) at the 10% level

(b) at the 5% level.

4. The number of bees arriving at a peach tree was recorded during 100 non-overlapping one-minute intervals. The observed frequencies were as follows:

Number of bees	0	1	2	3	4	5	6
Frequencies	21	30	27	16	3	2	1

Test the null hypothesis, at the 5% level of significance, that this is a random sample from a Poisson distribution

(a) with mean $\lambda = 2$

(b) with λ not specified.

(For ease of computation, round off the expected frequencies to the nearest integer.)

Hint: $e^{-1.6} = 0.2019$.

5. On the assumption that the lifetime of a product is normally distributed with mean 32 months and standard deviation eight months, a guarantee was determined. The following data were subsequently collected:

Lifetime (months)	Frequency
Less than 16	6
16 to 20	9
20 to 24	12
24 to 28	16
28 to 32	20
32 to 36	22
36 to 40	10
more than 40	5

Test the assumption of normality with mean 32 and variance 64 at the 5% level of significance.

6. The following data have been observed in an experiment:

29	12	28	46	15	13	25	44	20	14
37	41	11	38	28	12	40	47	19	29
13	39	6	13	29	15	34	17	33	51

Test the null hypothesis that the sample comes from a $n(25; 12^2)$ distribution. Use **five classes** of equal probability to derive the intervals.

7. Test the following sample for kurtosis:

$$17; \quad 22; \quad 15; \quad 25; \quad 22; \quad 26; \quad 16; \quad 14; \quad 18; \quad 21; \quad 24$$

(10% level).

8. From a sample of 50 observations the following statistics were computed:

$$\overline{X} = 25; \quad \Sigma\left(X_i - \overline{X}\right)^2 = 200; \quad \Sigma\left(X_i - \overline{X}\right)^3 = -320; \quad \Sigma\left(X_i - \overline{X}\right)^4 = 4\,000$$

Would you regard this as a sample from a normal distribution? Use the 10% level (two-sided).

9. From a sample of $1\,000$ observations it was found that

$$\overline{X} = 50; \quad \frac{1}{n}\Sigma\left(X_i - \overline{X}\right)^2 = 16; \quad \frac{1}{n}\Sigma\left(X_i - \overline{X}\right)^3 = 6.4; \quad \frac{1}{n}\Sigma\left(X_i - \overline{X}\right)^4 = 819.2.$$

Test at the 10% level (two-sided) whether the sample comes from a normal distribution.

4.6 Learning outcomes

After studying unit 4 you should **understand and be able to apply and interpret** the following **tests for normality**:

- using *normal probability paper* to plot the *order statistics* Y_i against $100i/(n+1)$
- a *normal quantile graph* (using JMP)
- a *goodness-of-fit* test (ie χ^2-test) where the *expected frequencies* are obtained by
  - assuming equal probabilities and computing the corresponding set of intervals
  - computing the areas under the pdf of a normal distribution for a given set of intervals
- test for *skewness* and test for *kurtosis* which together form one test for normality

You should be able to perform a **goodness-of-fit** test (χ^2-test) for any other type of distribution (which will be specified).

STUDY UNIT 5

Statistical independence

5.1 The meaning of independence

The assumption of independence is an integral part of most statistical models. Thus, for example, independence forms part of the definition of a random sample. Sometimes it is possible to test whether certain observations are independent, but in most cases the independence, or lack thereof, must be deduced from the way in which the experiment was conducted. The formal definition of independence is: the random variables X_1 and X_2 are *independent* if their joint pdf is given by

$$f_{X_1;X_2}(x_1;\ x_2) = f_{X_1}(x_1)\, f_{X_2}(x_2) \text{ for all } x_1 \text{ and } x_2.$$

The definition of the independence of n random variables is given in unit 1. An equivalent definition, in terms of conditional distributions, is that X_1 and X_2 are independent if the conditional pdf of X_2, given that $X_1 = x_1$, for all values x_1, is not a function of x_1.

The question is: how do we know that this condition holds good for our experiment? The answer to this question is not easy, but the acid test is to ask the following question: does the outcome of one observation have any influence on the outcome of any other observation? We shall discuss a few examples of non-independence which may help you in answering this question.

To begin with, we have to point out that there is a difference in the definitions of a random sample for finite and infinite populations. The results of sampling with replacement from a finite population may be regarded as independent observations, but such samples are usually not desirable since one does not want to observe one number of the population more than once. On the other hand, if one draws a sample without replacement, the composition of the population changes after each draw and the consecutive observations are not independent. A random sample from a finite population requires only that each and every distinct sample of size n of the $\binom{N}{n}$ different samples must have the same probability $1/\binom{N}{n}$ of being selected. Mutual independence of the n observations in the sample is not part of the definition.

If the population is finite but very large, and the sample to be drawn from it is comparatively small, the population is regarded as an infinite population for practical purposes. The change in the composition of the population after each draw is then so small as to be negligible. In principle it is easy to draw a random sample from a finite population (small or large) provided each member of the population can

be identified uniquely by means of a number. In such a case one may draw a sample of the numbers, using numbers in a hat, tables, random numbers or random numbers generated by a computer.

We shall now concentrate on samples from an infinite population.

5.2 Examples of dependence

The problem with dependence is that one cannot really do analyses or applications without being able to quantify this dependency in a model, that is to set up a model for the dependence.

A. Repeated measurements on the same individual

The following type of experiment is often performed: an individual is subjected to a treatment and the result is observed at a number of specified times. Examples of this are a patient who consumes an amount of sugar and has his or her blood sugar tested every 30 minutes in order to determine his or her sugar curve; a learner who is taught arithmetic and whose arithmetic ability is tested every term; a pig that is placed on a certain diet and whose mass is determined every week.

The result of such an experiment is a number of observations $X_1, \ ..., \ X_n$. It is not safe to assume automatically that the observations are independent. If we select a patient at random from a population, measure his or her blood pressure X_1, administer a treatment and measure his or her blood pressure X_2, then X_2 will depend on X_1 because the response of the patient to the treatment will depend on his or her initial blood pressure. Given X_1, we cannot regard X_2 as the blood pressure after treatment of a patient selected at random from the population.

In repeated measurements there is also the possibility of a *carry-over effect*. If we administer one treatment to an individual and measure the result, then administer another treatment to the same individual and measure the effect again, there is a possibility that the effect of the first treatment has not "worn off" and had an effect on the second measurement. Think of an experiment to test the effect of two methods of teaching arithmetic. If we teach a learner by the one method and measure his or her ability, the knowledge acquired by the learner in the first phase of the experiment will not be forgotten, and the second measurement will not be independent of the first.

To summarise, the results $X_1, ..., \ X_n$ of n measurements on one individual may have to be analysed in a completely different way from the results $X_1, \ ..., \ X_n$ of one measurement on each of n individuals.

B. Paired observations

Very similar to the discussion in A above is the difference between the following two experiments to determine the effect of a treatment on a group of individuals.

One experiment is done by measuring every individual with respect to the variable being studied, administering the treatment and measuring every individual again.

The other experiment is done by dividing the individuals in a random manner into two groups. The one group, called the *control group*, is measured without treatment, and the other group is treated and then measured.

The results of these two experiments will be analysed differently. In the first case we have paired observations with dependence in each pair, and in the second case we have two independent samples.

This dependence between measurements on the same individual will of course hold good for measurements of different variables on the individual, like height and mass, as well.

C. Ordering of observations

Let $X_1,\ ...,\ X_n$ be a random sample; we know that $X_1,\ ...,\ X_n$ are mutually independent. Suppose we arrange the observations from the smallest to the largest, and call the result $Y_1,\ ...,\ Y_n$. Then $Y_1,\ ..., Y_n$ are called the *order statistics* of the sample. There is an ordering in these statistics:

$$Y_1 \leq Y_2 \leq ... \leq Y_n$$

Although X_1 and X_2 are independent, it is no longer true that Y_1 and Y_2 are independent. For one thing, Y_2 is bounded from below by Y_1, and, given Y_1, Y_2 cannot assume all possible values. The distribution of the order statistics is not the same as the distribution of $X_1,\ ...,\ X_n$.

D. Recognisable subsets

In many populations there are recognisable subsets of individuals who are more similar than the population as a whole. Children from the same family, piglets from the same litter and people living in the same suburb are examples of such subsets. If we select a number of individuals from the same subset they may be regarded as a random sample from that subset. However, regarded as a sample from the whole population, there is a definite dependence and the sample is not a random sample from that population.

E. Time dependence

In economic data especially, there is often a time dependence which may result in a special kind of mutual dependence between the observations. Consider an inflation rate which is computed monthly. One feels intuitively that the inflation rate in April will not be completely independent of the inflation rate in March of the same year, but will be less dependent on the inflation rate in April of the previous year. A curve which joins the points in the following graph will be fairly smooth:

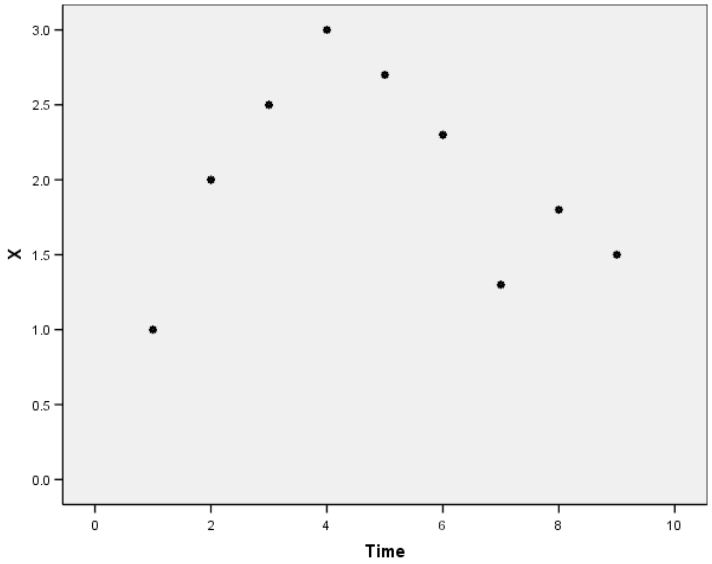


Figure 5.1

In a random sample one would expect all rearrangements of the data to be equally likely to occur. The following rearrangement *of the same points* will be less likely in this application, however.

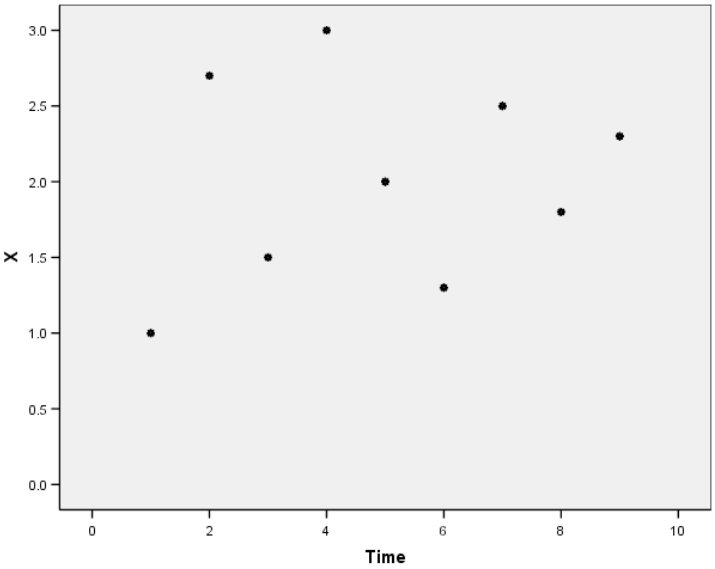


Figure 5.2

The data in the first graph are subject to *autocorrelation*: each observation depends in a specific way on the previous one. Such data must be analysed in a special way. (This is covered in STA2604 and STA3704.)

In the rest of this study unit we are going to look at a few types of analysis that test for dependence. When we talk about "tests of independence/dependence, we are usually interested in the possibility that one variable could affect or influence a second variable. This means we are moving into the field of studying the variables simultaneously (as opposed to studying them *one at a time*). This immediately alerts us to the type of variable involved in the analysis. We could have the situation where both variables are nominal, or one could be nominal and one continuous or both could be continuous!

In the next section we explain the technique of how to test for dependence when we have two *nominal* (or also called categorical) variables.

5.3 Contingency table analysis

Contingency tables generally consist of frequencies arranged into a *two-way table* according to two categorical variables (eg A and B). Sometimes the variables are truly categorical (eg gender, profession, city) and sometimes the variables are continuous, but are divided or forced into categories (for example age group).

In general we have frequencies $N_{ij}; \quad i = 1, ..., h; \quad j = 1, ..., k$ which are random variables. We use the following notation:

$$N_{i\cdot} = \sum_{j=1}^{k} N_{ij}; \qquad N_{\cdot j} \sum_{i=1}^{h} N_{ij}; \qquad N_{\cdot\cdot} = \sum_{i=1}^{h} \sum_{j=1}^{k} N_{ij}$$

where $N_{1\cdot}, ..., N_{h\cdot}$ are the *row totals*; $N_{\cdot 1}, ..., N_{\cdot k}$ are the *column totals* and $N_{\cdot\cdot}$ is the *grand total*.

$N_{i\cdot}$ and $N_{\cdot j}$ are called the *marginal totals*. The general $h \times k$ ("h by k") contingency table with h rows and k columns is as follows:

		Categories of variable A				
		1	2	$\cdots\cdots$	k	Total
Categories of variable B	1	N_{11}	N_{12}	$\cdots\cdots$	N_{1k}	$N_{1\cdot}$
	2	N_{21}	N_{22}	$\cdots\cdots$	N_{2k}	$N_{2\cdot}$
	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
	h	N_{h1}	N_{h2}	$\cdots\cdots$	N_{hk}	$N_{h\cdot}$
Total		$N_{\cdot 1}$	$N_{\cdot 2}$	$\cdots\cdots$	$N_{\cdot k}$	$N_{\cdot\cdot}$

The example below is a typical example of a contingency table.

Example 5.1

A monkey was fitted with a radio transmitter and its position was determined 100 times at various times of the day over a period of a few months. The observation times were classified into one of the following categories: Early morning (EM); Late morning (LM); Early afternoon (EA) and Late afternoon (LA). The monkey's distance from the river was computed every time, and these distances were classified as Close to, Near and Far from the river. Counting the number of times (frequencies) the observations fell into each of these categories, the results are as follows:

		Time				
		EM	LM	EA	LA	Total
Distance from river	Close	12	11	4	13	40
	Near	6	0	20	4	30
	Far	2	19	6	3	30
Total		20	30	30	20	100

(Eg of the 30 late morning observations, the monkey was close to the river on 11 occasions, near the river on 0 occasions and far from the river on 19 occasions.)

The question is: does the distance from the river depend on the time of day or are the two variables independent?

Contingency tables may be obtained in various ways, and we will discuss two. The method of analysis will be identical, but theoretically the hypotheses are not the same.

A. Fixed grand total

We assume a random sample of $N_{..}$ individuals was chosen, and two variables were recorded for each individual (eg home language and type of work). The problem is to test whether the two variables are independent. In this case $N_{i.}$ and $N_{.j}$ are random variables. Let

$\pi_{ij} = P\,(\text{individual falls into row } i \text{ and column } j)$

$\pi_{i.} = \sum_{j=1}^{k} \pi_{ij} = P\,(\text{individual falls into row } i)$

$\pi_{.j} = \sum_{i=1}^{h} \pi_{ij} = P\,(\text{individual falls into column } j)$

$\pi_{..} = \sum_{i=1}^{h} \sum_{j=1}^{k} \pi_{ij} = 1$

The null hypothesis of independence is

$$H_0 : \pi_{ij} = \pi_{i.}\pi_{.j}; \quad i = 1, ..., h; \quad j = 1, ..., k.$$

B. Fixed row (or column) totals

Assume we have k populations, and each individual from each population can be classified into one of h categories. We choose a random sample of size $N_{.j}$ from population j where $N_{.j}$ is not a random variable but a chosen sample size. In this case

$$\pi_{ij} = P\,(\text{individual from population } j \text{ falls into category } i)$$

and

$$\pi_{.j} = \sum_{i=1}^{h} \pi_{ij} = 1.$$

The null hypothesis of independence is that the probability of falling into category i is the same for all k populations:

$$H_0 : \pi_{i1} = \pi_{i2} = ... = \pi_{ik} \quad \text{for} \quad i = 1, ..., h.$$

(For example, in the case $h = k = 2$ we want to test whether two probabilities are equal.) Example 5.1 is an example of this kind since the experimenter presumably selected his or her observation times, and the column totals are therefore not random.

Analysis

Let $e_{ij} = \dfrac{N_{i.}N_{.j}}{N_{..}}$ (in other words the expected frequency for a cell equals the row total times the column total divided by the grand total).

The test statistic we use for testing the null hypothesis is

$$Y^2 = \sum_{i=1}^{h} \sum_{j=1}^{k} \frac{(N_{ij} - e_{ij})^2}{e_{ij}}.$$

(Does this look familiar?)

The distribution of Y^2 (under H_0) is given by the following theorem which we shall not prove here.

Theorem 5.1

Under the null hypothesis the distribution of Y^2 is approximately that of χ^2 with $(h-1)(k-1)$ degrees of freedom.

Please note the following:
We could have called the test statistic X, or W or whatever, but we stick to the notation Y^2 to have the connection with the "square" in the χ^2-variable. Please do not try to create your own test statistic by taking the square root of whatever you compute. Stick to Y^2 and report it as Y^2.

Why do you think we have $(h-1)(k-1)$ for the degrees of freedom? Note that if we fill in the marginal totals and choose $h-1$ rows and $k-1$ columns then the remaining row and column are fixed. We have the freedom to vary $(h-1)(k-1)$ of the cell frequencies.

Illustration: Example 5.1 (continued)

The null hypothesis of independence is that the probability of being close to the river $(i=1)$ is the same for any time of the day. It also means that we could say that the distance from the river is independent of the time of day. The alternative hypothesis is that the two factors are not independent.

The expected frequencies are as follows:

	EM	LM	EA	LA	Total
Close	$\frac{20\times 40}{100}=8$	12	12	8	40
Near	$\frac{20\times 30}{100}=6$	9	9	6	30
Far	$\frac{20\times 30}{100}=6$	9	9	6	30
Total	20	30	30	20	100

Therefore $Y^2 = \frac{(12-8)^2}{8}+\frac{(11-12)^2}{12}+...+\frac{(3-6)^2}{6}$

$= \frac{3\,595}{72}$

$= 49.9306.$

We have $(4-1)(3-1)=6$ degrees of freedom and $\chi^2_{0.01;6}=16.8119$. Since $49.9306>16.8119$ the null hypothesis is rejected at the 1% level of significance, and we conclude that there is a relationship between the time of day and the distance from the river. Comparing the observed and expected frequencies, we notice that the monkey was more often than expected close to the river in the early morning and late afternoon, more often than expected far from the river in the late morning and more often than expected near the river in the early afternoon.

A few important general notes:

1. Note that the statistic Y^2 is not changed if we exchange the roles of rows and columns (ie if N_{ij} is renamed N_{ji}) or if two rows (or two columns) are switched.
2. A significant Y^2 only indicates *association* between the two variables and **not a causal relationship**.
3. The expected frequencies $e_{ij}=N_{i.}N_{.j}/N_{..}$ need not be integers. The examples in this study guide have been chosen in such a way that the e_{ij} are integers to make the computations easier. Normally one would work with the e_{ij} correct to about two decimal digits.
4. In order for the χ^2 approximation to the distribution of Y^2 to be adequate, we should not have too many small e_{ij} otherwise we should pool rows and/or columns. An empirical rule (Cochran's rule) states that no e_{ij} should be smaller than 1 and not more than 20% of the e_{ij} should be smaller than 5. (A more stringent rule given in many textbooks is that no e_{ij} should be less than 5.)
5. When choosing categories or when deciding which rows or columns should be pooled, one must be careful not to choose categories deliberately in the way most favourable for rejection of H_0 (or acceptance, if that is what we want). The choice should be made objectively on external grounds or be based on expected frequencies – not observed frequencies.
6. Although the chi-square test looks like a one-sided test (because the critical value is on the right and we reject H_0 if $Y^2\geq\chi^2_{\alpha;\nu}$) it is in fact a test for two-sided alternatives! (A large numerical value for Y^2 can be obtained if the observed cell frequency is very small or very large.)
7. Alternative methods of analysing contingency tables are available, but based on advanced statistical and mathematical principles and therefore beyond the scope of this module. We mention two such techniques briefly which are covered in STA4806 (an honours course).

   (i) *The log-linear model*

   The null hypothesis of independence $\pi_{ij}=\pi_{i.}\pi_{.j};\quad i=1,...,k;\quad j=1,...,h$ may be written $\log(\pi_{ij})=\log(\pi_{i.})+\log(\pi_{.j})$ for all i and j.

If this null hypothesis is rejected, then it may be that alternative models are suitable for expressing the log-probability, $\log \pi_{ij}$, in terms of the logs of the marginal probabilities. The purpose of such an analysis is to find a model that will adequately explain the data. This model can be used for multidimensional contingency tables. The analysis cannot be done without a computer.

(ii) *Correspondence analysis*

Correspondence analysis is a technique that enables one to display a contingency table on a special graph. Rows in the table that are very similar are close to one another on the graph and likewise for columns that are very similar. If a given column is very close to a given row on the graph, then the frequency in that row and column is very large compared to the other frequencies in that row (and in that column).

Theorem 5.1 states that the distribution of Y^2 is approximately χ^2, and then only under certain conditions.
There is a special case where an exact test exists. This means we may compute probabilities 100% accurately. The exact test exists in the case of 2×2 tables.

C. Exact test for a 2×2 table

In the case of 2×2 contingency tables an exact test exists – the only problem is that **extensive tables** are needed to apply it. In fact, one such table fills a whole book:

Lieberman, GJ and Owen, DB: *Tables of the hypergeometric probability distribution*, Stanford, California, Stanford University Press, 1961.

It falls beyond the scope of this module to explain how the *exact probabilities* for a 2×2 table can be computed. You only need to know how to apply the table to perform a hypothesis test. For a 2×2 table we have four cells.

Consider the following notation:

Let x be any one of the four cell frequencies
k the column total corresponding to that cell
n the row total corresponding to that cell
N the total number of observations.

Thus we have

		Attribute A		
		A_1	A_2	Total
Attribute B	B_1	x		n
	B_2			
	Total	k		N

The remainder of the table can now be completed in by subtraction:

	A_1	A_2	Total
B_1	x	$n-x$	n
B_2	$k-x$	$N-k-n+x$	$N-n$
Total	k	$N-k$	N

For fixed $\mathbf{N}$, $\mathbf{n}$ and $\mathbf{k}$, x can be regarded as a value assumed by a random variable which has a *hypergeometric* distribution, denoted by $H\left(N;n;k\right)$, and this is the distribution tabulated in table D for the special case $N=12$.

Luckily JMP can compute these probabilities for any value of N and you need to be able to interpret the output for JMP which is explained in the workbook. In order to see how the hypothesis test works and how you have to use the table, we present only a very small part of the thick book of tables on the hypergeometric distribution. Table D below is the special case where the total sample size is $N=12$ and the possible combinations of n and k go from 1 to 6.

Table D:
The hypergeometric probability distribution: $P(X \leq x)$ for $N = 12$

n	k	x	P	n	k	x	P	n	k	x	P
1	1	0	0.917	4	4	0	0.141	6	2	0	0.227
1	1	1	1.000	4	4	1	0.594	6	2	1	0.773
				4	4	2	0.933	6	2	2	1.000
2	1	0	0.833	4	4	3	0.998				
2	1	1	1.000	4	4	4	1.000	6	3	0	0.091
								6	3	1	0.500
2	2	0	0.682	5	1	0	0.583	6	3	2	0.909
2	2	1	0.985	5	1	1	1.000	6	3	3	1.000
2	2	2	1.000								
				5	2	0	0.318	6	4	0	0.030
3	1	0	0.750	5	2	1	0.848	6	4	1	0.273
3	1	1	1.000	5	2	2	1.000	6	4	2	0.727
								6	4	3	0.970
3	2	0	0.545	5	3	0	0.159	6	4	4	1.000
3	2	1	0.955	5	3	1	0.636				
3	2	2	1.000	5	3	2	0.955	6	5	0	0.008
				5	3	3	1.000	6	5	1	0.121
3	3	0	0.382					6	5	2	0.500
3	3	1	0.873	5	4	0	0.071	6	5	3	0.879
3	3	2	0.995	5	4	1	0.424	6	5	4	0.992
3	3	3	1.000	5	4	2	0.848	6	5	5	1.000
				5	4	3	0.990				
4	1	0	0.667	5	4	4	1.000	6	6	0	0.001
4	1	1	1.000					6	6	1	0.040
				5	5	0	0.027	6	6	2	0.284
4	2	0	0.424	5	5	1	0.247	6	6	3	0.716
4	2	1	0.909	5	5	2	0.689	6	6	4	0.960
4	2	2	1.000	5	5	3	0.955	6	6	5	0.999
				5	5	4	0.999	6	6	6	1.000
4	3	0	0.255	5	5	5	1.000				
4	3	1	0.764								
4	3	2	0.982	6	1	0	0.500				
4	3	3	1.000	6	1	1	1.000				

In real life there could be any possible value for the total sample space!

The null hypothesis is the same as for the $h \times k$ contingency table.
H_0 : There is no association between attribute A and attribute B.

However, unlike the chi-square test, which is a test for a two-sided alternative, the exact 2×2 test can be applied for one or two-sided alternatives.

To use table D, first find the smallest marginal total (row or column) or if there is more than one marginal total equal to the smallest value, choose any one of these, and call it k. If k is a row total, choose n the smallest column total (or any of the two column totals if they are equal). Then x is the cell frequency corresponding to the row and column with marginal totals n and k. If k is a column total then n is the smallest row total.

For example, in the case

$$\begin{array}{cc|cc} & x & & \\ & \uparrow & & \\ 2 & 1 & 3 & \rightarrow \quad k \\ 4 & 5 & 9 & \\ \hline 6 & 6 & 12 & \\ & \uparrow & & \\ & n & & \end{array}$$

choose $k = 3$ (the smallest marginal total, in this case a row total) and $n = 6$ (the smallest column total; suppose for argument's sake we choose the second column) **then** $\mathbf{x = 1}$. We now use the symbol X to denote the random variable which has outcome x (ie $X = 1$) in the table. If $k = 3$ then X can assume the values 0; 1; 2 or 3, that is if $\mathcal{A}$ is the set of discrete points of X, then $\mathcal{A} = \{0; 1; 2; 3\}$ and

$$\underbrace{P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) = 1.}$$

These probabilities are not given individually, but cumulatively in table D.

The second block from the top, in the last column of the table, gives
$P(X = 0,\ k = 3,\ n = 6,\ N = 12) = 0.091.$

In that same block, $P(X \leq 1) = 0.500$
$P(X \leq 2) = 0.909$
$P(X \leq 3) = 1.000.$

Next we have to **figure out** the alternative hypothesis for x. (The null hypothesis is of course that there is no association between the two attributes A and B.)

The wording "figure out" is exactly what it says! From the given problem and the cell you chose for x,you have to figure out whether the one-sided alternative would mean small values for x to favour the alternative or large values for x to favour the alternative.

If the alternative implies small values of x: reject H_0 at the α level if $P(X \leq x) \leq \alpha$.

If the alternative implies large values of x: reject H_0 at the α level if $P(X \geq x) \leq \alpha$, that is $1 - P(X \leq x - 1) \leq \alpha$.

If the alternative is two-sided, in other words we want to reject H_0 if x is too large or too small, compute $P(X \leq x)$ and $P(X \geq x) = 1 - P(X \leq x - 1)$. If the smaller of these two probabilities is $\leq \frac{1}{2}\alpha$, reject the null hypothesis at the α level.

Example 5.2

We want to test whether there is an association between smoking and preference for coffee (whether smokers tend to prefer coffee or equivalently whether coffee drinkers tend to smoke). A random sample of 12 people yielded the following table:

	Coffee	Tea	Total
Smokers	4	1	5
Nonsmokers	4	3	7
Total	8	4	12

Solution

H_0: There is no association between smoking and preference for coffee.

H_1: Smokers tend to prefer coffee to tea.

For this 2×2 table we have to choose $k = 4$ and $n = 5$. (Our table D will not allow us to work with an $n > 6$ or a $k > 6$.)

As soon as you choose k and n it "fixes the class" for x. In this example $x = 1$ and it means there was one person in the class of people who smoke and do not drink coffee.

Now comes the "figuring out" of the alternative hypothesis!

		Tea			
		x			
		$\uparrow$			
Smokers	4	1	5	$\longleftarrow$	n
	4	3	7		
	8	4	12	$\longrightarrow$	N
		$\uparrow$			
		k			

The alternative (smokers prefer coffee) would imply a small value of x to reject H_0, that is so small that $\mathbf{P}(\mathbf{X} \leq \mathbf{x}) \leq \boldsymbol{\alpha}$.

Now $x = 1$ and $P(X \leq 1) = 0.424$ (from table D)
$> 0.05 = \alpha$.

The small p-value or exceedance probability is therefore larger than α so that the test statistic is not significant.

The null hypothesis therefore cannot be rejected at the 5% level (or any of the usual significance levels).

Example 5.3

Suppose we want to test for association (two-sided alternative) in the following table:

	A_1	A_2	Total
B_1	6	1	7
B_2	0	5	5
Total	6	6	12

Solution

H_0: There is no association between attribute A and attribute B.

H_1: There is an association. (Since the direction cannot be specified a two-sided test has to be done.)

Choose $k = 5$, $n = 6$ and $x = 5$. Under H_0 X has a $H\,(12;\ 6;\ 5)$ distribution and

$$P\,(X = 0) + P\,(X = 1) + P\,(X = 2) + P\,(X = 3) + P\,(X = 4) + P\,(X = 5) = 1.$$

To show that this looks like a proper discrete distribution (and for illustration purposes) we draw the following probability distribution. The individual probabilities are obtained from table D by subtraction. (See the second part of activity 5.6 in the workbook for a similar example.)

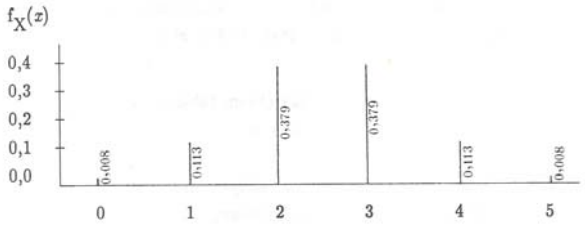


Figure 5.3: Probability distribution of $X \sim H\,(N = 12;\ \ n = 6;\ \ k = 5)$

We can only reject H_0 in favour of the two-sided alternative if x is too large or too small and if it represents a "rare event", in other words only if

$$P\,(X \leq x) \leq \frac{\alpha}{2} \text{ or if } P\,(X \geq x) \leq \frac{\alpha}{2}.$$

For any 2×2 table the question of hypothesis testing actually means: given the values of N, n and k, is the value of x unusual (too large or too small) to be ascribed to chance?

Suppose we choose $\alpha = 0.05 \Rightarrow \frac{\alpha}{2} = 0.025$.

The observed value of x is 5. From table D we find that

$$P(X \leq 5) = 1 \quad \text{and} \quad P(X \geq 5) = 1 - P(X \leq 4) = 0.008.$$

Since the smaller of these two probabilities is $P(X \geq 5) = 0.008 < \frac{\alpha}{2}$ we reject H_0 and conclude that there is an association between the two attributes A and B.

Final remarks

We introduced you to contingency table analysis by stating that it is the simultaneous study of two nominal variables. How will you capture two nominal variables in a JMP data set? Say for example you extend example 5.2 to capture the smoking habit and preference for coffee/tea for all the students taking STA2601? How will you then go a step further to create a cross-tabulation and test for independence? Please see activities 5.8 and 5.9 of the workbook.

5.4 Correlation

A. Correlation and independence

The concepts "independent" and "uncorrelated" are confused very often. We repeat two results from unit 1.

Theorem 5.2

Let X and Y be two random variables with correlation coefficient ρ.

If X and Y are independent then $\rho = 0$ (ie X and Y are uncorrelated).

Theorem 5.3

Let X and Y have a bivariate normal distribution with correlation coefficient ρ.

Then X and Y are independent if and only if $\rho = 0$.

Thus, if X and Y do not have a bivariate normal distribution then $\rho = 0$ does not necessarily imply independence.

Up to this point we did not explicitly mention that we are busy studying the simultaneous behaviour of two continuous variables. So, in contrast to the previous section where we had two categorical variables we are now interested in the independence/dependence of two interval-measured variables.

The sample correlation coefficient is

$$R = \frac{\Sigma \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\Sigma \left(X_i - \overline{X}\right)^2 \Sigma \left(Y_i - \overline{Y}\right)^2}}.$$

An alternative formula is

$$R = \frac{\Sigma X_i Y_i - \dfrac{\left(\Sigma x_i\right)\left(\Sigma Y_i\right)}{n}}{\sqrt{\left(\Sigma X_i^2 - \dfrac{\left(\Sigma X_i\right)^2}{n}\right)\left(\Sigma Y_i^2 - \dfrac{\left(\Sigma Y_i\right)^2}{n}\right)}}.$$

The first formula is "better" when using a computer. R, which is computed from a random sample $\left(X_i;\ Y_i\right),\ i = 1, ..., n$ where $\overline{X}$ and $\overline{Y}$ are the sample means, is used as an estimator for ρ. If X and Y follow a bivariate normal distribution then R is the MLE for ρ. If X and Y do not follow a bivariate normal distribution, there is usually not a parameter ρ which is to be estimated. **However, R is still used as a measure of the strength of the relationship between X and Y.** It must be remembered, however, that R is a measure of *linear* relationship. A small value of R may mean either that there is not a strong relationship between X and Y or that the relationship is not linear. It is always necessary to draw a graph on which each observation $\left(X_i;\ Y_i\right)$ is represented as a dot on the $\left(X;\ Y\right)$ plane in order to find out whether the relationship, if it exists, is linear.

> If we intend to construct confidence intervals for ρ or test hypotheses about ρ, we must assume that X and Y follow a bivariate normal distribution.

How do we know that we have a bivariate normal distribution in a practical application? This is a difficult problem. There are indicators which are necessary but not sufficient. If certain conditions are not satisfied, we do not have bivariate normality. If the conditions are satisfied, we can feel a bit more confident (but not certain) of bivariate normality. The indicators are as follows:

(a) *Marginal normality*

If the joint distribution of X and Y is bivariate normal then the marginal distributions of X and Y are univariate normal. We plot a histogram of each variable. These histograms should have a bell shape. We may of course perform a goodness-of-fit test for normality for the marginal distributions if we have a large number of observations.

(b) *Linearity*

The product moment correlation coefficient is a measure of linear correlation. If X and Y have a bivariate normal distribution then the relationship between them must be linear. We may plot a scatter diagram of the observations. The points should be scattered around a straight line, otherwise we can be sure that the joint distribution is not bivariate normal.

A *necessary and sufficient condition* for X and Y to have a bivariate normal distribution is that all linear combinations of the form $aX + bY$ (for all possible choices of a and b) should be univariate normal. You may use your imagination to think how one would use this fact to test whether X and Y have a bivariate normal distribution.

A further point could not be stressed often enough: if X and Y are correlated then it does not imply that there is a *causal* relationship. A famous case in point is a study of the relationship between smoking and the incidence of lung cancer. Although a definite correlation was observed, it was not a proof that smoking *causes* lung cancer. A causal relationship could only be demonstrated by means of carefully controlled experiments in which external factors which may contribute towards cancer can be "held constant". An example which is often cited is that there is a high correlation between the salary of the minister of a certain church in Xville and the price of rum in Jamaica. The question is, which is cause and which is effect?

B. Testing for zero correlation

Theorem 5.4

Let $(X_i;\ Y_i)\,, i = 1, ..., n$ be a random sample of size n from a bivariate normal distribution, and let R be the sample correlation coefficient. If $\rho = 0$ the statistic

$$T = \frac{\sqrt{n-2}R}{\sqrt{1-R^2}}$$

has a Student's t-distribution with $(n-2)$ degrees of freedom.

We give this theorem without proof and use this result to test

$H_0 : \rho = 0$ against the alternatives

$H_1 : \rho < 0$ or

$H_1 : \rho > 0$ or

$H_1 : \rho \neq 0.$

In the latter case, for example, H_0 is rejected if $|T| > t_{\frac{1}{2}\alpha;\, n-2}$.

Table IX of Stoker makes it unnecessary for us to compute T, since this table gives critical values for R itself. To see that this table is based on theorem 5.4, consider the case of testing $H_0 : \rho = 0$ against $H_1 : \rho \neq 0$. Let $t = t_{\frac{1}{2}\alpha;\, n-2}$ then H_0 is accepted if

$$\begin{aligned}
\sqrt{n-2}\,|R| / \sqrt{1-R^2} &< t \\
\therefore (n-2)\, R^2 / \left(1-R^2\right) &< t^2 \\
\therefore (n-2)\, R^2 &< t^2 - t^2 R^2 \\
\therefore \left(n-2+t^2\right) R^2 &< t^2 \\
\therefore R^2 &< t^2 / \left(n-2+t^2\right) \\
\therefore |R| &< t/\sqrt{n-2+t^2}.
\end{aligned}$$

Choose, for example, $\alpha = 0.05$ and $n = 20$. Then $t = t_{0.025;18} = 2.101$ (table III).

$\therefore t/\sqrt{n-2+t^2} = 2.101/\sqrt{18+4.414} = 0.4438$ which is the same as the critical value in table IX.

Example 5.4

At a certain university the 18 students who enrolled for a specific course were subjected to an aptitude test at the beginning of the year. Their scores in the aptitude test (X) and their marks in the final examinations (Y) were as follows:

Student	X_i	Y_i	Student	X_i	Y_i	Student	X_i	Y_i
1	13	65	7	16	45	13	5	65
2	11	75	8	11	35	14	2	45
3	5	60	9	12	50	15	7	40
4	15	70	10	8	40	16	9	60
5	10	75	11	10	80	17	12	80
6	6	60	12	14	75	18	14	60

We wish to test at the 10% level whether X and Y are correlated.

Solution

The first step is drawing a scatter diagram of X and Y to determine whether the relationship, if there is one, is indeed linear.

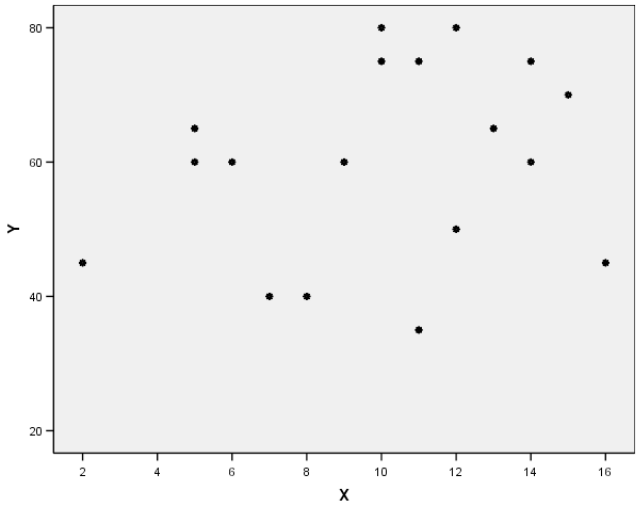


Figure 5.4: Scatter diagram of X and Y

From figure 5.4 it seems as though there is no strong linear relationship between X and Y, but this is a subjective conclusion. We formally test

$H_0 : \rho = 0$ against

$H_1 : \rho \neq 0$

by computing the test statistic R (or T).

We compute the sample correlation coefficient in tabular form.

X	Y	$X-\overline{X}$	$\left(X-\overline{X}\right)^2$	$Y-\overline{Y}$	$\left(Y-\overline{Y}\right)^2$	$\left(X-\overline{X}\right)\left(Y-\overline{Y}\right)$
13	65	3	9	5	25	15
11	75	1	1	15	225	15
5	60	−5	25	0	0	0
15	70	5	25	10	100	50
10	75	0	0	15	225	0
6	60	−4	16	0	0	0
16	45	6	36	−15	225	−90
11	35	1	1	−25	625	−25
12	50	2	4	−10	100	−20
8	40	−2	4	−20	400	40
10	80	0	0	20	400	0
14	75	4	16	15	225	60
5	65	−5	25	5	25	−25
2	45	−8	64	−15	225	120
7	40	−3	9	−20	400	60
9	60	−1	1	0	0	0
12	80	2	4	20	400	40
14	60	4	16	0	0	0
180	1 080	0	256	0	3 600	240

$\overline{X} = 10; \quad \overline{Y} = 60$

$$R = \frac{240}{\sqrt{256 \times 3600}} = \frac{240}{16 \times 60} = 0.25$$

In table IX we find the 10% two-sided critical value (which is the same as the 5% one-sided critical value) which is equal to 0.4. Since $0.25 < 0.4$ H_0 is not rejected in favour of H_1 at the 10% level.

Alternatively

$$T = \sqrt{n-2}\frac{R}{\sqrt{1-R^2}} = \sqrt{16}\frac{0.25}{\sqrt{0.9375}} = \frac{1}{0.9682} = 1.0328$$

in other words we do not reject $H_0 : \rho = 0$ at the **10% level** $(t_{\mathbf{0.05};16} = 1.746)$.

Notes about the computations of $\mathbf{R}$

1. In the above example the data were chosen in such a way that $\overline{X}$ and $\overline{Y}$ are integers. In practice this would happen only rarely. If $\overline{X}$ and $\overline{Y}$ are not integers or have more than a few decimal digits then it would be preferable to compute the covariance and variances by means of the alternative formulae

$$\Sigma\left(X_i - \overline{X}\right)^2 = \Sigma X_i^2 - \left(\Sigma X_i\right)^2/n$$
$$\Sigma\left(Y_i - \overline{Y}\right)^2 = \Sigma Y_i^2 - \left(\Sigma Y_i\right)^2/n$$
$$\Sigma\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right) = \Sigma X_i Y_i - \left(\Sigma X_i\right)\left(\Sigma Y_i\right)/n.$$

2. The correlation coefficient between X and Y is identical to the correlation coefficient between $a_1X + b_1$ and $a_2Y + b_2$ provided $a_1 > 0$ and $a_2 > 0$. For ease of computation one may subtract a constant near the mean of each variable and divide by another suitably chosen constant to reduce the observations to smaller numbers. For example in example 5.4 we could replace X_i by $X_i - 10$ and Y_i by $(Y_i - 50)/5$ where the constants 10; 50 and 5 were chosen by inspecting the data.

   (This type of transformation is nowadays seldom done because of the availability of calculators and the use of computers.)

C. Testing other hypotheses about the correlation coefficient

A famous British statistician, Sir Ronald Fisher, found an approximation to the distribution of the correlation coefficient (the distribution itself is much less manageable when $\rho \neq 0$). We state this as a theorem which we give without proof.

Theorem 5.5

Let R be the sample correlation coefficient of a random sample from a bivariate normal distribution.

Let $U = \frac{1}{2}\log_e \frac{1+R}{1-R}$ and $\eta = \frac{1}{2}\log_e \frac{1+\rho}{1-\rho}$.

Then, for large samples, $Z = \sqrt{n-3}\,(U-\eta)$ is approximately a $n\,(0;\ 1)$ variate.

Table X lists this transformation, usually called Fisher's Z-transformation. Note that

$$\frac{1}{2}\log_e \frac{1+R}{1-R} = -\frac{1}{2}\log_e \frac{1+(-R)}{1-(-R)}$$

so that for negative $R,$ one must look up the transformation of $|R|$ and add a negative sign.

Suppose that the notation ρ_0 implies a known value (other than zero) specified under the null hypothesis.

In order to test $H_0 : \rho = \rho_0$ against
$H_1 : \rho > \rho_0$ or
$H_1 : \rho < \rho_0$ or
$H_1 : \rho \neq \rho_0$ we compute

$$Z = \sqrt{n-3}\,(U-\eta_0) \text{ where } \eta_0 = \frac{1}{2}\log_e \left[(1+\rho_0)\,/\,(1-\rho_0)\right]$$

and reject H_0 if this quantity exceeds a critical value of the $n\,(0;\ 1)$ distribution.

Example 5.5

In a sample of 28 observations it is found that $R = 0.2$. We wish to test $H_0 : \rho = 0.5$ against $H_1 : \rho \neq 0.5$ at the 5% level.

Solution

We look up in table X (Stoker):

$$u = \frac{1}{2}\log_e \frac{1+0.2}{1-0.2} = 0.2027$$

$$\eta_0 = \frac{1}{2}\log_e \frac{1+0.5}{1-0.5} = 0.5493$$

$$\therefore z = \sqrt{25}\,(0.2027 - 0.5493) = -1.733$$

In table II we find $z_{0.025} = -1.96$.

Since $-1.96 < -1.733 < 1.96$ we do not reject $H_0 : \rho = 0.5$ at the 5% level.

D. Confidence interval for ρ

As has been said, the distribution of R when $\rho \neq 0$, is very complicated. We could use theorem 5.5 to construct a confidence interval for ρ.

For a 95% confidence interval, for example, we use the fact that

$$\begin{aligned} 0.95 &= P\,(-1.96 < Z < 1.96) \\ &\approx P\left(-1.96 < \sqrt{n-3}\,(U-\eta) < 1.96\right) \\ &= P\left(U - \frac{1.96}{\sqrt{n-3}} < \eta < U + \frac{1.96}{\sqrt{n-3}}\right) \end{aligned}$$

$$\begin{aligned} \text{where } U &= \frac{1}{2}\log_e \frac{1+R}{1-R} \\ \eta &= \frac{1}{2}\log_e \frac{1+\rho}{1-\rho} \end{aligned}$$

and then we have to transform the confidence limits for η back to confidence limits for ρ by means of the formula

$$\rho = \frac{e^{\eta} - e^{-\eta}}{e^{\eta} + e^{-\eta}} = \tan(\eta)$$

or by using table X inversely.

Example 5.6

From a sample of 25 observations from a bivariate normal distribution $R = -0.3$ was found. To find a 95% confidence interval for ρ. we look up for $R = 0.3$ in table X to find $U = 0.3095$; thus for $R = -0.3$ we have $U = -0.3095$. The 95% confidence interval for η is

$$\begin{aligned} -0.3095 - \frac{1.96}{\sqrt{22}} &\leq \eta \leq -0.3095 + \frac{1.96}{\sqrt{22}} \\ \therefore -0.3095 - 0.4179 &\leq \eta \leq -0.3095 + 0.4179 \\ \therefore -0.7274 &\leq \eta \leq 0.1084. \end{aligned}$$

Now $\dfrac{e^{-0.7274} - e^{0.7274}}{e^{-0.7274} + e^{0.7274}} = \dfrac{0.4832 - 2.0697}{0.4832 + 2.0697} = -0.62$

and $\dfrac{e^{0.1084} - e^{-0.1084}}{e^{0.1084} + e^{-0.1084}} = \dfrac{1.1145 - 0.8973}{1.1145 + 0.8973} = 0.11$

that is a 95% confidence interval for ρ is $(-0.62;\ 0.11)$. Using table X we have:

for $\eta = -0.7250 : \rho = -0.62$
for $\eta = -0.7414 : \rho = -0.63.$

Using linear interpolation:

For $\eta = -0.7274 :\ \rho = -0.62 + \dfrac{0.7274 - 0.7250}{0.7414 - 0.7250}(0.62 - 0.63)$

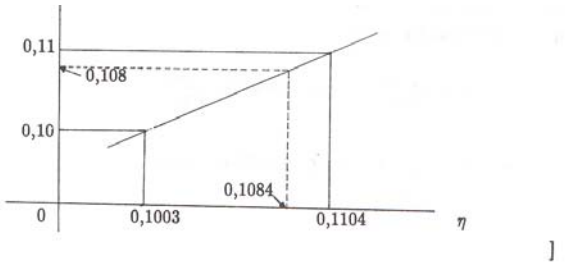
$$= -0.62 - 0.0015$$

$$\approx -0.62$$

for $\eta = 0.1104 : \rho = 0.11$
for $\eta = 0.1003 : \rho = 0.10.$

Once more using linear interpolation:

If you are fond of graphs, make the following sketch:



]

$$\therefore \text{ for } \eta = 0.1084: \quad \rho = 0.10 + \frac{0.1084 - 0.1003}{0.1104 - 0.1003}(0.11 - 0.10)$$

$$= 0.10 + 0.008$$

$$\approx 0.11$$

which is the same result.

E. Testing the equality of two correlation coefficients

Let R_1 be a correlation coefficient computed from a random sample of size n_1 from a distribution with population correlation coefficient ρ_1. Let R_2 be a correlation coefficient computed from a sample of size n_2 from a distribution with correlation coefficient ρ_2. We assume R_1 and R_2 are based on independent samples, in other words R_1 and R_2 are independent. We wish to test $H_{\mathbf{0}}$:$\rho_{\mathbf{1}} = \rho_{\mathbf{2}}$.

Let $U_1 = \frac{1}{2}\log_e(1 + R_1)/(1 - R_1)$

$$U_2 = \frac{1}{2}\log_e(1 + R_2)/(1 - R_2)$$

$$\eta_i = \frac{1}{2}\log_e(1 + \rho_i)/(1 - \rho_i). \quad i = 1; 2$$

If $H_0 : \rho_1 = \rho_2$ is true then $\eta_1 = \eta_2$.

For large values of n_1 and n_2

$$U_1 - \eta_1 \text{ is approximately } n\left(0;\ \frac{1}{n_1 - 3}\right)$$

$$U_2 - \eta_2 \text{ is approximately } n\left(0;\ \frac{1}{n_2 - 3}\right).$$

Therefore $(U_1 - U_2) - (\eta_1 - \eta_2)$ is approximately $n\left(0;\ \frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}\right)$ provided U_1 and U_2 are independent.

Theorem 5.6

If $\rho_1 = \rho_2$, that is $\eta_1 = \eta_2$, then $Z = \dfrac{U_1 - U_2}{\sqrt{\dfrac{1}{n_1 - 3} + \dfrac{1}{n_2 - 3}}}$ is approximately $n\,(0;\ 1)$.

We use this result in the usual manner to test H_0.

Example 5.7

In a sample of 111 schoolboys the correlation coefficient between their scores in an intelligence test and their scores in the final examinations was found to be 0.25. In a sample of 57 girls the correlation coefficient between the same two scores is 0.35. We wish to test at the 10% level whether the difference between the two sample correlation coefficients is significant.

Solution

We want to test $H_0 : \rho_1 = \rho_2$ against
$H_1 : \rho_1 \neq \rho_2$.

We compute

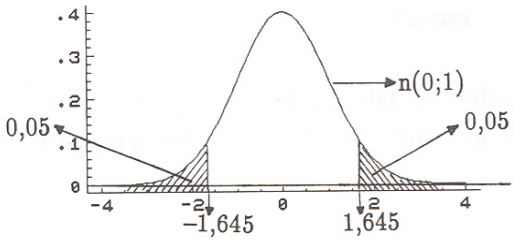
$$U_1 = \frac{1}{2}\log_e \frac{1.25}{0.75} = 0.2554$$

$$U_2 = \frac{1}{2}\log_e \frac{1.35}{0.65} = 0.3654$$

$$\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3} = \frac{1}{108} + \frac{1}{54} = \frac{1}{36}$$

$$\therefore Z = \frac{U_1 - U_2}{\sqrt{\dfrac{1}{n_1 - 3} + \dfrac{1}{n_2 - 3}}} = 6\,(0.2554 - 0.3654) = -0.66.$$

We will reject H_0 if $|Z| \geq 1.645$ in other words if $Z \leq -1.645$ or if $Z \geq 1.645$.



Since $-1.645 < -0.66 < 1.645$ we cannot reject H_0 at the 10% level.

Exercise 5.1

1. In a random sample of 50 men it was found that 26 smoked. In a random sample of 50 women, 14 smoked. Is there a relationship between gender and smoking? (Use the $2\frac{1}{2}\%$ level.)

2. One hundred students were classified with respect to their appearance ($A =$ attractive; $O =$ ordinary; $U =$ unattractive) and their intelligence (very high; high; average; low). The frequencies are as follows:

	VH	H	A	L
A	9	12	7	2
O	8	11	14	7
U	3	7	9	11

   Test at the 10% level whether there is an association between appearance and intelligence. Discuss the relationship between the two variables.

3. The following table gives the high school score (X) (on a five-point scale) and the first-year university score (Y) of 30 students:

X	Y	X	Y	X	Y
2.9	2.3	2.9	1.9	3.1	2.8
2.3	2.5	2.7	2.2	3.3	3.2
3.6	2.9	3.7	3.1	2.7	1.8
3.5	3.8	2.7	2.6	3.5	2.7
3.7	3.5	3.3	2.8	2.9	2.1
2.8	2.9	2.8	2.7	2.7	1.7
3.5	3.0	3.1	2.4	2.9	1.7
3.0	2.7	2.8	3.0	3.2	2.3
2.3	2.1	3.0	3.3	3.4	2.6
3.0	2.9	2.2	1.8	2.5	2.7

Draw a graph of the data to decide whether the relationship is linear or not. Compute the sample correlation coefficient and test $H_0 : \rho = 0$ against $H_1 : \rho > 0$ at the 1% level.

4. In a random sample of 39 observations the sample correlation coefficient was -0.35. Test $H_0 : \rho = -0.2$ against $H_1 : \rho \neq -0.2$ at the 5% level.

5. In a random sample of 33 observations $R = -0.6$ was found, and in a second random sample of 153 observations $R = -0.8$. Test at the 5% level $H_0 : \rho_1 = \rho_2$ against $H_1 : \rho_1 \neq \rho_2$.

6. In a random sample of 10 observations $R = 0.7$ was found. Find 95% confidence limits for ρ.

7. For the case $N = 12, n = 6, k = 5$ construct a 2×2 contingency table for which the null hypothesis would be rejected at the 1% level of significance in favour of a one-sided alternative.

8. In an experiment to test whether white mice are more susceptible to influenza than brown mice, six mice of each colour were exposed to an influenza virus. One of the six brown mice contracted influenza, compared to five of the six white mice. Construct a contingency table and test the hypothesis that the two strains are equally susceptible, against the alternative that white mice are more susceptible, at the 5% level of significance.

9. A random sample of 200 elderly men were classified according to level of training and number of children:

	Number of children			
Training	0	1	2	more than 2
Primary school	18	22	30	30
Secondary school	6	24	15	15
College	2	0	13	15
University	4	4	2	0

Test at the 5% level whether the number of children is independent of the father's level of training.

10. In a random sample of 19 observations from a bivariate normal distribution, the correlation coefficient was $R = 0.5$.

    (a) Test $H_0 : \rho = 0.2$ against $H_1 : \rho < 0.2$ at the 5% level of significance.

    (b) Construct a 95% (two-sided) confidence interval for ρ.

11. Two independent random samples from bivariate normal distributions yielded the following correlation coefficients:

    Sample 1: $R_1 = 0.6 \quad n_1 = 53$

    Sample 2: $R_2 = 0.9 \quad n_2 = 53$

    Test $H_0 : \rho_1 = \rho_2$ against $H_1 : \rho_1 < \rho_2$ at the 5% level of significance.

12. A certain agricultural product is produced in ten districts. The rainfall (cm) and yield (tons per ha) were recorded on one farm in each district:

District	1	2	3	4	5	6	7	8	9	10
Rainfall	60	48	34	46	58	70	26	44	62	52
Yield	17	22	19	26	32	12	10	21	16	25

    Compute the correlation coefficient and test at the 5% level of significance whether rainfall and yield are correlated. Discuss your assumptions and your conclusions.

5.5 Learning outcomes

Use the following learning outcomes as a checklist after you have completed this study unit to evaluate the knowledge you have acquired.

After studying study unit 5, you should **be able to**

- define statistical independence
- check for independent observations
- explain the dependence of five classical examples
- explain what is meant by a contingency table
- perform and interpret the chi-square test of independence for an $h \times k$ contingency table
- perform an exact test for a 2×2 contingency table
- define the terms *sample covariance* and *sample correlation coefficient*
- perform and interpret the hypothesis test $H_0 : \rho = 0$
- compute a confidence interval for ρ
- perform and interpret the hypothesis test $H_0 : \rho = \rho_0$
- perform and interpret the hypothesis test for the equality of two correlation coefficients, $H_0 : \rho_1 = \rho_2$

STUDY UNIT 6

Inference on variances

We discuss four problems: inference on the variance of a normal distribution, inference on the ratio of the variances of two normal distributions based on independent samples, testing the equality of two variances based on paired observations and testing the equality of more than two variances. In the first two instances we have to distinguish between problems involving known means and problems involving unknown means.

6.1 One-sample problem

Example 6.1

A tyre manufacturer claims that a certain type of tyre will last an average of 50 000 km on a certain make of car, and that the standard deviation is no more than 3 000 km. Eight tyres were tested by an inspector, and they lasted the following distances (in thousands of km):

47; 48; 50; 51; 52; 55; 55; 58

Would you believe the claim that the standard deviation is at most 3 000 km

(a) if you accept that the population mean is 50 000 km?

(b) if you do not accept the specified mean?

How shall we test this?

To derive a test statistic and a proper hypothesis test, we combine result 1.2 and result 1.3 of study unit 1 into the following theorem:

Theorem 6.1

Let $X_1,\ X_2,\ ...,\ X_n$ be a random sample from a $n\left(\mu;\ \sigma^2\right)$ distribution. Then

(a) $\sum_{i=1}^{n}\left(X_i-\mu\right)^2/\sigma^2 \sim \chi^2_n$

(b) $\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2/\sigma^2 \sim \chi^2_{n-1}$ where $\overline{X}=\frac{1}{n}\Sigma X_i$.

Hypothesis testing

We want to test the null hypothesis $\mathbf{H_0}$: $\boldsymbol{\sigma^2}=\mathbf{c}$.

(a) $\boldsymbol{\mu}$ **known**

The procedure is based on the statistic $U=\sum_{i=1}^{n}\left(X_i-\mu\right)^2/c$ which, if H_0 is true, is a χ^2_n variate.

If $\Sigma\left(X_i-\mu\right)^2$ is small, it is an indication that σ^2 is small and vice versa. We reject $H_0:\sigma^2=c$ against the alternatives

(i) $H_1:\sigma^2\neq c$ if $U<\chi^2_{1-\frac{1}{2}\alpha;n}$ or $U>\chi^2_{\frac{1}{2}\alpha;n}$

(ii) $H_1:\sigma^2<c$ if $U<\chi^2_{1-\alpha;n}$

(iii) $H_1:\sigma^2>c$ if $U>\chi^2_{\alpha;n}$.

(The explanation and the application of the critical values of the chi-square distribution are shown in figure 6.1.)

(b) $\boldsymbol{\mu}$ **unknown**

The procedure is based on the statistic $U=\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2/c$ which, if H_0 is true, is a χ^2_{n-1} variate.

We reject $H_0:\sigma^2=c$ against the alternatives

(i) $H_1:\sigma^2\neq c$ if $U<\chi^2_{1-\frac{1}{2}\alpha;n-1}$ or $U>\chi^2_{\frac{1}{2}\alpha;n-1}$

(ii) $H_1:\sigma^2<c$ if $U<\chi^2_{1-\alpha;n-1}$

(iii) $H_1:\sigma^2>c$ if $U>\chi^2_{\alpha;n-1}$.

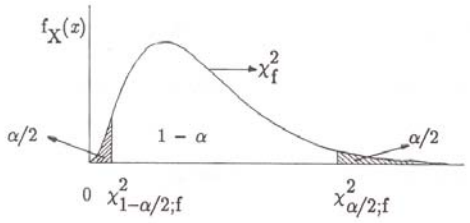
Suppose we use the notation of table IV (Stoker) for the critical values of the χ^2 distribution except that we interchange ν and P. This means we first write down the **tail-to-the-right area** and follow it by the degrees of freedom. [Please see activities 6.2 and 6.3 for concrete examples of this notation.]

We define the use and the notation of the critical values of the chi-square distribution such that, if $U \sim \chi^2_f$ then

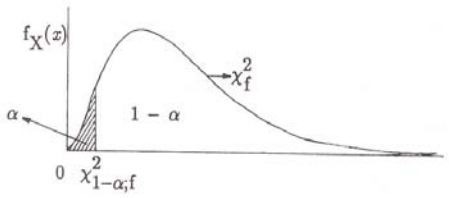
(a) $P\left(\chi^2_{1-\alpha/2;f} < U < \chi^2_{\alpha/2;f}\right) = 1 - \alpha$

(b) $P\left(U > \chi^2_{1-\alpha;f}\right) = 1 - \alpha$

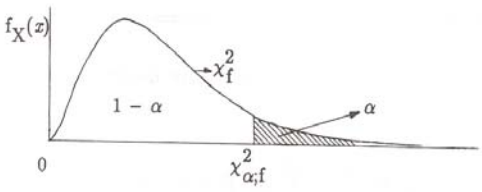
(c) $P\left(U < \chi^2_{\alpha;f}\right) = 1 - \alpha.$



(a) $P(\chi^2_{1-\alpha/2\,;\,f} < U < \chi^2_{\alpha/2\,;\,f}) = 1 - \alpha$



(b) $P(U > \chi^2_{1-\alpha\,;\,f}) = 1 - \alpha$



(c) $P(U < \chi^2_{\alpha\,;\,f}) = 1 - \alpha$

Figure 6.1: Critical values of the χ^2_f distribution

Example 6.1 (continued)

This example is a one-sided problem. That means we have to test

$H_0 : \sigma^2 = 9$ against
$H_1 : \sigma^2 > 9.$

To calculate the value of the test statistic, it depends on what we assume about the population mean.

(a) **Suppose we assume that the population mean is 50 000 km**. How will we utilise this information?

The observations, $X_1,\ X_2,\ ...,\ X_8$ are given in *thousands of km* which means we must do the same with μ. Hence we write: **Assume that $\boldsymbol{\mu = 50}$.**

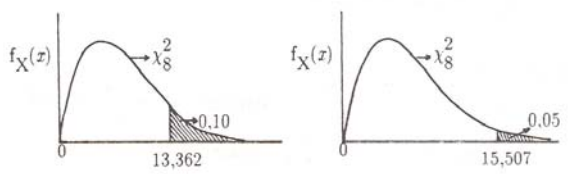
Now we are able to compute $\sum_{i=1}^{8} (X_i - 50)^2 = 132.$

We use the test statistic $U = \dfrac{\Sigma\,(X_i - \mu)^2}{c}$ which has a χ^2_n distribution.

So, $U = \dfrac{132}{9} = 14.6667.$

Since we have one-sided testing (to the right) we will reject H_0 if $U \geq \chi^2_{\alpha;8}$.

α was not specified and we will look up the critical values for both $\alpha = 0.05$ and $\alpha = 0.10.$



Now $\chi^2_{0.10;8} = 13.3616$ and since $14.6667 > 13.362$ we reject H_0 at the 10% level in favour of H_1. (We do not reject H_0 at the 5% level since $\chi^2_{0.05;8} = 15.5073$.)

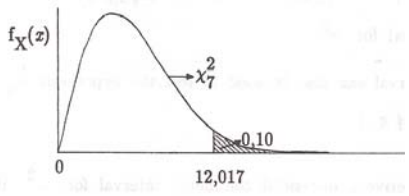
(b) **Suppose we do not know that $\boldsymbol{\mu = 50}$.**

Now we have to estimate the "unknown population mean" as $\hat{\mu} = \overline{X}$ and we have to compute $\sum_{i=1}^{8} \left(X_i - \overline{X}\right)^2$.

$\overline{X} = \dfrac{\Sigma X_i}{8} = \dfrac{416}{8} = 52$ and $\sum_{i=1}^{8} \left(X_i - \overline{X}\right)^2 = 100.$

We use the test statistic $U = \dfrac{\Sigma\left(X_i - \overline{X}\right)^2}{c}$ which has a χ^2_{n-1} distribution.

So, now $U = \dfrac{100}{9} = 11.1111.$



Now $\chi^2_{0.10;7} = 12.017$ and since $11.1111 < 12.017$ we cannot reject H_0 in favour of H_1 at the 10% level.

Confidence intervals

We want to derive a confidence interval for σ^2.

How shall we derive a **two-sided** $100\,(1-\alpha)\,\%$ confidence interval for an unknown variance?

From theorem 6.1 we know that if $U = \dfrac{\Sigma\left(X_i - \overline{X}\right)^2}{\sigma^2}$ then $U \sim \chi^2_{n-1}$. From this we may derive the probability expression

$$\begin{aligned}
1-\alpha &= P\left(\chi^2_{1-\frac{1}{2}\alpha;n-1} < U < \chi^2_{\frac{1}{2}\alpha;n-1}\right) \\
&= P\left[\chi^2_{1-\frac{1}{2}\alpha;n-1} < \frac{\Sigma\left(X_i - \overline{X}\right)^2}{\sigma^2} < \chi^2_{\frac{1}{2}\alpha;n-1}\right] \\
&= P\left[\frac{1}{\chi^2_{\frac{1}{2}\alpha;n-1}} < \frac{\sigma^2}{\Sigma\left(X_i - \overline{X}\right)^2} < \frac{1}{\chi^2_{1-\frac{1}{2}\alpha;n-1}}\right] \\
&= P\left[\frac{\Sigma\left(X_i - \overline{X}\right)^2}{\chi^2_{\frac{1}{2}\alpha;n-1}} < \sigma^2 < \frac{\Sigma\left(X_i - \overline{X}\right)^2}{\chi^2_{1-\frac{1}{2}\alpha;n-1}}\right].
\end{aligned}$$

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 $100\,(1-\alpha)\,\%$ confidence interval for σ^2.

Please note that this interval can also be used to test the hypothesis $H_0 : \sigma^2 = c$ against $H_1 : \sigma^2 \neq c$.

The interval above was derived under the (b)-assumption of theorem 6.1 and for a two-sided confidence of $1 - \alpha$.

As another example, we shall now derive a one-sided upper confidence interval for σ^2 if μ is assumed to be known.

Let $U = \sum_{i=1}^{n} (X_i - \mu)^2 / \sigma^2$, then $U \sim \chi^2_n$

$$\begin{aligned} \therefore 1 - \alpha &= P\left[\chi^2_{1-\alpha;n} < U\right] \\ &= P\left[\chi^2_{1-\alpha;n} < \frac{\Sigma (X_i - \mu)^2}{\sigma^2}\right] \\ &= P\left[\frac{1}{\chi^2_{1-\alpha;n}} > \frac{\sigma^2}{\Sigma (X_i - \mu)^2}\right] \\ &= P\left[\sigma^2 < \frac{\Sigma (X_i - \mu)^2}{\chi^2_{1-\alpha;n}}\right]. \end{aligned}$$

Therefore $\left[0;\ \frac{\Sigma (X_i - \mu)^2}{\chi^2_{1-\alpha;n}}\right]$ is a $100(1-\alpha)\,\%$ one-sided confidence interval for σ^2 which tests the hypothesis $H_0 : \sigma^2 = c$ against $H_1 : \sigma^2 < c$.

It is better (and safer) to understand how to derive these intervals than to try and memorise the results!

You must be able to derive the other one and two-sided intervals if μ is known or unknown.

Example 6.2

In a sample of size 20 from a $n(\mu;\ \sigma^2)$ distribution it was found that $\Sigma X_i = 30$ and $\Sigma X_i^2 = 60$. Construct 90% two-sided confidence limits for σ^2,

(a) assuming $\mu = 2$

(b) assuming μ is unknown.

Would you accept $H_0 : \sigma^2 = 1$ against $H_1 : \sigma^2 \neq 1$ in each case?

Solution

$$\begin{aligned}\Sigma\left(X_i-\overline{X}\right)^2 &= \Sigma X_i^2 - n\overline{X}^2\\ &= \Sigma X_i^2 - \left(\Sigma X_i\right)^2/n\\ &= 60-(30)^2/20\\ &= 60-45\\ &= 15\end{aligned}$$

$$\begin{aligned}\Sigma\left(X_i-\mu\right)^2 &= \Sigma X_i^2 - 2\mu\Sigma X_i + n\mu^2\\ &= 60-2\,(2)\,(30)+20\,(2)^2\\ &= 60-120+80\\ &= 20\end{aligned}$$

(a) **If μ is known:**

A 90% two-sided confidence interval for σ^2 is $\left(\frac{20}{\chi^2_{0.05;20}};\quad \frac{20}{\chi^2_{0.95;20}}\right)$ that is $\left(\frac{20}{31.4104};\quad \frac{20}{10.8508}\right)$ that is $(0.64;\quad 1.84)$.

A variation on the theme is where it is required to derive a confidence interval for the *standard deviation*. What will a 90% two-sided confidence interval for σ be?

We simply take the square root on both sides and a 90% confidence interval for σ is therefore $(0.80;\quad 1.36)$.

(b) **If μ is unknown:**

The 90% two-sided confidence interval for σ^2 now becomes $\left(\frac{15}{\chi^2_{0.05;19}};\quad \frac{15}{\chi^2_{0.95;19}}\right)$ that is $\left(\frac{15}{30.1435};\quad \frac{15}{10.117}\right)$ that is $(0.50;\quad 1.48)$.

A two-sided confidence interval may be used to test a two-sided alternative. In both cases $\sigma^2=1$ falls inside the interval and we do not reject $H_0:\sigma^2=1$. If $\sigma^2=1$ did not fall inside the interval we would have come to the conclusion that we reject H_0 at the 10% level of significance.

Example 6.3

Suppose we wish to construct a 95% two-sided confidence interval for σ^2 based on a random sample of 16 observations from a $n\left(\mu;\ \sigma^2\right)$ distribution, assuming μ is unknown.

(a) What is the expected length of the confidence interval?

(b) What is the expected length of the confidence interval if the sample size is 30?

Solution

The crux of this question is that we have to bring "mathematical expectation" somewhere into the picture. This is where we have to fall back on theoretical knowledge which always takes us back to study unit 1.

We will use the fact that $\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2/\sigma^2 \sim \chi^2_{n-1}$.

$\therefore E\left[\Sigma\left(X_i-\overline{X}\right)^2/\sigma^2\right]=n-1$ (see result 1.1)

$\therefore E\left[\Sigma\left(X_i-\overline{X}\right)^2\right]=\sigma^2(n-1)$

(a) The 95% confidence interval for σ^2 when $n=16$ is

$$\left[\frac{\Sigma\left(X_i-\overline{X}\right)^2}{\chi^2_{\frac{1}{2}\alpha;n-1}};\quad \frac{\Sigma\left(X_i-\overline{X}\right)^2}{\chi^2_{1-\frac{1}{2}\alpha;n-1}}\right]=\left[\frac{\Sigma\left(X_i-\overline{X}\right)^2}{27.4884};\quad \frac{\Sigma\left(X_i-\overline{X}\right)^2}{6.2614}\right].$$

The length of the interval is the difference between the upper bound and the lower bound

$=\Sigma\left(X_i-\overline{X}\right)^2\left(\frac{1}{6.2614}-\frac{1}{27.4884}\right)=0.1233\Sigma\left(X_i-\overline{X}\right)^2$, which is a random variable.

The *expected length* is therefore $0.1233E\left[\Sigma\left(X_i-\overline{X}\right)^2\right]=(0.1233)(15)\sigma^2=1.8495\sigma^2$.

(b) Now $n=30$ and the length of the interval is

$$\Sigma\left(X_i-\overline{X}\right)^2\left(\frac{1}{\chi^2_{0.975;29}}-\frac{1}{\chi^2_{0.025;29}}\right)=\Sigma\left(X_i-\overline{X}\right)^2\left(\frac{1}{16.0471}-\frac{1}{45.7222}\right)$$

$$=\Sigma\left(X_i-\overline{X}\right)^2(0.0404).$$

The expected length is therefore $0.0404(29)\sigma^2=1.1716\sigma^2$.

[NB The intervals become **narrower** as n **gets larger** even though

$E\left[\Sigma\left(X_i-\overline{X}\right)^2\right]=(n-1)\sigma^2$. Convince yourself by calculating the expected length of a sample

size of $n = 10$.]

6.2 Two independent samples

Example 6.4

Two operators are asked to analyse 10 samples each of a mixture which contains exactly 14.6% iron. The one operator broke two test tubes, with the result that he has only eight analyses. Their determinations of the iron content were as follows:

Operator A:	14.6	14.5	14.8	14.4
	14.2	14.8	14.7	14.6
Operator B:	14.3	14.6	15.0	14.6
	14.1	15.1	15.0	14.6
	14.3	14.6		

Is there reason to believe that the two operators differ with respect to precision, in other words does one operator show greater variation in his determination than the other?

Example 6.5

A manufacturer of prestige cars has a choice between two makes of fan belts to install in the cars. He wants to use the make with the least variation in lifetime, because the lifetime of any given fan belt can then be predicted accurately and the belt replaced before it breaks. (The mean lifetime is not of prime importance.) However, make A is cheaper, and he wants to use B only if its standard deviation is less than 80% of the standard deviation of A. He tests a number of fan belts of each make, and the results (in thousands of km) are as follows:

Make A:	44;	44;	49;	38;	46;	41;	50;	46;	50;	42
Make B:	50;	51;	50;	48;	53;	48;	50			

Do these observations confirm that the standard deviation of B is less than 0.8 times that of A?

We use the following model for this type of problem:

Let $(X_{11},\ ...,\ X_{1n_1})$ and $(X_{21},\ ...,\ X_{2n_2})$ be two *independent* random samples from $n\left(\mu_1;\ \sigma_1^2\right)$ and

$n\left(\mu_2;\ \sigma_2^2\right)$ distributions respectively. Thus in total $X_{11},\ ...,\ X_{1n_1},\quad X_{21},\ ...,\ X_{2n_2}$ are (n_1+n_2) independent random variables.

From result 1.3 we know that $\Sigma\left(X_{1i}-\overline{X}_1\right)^2/\sigma_1^2$ and $\Sigma\left(X_{2i}-\overline{X}_2\right)^2/\sigma_2^2$ are independent $\chi^2_{n_1-1}$ and $\chi^2_{n_2-1}$ variates respectively. Thus it follows from definition 1.21 that $F=\dfrac{\left(\chi^2_{n_1-1}\right)/(n_1-1)}{\left(\chi^2_{n_2-1}\right)/(n_2-1)}$ has an $F_{(n_1-1);(n_2-1)}$ distribution.

In other words

$$\begin{aligned} F &= \frac{\Sigma\left(X_{1i}-\overline{X}_1\right)^2}{\sigma_1^2(n_1-1)}\Big/\frac{\Sigma\left(X_{2i}-\overline{X}_2\right)^2}{\sigma_2^2(n_2-1)} \\ &= \frac{S_1^2}{\sigma_1^2}\Big/\frac{S_2^2}{\sigma_2^2} \\ &= \frac{\sigma_2^2}{\sigma_1^2}\cdot\frac{S_1^2}{S_2^2} \end{aligned}$$

has an $F_{n_1-1;n_2-1}$ distribution, where

$$\overline{X}_1=\frac{1}{n_1}\sum_1^{n_1}X_{1i};\quad \overline{X}_2=\frac{1}{n_2}\sum_1^{n_2}X_{2i};$$

$$S_1^2=\frac{1}{n_1-1}\Sigma\left(X_{1i}-\overline{X}_1\right)^2;\quad S_2^2=\frac{1}{n_2-1}\Sigma\left(X_{2i}-\overline{X}_2\right)^2.$$

Aha, and here we have the key to a test statistic.

> We use as test statistic
>
> $$F=\frac{\sigma_2^2}{\sigma_1^2}\cdot\frac{S_1^2}{S_2^2}\sim F_{n_1-1;n_2-1.}$$

Since this follows a "standard distribution" which has been studied and for which we have tables with critical values, the last problem to solve is to express the null hypothesis in such a way that $\left[\dfrac{\sigma_2^2}{\sigma_1^2}\right]$ will "disappear" and thus we will only have to compute a statistic based on the sample outcomes of two independent samples.

We wish to test $H_0:\sigma_1^2=\sigma_2^2$ against three possible alternatives.

(A) $H_1:\sigma_1^2\neq\sigma_2^2$

(B) $H_1:\sigma_1^2>\sigma_2^2$

(C) $H_1 : \sigma_1^2 < \sigma_2^2$

The trick is to "manipulate" $H_0 : \sigma_1^2 = \sigma_2^2$ and rewrite it as the equivalent expression $H_0 : \frac{\sigma_2^2}{\sigma_1^2} = 1.$

This means that the three possible alternatives will change accordingly to

(A) $H_1 : \frac{\sigma_2^2}{\sigma_1^2} \neq 1$

(B) $H_1 : \frac{\sigma_2^2}{\sigma_1^2} < 1$

(C) $H_1 : \frac{\sigma_2^2}{\sigma_1^2} > 1.$

We can even take all the possible hypotheses a step further (to a more general expression) by replacing the 1 with a *known constant* (call it c).

What does this imply?

$H_0 : \frac{\sigma_2^2}{\sigma_1^2} = c$ means we are actually testing

$H_0 : \sigma_2^2 = c\sigma_1^2$ (or $H_0 : \sigma_1^2 = \left(\frac{1}{c}\right)\sigma_2^2$ if you prefer!)

Before we apply this to our two examples (which were introduced at the beginning of this section) we need to clarify how to obtain the critical values. Do you recall that if $F \sim F_{f;g}$ then $\frac{1}{F} \sim F_{g;f}$? (See result 1.4.)

We assume the shorthand notation $F_{\alpha;f;g}$ to mean $P\left[F \geq F_{\alpha;f;g}\right] = \alpha$. Thus, to obtain the *lower critical value* from tables V, VI and VII, we use the fact that $F_{1-\alpha;f;g} = \frac{1}{F_{\alpha;g;f}}$.

Please revisit exercise 1.2 and example 1.11 of study unit 1.

For example if $F \sim F_{3;6}$ then $P\left[F_{0.975;3;6} < F < F_{0.025;3;6}\right] = 0.95.$

Thus $P\left[\frac{1}{F_{0.025;6;3}} < F < F_{0.025;3;6}\right] = 0.95.$

If we use table VI it follows that $P\left[\frac{1}{14.7} < F < 6.6\right] = 0.95.$

It is important to realise that there will always be a connection between a hypothesis test and the derivation of a confidence interval. The one is only a different algebraic manipulation of the other.

Back to the critical values, we may find critical values such that

$$P\left(F_{1-\frac{\alpha}{2};n_1-1;n_2-1} < \frac{\sigma_2^2}{\sigma_1^2}\frac{S_1^2}{S_2^2} < F_{\frac{\alpha}{2};n_1-1;n_2-1}\right) = 1-\alpha \ \ldots\ldots\ldots\ldots\ldots\ldots (A).$$

Please note that instead of writing $P\,(\text{lower value } < F < \text{ upper value}) = 1 - \alpha$ we have replaced "lower value" with the correct notation; replaced F with $\frac{\sigma_2^2}{\sigma_1^2} \cdot \frac{S_1^2}{S_2^2}$ and replaced "upper value" with the correct notation. What is the point I am trying to make? Expression (A) is really simple and the replacements make it look very complicated.

For one-sided critical values we use the following two expressions:

$$P\left[\frac{\sigma_2^2}{\sigma_1^2} \cdot \frac{S_1^2}{S_2^2} < F_{\alpha;n_1-1;n_2-1}\right] = 1 - \alpha \ldots\ldots\ldots\ldots\ldots\ldots (B)$$

and

$$P\left[F_{1-\alpha;n_1-1;n_2-1} < \frac{\sigma_2^2}{\sigma_1^2} \cdot \frac{S_1^2}{S_2^2}\right] = 1 - \alpha \ldots\ldots\ldots\ldots\ldots\ldots (C)\,.$$

The critical values given in expression (A) can be used to test

$H_0 : \frac{\sigma_2^2}{\sigma_1^2} = c$ (where c is a specified positive number) against

$H_1 : \frac{\sigma_2^2}{\sigma_1^2} \neq c$, because we will reject H_0 in favour of H_1 if

$$F \leq F_{1-\alpha/2;n_1-1;n_2-1} \quad \text{or if} \quad F \geq F_{\alpha/2;n_1-1;n_2-1}.$$

Expression (A) is also the first step in the derivation of a two-sided confidence interval for $\frac{\sigma_2^2}{\sigma_1^2}$.

This we find as $\left(\frac{\text{lower value}}{S_1^2/S_2^2};\quad \frac{\text{upper value}}{S_1^2/S_2^2}\right)$.

In other words, we are $(1-\alpha)\,100\%$ confident that the ratio $\frac{\sigma_2^2}{\sigma_1^2}$ will fall between

$$\left[\frac{F_{1-\alpha/2;n_1-1;n_2-1}}{S_1^2/S_2^2};\quad \frac{F_{\alpha/2;n_1-1;n_2-1}}{S_1^2/S_2^2}\right].$$

We will use expression (B) to test H_0 against $H_1 : \frac{\sigma_2^2}{\sigma_1^2} < c$ and we will reject H_0 in favour of H_1 if $F = \frac{\sigma_2^2}{\sigma_1^2} \cdot \frac{S_1^2}{S_2^2} \geq F_{\alpha;n_1-1;n_2-1}$.

Similarly, we will use expression (C) to test H_0 against $H_1 : \frac{\sigma_2^2}{\sigma_1^2} > c$ and we will reject H_0 in favour of H_1 if $F = \frac{\sigma_2^2}{\sigma_1^2} \cdot \frac{S_1^2}{S_2^2} \leq F_{1-\alpha;n_1-1;n_2-1}$.

Are you able to derive the relevant one-sided confidence intervals for $\frac{\sigma_2^2}{\sigma_1^2}$ and how will you interpret and apply them?

Please note:
Do you recall from section 6.1 that the χ^2-test statistic changed depending on whether μ was known or unknown? Exactly the same could happen with the two-sample problem resulting in an F-test.

1. If μ_1 and μ_2 are known, we simply replace the $\overline{X}_i$ in S_i^2 by μ_i (for $i = 1$ and 2) and then the F-statistic has n_1 and n_2 degrees of freedom.

$$\Rightarrow F = \frac{\sigma_2^2}{\sigma_1^2} \cdot \frac{\sum_{i=1}^{n_1} (X_{1i} - \mu_1)^2 / n_1}{\sum_{i=1}^{n_2} (X_{2i} - \mu_2)^2 / n_2} \sim F_{n_1;n_2}$$

2. Either of the two samples may be regarded as the "first" sample, provided that the null and alternative hypotheses correspond with this. (This is a way to avoid the "awkward" lower critical value when using expression (C) for one-sided testing by switching to expression (B).)

Example 6.4 (continued)

We assume that both the analyses represent normal distributions and that they are independent (which seems logical). We have to test

$H_0 : \sigma_2^2 = \sigma_1^2$, written as $\frac{\sigma_2^2}{\sigma_1^2} = 1$ (ie $c = 1$) against

$H_1 : \sigma_2^2 \neq \sigma_1^2$, written as $\frac{\sigma_2^2}{\sigma_1^2} \neq 1$.

We use the test statistic $\Rightarrow F = \frac{\sigma_2^2}{\sigma_1^2} \cdot \frac{\sum_{i=1}^{n_1} (X_{1i} - \mu_1)^2 / n_1}{\sum_{i=1}^{n_2} (X_{2i} - \mu_2)^2 / n_2}$ because we assume that $\mu_1 = \mu_2 = 14.6$ (a known value).

Computation of F:

$$\sum_{i=1}^{8} (X_{1i} - 14.6)^2 = 0.30; \quad \sum_{i=1}^{10} (X_{2i} - 14.6)^2 = 1.00 \quad n_1 = 8; \quad n_2 = 10$$

$$\therefore F = \frac{\sigma_2^2}{\sigma_1^2} \cdot \frac{\sum_{i=1}^{8} (X_{1i} - 14.6)^2 / 8}{\sum_{i=1}^{10} (X_{2i} - 14.6)^2 / 10}$$

$$= (1) \frac{(0.30)/8}{(1.00)/10}$$

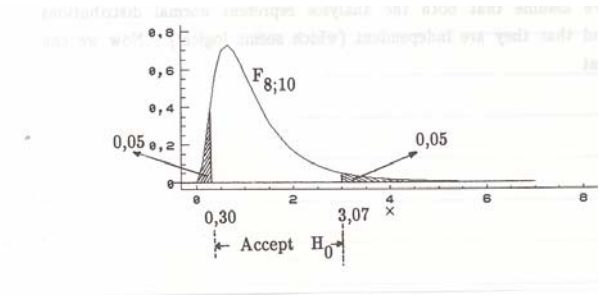
$$= \frac{0.0375}{0.1}$$

$$= 0.375$$

Critical values

Under H_0 the test statistic has an $F_{8;10}$ distribution. We choose $\alpha = 0.10$ and find from table V that $F_{0.05;8;10} = 3.07$

$F_{0.95;8;10} = 1/F_{0.05;10;8} = 1/3.35 = 0.30.$



Since $0.30 < F < 3.07$ we cannot reject H_0. The two operators do not differ with respect to precision.

Example 6.5 (continued)

Let us call *make A* the first sample and *make B* the second sample. If we once more assume that we have two independent random samples from $n\left(\mu_1;\ \sigma_1^2\right)$ and $n\left(\mu_2;\ \sigma_2^2\right)$ distributions respectively, we can use the F-test to test whether "the standard deviation of B is less than 0.8 times that of A".

$H_0 : \sigma_2/\sigma_1 = 0.8$ that is

$$H_0 : \frac{\sigma_2^2}{\sigma_1^2} = 0.64 \text{ against}$$

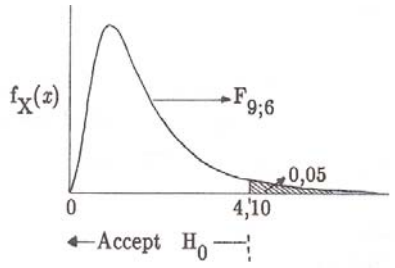
$$H_1 : \frac{\sigma_2^2}{\sigma_1^2} < 0.64.$$

We compute

$$\overline{X}_1 = 45; \quad \Sigma\left(X_{1i} - \overline{X}_1\right)^2 = 144; \quad n_1 - 1 = 9; \quad S_1^2 = 144/9 = 16$$

$$\overline{X}_2 = 50; \quad \Sigma\left(X_{2i} - \overline{X}_2\right)^2 = 18; \quad n_2 - 1 = 6; \quad S_2^2 = 18/6 = 3$$

Since we have a one-sided test we use expression (B) to find the critical value.



In table V we find $F_{0.05;9;6} = 4.10$. We reject H_0 if $F > 4.10$.

Now

$$\begin{aligned} F &= \frac{\sigma_2^2}{\sigma_1^2} \cdot \frac{S_1^2}{S_2^2} \\ &= 0.64 \times \frac{16}{3} \\ &\approx 3.4133. \end{aligned}$$

Since $F = 3.4133 < 4.10$ we do not reject H_0. There is insufficient evidence that the standard deviation of make B is less than 0.8 times that of make A. The manufacturer will probably decide to use make A.

It also follows from expression (B) that

$$P\left[\frac{\sigma_2^2}{\sigma_1^2} \cdot \frac{S_1^2}{S_2^2} < F_{0.05;9;6}\right] = 1 - 0.05$$

$$\therefore P\left[\frac{\sigma_2^2}{\sigma_1^2} < \frac{4.10}{S_1^2/S_2^2}\right] = 0.95.$$

This gives an upper bound for a 95% one-sided confidence interval:

$$\frac{4.10}{S_1^2/S_2^2} = \frac{4.10}{16/3} = 0.77$$

A 95% one-sided confidence interval for $\frac{\sigma_2^2}{\sigma_1^2}$ is therefore $[0;\ 0.77)$ which shows that 0.64 is just inside the interval. We notice that 1 is not inside the interval, in other words the two sample variances are significantly different ($H_0 : \sigma_2^2/\sigma_1^2 = 1$ is rejected) at the 5% level.

6.3 Paired observations

Example 6.6

A factory produces shafts which must have very high precision. An engineer has designed a device which, he claims, can reduce the variability of the product. (Although this device also reduces the mean diameter, this does not matter because the machine which produces the shafts can be readjusted to produce slightly thicker shafts.) To test the idea, six shafts were produced and measured, and then passed through the device. The results are as follows:

Shaft no	1	2	3	4	5	6
Diameter (mm) before treatment	99.78	100.02	99.90	99.86	99.99	99.85
Diameter (mm) after treatment	99.14	99.02	99.10	99.10	99.11	99.13

Do these observations indicate that the variance was decreased by the treatment?

We may again postulate the model $X_{ij} \sim n\left(\mu_i;\ \sigma_i^2\right), \quad j = 1, ..., n_i; \quad i = 1; 2$ and assert that

$$\sum_{j=1}^{n_i} \left(X_{ij} - \overline{X}_i\right)^2 / \sigma_i^2 \sim \chi^2_{n_i - 1}; \quad i = 1; 2$$

but unfortunately these two chi-square variates are *not independent* and their ratio does not give rise to an F-distribution. The distribution of their ratio depends on the *joint* distribution of each pair $(X_{1j},\ X_{2j})$. If we are prepared to assume that each pair is a random observation from a *bivariate normal distribution* (cf unit 1), there is a method for dealing with the problem.

NB The assumption of bivariate normality is not a trivial assumption.

The method of analysing the variances of the joint distribution is developed in theorems 6.2 and 6.3.

Theorem 6.2

Let X_1 and X_2 have a bivariate normal distribution. Then $Var(X_1) = Var(X_2)$ if and only if $(X_1 - X_2)$ and $(X_1 + X_2)$ are uncorrelated.

Proof

Let $E(X_1) = \mu_1;\quad E(X_2) = \mu_2;\ Var(X_1) = \sigma_1^2;$
$Var(X_2) = \sigma_2^2;\quad Cov(X_1;\ X_2) = \rho\sigma_1\sigma_2.$

Then
$$\begin{aligned} Cov(X_1 - X_2;\ X_1 + X_2) &= E[(X_1 - X_2) - (\mu_1 - \mu_2)][(X_1 + X_2) - (\mu_1 + \mu_2)] \\ &= E[(X_1 - \mu_2) - (X_2 - \mu_2)][(X_1 - \mu_1) + (X_2 - \mu_2)] \\ &= E\left[(X_1 - \mu_1)^2 - (X_2 - \mu_2)^2\right] \\ &= Var(X_1) - Var(X_2) = \sigma_1^2 - \sigma_2^2. \end{aligned}$$

It follows that $Cov(X_1 - X_2;\ X_1 + X_2) = 0$ (ie $(X_1 - X_2)$ and $(X_1 + X_2)$ are uncorrelated) if and only if $\sigma_1^2 = \sigma_2^2$.

Please note: If X_1 and X_2 have a bivariate normal distribution, and we define $Y_1 = X_1 - X_2$ and $Y_2 = X_1 + X_2$, then Y_1 and Y_2 have a bivariate distribution as well. For the newly created bivariate distribution we will have

$$E(Y_1) = E(X_1) - E(X_2);\quad E(Y_2) = E(X_1) + E(X_2),$$

variances

$$Var(Y_1) = Var(X_1) + Var(X_2) - 2Cov(X_1;\ X_2)$$
$$Var(Y_2) = Var(X_1) + Var(X_2) + 2Cov(X_1;\ X_2)$$

and covariance

$$Cov(Y_1;\ Y_2) = Var(X_1) - Var(X_2).$$

Now consider a random sample $(X_{1j};\ X_{2j}),\quad j = 1, ..., n$ from the bivariate normal distribution of X_1 and X_2. Let us define the following:

$$\overline{X}_1 = \frac{1}{n}\Sigma X_{1j};\quad \overline{X}_2 = \frac{1}{n}\Sigma X_{2j};\quad U_{11} = \Sigma\left(X_{1j} - \overline{X}_1\right)^2;$$

$$U_{22} = \Sigma \left(X_{2j} - \overline{X}_2\right)^2; \quad U_{12} = \Sigma \left(X_{1j} - \overline{X}_1\right)\left(X_{2j} - \overline{X}_2\right).$$

Also let

$$Y_{1j} = X_{1j} - X_{2j}; \quad Y_{2j} = X_{1j} + X_{2j}, \quad j = 1, ..., n.$$

Then

$$\overline{Y}_1 = \overline{X}_1 - \overline{X}_2; \quad \overline{Y}_2 = \overline{X}_1 + \overline{X}_2.$$

Now $(Y_{1j};\ Y_{2j})\,, j = 1, ..., n$ may be regarded as a random sample from a bivariate normal distribution, and we may use theorem 5.4, which states that if R is the sample correlation coefficient of Y_1 and Y_2, then

$$T = \frac{\sqrt{n-2}R}{\sqrt{1-R^2}}$$

has a t_{n-2} distribution, provided Y_1 and Y_2 are uncorrelated.

Theorem 6.3

Let R be the sample correlation coefficient between Y_1 and Y_2 as defined above. Then $\sqrt{n-2}R/\sqrt{1-R^2} = \sqrt{n-2}\,(U_{11}-U_{22})\,/2\sqrt{U_{11}U_{22}-U_{12}^2}$.

Proof

$$R = \Sigma\left(Y_{1j}-\overline{Y}_1\right)\left(Y_{2j}-\overline{Y}_2\right)/\sqrt{\Sigma\left(Y_{1j}-\overline{Y}_1\right)^2\Sigma\left(Y_{2j}-\overline{Y}_2\right)^2}$$

But
$$\begin{aligned}\Sigma\left(Y_{1j}-\overline{Y}_1\right)\left(Y_{2j}-\overline{Y}_2\right) &= \Sigma\left[\left(X_{1j}-\overline{X}_1\right)-\left(X_{2j}-\overline{X}_2\right)\right]\left[\left(X_{1j}-\overline{X}_1\right)+\left(X_{2j}-\overline{X}_2\right)\right]\\ &= \Sigma\left(X_{1j}-\overline{X}_1\right)^2-\Sigma\left(X_{2j}-\overline{X}_2\right)\\ &= U_{11}-U_{22}.\end{aligned}$$

$$\begin{aligned}\Sigma\left(Y_{1j}-\overline{Y}_1\right)^2 &= \Sigma\left[\left(X_{1j}-\overline{X}_1\right)-\left(X_{2j}-\overline{X}_2\right)\right]^2\\ &= \Sigma\left(X_{1j}-\overline{X}_1\right)^2-2\Sigma\left(X_{1j}-\overline{X}_1\right)\left(X_{2j}-\overline{X}_2\right)+\Sigma\left(X_{2j}-\overline{X}_2\right)^2\\ &= U_{11}-2U_{12}+U_{22}.\end{aligned}$$

Likewise $\Sigma\left(Y_{2j}-\overline{Y}_2\right)^2 = U_{11}+2U_{12}+U_{22}$

$$\begin{aligned}\therefore \Sigma\left(Y_{1j}-\overline{Y}_1\right)^2\Sigma\left(Y_{2j}-\overline{Y}_2\right)^2 &= (U_{11}-2U_{12}+U_{22})\,(U_{11}+2U_{12}+U_{22})\\ &= (U_{11}+U_{22})^2-4U_{12}^2\end{aligned}$$

$$\therefore R = \frac{U_{11}-U_{22}}{\sqrt{(U_{11}+U_{22})^2-4U_{12}^2}}$$

$\therefore R^2/\left(1-R^2\right) = (U_{11}-U_{22})^2/4\left(U_{11}U_{22}-U_{12}^2\right)$ (after some manipulation)

from which the theorem follows.

Result 6.1

If we apply these two theorems, we see that

$$T = \sqrt{n-2}\frac{U_{11}-U_{22}}{2\sqrt{U_{11}U_{22}-U_{12}^2}}$$

has a t_{n-2} distribution provided $H_0 : \sigma_1^2 = \sigma_2^2$ is true.

This result may be used to perform one or two-sided tests of H_0.

Example 6.6 (continued)

We want to test $H_0 : \sigma_1^2 = \sigma_2^2$ against $H_1 : \sigma_1^2 > \sigma_2^2$ (ie one-sided testing). For the critical value, we choose the 5% level and find $t_{0.05;4} = 2.132$. Reject H_0 if $T \geq 2.132$.

We perform the calculations of the data in tabular form as follows:

X_1	X_2	$X_1 - \overline{X}_1$	$(X_1 - \overline{X}_1)^2$	$X_2 - \overline{X}_2$	$(X_2 - \overline{X}_2)^2$	$(X_1 - \overline{X}_1)(X_2 - \overline{X}_2)$
99.78	99.14	-0.12	0.0144	0.04	0.0016	-0.0048
100.02	99.02	0.12	0.0144	-0.08	0.0064	-0.0096
99.90	99.10	0.00	0.0000	0.00	0.0000	0.0000
99.86	99.10	-0.04	0.0016	0.00	0.0000	0.0000
99.99	99.11	0.09	0.0081	0.01	0.0001	0.0009
99.85	99.13	-0.05	0.0025	0.03	0.0009	-0.0015
599.40	594.60	0.00	0.0410	0.00	0.0090	-0.0150

From this it follows that

$$\overline{X}_1 = \frac{599.40}{6} = 99.9$$

$$\overline{X}_2 = \frac{594.60}{6} = 99.1$$

$$U_{11} = 0.041; \quad U_{12} = -0.015; \quad U_{22} = 0.009$$

$$T = \frac{\sqrt{4}\,(0.041 - 0.009)}{2\sqrt{(0.041)\,(0.009) - (0.015)^2}} = \frac{0.032}{\sqrt{0.000144}} = \frac{0.032}{0.012} = 2.6667.$$

Since $2.6667 > 2.132$ we reject H_0 at the 5% level. We are inclined to agree with the engineer that this device reduces the variance.

6.4 More than two independent samples

We now consider the following model:

Let $X_{ij}, \quad j = 1, ..., n; \quad i = 1, ..., k$ be independent random variables with $X_{ij} \sim n\left(\mu_i;\ \sigma_i^2\right)$. We wish to test the null hypothesis

$H_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2$ against the alternative $H_1 : \sigma_p^2 \neq \sigma_q^2$ for at least one $p \neq q$.

Let $\overline{X}_i$ be the sample mean and S_i^2 the sample variance of the i-th sample $(i = 1, ..., k)$, that is

$$\overline{X}_i = \frac{1}{n}\sum_{j=1}^{n} X_{ij}; \qquad S_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(X_{ij} - \overline{X}_i\right)^2.$$

Thus we have k sample variances $S_1^2, ..., S_k^2$ and we want to test whether they differ significantly at the $100\alpha\%$ level. If we select two of the sample variances at random, say S_p^2 and S_q^2, then we know that S_p^2/S_q^2 will have an $F_{n-1;n-1}$ distribution. However, if we arrange $S_1^2, ..., S_k^2$ from the smallest to the largest, the distribution of the ratio

$$U = \max_i S_i^2 / \min_i S_i^2$$

will not resemble the F-distribution at all. The distribution of U has been studied by statisticians in the past, and critical values are given in table E. Using this table is easy enough. For example, if six sample variances are computed from six independent samples of size 11 each, then each sample variance has 10 degrees of freedom. If the ratio of the largest to the smallest exceeds 6.92 H_0 is rejected at the 5% level.

In order to use table E the sample sizes must be equal. In the case of unequal sample sizes one may use another test known as Bartlett's test, but you will not be required to know that test for examination purposes.

Table E:
Percentage points of the ratio, $S^2_{\max}/S^2_{\min}$
Upper 5% points

v	$k=2$	3	4	5	6
2	39.0	87.5	142	202	266
3	15.4	27.8	39.2	50.7	62.0
4	9.60	15.5	20.6	25.2	29.5
5	7.15	10.8	13.7	16.3	18.7
6	5.82	8.38	10.4	12.1	13.7
7	4.99	6.94	8.44	9.70	10.8
8	4.43	6.00	7.18	8.12	9.03
9	4.03	5.34	6.31	7.11	7.80
10	3.72	4.85	5.67	6.34	6.92
12	3.28	4.16	4.79	5.30	5.72
15	2.86	3.54	4.01	4.37	4.68
20	2.46	2.95	3.29	3.54	3.76
30	2.07	2.40	2.61	2.78	2.91
60	1.67	1.85	1.96	2.04	2.11
∞	1.00	1.00	1.00	1.00	1.00

$k =$ number of samples
$v =$ degrees of freedom for each sample variance

Example 6.7

Four independent samples of size $n = 5$ from assumed $n\left(\mu, \sigma^2\right)$ distributions yield the following results:

Sample 1	16	16	15	14	14
Sample 2	20	17	17	16	15
Sample 3	20	20	19	18	18
Sample 4	22	22	21	21	19

Test $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$ at the 5% level of significance.

Solution

We have to 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 $p \neq q$

$$\begin{array}{lll} \overline{X}_1 = 15 & \sum X_{1i} = 75 & \sum X_{1i}^2 = 1\,129 \\ \overline{X}_2 = 17 & \sum X_{2i} = 85 & \sum X_{2i}^2 = 1\,459 \\ \overline{X}_3 = 19 & \sum X_{3i} = 95 & \sum X_{3i}^2 = 1\,809 \\ \overline{X}_4 = 21 & \sum X_{4i} = 105 & \sum X_{4i}^2 = 2\,211 \\ n = 5 & & \end{array}$$

$$\begin{aligned} S_1^2 &= \frac{1}{n-1}\left(\sum X_{1i}^2 - \frac{\left(\sum X_{1i}\right)^2}{n}\right) \\ &= \frac{1}{5-1}\left(1\,129 - \frac{(75)^2}{5}\right) \\ &= \frac{1}{4}(4) \\ &= 1 \end{aligned} \qquad \begin{aligned} S_2^2 &= \frac{1}{n-1}\left(\sum X_{2i}^2 - \frac{\left(\sum X_{2i}\right)^2}{n}\right) \\ &= \frac{1}{5-1}\left(1\,459 - \frac{(85)^2}{5}\right) \\ &= \frac{1}{4}(14) \\ &= 3.5 \end{aligned}$$

$$\begin{aligned} S_3^2 &= \frac{1}{n-1}\left(\sum X_{3i}^2 - \frac{\left(\sum X_{3i}\right)^2}{n}\right) \\ &= \frac{1}{5-1}\left(1\,809 - \frac{(95)^2}{5}\right) \\ &= \frac{1}{4}(4) \\ &= 1 \end{aligned} \qquad \begin{aligned} S_4^2 &= \frac{1}{n-1}\left(\sum X_{4i}^2 - \frac{\left(\sum X_{4i}\right)^2}{n}\right) \\ &= \frac{1}{5-1}\left(2\,211 - \frac{(105)^2}{5}\right) \\ &= \frac{1}{4}(6) \\ &= 1.5 \end{aligned}$$

The test statistic is

$$\begin{aligned} U &= \frac{\max\limits_i S_i^2}{\min\limits_i S_i^2} \\ &= \frac{3.5}{1} \\ &= 3.5. \end{aligned}$$

The critical value is 20.6. H_0 is rejected if $U > 20.6$.

Since $3.5 < 20.6$, we do not reject H_0 at the 5% level and conclude that the variances of the four populations are equal.

6.5 Computers and testing for homogeneity of variance

Most statistical software packages will automatically include a test for the equality of variances when you request to do a *test for means*. This also happens when you request to do an ANOVA test for means. (Both these "tests for means" will be dealt with in the next study unit.)

In statistical software jargon, the testing of equality of variances is referred to as "testing for homogeneity of variance". Usually these tests are not treated on their own, in other words as separate tests, but are considered to be part of "testing the assumptions" for other tests!

The output below in figure 6.2 shows the output for a test for the difference between two means (which you need not worry about at this stage because it will be dealt with in the next study unit) and you must please take note of the first two lines. The output was produced by using the statistical package SPSS.

The results for the test for the equality of variances is a so-called F-test. It is not computed in the way we computed F in section 6.2 and the definition of Levene's test falls beyond the scope of this module. However, you need to be able to interpret the first two lines of the output.

The computed value of the F-statistic is 0.218 and the p-value associated with this specific value is 0.641. Since p-value$> \alpha \Rightarrow$ we cannot reject H_0 and for this specific data set we may assume that the variances of the two groups are the same.

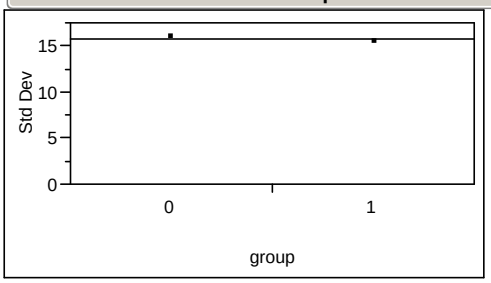
Independent Samples Test

			normal	
			Equal variances assumed	Equal variances not assumed
Levene's Test for Equality of Variances	F		.218	
	Sig.		.641	
t-test for Equality of Means	t		1.145	1.145
	df		198	197.744
	Sig. (2-tailed)		.254	.254
	Mean Difference		2.5419413	2.5419413
	Std. Error Difference		2.2206805	2.2206805
	95% Confidence Interval of the Difference	Lower	-1.8372795	-1.8373144
		Upper	6.9211620	6.9211970

Figure 6.2: SPSS output

If you compare the above with the output in figure 6.3 it shows that JMP provides more than one test that the variances are equal. Using the same data set, JMP also computed Levene's F as 0.2177 with a p-value of 0.6413 but it gives four other tests as well. This output is again obtained as part of the output when we *test for means*. JMP computes the F-test as we defined it in section 6.2 in the study guide. (The last line of the group of F Ratio tests.)

Tests that the Variances are Equal



Level	Count	Std Dev	MeanAbsDif to Mean	MeanAbsDif to Median
0	100	15.98282	12.58769	12.54831
1	100	15.41725	11.94631	11.94125

Test	F Ratio	DFNum	DFDen	p-Value
O'Brien[.5]	0.1312	1	198	0.7176
Brown-Forsythe	0.1921	1	198	0.6617
Levene	0.2177	1	198	0.6413
Bartlett	0.1278	1	.	0.7207
F Test 2-sided	1.0747	99	99	0.7207

Welch Anova testing Means Equal, allowing Std Devs Not Equal

F Ratio	DFNum	DFDen	Prob > F
1.3103	1	197.74	0.2537

t Test
1.1447

Figure 6.3: JMP output

Refer to activity 6.9 to produce output to test $H_0 : \sigma_1^2 = \sigma_2^2$.

Refer to activity 6.14 to produce output to test $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2$.

Exercise 6.1

1. Explain why and under what conditions you will use the following confidence intervals for σ^2:

$$\left[\frac{\Sigma\left(X_i-\overline{X}\right)^2}{\chi^2_{\alpha;n-1}};\quad\infty\right] \quad \text{or} \quad \left[0;\quad\frac{\Sigma\left(X_i-\overline{X}\right)^2}{\chi^2_{1-\alpha;n-1}}\right] \quad \text{or} \quad \left[\frac{\Sigma\left(X_i-\mu\right)^2}{\chi^2_{\frac{1}{2}\alpha;n}};\quad\frac{\Sigma\left(X_i-\mu\right)^2}{\chi^2_{1-\frac{1}{2}\alpha;n}}\right] \quad \text{or}$$

$$\left[\frac{\Sigma\left(X_i-\mu\right)^2}{\chi^2_{\alpha;n}};\quad\infty\right]$$

2. Consider the following sample from a $n\left(\mu;\ \sigma^2\right)$ distribution:

$$6 \quad 10 \quad 14 \quad 12 \quad 4 \quad 11 \quad 15 \quad 8 \quad 7 \quad 10 \quad 13$$

   (a) Test $H_0 : \sigma = 5$ against the alternative $H_1 : \sigma < 5$ at the 5% level assuming

      (i) μ is unknown

      (ii) $\mu = 9$.

   (b) Find a 95% one-sided confidence interval of the form $(0;\ \mu)$ for σ assuming

      (i) μ is unknown

      (ii) $\mu = 9$.

3. A 90% two-sided confidence interval for σ^2 is constructed from a sample of 10 observations from a $n\left(\mu;\ \sigma^2\right)$ distribution. What is the expected length of the interval in the following cases?

   (a) μ is known

   (b) μ is unknown

4. Suppose a 95% confidence interval is to be constructed for the variance of a normal distribution with unknown mean. What is the smallest sample size n which would ensure that the expected length of the confidence interval is at most $2.5\sigma^2$?

5. At a certain factory a product is produced in two identical plants. A modification of the process is suggested for increasing the daily yield. The one plant was then modified and the yields of the two plants were recorded on six consecutive days:

Unmodified:	$24;$	$35;$	$30;$	$28;$	$31;$	32	metric tons
Modified:	$29;$	$35;$	$32;$	$28;$	$36;$	32	metric tons

Treating the data as $n\left(\mu_1;\ \sigma_1^2\right)$ and $n\left(\mu_2;\ \sigma_2^2\right)$ samples respectively, test $H_0 : \sigma_1^2 = \sigma_2^2$ (two-sidedly) at the 10% level. Also find a 90% confidence interval for σ_1^2/σ_2^2.

6. 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: $n_1 = 10 \quad \Sigma X_{1i} = 20 \quad \Sigma X_{1i}^2 = 148$

Sample 2: $n_2 = 12 \quad \Sigma X_{2i} = 36 \quad \Sigma X_{2i}^2 = 152$

(a) Test the claim that the standard deviation of the first population is more than twice the standard deviation of the second population (5% level of significance).

(b) Compute a 95% one-sided confidence interval for σ_1/σ_2.

7. Consider the following 11 observations from a bivariate normal distribution:

X_1	29	37	23	42	14	36	39	25	31	38	16
X_2	27	31	25	34	22	28	37	31	33	34	28

Test $H_0 : \sigma_1^2 = \sigma_2^2$ against $H_1 : \sigma_1^2 > \sigma_2^2$ at the 10% level.

8. A random sample of 10 students were subjected to an arithmetic test, the result of which is denoted by X. The students were then given remedial training and tested again, the result being denoted by Y:

X	25	26	27	29	30	31	32	33	33	34
Y	47	51	49	50	50	53	48	49	52	53

Regard these results as a random sample from a bivariate normal distribution, and test, at the 10% level of significance, whether the students were more uniform after the remedial training than before..3

9. In order to test whether four operators maintain the same uniformity in determining the sodium content of a mixture, each operator was given six samples containing exactly 20% sodium. Their determinations were as follows:

Operator 1	20.0	20.4	19.7	19.5	20.7	20.3
Operator 2	19.4	20.4	19.2	20.2	19.7	20.5
Operator 3	19.0	19.2	20.7	21.4	21.1	19.8
Operator 4	20.1	19.9	20.3	19.5	20.6	19.6

Test at the 5% level whether there is a difference in the variances of the four populations.

10. Three independent random samples of size $n = 10$ from $n\left(\mu_j;\, \sigma^2\right)$ distributions, with $\mu_1 = 5$, $\mu_2 = 7$ and $\mu_3 = 8$ (known) yielded the following statistics:

$$\Sigma X_{1i}^2 = 390 \qquad \Sigma X_{2i}^2 = 730 \qquad \Sigma X_{3i}^2 = 740$$

$$\Sigma X_{1i} = 60 \qquad \Sigma X_{2i} = 80 \qquad \Sigma X_{3i} = 85$$

Test $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2$ at the 5% level of significance. (Be careful with your definition of S_j^2 since μ_j is known!)

6.6 Learning outcomes

After studying study unit 6, you should **be able to**

- perform hypothesis tests concerning the variance of a single sample
- derive one or two-sided confidence intervals for the variance of a single sample
- perform hypothesis tests concerning the equality of the variances of two independent samples
- derive one or two-sided confidence intervals for the ratio $\frac{\sigma_2^2}{\sigma_1^2}$ of two independent samples
- perform hypothesis tests concerning the equality of the variances of paired observations
- perform hypothesis tests concerning the equality of the variances of more than two independent samples
- interpret the computer output of JMP concerning the homogeneity of variance tests

STUDY UNIT 7

Inference on means

7.1 One-sample problem

Let $X_1,\ ...,\ X_n$ be independent random variables with $X_i \sim n\left(\mu;\ \sigma^2\right)$. In previous study units we have seen how one would investigate the two basic assumptions, *independence* and *normality*, and how one would find out more about the variance, σ^2. We now turn our attention to μ.

We already know that $\overline{X} = \frac{1}{n}\Sigma X_i$ is an unbiased estimator for μ, irrespective of whether the underlying distribution is normal or not. The assumption of normality enables us to do more than just estimate μ. The basic result, which you learned in first-year statistics and will have gathered by now is of prime importance in statistical inference, is repeated here.

Theorem 7.1

Let $X_1,\ ...,\ X_n$ be independent $n\left(\mu;\ \sigma^2\right)$ variates and let $\overline{X} = \frac{1}{n}\Sigma X_i;\quad S^2 = \frac{1}{n-1}\Sigma\left(X_i - \overline{X}\right)^2$. Then

(a) $\overline{X}$ is a $n\left(\mu;\ \sigma^2/n\right)$ variate, that is $\sqrt{n}\left(\overline{X} - \mu\right)/\sigma$ is a $n\left(0;\ 1\right)$ variate;

(b) $(n-1)\,S^2/\sigma^2$ is a χ^2_{n-1} variate;

(c) $\overline{X}$ and S^2 are independent;

(d) $T = \sqrt{n}\left(\overline{X} - \mu\right)/S$ is a t_{n-1} variate.

This theorem is used in various ways to test one or two-sided hypotheses about μ or to find one or two-sided confidence intervals for μ. If σ^2 is known (*a very rare occurrence in practice*) we use (a). If σ^2 is unknown, we use (d).

In result (a) we have $Z \sim n(0;\ 1)$ which implies that we use table II (Stoker) to obtain the critical value and in result (d) we have $T \sim t_{n-1}$ which implies that we use table III (Stoker) to obtain the critical value.

To test $H_0 : \mu = \mu_0$ against

$H_1 : \mu > \mu_0$ or

$H_1 : \mu < \mu_0$ or

$H_1 : \mu \neq \mu_0$

we summarise the application of theorem 7.1 in the following flow chart (which is a revision of first-year statistics!)

(a) σ^2 is known ⋮ σ^2 is unknown (b)

⋮

We use as a test statistic

↙ ↘

$Z = \dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$ ⋮ $T = \dfrac{\overline{X} - \mu_0}{S/\sqrt{n}}$

⋮

We use the critical value

↙ ↘

		⋮	
One-sided:	z_α	⋮	$t_{\alpha;n-1}$
Two-sided:	$z_{\frac{a}{2}}$	⋮	$t_{\alpha/2;n-1}$

⋮

A $(1-\alpha)\,100\%$ confidence interval for μ

↙ ↘

$\overline{X} - z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$ ⋮ $\overline{X} - t_{\alpha/2;n-1}\dfrac{S}{\sqrt{n}} \leq \mu \leq \overline{X} + t_{\alpha/2;n-1}\dfrac{S}{\sqrt{n}}$

For a lower $(1-\alpha)\,100\%$ one-sided confidence interval, the probability statement $P(Z \leq z_\alpha) = 1-\alpha$ is reorganised to obtain

$$\mu \geq \overline{X} - z_\alpha \frac{\sigma}{\sqrt{n}} \qquad \left(\text{ie the interval } \left(\overline{X} - z_\alpha \frac{\sigma}{\sqrt{n}};\ \infty\right)\right).$$

(This confidence interval may be used to test the alternative $H_1 : \mu > \mu_0$.)

For an upper $(1-\alpha)\,100\%$ one sided confidence interval, the probability statement $P(Z \geq -z_\alpha) = 1-\alpha$ is reorganised to obtain

$$\mu \leq \overline{X} + z_\alpha \frac{\sigma}{\sqrt{n}} \qquad \left(\text{ie the interval } \left(-\infty;\ \overline{X} + z_\alpha \frac{\sigma}{\sqrt{n}}\right)\right).$$

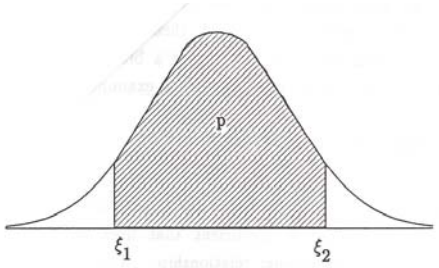
(This confidence interval may be used to test the alternative $H_1 : \mu < \mu_0$.)

Tolerance intervals

If $X_1,\ X_2,\ ...,\ X_n$ is a random sample from a distribution, say a $n\left(\mu;\ \sigma^2\right)$ distribution, then we have already seen that a confidence interval for μ is given by

$$\overline{X} - t_{\frac{1}{2}\alpha;n-1}\frac{S}{\sqrt{n}} \leq \mu \leq \overline{X} + t_{\frac{1}{2}\alpha;n-1}\frac{S}{\sqrt{n}}.$$

As $n \to \infty$ the width of this interval tends towards zero. Suppose, for example, the random variable X represents the breaking strength of a beam selected at random from a population of beams to be used in constructing house roofs. If a random sample of these beams is selected to construct the roof of my house, and if the roof were to cave in later, it would be small consolation to me knowing that the mean μ of all the beams conformed to tight specifications. A tolerance interval would be more appropriate. Define two percentiles ξ_1 and ξ_2 such that $P\left(\xi_1 < X < \xi_2\right) = p$.



Then a *tolerance* interval is of the form $\left(\overline{X} - K_\alpha S;\quad \overline{X} + K_\alpha S\right)$ where K_α is read from a table, and where

$$P\left(\overline{X} - K_\alpha S \leq \xi_1 \leq \xi_2 \leq \overline{X} + K_\alpha S\right) = 1 - \alpha.$$

For example if $p = 0.9$ and $\alpha = 0.05$ then we would be 95% sure that at least 90% of all the individuals in the population lie between $\overline{X} - K_\alpha S$ and $\overline{X} + K_\alpha S$.

Tolerance intervals are generally wider than confidence intervals for the mean, and as $n \to \infty$

$$\overline{X} - K_\alpha S \to \xi_1$$

$$\overline{X} + K_\alpha S \to \xi_2.$$

We would not expect you to compute a tolerance interval manually but only electronically using JMP. Please see activity 7.5 in the workbook.

7.2 The power of the test and the noncentral t-distribution

Something that was not discussed in detail in your first-year modules is the **power of the test**.

In definition 2.7 of section 2.5 of study unit 2 we defined the power of the test as the probability that H_0 is rejected when H_1 is true. We actually defined the power as $1-\beta$ where β is the probability of a type II error.

How will we compute the power for situation (a) of theorem 7.1?

We know that $Z_0 = \dfrac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}$ is a $n\,(0;\ 1)$ variate provided that $H_0 : \mu = \mu_0$ is true $\left(\Rightarrow \overline{X} \sim n\left(\mu_0;\ \dfrac{\sigma^2}{n}\right)\right)$.

What is the distribution of Z_0 if H_0 is not true?

Suppose $H_1 : \mu = \mu_1$ is true. Then $Z_1 = \dfrac{\overline{X}-\mu_1}{\sigma/\sqrt{n}} \sim n\,(0;\ 1) \Rightarrow \overline{X} \sim n\left(\mu_1;\ \dfrac{\sigma^2}{n}\right)$.

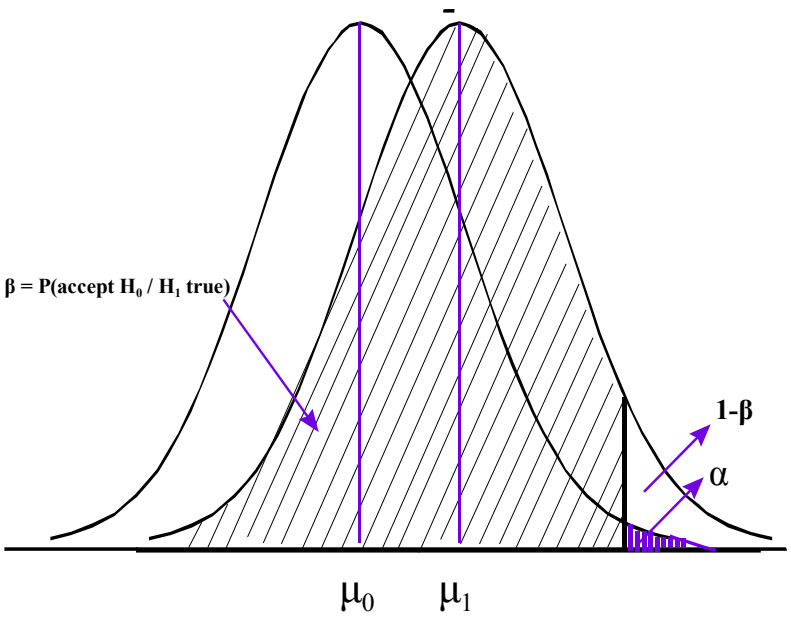


Figure 7.1: Illustration of α and β for right-sided testing

It is a laborious process, but β can be computed for different values of μ_1 where $\mu_1 > \mu_0$. The closer μ_1 lies to μ_0, the bigger a type II error becomes, and the further μ_1 moves to the right the smaller β becomes. (See activity 7.1 of the workbook.)

The authors of the textbook say that "statisticians are often unaware that they use certain words in a completely different way than other professionals" [p. 102]. They give a list of definitions for model; parameters; hypotheses, et cetera and you can read at the bottom of page 103 how they define "**Power, β level**" in general.

What is now very ironic and confusing, is that Sall, Creighton and Lehman use exactly the opposite symbols than we do! They define $1-\beta=$ probability of type II error and thus $\beta=$ power of the test. If you click on the help function of JMP, you will see that JMP uses the same symbols as our study guide.

READ THROUGH

Sall, Creighton and Lehman, **Chapter 7** **Univariate distributions: one variable, one sample**

Pages 138 - 139 *Testing hypotheses: Terminology*

This does not matter! As long as we define the concept "type II error" and "the complement of a type II error" the same! What they define as "power" is exactly the same as what we define as power.

How will we compute the power for situation (b) of theorem 7.1?

We know that $T_0=\dfrac{\overline{X}-\mu_0}{S/\sqrt{n}}=\sqrt{n}\left(\overline{X}-\mu_0\right)/S$ is a t_{n-1} variate provided $H_0:\mu=\mu_0$ is true. The t-distribution is symmetric about zero and has about the same shape as the normal distribution, except that it is more peaked and has more probability in the tails. What is the distribution of T_0 if H_0 is not true?

The noncentral t-distribution

If we know that $\mu\neq\mu_0$ then $T_0=\dfrac{\sqrt{n}\left(\overline{X}-\mu_0\right)}{S}$ has a so-called **non-central t-distribution** with *noncentrality parameter* $\delta=\sqrt{n\left(\mu-\mu_0\right)}/\sigma.$

It is not necessary for our purposes to derive an expression for the pdf of the distribution. It is sufficient to know that the distribution is not symmetric and lies more to the right of zero if $\delta>0$ and more to the left of zero if $\delta<0.$

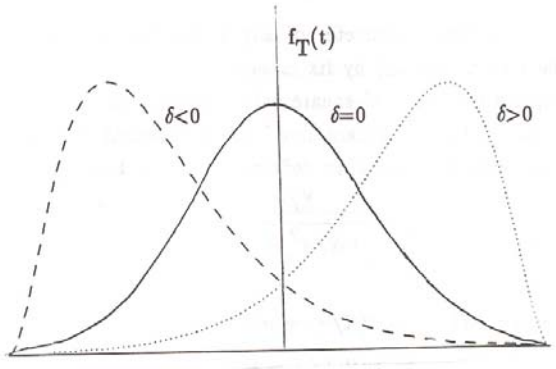


Figure 7.2

This is a situation where a computer can be a marvellous educational tool!

READ THROUGH

Sall, Creighton and Lehman, **Chapter 7** **Univariate distributions: one variable, one sample**

Pages 148-149 *Power of the t-test*

See activity 7.2 of the workbook on how to do a "Power animation" with JMP.

Formally, we may define a noncentral variate $t_{f;\delta}$ as follows:

Definition 7.1

Let X and W be independent with $X \sim n\left(\mu;\ \sigma^2\right)$ and $\frac{W}{\sigma^2} \sim \chi^2_f$. Then

$$T = X/\sqrt{W/f} = \frac{X/\sigma}{\sqrt{\left(W/\sigma^2\right)/f}}$$

is a noncentral t-variate with f degrees of freedom and noncentrality parameter $\delta = \frac{\mu}{\sigma}$.

To find the noncentrality parameter of any t-statistic, we replace the numerator (the normal variate) by its expected value and the denominator, which is the square root of a chi-square variate divided by its degrees of freedom, by the square root of the expected value of the square of the denominator. Thus, in definition 7.1 we have

$$T = \frac{X/\sigma}{\sqrt{\left(W/\sigma^2\right)/f}}.$$

Now

$$X/\sigma \sim n\left(\mu/\sigma; 1\right) \qquad \therefore E\left(X/\sigma\right) = \mu/\sigma.$$

$$W/\sigma^2 \sim \chi^2_f \qquad \therefore E\left(W/\sigma^2\right) = f$$

$$\therefore E\left(W/\sigma^2\right)/f = 1$$

$$\therefore \delta = \frac{\mu/\sigma}{\sqrt{1}} = \frac{\mu}{\sigma}.$$

In the expression

$$T_0 = \frac{\sqrt{n}\left(\overline{X} - \mu_0\right)}{S} = \frac{\sqrt{n}\overline{X} - \sqrt{n}\mu_0}{\sqrt{S^2}} \qquad \begin{array}{l} \rightarrow \text{numerator} \\ \rightarrow \text{denominator} \end{array}$$

we replace "numerator" by "E(numerator)" $= \sqrt{n}E\left(\overline{X}\right) - \sqrt{n\mu_0}$

$$= \sqrt{n}\mu - \sqrt{n\mu_0}$$

$$= \sqrt{n}\left(\mu - \mu_0\right)$$

(because we know that $E\left(\overline{X}\right) = \mu$). Similarly $E\,(\text{squared denominator}) = E\left(S^2\right) = \sigma^2$.

So that $\delta = \dfrac{E\,(\text{numerator})}{\sqrt{E\,(\text{squared denominator})}}$

$$= \frac{\sqrt{n}\left(\mu - \mu_0\right)}{\sqrt{\sigma^2}}$$

$$= \frac{\sqrt{n}\left(\mu - \mu_0\right)}{\sigma}.$$

It is important to note that δ is a function of three different quantities: $\sqrt{n}$; difference $(\mu - \mu_0)$ and σ.

The figure below illustrates the connection between α, β and δ, and it is apparent from the figure that β will *decrease* as δ *increases*.

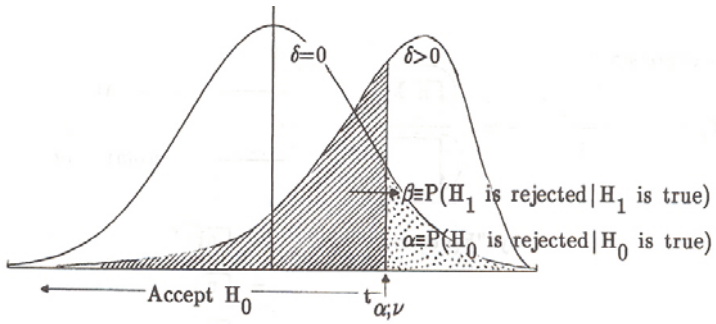


Figure 7.3

Since the power of the test $= 1 - \beta$, the power will increase as δ increases.

> Table F contains the power of the two-sided t-test. In order to use the table one has to compute $\phi = \delta/\sqrt{2}$; v represents the degrees of freedom as usual. The table gives $100 \times (\text{power})$ to the nearest integer.

Example 7.1

It is desired to test $H_0 : \mu = 20$ against $H_1 : \mu \neq 20$ using a sample of size $n = 8$ from a $n\left(\mu;\ \sigma^2\right)$ distribution. What will the power of the test be if $\mu = 20 + 1.5\sigma$ (ie if the true mean is $1\frac{1}{2}$ standard deviations away from the hypothesised value)?

Solution

We know that H_0 is tested against H_1 with the test statistic $T_0 = \dfrac{\sqrt{n}\left(\overline{X} - \mu_0\right)}{S} \sim t_{n-1}$

where $\delta = \dfrac{\sqrt{n}\left(\mu - \mu_0\right)}{\sigma}$.

We have $v = n - 1 = 7$; $\quad \delta = \dfrac{\sqrt{8}\left[(20 + 1.5\sigma) - 20\right]}{\sigma}$

$$= \sqrt{8}\left(1.5\right)$$

and $\phi = \dfrac{1}{\sqrt{2}}\delta = \sqrt{\dfrac{8}{2}} \times 1.5 = 3.$

From table F we read off the power, namely (approximately) 0.95 if $\alpha = 0.05$ (or 0.75 if $\alpha = 0.01$).

Table F:
$100 \times$ (power) of the two-sided t-test with level α

ϕ	6	7	8	9	10	12	15	20	30	60	∞	$v =$ degrees of freedom
1.2	30	31	32	33	34	35	36	37	38	39	40	
1.3	35	36	37	38	39	40	41	42	43	44	45	
1.4	39	40	41	42	43	45	46	47	49	50	51	
1.5	43	45	46	47	48	50	51	52	54	55	56	
1.6	48	50	52	53	54	55	57	58	59	61	62	
1.7	52	55	57	58	59	60	62	64	65	66	67	
1.8	57	60	62	63	64	65	67	69	70	71	72	
1.9	62	64	65	67	68	69	71	73	74	76	77	
2.0	66	68	70	71	72	74	75	77	78	80	81	
2.1	70	72	74	75	77	78	79	81	82	83	85	
2.2	74	76	78	79	80	81	83	84	86	87	88	
2.3	77	80	81	83	84	85	86	87	88	89	90	
2.4	81	83	85	86	87	88	89	90	91	92	93	$\alpha = 0.05$
2.5	84	86	87	88	89	90	91	92	93	94	94	
2.6	86	88	90	91	91	92	93	94	95	95	96	
2.7	89	90	92	93	93	94	95	95	96	96	97	
2.8	91	92	93	94	95	95	96	96	97	97	98	
2.9	92	94	95	95	96	96	97	97	98	98	98	
3.0	94	95	96	96	97	97	98	98	98	99	99	
3.1	95	96	97	97	98	98	98	99	99	·	·	
3.2	96	97	98	98	98	99	99	·	·	·	·	
3.3	97	98	98	99	99	·	·	·	·	·	·	
3.4	98	98	99	·	·	·	·	·	·	·	·	
3.5	98	99	·	·	·	·	·	·	·	·	·	
2.0	31	33	37	40	42	45	48	50	54	57	60	
2.2	39	42	46	49	51	54	58	61	64	67	70	
2.4	47	51	55	58	60	63	67	70	74	77	80	
2.6	55	60	63	67	69	72	76	79	82	85	87	
2.8	62	68	71	74	77	80	83	86	88	90	92	
3.0	69	75	78	81	83	86	89	91	92	94	95	
3.2	75	81	84	87	88	90	93	94	96	97	97	
3.4	81	86	88	91	92	94	95	97	98	98	99	
3.6	86	90	92	94	95	96	97	98	99	99	·	$\alpha = 0.01$
3.8	90	93	95	96	97	98	99	99	·	·	·	
4.0	93	95	97	98	98	99	·	·	·	·	·	
4.2	95	97	98	99	99	·	·	·	·	·	·	
4.4	96	98	99	·	·	·	·	·	·	·	·	
4.6	97	99	·	·	·	·	·	·	·	·	·	
4.8	98	·	·	·	·	·	·	·	·	·	·	
5.0	99	·	·	·	·	·	·	·	·	·	·	

Notes on the use of table F

(a) If a two-sided test is performed, the power does not depend on the sign of δ. Due to the symmetry of the problem, $H_0 : \mu = \mu_0$ is equally likely to be rejected if $\mu = \mu_0 + k\sigma$ or $\mu = \mu_0 - k\sigma$; in the one case $\delta = \sqrt{n}k$ and in the other $\delta = -\sqrt{n}k$. Thus, when dealing with a two-sided test, the definition of ϕ should actually be $\phi = \frac{|\delta|}{\sqrt{2}}$.

(b) Table F enables one to decide on the sample size required to ensure a chosen power (eg $1 - \beta = 0.99$) when μ is a specified multiple of σ away from μ_0. This is done by reading off the power of a number of sample sizes, and selecting the smallest n such that $1 - \beta \geq 0.99$.

(c) δ (or ϕ) contains two unknown parameters: $\mu - \mu_0$ and σ. If the problem is stated as in (a) and (b), this does not complicate the problem, since δ is actually a function of $\frac{(\mu - \mu_0)}{\sigma}$. Sometimes a small pilot sample is drawn to estimate μ and σ and these estimated values are used to estimate δ. This estimate is subject to a random variation, but does give a rough idea of the sample size required. Sample size tables exist which make it unnecessary to compute ϕ for a number of sample sizes and find the sample size by trial and error, but since such tables are not included in our book of tables they will not be dealt with here.

(d) From table F it is obvious that the power of the t-test increases as ϕ increases. The definition of ϕ (as amended) is $\phi = \sqrt{\frac{n}{2}}\frac{|\mu - \mu_0|}{\sigma}$ and it is clear that ϕ, and thus the power, increases as
   (i) n increases
   (ii) $|\mu - \mu_0|$ increases
   (iii) σ decreases

7.3 Two-sample problem; independent samples

We now consider the following problem: we have two independent random samples of sizes n_1 and n_2 respectively, and we want to test whether the population means are equal. We use the notation $(X_{11},\ ...,\ X_{1n_1})$ and $(X_{21},\ ...,\ X_{2n_2})$ for the two samples. A model which is generally used for this problem is the following:

Assume that $X_{ij};\quad j = 1, ..., n_i;\quad i = 1, 2$ are independent random variables with $X_{ij} \sim n\left(\mu_i;\ \sigma^2\right)$.

We wish to test $H_0 : \mu_1 = \mu_2$ or find a confidence interval for $\mu_1 - \mu_2$. Note the assumptions:

(a) Not only are the observations in each sample independent, but the two samples are mutually independent.

(b) The observations are normally distributed.

(c) The two population variances are equal, that is the variance of X_{ij} does not depend on i. This is rather important. If we think the two variances are equal and they are unequal, it could have a serious effect on the significance level and the power of the test or on the confidence level of the confidence interval.

Luckily we already know how to verify (or at least investigate) these assumptions!

In order to make probability statements about $\mu_1 - \mu_2$, we simply use the results already known:

Let

$$\overline{X}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}; \qquad S_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}\left(X_{ij} - \overline{X}_i\right)^2; \qquad i = 1; 2$$

Then:

(a) $\overline{X}_1,\ \overline{X}_2,\ S_1^2$ and S_2^2 are independent;

(b) $\overline{X}_1 \sim n\left(\mu_1;\ \frac{\sigma^2}{n_1}\right); \qquad \overline{X}_2 \sim n\left(\mu_2;\ \frac{\sigma^2}{n_2}\right)$

$$\therefore \overline{X}_1 - \overline{X}_2 \sim n\left(\mu_1 - \mu_2;\ \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}\right)$$

(*Question*: Would this be true if $\overline{X}_1$ and $\overline{X}_2$ were not independent?)

so that $U = \dfrac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{\sigma\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \sim n\,(0;\ 1)\,;$

(c) $\dfrac{(n_1 - 1)\,S_1^2}{\sigma^2} \sim \chi^2_{n_1 - 1}$ and $\dfrac{(n_2 - 1)\,S_2^2}{\sigma^2} \sim \chi^2_{n_2 - 1}$ (see result 1.3)

so that $W = \dfrac{\left[(n_1 - 1)\,S_1^2 + (n_2 - 1)\,S_2^2\right]}{\sigma^2} \sim \chi^2_{n_1 + n_2 - 2}.$

(*Question*: Would this be true if S_1^2 and S_2^2 were not independent?)

From (a), (b) and (c) and using the notation defined above, we rewrite theorem 1.4 as follows:

Theorem 7.2

$$T = \frac{U}{\sqrt{\frac{W}{(n_1+n_2-2)}}} \sim t_{n_1+n_2-2}$$

where $$T = \frac{\left[\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)\right] / \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}{\sqrt{\left[(n_1-1)\,S_1^2 + (n_2-1)\,S_2^2\right] / (n_1+n_2-2)}}$$

$$= \frac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

and $$S_p^2 = \frac{\left[(n_1-1)\,S_1^2 + (n_2-1)\,S_2^2\right]}{(n_1+n_2-2)}$$

$$= \frac{\left[\sum_{j=1}^{n_1}\left(X_{1j} - \overline{X}_1\right)^2 + \sum_{j=1}^{n_2}\left(X_{2j} - \overline{X}_2\right)^2\right]}{(n_1+n_2-2)}$$

S_p^2 is the (unbiased) estimator of σ^2 and is called a "pooled" (hence the subscript "p") estimator, since we pool the sums of squares of deviations from the sample means of the two samples.

This is the well-known t-statistic you used in first-year modules to test for the difference between two means.

This t-statistic is used in the usual way to test

$H_0 : \mu_1 - \mu_2 = 0$ (or $\mu_1 - \mu_2 = c$ for that matter) against

$H_1 : \mu_1 - \mu_2 \neq 0$ (ie $\mu_1 \neq \mu_2$) or against

$H_1 : \mu_1 - \mu_2 < 0$ (ie $\mu_1 < \mu_2$) or against

$H_1 : \mu_1 - \mu_2 > 0$ (ie $\mu_1 > \mu_2$).

We simply replace $(\mu_1 - \mu_2)$ by 0 (or c) and compare T with $t_{\frac{\alpha}{2};n_1+n_2-2}$ for two-sided testing or with $t_{\alpha;n_1+n_2-2}$ for one-sided testing.

A two-sided confidence interval for $\mu_1 - \mu_2$ is given by the probability statement

$$P\left(\overline{X}_1 - \overline{X}_2 - t_{\frac{\alpha}{2};n_1+n_2-2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq \mu_1 - \mu_2 \leq \overline{X}_1 - \overline{X}_2 + t_{\frac{\alpha}{2};n_1+n_2-2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right) = 1 - \alpha.$$

Are you able to derive the confidence limits?

What is the power of the test? If $H_0 : \mu_1 - \mu_2 = 0$ is not true, the distribution of T is noncentral t with noncentrality parameter

$$\delta = \frac{\mu_1 - \mu_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

We may use table F as before to compute the power.

Example 7.2

Suppose we have two independent samples of size 16 each, and we wish to test $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$. What will the power of the test be if $\mu_1 - \mu_2 = 1.5\sigma$?

Solution

We know that H_0 is tested against H_1 using the test statistic T where $\delta = \dfrac{\mu_1 - \mu_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$.

We have $v = n_1 + n_2 - 2 = 30$; $\quad \delta = \dfrac{1.5\sigma}{\sigma\sqrt{\frac{1}{16} + \frac{1}{16}}} = 1.5\sqrt{8}$ so that $\phi = \dfrac{1}{\sqrt{2}}|\delta| = 3$.

From table F we see that the power will be 0.98 at the 5% level and 0.92 at the 1% level.

7.4 Paired observations

In certain problems, as also illustrated in study unit 6, we do not have two independent samples, but rather n pairs of observations $(X_{1i};\ X_{2i})\,,\ i=1,...,n$ such that

$$E\,(X_{1i})=\mu_1,\quad E\,(X_{2i})=\mu_2;\quad Var\,(X_{1i})=\sigma_1^2;\quad Var\,(X_{2i})=\sigma_2^2;\quad Cov\,(X_{1i};\ X_{2i})=\rho\sigma_1\sigma_2.$$

The problem is to test $H_0:\mu_1=\mu_2$.

In this case S_1^2 and S_2^2 (as defined in the previous section) are not independent, and it is not possible to construct a statistic with a t-distribution in a similar manner to that of the previous section. There is a simple solution, however.

Let

$$Y_i=X_{1i}-X_{2i},\quad i=1,...,n.$$

Then $Y_1,\ ...,\ Y_n$ form a random sample such that $Y_i\sim n\left(\mu_1-\mu_2;\ \sigma_1^2+\sigma_2^2-2\rho\sigma_1\sigma_2\right)$.

Since we are not interested in $\sigma_1^2,\ \sigma_2^2$ and ρ, we set $\sigma_1^2+\sigma_2^2-2\rho\sigma_1\sigma_2=\sigma^2$ and also let $\mu=\mu_1-\mu_2$.

> We consider $Y_1,\ ...,\ Y_n$ to be a random sample from a $n\left(\mu;\ \sigma^2\right)$ distribution and we wish to test $H_0:\mu=0$. This is exactly the one-sample problem dealt with in section 7.1.

This means we "transform" the paired observations to a one-sample problem by means of subtraction.

Example 7.3

Twelve people are randomly chosen and their pulse rate measured before and after being given a specific dosage of a new drug. Do the results confirm a researcher's theory that the drug quickens heartbeat?

Patient	1	2	3	4	5	6	7	8	9	10	11	12
Pulse rate before	90	70	68	68	75	80	75	74	70	88	65	64
Pulse rate after	80	59	80	77	87	70	82	62	61	79	58	75

Solution

Let $Y_i =$ pulse rate before $-$ pulse rate after, for $i = 1,\ 2,\ ...,\ 12.$

Patient (i)	1	2	3	4	5	6	7	8	9	10	11	12
Y_i	10	11	-12	-9	-12	10	-7	12	9	9	7	-11

We want to test $H_0 : \mu = 0$ against $H_1 : \mu < 0.$ (If the drug increases heartbeat, the difference of before minus after will be negative.)

$$T = \frac{\sqrt{n}\left(\overline{Y} - \mu\right)}{S_Y} \sim t_{n-1}$$

where

$$\overline{Y} = \sum_{i=1}^{12} \frac{Y_i}{12} = \frac{17}{12} = 1.4167;$$

$$\begin{aligned} S_Y^2 &= \frac{1}{11} \sum_{i=1}^{12} \left(Y_i - \overline{Y}\right)^2 \\ &= \frac{1}{11}\left(\Sigma Y_i^2 - \frac{\left(\Sigma Y_i\right)^2}{12}\right) \\ &= \frac{1}{11}\left(1\,215 - \frac{(17)^2}{12}\right) \\ &= \frac{1\,215 - 24.083333333}{11} \\ &= 108.2652 \end{aligned}$$

$$S = 10.4051$$

$$\therefore T = \frac{\sqrt{12}\,(1.4167 - 0)}{10.4051} \approx 0.4717$$

From table III we find $t_{0.05;11} = 1.796.$ Reject H_0 if $T > 1.796.$ Since $-1.796 < 0.4717,$ we cannot reject H_0 at the 5% level of significance. The drug does not increase heartbeat.

JMP offers a special platform for the analysis of paired data called "Matched Pairs".

READ THROUGH

Sall, Creighton and Lehman, **Chapter 8** **The difference between two means**

Pages 186-193 *Testing means for matched pairs*

7.5 Independent samples with unequal variances

As was said before, **if the two population variances are unequal** and we nevertheless proceed as if they are equal, the significance level and power will be affected. Suppose $X_{1j},\ j = 1, ..., n_1$ and $X_{2j},\ j = 1, ..., n_2$ are two independent random samples such that

$$X_{ij} \sim n\left(\mu_i,\ \sigma_i^2\right) \qquad j = 1,\ ...,\ n_i; \qquad i = 1,\ 2.$$

Then we could have constructed a t-statistic using the fact that $\overline{X}_1, \overline{X}_2,\ S_1^2$ and S_2^2 (defined before) are independent with $\overline{X}_i \sim n\left(\mu_i;\ \frac{\sigma_i^2}{n_i}\right)$ and $\frac{(n_i - 1)\, S_i^2}{\sigma_i^2} \sim \chi_{n_i - 1}, \quad i = 1; 2,$ but unfortunately the σ_i^2 will not "cancel out" as σ^2 did in section 7.3. The result is that the t-statistic will contain unknown parameters (except in the unlikely event that $\frac{\sigma_1^2}{\sigma_2^2}$ is known). This problem is known as the Behrens-Fisher problem, named after the two people who studied it in the previous century. A completely satisfactory solution does not exist, but certain practical solutions have evolved.

The Welch solution is as follows:

We want to test $H_0: \ \mu_1 - \mu_2 = c$ (with c specified) against
$H_1: \ \mu_1 - \mu_2 \neq c$ or
$H_1: \ \mu_1 - \mu_2 < c$ or
$H_1: \ \mu_1 - \mu_2 > c.$

We use the statistic

$$T = \frac{\overline{X}_1 - \overline{X}_2 - c}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

where

$$\overline{X}_i = \frac{1}{n_i} \sum_j X_{ij} \qquad i = 1 \text{ or } 2$$

$$S_i^2 = \frac{1}{n_i - 1} \sum_j \left(X_{ij} - \overline{X}_i\right)^2 \qquad i = 1 \text{ or } 2.$$

Under H_0, this statistic has an approximate Student's t-distribution for large samples. However, the degrees of freedom are not $n_1 + n_2 - 2$ as was the case in section 7.3 but the approximate degrees of freedom are

$$v = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{S_1^4}{n_1^2\,(n_1 - 1)} + \frac{S_2^4}{n_2^2\,(n_2 - 1)}}.$$

Since v is usually not an integer, one would have to interpolate in table III (Stoker). If $v = f + p$ where f is an integer and $0 \leq p < 1,$ then

$$t_{\alpha;v} \approx (1-p)\, t_{\alpha;f} + p t_{\alpha;f+1}.$$

Example 7.4

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:

$$n_1 = 11 \qquad \Sigma X_{1i} = 330 \qquad \Sigma X_{1i}^2 = 9\,950$$

$$n_2 = 16 \qquad \Sigma X_{2i} = 560 \qquad \Sigma X_{2i}^2 = 19\,720$$

Test $H_0 : \mu_1 = \mu_2 - 3$ against
$H_1 : \mu_1 < \mu_2 - 3$ at the 2.5% level of significance.

Solution

We compute

$$\overline{X}_1 = 30; \qquad\qquad \overline{X}_2 = 35;$$

$$S_1^2 = \frac{1}{10}\left[9\,950 - (11)\,(30)^2\right] = 5 \qquad S_2^2 = \frac{1}{15}\left[19\,720 - (16)\,(35)^2\right] = 8.$$

The null hypothesis implies that $\mu_1 - \mu_2 = -3 \Rightarrow c = -3$

$$\begin{aligned} \therefore T &= \frac{(30-35)-(-3)}{\sqrt{\dfrac{5}{11}+\dfrac{8}{16}}} \\ &= \frac{-2}{0.977000842} \\ &\approx -2.0471 \end{aligned}$$

We compute the approximate degrees of freedom as

$$\begin{aligned} v &= \frac{\left[\dfrac{5}{11}+\dfrac{8}{16}\right]^2}{\dfrac{25}{11^2\times 10}+\dfrac{64}{16^2\times 15}} \\ &= \frac{0.911157024}{0.037327823} \\ &\approx 24.41. \end{aligned}$$

Since table III (Stoker) only gives integer values for v we need to interpolate between $v = 24$ and $v = 25$

$$\therefore t_{0.025;24;41} \approx 2.064 + 0.41\,(2.060 - 2.064) = 2.062.$$

Thus our critical value is 2.062 and we will reject H_0 at the 2.5% level (one-sided) if $T < -t_{0.025;v}$, that is if $T < -2.062$.

Since $T = -2.0471 > -2.062$ we do not reject H_0 at the $2\frac{1}{2}$% level of significance. We cannot conclude that $\mu_1 < \mu_2 - 3$.

7.6 More than two independent samples

(One-way analysis of variance)

We discuss the problem of comparing k sample means with *equal sample sizes*. In a more advanced module it will be shown how the test can be modified if the sample sizes are unequal.

Thus we suppose that we have k independent random samples $(X_{11},\ ...,\ X_{1n})\,;\,(X_{21},\ ...,\ X_{2n})\,;\ \ ...;$ $(X_{k1},\ ...,\ X_{kn})$, such that the i-th sample comes from a normal distribution with mean μ_i and variance σ^2.

From this it follows that the essential assumptions are

(a) independence

(b) normality

(c) equal variances

The assumption of equal sample sizes is made here *for convenience* and is *not essential*. The model is therefore as follows:

$$X_{ij};\quad j = 1,\ ...,\ n;\quad i = 1,\ ...,\ k$$

are independent random variables such that $X_{ij} \sim n\left(\mu_i;\ \sigma^2\right)$. We wish to test the null hypothesis $H_0 : \mu_1 = \mu_2 = ...\mu_k$ against the general alternative $H_1 : \mu_p \neq \mu_q$ for at least one pair $p \neq q$.

We shall derive our test statistic for H_0 from the following results, stated here as a theorem and which is a summary of results from study unit 1.

Theorem 7.3

Let $\overline{X}_i = \frac{1}{n}\sum_{j=1}^{n} X_{ij};\quad S_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(X_{ij} - \overline{X}_i\right)^2$. Then

(a) $\overline{X}_1,\ ...,\ \overline{X}_k,\ \ S_1^2,\ ...,\ S_k^2$ are independent

(b) $\overline{X}_i \sim n\left(\mu_i;\ \frac{\sigma^2}{n}\right) \Rightarrow \frac{\sqrt{n}\left(\overline{X}_i - \mu_i\right)}{\sigma} \sim n\,(0;\ 1)$

(c) $\frac{(n-1)\,S_i^2}{\sigma^2} \sim \chi^2_{n-1}$.

Let $\overline{X} = \frac{1}{k}\sum_{1}^{k}\overline{X}_i = \frac{1}{nk}\sum_{i=1}^{k}\sum_{j=1}^{n} X_{ij}$ be the **overall mean** of all the observations.

We need to study the following two random variables:

$$U = \frac{n\sum_{i=1}^{k}\left(\overline{X}_i - \overline{X}\right)^2}{\sigma^2} \quad \text{and} \quad V = \frac{\sum_{i=1}^{k}(n-1)S_i^2}{\sigma^2} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n}\left(X_{ij} - \overline{X}_i\right)^2}{\sigma^2}$$

What are the distributions? Are they independently distributed?

Theorem 7.4

(a) $V \sim \chi^2_{kn-k}$.

(b) $U \sim \chi^2_{k-1}$ if $H_0 : \mu_1 = \mu_2 = ... = \mu_k$ is true.

(b) U and V are independent.

Proof

(a) Follows directly from theorem 7.3 and property (ii) of result 1.1.

(b) Suppose $\mu_1 = \mu_2 = ... = \mu_k = \mu$, Then $\overline{X}_1,\ ...,\ \overline{X}_k$ are independent $\left(\mu;\ \frac{\sigma^2}{n}\right)$ variates. (The k means can be considered to be a single sample of size k.) From study unit 1 it follows that

$$U = \frac{\Sigma\left(\overline{X}_i - \overline{X}\right)^2}{(\sigma^2/n)} \sim \chi^2_{k-1}.$$

(If H_0 is not true, the distribution of U is called noncentral chi-square.)

(c) Since $\overline{X}_1,\ ...,\ \overline{X}_k$ are independent of $S_1^2,\ ...,\ S_k^2$, any function of $\overline{X}_1,\ ...,\ \overline{X}_k$, such as U, is independent of any function of $S_1^2,\ ...,\ S_k^2$, such as V.

Theorem 7.5

Let $S^2 = \frac{\sum_i\sum_j\left(X_{ij} - \overline{X}_i\right)^2}{(kn-k)}$. Then $E\left(S^2\right) = \sigma^2$.

Proof

$S^2 = \frac{\sigma^2 V}{(kn-k)}$. Since $V \sim \chi^2_{kn-k},\ E(V) = kn - k$

$\therefore E\left(S^2\right) = \sigma^2$.

Thus S^2 is an unbiased estimator of σ^2. How do we interpret S^2?

We may write $S^2 = \frac{1}{k}\left(S_1^2 + ... + S_k^2\right)$, which is an ordinary average. So, S^2 is the average of all the sample variances. Now S_i^2 is a measure of the variation within the i-th sample. The only reason why $S_i^2 \neq 0$, in other words why $X_{i1},\ ...,\ X_{in}$ are not identical, is random variation which is called "error" (not to be confused with "mistake"). Any variation which cannot be explained except as random variation is called *variation due to error*. (NB This does not imply that someone erred.)

Definition 7.2

$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(X_{ij} - \overline{X}_i\right)^2$$

is called the *sum of squares due to error* or *error sum of squares* and

$$MSE = \frac{SSE}{(kn-k)}$$

is called the *mean square error*.

Definition 7.3

$$SSTr = n\sum_{i=1}^{k}\left(\overline{X}_i - \overline{X}\right)^2 = \sigma^2 U$$

measures the variation between samples and is called the *sum of squares due to treatments*.

$$MSTr = \frac{n\sum_{i=1}^{k}\left(\overline{X}_i - \overline{X}\right)^2}{k-1} = \frac{\sigma^2 U}{(k-1)}$$

is called the *mean square treatment*.

The k samples may be regarded as the result of k treatments, and the reason why $SSTr \neq 0$ is

(a) random variation and

(b) the fact that $\mu_1,\ ...,\ \mu_k$ may differ.

Let $\mu = \frac{1}{k}\sum_{1}^{k}\mu_i$.

Theorem 7.6

$$E\left(MSTr\right) = \sigma^2 + \frac{n\Sigma\left(\mu_i - \mu\right)^2}{(k-1)}$$

Proof

Let $Y_i = \sqrt{n}\left(\overline{X}_i - \mu_i\right), \quad i = 1, ..., k$

$$\therefore \overline{Y} = \frac{1}{k}\sum_{i=1}^{k} Y_i = \frac{1}{k}\Sigma\sqrt{n}\left(\overline{X}_i - \mu_i\right) = \sqrt{n}\left(\overline{X} - \mu\right).$$

Then $Y_1, \ ..., Y_k$ are independent $n\left(0;\ \sigma^2\right)$ variates, and from study unit 1 (see result 1.3) it follows that

$$\sum_{i=1}^{k} \frac{\left(Y_i - \overline{Y}\right)^2}{\sigma^2} \sim \chi^2_{k-1}$$

$$\therefore E\left(\Sigma\left(Y_i - \overline{Y}\right)^2\right) = \sigma^2\left(k-1\right) \text{ (see property (i) of result 1.1).}$$

$$\begin{aligned}
\text{Consider} \quad SSTr &= n\sum_{i=1}^{k}\left(\overline{X}_i - \overline{X}\right)^2 \\
&= \Sigma\left[\sqrt{n}\left(\overline{X}_i - \overline{X}\right)\right]^2 \\
&= \Sigma\left[\sqrt{n}\left(\overline{X}_i - \mu_i - \overline{X} + \mu + \mu_i - \mu\right)\right]^2 \\
&= \Sigma\left[\left(Y_i - \overline{Y}\right) + \sqrt{n}\left(\mu_i - \mu\right)\right]^2 \\
&= \Sigma\left(Y_i - \overline{Y}\right)^2 + n\Sigma\left(\mu_i - \mu\right)^2 + 2\sqrt{n}\Sigma\left(\mu_i - \mu\right)\left(Y_i - \overline{Y}\right)
\end{aligned}$$

$$\begin{aligned}
\therefore E\left(SSTr\right) &= E\Sigma\left(Y_i - \overline{Y}\right)^2 + n\Sigma\left(\mu_i - \mu\right)^2 + 2\sqrt{n}\Sigma\left(\mu_i - \mu\right)E\left(Y_i - \overline{Y}\right) \\
&= \sigma^2\left(k-1\right) + n\Sigma\left(\mu_i - \mu\right)^2 + 0
\end{aligned}$$

since $E\left(Y_i\right) = E\left(\overline{Y}\right) = 0$

$$\therefore E\left(MSTr\right) = \frac{E\left(SSTr\right)}{(k-1)} = \sigma^2 + \frac{n\Sigma\left(\mu_i - \mu\right)^2}{(k-1)}.$$

Note that, if $\mu_1 = \mu_2 = ... = \mu_k = \mu$, then $E(MSTr) = \sigma^2$ and $MSTr$ is then also an unbiased estimator for σ^2. However, if the means are not equal, $E(MSTr) > \sigma^2$.

Aha! Here we have the beginnings of a test statistic.

Theorem 7.7

Let $F = \dfrac{MSTr}{MSE} = \dfrac{\sigma^2 U/(k-1)}{\sigma^2 V/(kn-k)} = \dfrac{n\sum_{i=1}^{k}\left(\overline{X}_i - \overline{X}\right)^2/(k-1)}{\sum_{i=1}^{k}\sum_{j=1}^{n}\left(X_{ij} - \overline{X}_i\right)^2/(kn-k)}$.

Then $F \sim F_{k-1;kn-k}$ if $H_0 : \mu_1 = \mu_2 = ... = \mu_k$ is true.

The theorem follows directly from theorem 7.4.

This result is used to test H_0. The F-statistic is computed and compared to $F_{\alpha;k-1;kn-k}$. If H_0 is true $F \sim F_{k-1;kn-k}$ and if H_0 is not true we expect $MSTr$, and therefore F, to have a large value. H_0 is rejected if $F > F_{\alpha;k-1;kn-k}$.

Example 7.5

A company manufacturing medicine is screening various chemicals for possible use against cancer. They have three possible chemicals which they wish to test. Twenty mice are selected, and cancer cells are implanted into each. The mice are then divided at random (eg by lottery) into four groups of five mice each; three groups are treated by means of 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;	1.50;	1.80;	1.30;	1.80
Chemical B:	1.70;	2.05;	1.80;	2.15;	1.80
Chemical C:	1.70;	1.75;	1.50;	1.40;	1.90
Control:	1.90;	2.05;	2.35;	1.85;	2.10

Do these results indicate that the tumors respond differently to the treatments?

Solution

We want to test $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ against
$H_1 : \mu_p \neq \mu_q$ for at least one pair $p \neq q$.

We assume the tumour masses $(X_{ij}) \sim n\left(\mu_i;\ \sigma^2\right)$ for $i = 1, 2, ..., 4$.
We choose $\alpha = 0.05$ (ie rather large) in order for β to be smaller. We would like to keep the probability small for potential medicines to be discarded (ie to reject H_1 when it is true $=$ type II error $= \beta$.) If a type I error is committed, it only means that further tests will be performed on a useless chemical.

We have

$$k = 4; \qquad n = 5; \qquad kn - k = 16; \qquad k - 1 = 3;$$

$$\overline{X}_1 = \frac{8.0}{5} = 1.6; \qquad SS_1 = \Sigma \left(X_{1i} - \overline{X}_1\right)^2 = 0.18$$

$$\overline{X}_2 = 1.9 \qquad SS_2 = 0.145$$

$$\overline{X}_3 = 1.65 \qquad SS_3 = 0.16$$

$$\overline{X}_4 = 2.05 \qquad SS_4 = 0.155$$

$$\overline{X} = \frac{7.2}{4} = 1.8 \qquad SSE = SS_1 + ... + SS_4 = 0.64$$

$$MSE = S^2 = \frac{0.64}{16} = 0.04$$

Furthermore

$$\sum_{i=1}^{4} \left(\overline{X}_i - \overline{X}\right)^2 = (1.96 - 1.8)^2 + \cdots + (2.05 - 1.8)^2 = 0.135$$

$$SSTr = n\Sigma \left(\overline{X}_i - \overline{X}\right)^2 = 6(0.135) = 0.675$$

$$MSTr = \frac{n\Sigma \left(\overline{X}_i - \overline{X}\right)^2}{(k-1)} = \frac{0.675}{3} = 0.225$$

$$F = \frac{MSTr}{MSE} = \frac{0.225}{0.04} = 5.625$$

From table V we find $F_{0.05;3;16} = 3.24$. Since $F > F_{0.05;3;16}$ we reject H_0.

The analysis is often summarised in tabular form called an ANOVA table.

ANOVA table

Source of variation	Sum of squares	Degrees of freedom	Mean square	F
Treatments	0.675	3	0.225	5.625
Error	0.640	16	0.04	
Total	1.315	19		

The "*total sum of squares*" is

$$SST = \sum_{i=1}^{k} \sum_{j=1}^{n} \left(X_{ij} - \overline{X}\right)^2$$

which measures the total variation in all the observations. If we did not know that the groups of mice had received different treatments, this would have been used to estimate σ^2; its degrees of freedom are, in general terms, $kn - 1 = (k - 1) + (kn - k)$.

Multiple comparisons

The F-test we have just discussed, leads to one of two decisions: either H_0 is accepted and we believe the k population means are equal, or H_0 is rejected and we believe they are unequal. However, the latter decision includes many possibilities, for example

$\mu_1 = \mu_2 \neq \mu_3 \neq \mu_4$ or $\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$ or $\mu_1 = \mu_2 = \mu_3 \neq \mu_4$, et cetera,

and we often want to know which of these alternatives is a likely representation of the truth. We may compute, for each pair of means $\overline{X}_p$ and $\overline{X}_q$, a t-statistic

$$T_{pq} = \frac{\overline{X}_p - \overline{X}_q}{S\sqrt{\frac{1}{n} + \frac{1}{n}}} = \frac{\sqrt{n}\left(\overline{X}_p - \overline{X}_q\right)}{\sqrt{2}S}$$

and reject $H_0\left(p;q\right): \mu_p = \mu_q$ in favour of

$H_1\left(p;q\right): \mu_p \neq \mu_q$ if $|T_{pq}|$ exceeds a critical value.

However, this means that $\binom{k}{2}$ different hypotheses are tested on the same data set, and the overall significance level would be much larger than we think. (Please refer to section 2.7 of study unit 2.) However, it can be proved that, for all p and q, $T_{pq}^2 \leq (k-1)F$ where $F = \frac{MSTr}{MSE}$. Since H_0 is rejected if $F > F_{\alpha;k-1;kn-k}$, our significance level would remain α if we reject $H_0\left(p;q\right)$ if $T_{pq}^2 > (k-1)F_{\alpha;k-1;kn-k}$, that is if

$$|T_{pq}| > \sqrt{(k-1)F_{\alpha;k-1;kn-k}}.$$

Example 7.5 (continued)

$(k-1)F_{\alpha;k-1;kn-k} = 3F_{0.05;3;16} = 3\left(3.24\right) = 9.72$

$$T_{pq} = \frac{\sqrt{n}\left(\overline{X}_p - \overline{X}_q\right)}{\sqrt{2}S} = \frac{\sqrt{5}\left(\overline{X}_p - \overline{X}_q\right)}{\sqrt{2}\sqrt{0.04}} = \sqrt{62.5}\left(\overline{X}_p - \overline{X}_q\right)$$

We reject $H_0\left(p;q\right): \mu_p = \mu_q$ if

$$|T_{pq}| > \sqrt{9.72}$$

$$\therefore \left|\overline{X}_p - \overline{X}_q\right| > \sqrt{\frac{9.72}{62.5}} \approx 0.3944$$

Now $\overline{X}_4 - \overline{X}_1 = 0.45$ (the largest observed difference) and $\overline{X}_4 - \overline{X}_3 = 0.4$ are both significant. We note, however, that $\overline{X}_1 = 1.6$ and $\overline{X}_3 = 1.65$ are rather close together, that $\overline{X}_2 = 1.9$ and $\overline{X}_4 = 2.05$ are close together, but that the two pairs are comparatively more different.

We therefore assume that $\mu_1 = \mu_3 \neq \mu_2 = \mu_4$, which would imply that further research could be done on chemicals A and C as potential remedies for cancer.

Exercise 7.1

1. A machine is set to produce washers with a thickness of 0.50 mm. To test whether the machine is working properly, 11 washers are chosen at random and their thickness measured. The results are

$$0.53 \quad 0.52 \quad 0.60 \quad 0.45 \quad 0.55 \quad 0.53 \quad 0.63 \quad 0.48 \quad 0.49 \quad 0.62 \quad 0.43$$

   (a) Test $H_0 : \mu = 0.5$ at the 10% significance level against $H_1 : \mu \neq 0.5$.

   (b) Find a 90% (two-sided) confidence interval for μ.

2. The following is the yield (kg) per plant of a certain tomato cultivar:

$$1.54 \quad 1.60 \quad 1.42 \quad 1.36 \quad 1.48 \quad 1.60$$

   (a) Test $H_0 : \mu = 1.6$ against $H_1 : \mu < 1.6$ at the 5% level.

   (b) Find a 95% upper confidence limit for μ.

3. $H_0 : \mu = 100$ is tested against $H_1 : \mu \neq 100$ using a sample of size $n = 16$. Find the power of the test if $\mu = 100 - \sqrt{0.72\sigma^2}$ at the level

   (a) $\alpha = 0.05$

   (b) $\alpha = 0.01$.

4. An aptitude test based on spatial orientation was given to 10 students studying for a diploma in engineering and to 12 students studying for a diploma in graphical design. The following results were computed:

$$\text{Engineering:} \quad n = 10; \quad \Sigma X_{1i} = 1\,070; \quad \Sigma X_{1i}^2 = 115\,990$$

$$\text{Graphical design:} \quad n = 12; \quad \Sigma X_{2i} = 1\,344; \quad \Sigma X_{2i}^2 = 152\,328$$

   (a) Test $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 < \mu_2$ at the 5% level.

   (b) Find a 90% two-sided confidence interval for $\mu_1 - \mu_2$.

5. Suppose we wish to draw two random samples in order to test $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$. and let the total sample size $n_1 + n_2$ be fixed, say $n_1 + n_2 = 20$. Which sample sizes n_1 and n_2 (subject to $n_1 + n_2 = 20$) will yield the highest power?

6. Two samples of sizes $n_1 = 3$ and $n_2 = 9$ are used to test $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$. What will the power of the test be if $\mu_1 = \mu_2 + 1.8\sigma\sqrt{2}$ and the significance level is as follows?

   (a) $\alpha = 0.05$

   (b) $\alpha = 0.01$

7. Derive a t-statistic to test $H_0 : \mu_1 = 2\mu_2$ against $H_1 : \mu_1 > 2\mu_2$. Base your t-statistic on $\overline{X}_1 - 2\overline{X}_2$ which is an estimator for $\mu_1 - 2\mu_2$.

8. The blood sugar content of eight patients was measured, each patient was given a fixed amount of glucose and the blood sugar content measured again after one hour. The results were as follows:

Patient:	1	2	3	4	5	6	7	8
Blood sugar (Before):	60	75	69	63	64	72	68	73
Blood sugar (After):	68	81	76	66	76	79	72	82

   (a) Test the hypothesis that the expected blood sugar content increases by more than five units after dosage (5% level).

   (b) Find a 95% lower confidence limit for the increase in blood sugar after dosage.

9. Suppose in two samples from normal distributions with unequal variances, it is found that:

$$\overline{X}_1 = 110; \quad \overline{X}_2 = 120; \quad S_1^2 = 180; \quad S_2^2 = 55;$$
$$n_1 = 9; \quad n_2 = 11$$

   Test $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$ at the 10% level.

10. Suppose we have two independent samples as before, with $X_{ij} \sim n\left(\mu_i;\ \sigma_i^2\right)$ but with $\sigma_1^2 = 2\sigma_2^2$ where σ_1^2 and σ_2^2 are unknown. Construct a t-statistic for testing $H_0 : \mu_1 = \mu_2$.

11. Two independent random samples of sizes 10 and 12 respectively from $n\left(\mu_1;\ \sigma_1^2\right)$ and $n\left(\mu_2;\ \sigma_2^2\right)$ distributions yielded the following results:

$$\overline{X}_1 = 40; \quad \overline{X}_2 = 60; \quad S_1^2 = 400; \quad S_2^2 = 720$$

Find a 95% confidence interval for $\mu_1 - \mu_2$.

12. Twenty-one babies were weighed, and divided at random into three groups of seven each. Each group of seven babies was fed a different kind of baby food, and their increase in body mass (kg) determined after a fixed period:

Food A:	2.2;	1.8;	2.4;	1.5;	1.9;	2.1;	1.4
Food B:	1.9;	2.1;	1.5;	1.8;	2.3;	1.4;	1.6
Food C:	2.7;	2.1;	2.6;	2.3;	2.0;	1.8;	2.6

Test at the 5% level whether there is a difference in the mean response to the three baby foods. (Note: normally such an experiment would be performed on a much larger scale. The present problem could be a preliminary trial.)

13. Each of four brands of feed was fed to eight animals selected at random. The following mass gains (kg) were obtained:

Brand A:	13.5	12.6	14.0	15.0	13.2	14.4	11.0	10.3
Brand B:	10.5	9.0	11.1	9.2	10.1	11.2	11.0	7.9
Brand C:	13.5	14.0	12.2	11.5	11.0	13.2	11.7	8.9
Brand D:	9.0	10.2	9.2	8.9	8.5	9.2	9.6	7.4

(a) Assuming the population variances to be equal, test at the 5% level whether the means differ.

(b) Perform multiple comparisons on all pairs of means. Discuss your results.

7.7 Learning outcomes

After studying study unit 7, you should **be able to**

- perform hypothesis tests concerning the mean of a single sample
- derive one or two-sided confidence intervals for the mean of a single sample
- interpret the computer output of JMP concerning the power of the test for means
- perform hypothesis tests concerning the difference between means of two independent samples
- derive one- or two-sided confidence intervals for the difference between means of two independent samples
- perform hypothesis tests concerning the equality of the means of paired observations
- perform hypothesis tests concerning the equality of the means of more than two independent samples
- interpret the computer output of JMP concerning the tests of means

STUDY UNIT 8

Regression

8.1 Correlation and regression

Correlation problems occur when we have two (or more) random variables, and we want to know whether the variables are related in the sense that they tend to vary together – the conditional expectation of one variate, given values of the other variate, is a function of these values:

$$E\left(Y_1 | \, Y_2 = y_2\right) = f\left(y_2\right)$$

If two variates are correlated, it does not necessarily mean that a change in the value of one variate *causes* the other variate to change. It may happen that two random variables are correlated because there is an unknown factor that causes both variates to vary.

In such experiments it is often desired to estimate the conditional expectation of Y_1 given Y_2, in other words the regression of Y_1 on Y_2 in order to be able to predict Y_1 when Y_2 is given. We shall not discuss this type of problem in this module.

Regression problems as discussed here, that is causal relationships, occur when we have a random variable and one (or more) mathematical variables which are not random. The mathematical variables are called "control variables", "predictors" or "independent variables" and the random variable is called the "response variable", "predictand" or "dependent variable". An ideal (causal) regression experiment is performed as follows:

A number of values of the control variables are chosen (eg 100°C; 120°C; 140°C and 160°C if the control variable is temperature) and the response (eg hardness of the product) is observed at each setting of the control variable. Usually the experiment is repeated a few times at each setting.

A correlation or non-causal regression study, on the other hand, involves taking pairs of observations (eg height and body mass) on a number of individuals (not necessarily people). The correlation study is passive – individuals are selected at random and the two variables are measured on each individual. Regression studies in the sense discussed here, are active – the control variable is changed deliberately in order to observe what effect the change has on the response variable.

The distinction between the two models is not always observed. Regression lines are computed from correlation data and correlation coefficients are computed from regression data. One must be careful when inference on correlation or regression coefficients is the object of the study. We shall show that there is some justification for the practice of mixing the models but only to a certain extent.

8.2 The simple linear regression model

We consider here the simple linear regression model

$$Y = \beta_0 + \beta_1 X + E$$

where Y is the response variable, X the non-random control variable, β_0 and β_1 are the unknown regression coefficients and E is a random variable (called the "error" or "random component") with a $n\left(0;\ \sigma^2\right)$ distribution where σ^2 is unknown. In order to estimate $\beta_0,\ \beta_1$ and σ^2 we choose a number of values of X (at least two different values of X) and observe the response one or more times at each setting of X (at least three observations are needed but we should preferably have more). The experiments should be run in random order. We may for example write each setting of X on a piece of paper with as many repetitions of the same X as we intend to repeat the experiment at that value of X. The pieces of paper are thrown into a hat, shuffled thoroughly and retrieved one by one to give the order in which the experiment should be run.

Our first task after obtaining the data is to make sure that the simple linear regression model is appropriate. To do this we plot the data on graph paper with Y on the vertical axis and X on the horizontal axis. We inspect the graph to see whether the data cluster around a straight line and whether the variance remains about the same for all values of X.

If the data show curvature we can try transforming the data into a straight line by plotting Y versus $\log X,\ \log Y$ versus $X,\ \log Y$ versus $\log X,\ \frac{1}{Y}$ versus X et cetera Sometimes we succeed in straightening out the data in this way and change our model accordingly. If, for example, we find $\log Y$ versus $\log X$ to form a straight line then our model becomes

$$\log Y = \beta_0 + \beta_1 \log X + E.$$

If we do not succeed in finding a suitable transformation we may consider a *polynomial model*:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + ... + \beta_p X^p + E,$$

and if this fails we have a *non-linear* regression problem which is much more difficult to solve. Ideally the model should be chosen on theoretical grounds without looking at the data, if this is at all possible.

Example 8.1

In order to evaluate the effect of temperature (X) on the yield (Y) of a chemical process, an experiment was run in the plant with the following results:

X	Y			
205	12;	13;	16;	16
210	18;	19;	20;	18
215	22;	26;	24;	28
220	35;	31;	33;	34
225	53;	44;	46;	43
230	64;	62;	59;	67

Solution

Start by plotting X against Y as in figure 8.1.

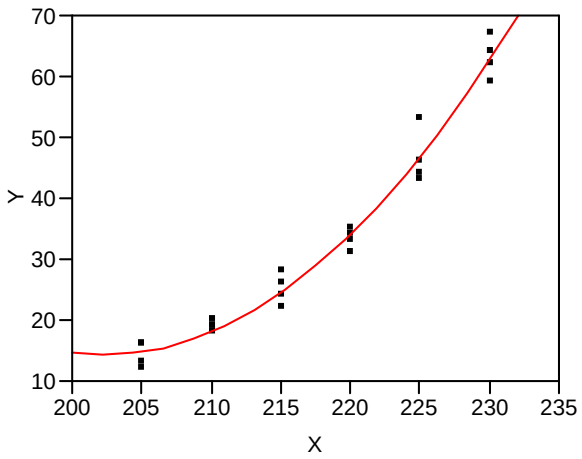


Figure 8.1

It is obvious from the graph that there is no linear relationship between X and Y.
Transform Y by finding $\log_{10} Y$ and then plot X against $\log_{10} Y$ (see figure 8.2).

X	$\log_{10} Y$			
205	1.079	1.114	1.204	1.204
210	1.255	1.279	1.301	1.255
215	1.342	1.415	1.380	1.447
220	1.544	1.491	1.519	1.531
225	1.724	1.643	1.662	1.633
230	1.806	1.792	1.771	1.826

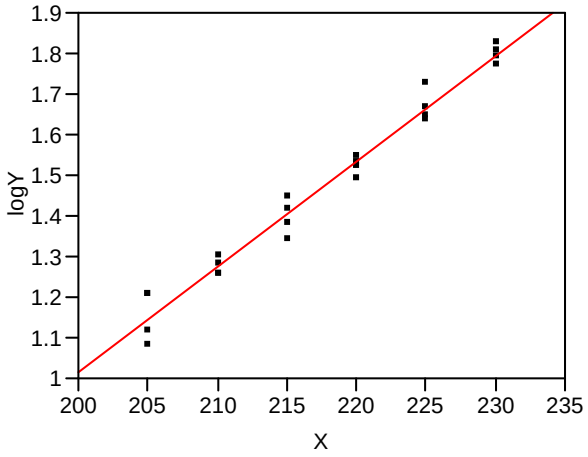


Figure 8.2

It is obvious from figure 8.2 that this transformation gives a good linear relationship. The "best" model seems to be

$$\log_{10} Y = \beta_0 + \beta_1 X + E.$$

If the variance does not remain constant then we should apply weighted regression. This can sometimes also be done by means of transformation. Suppose the model is

$$Y = \beta_0 + \beta_1 X + E$$

where $Var\,(E) = \sigma^2 f\,(X)$ and where $f\,(X)$ is a known function of X. Then

$$\frac{Y}{\sqrt{f\,(X)}} = \beta_0 \frac{1}{\sqrt{f\,(X)}} + \beta_1 \frac{X}{\sqrt{f\,(X)}} + \frac{E}{\sqrt{f\,(X)}}$$

where $\dfrac{E}{\sqrt{f\,(X)}}$ is $n\left(0;\ \sigma^2\right)$. This method is especially useful if $f\,(x) = X^2$.

The scatter diagram for this model is typically as follows:

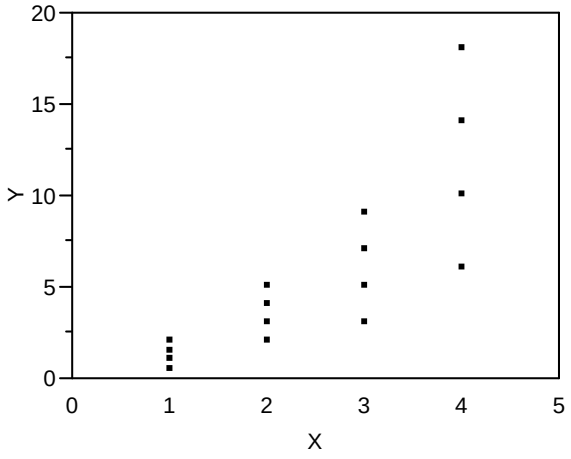


Figure 8.3

The model becomes

$$\frac{Y}{X} = \beta_0 \frac{1}{X} + \beta_1 + \frac{E}{X}.$$

If we plot $\frac{Y}{X}$ versus $\frac{1}{X}$ we should obtain a straight line with constant variance.

(This is left to you as an exercise – see activity 8.2 in the workbook.)

8.3 Estimation

We now assume the following model: $Y_1,\ ...,\ Y_n$ are independent random variables with

$$Y_i \sim n\left(\beta_0 + \beta_1 X_i;\ \sigma^2\right), \qquad i = 1, ..., n.$$

The problem is to estimate $\beta_0;\ \beta_1$ and σ^2. We use the method of maximum likelihood:

$$\begin{aligned} L &= \prod_{i=1}^{n} f_Y\left(y_i;\ \beta_0, \beta_1, \sigma^2\right) \\ \therefore L &= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(Y_1-\beta_0-\beta_1 X_1)^2}{\sigma^2}} ... \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(Y_n-\beta_0-\beta_1 X_n)^2}{\sigma^2}} \\ &= \left(\frac{1}{\sigma}\right)^n (2\pi)^{-\frac{1}{2}n} e^{-\frac{1}{2}\frac{\Sigma(Y_i-\beta_0-\beta_1 X_i)^2}{\sigma^2}} \\ \therefore \ln L &= -n\ln\sigma - \frac{1}{2}n\ln(2\pi) - \frac{1}{2}\frac{\Sigma\left(Y_i-\beta_0-\beta_1 X_i\right)^2}{\sigma^2} \end{aligned}$$

In order to maximise $\ln L$ (and therefore L), the partial derivatives with respect to $\beta_0,\ \beta_1$ and σ are equated to zero:

$$\begin{aligned}\frac{\partial \ln L}{\partial \beta_0} &= \frac{\Sigma\left(Y_i-\beta_0-\beta_1 X_i\right)}{\sigma^2}\\ &= \frac{\left(\Sigma Y_i-n\beta_0-\beta_1\Sigma X_i\right)}{\sigma^2}\\ &= 0 \quad \text{if } \ n\beta_0+\beta_1\Sigma X_i=\Sigma Yi \end{aligned} \tag{1}$$

$$\begin{aligned}\frac{\partial \ln L}{\partial \beta_1} &= \frac{\Sigma X_i\left(Y_i-\beta_0-\beta_1 X_i\right)}{\sigma^2}\\ &= \frac{\left(\Sigma X_iY_i-\beta_0\Sigma X_i-\beta_1\Sigma X_i^2\right)}{\sigma^2}\\ &= 0 \quad \text{if } \ \beta_0\Sigma X_i+\beta_1\Sigma X_i^2=\Sigma X_iYi \end{aligned} \tag{2}$$

From (1) and (2) follows:

$$\begin{aligned}\beta_1 &= \frac{n\Sigma X_iY_i-\Sigma Y_i\Sigma X_i}{n\Sigma X_i^2-\left(\Sigma X_i\right)^2}\\ &= \frac{\Sigma X_iY_i-\overline{X}\Sigma Y_i}{\Sigma X_i^2-n\overline{X}^2}\\ &= \frac{\Sigma Y_i\left(X_i-\overline{X}\right)}{\Sigma\left(X_i-\overline{X}\right)^2}\end{aligned}$$

$$\beta_0=\frac{1}{n}\Sigma Y_i-\beta_1\frac{1}{n}\Sigma X_i=\overline{Y}-\beta_1\overline{X}$$

$$\begin{aligned}\frac{\partial \ln L}{\partial \sigma} &= -\frac{n}{\sigma}+\frac{\Sigma\left(Y_i-\beta_0-\beta_1X_i\right)^2}{\sigma^3}\\ &= 0 \quad \text{if } \ \sigma^2=\frac{1}{n}\Sigma\left(Y_i-\beta_0-\beta_1X_i\right)^2\end{aligned}$$

We have therefore derived the following result:

Result 8.1

The maximum likelihood estimators for $\beta_0,\ \beta_1$ and σ^2 are

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

$$\hat{\beta}_1 = \frac{\Sigma Y_i \left(X_i - \overline{X}\right)}{\Sigma \left(X_i - \overline{X}\right)^2}$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)^2.$$

These estimators are also the least squares estimators. (Under the assumption of normality they are the MLEs.)

Only $\hat{\sigma}^2$ is biased; it may be shown that

$$E\left(\hat{\sigma}^2\right) = \frac{n-2}{n}\sigma^2$$

so that

$$S^2 = \frac{1}{n-2}\Sigma \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)^2.$$

is an unbiased estimator for σ^2.

Keep in mind that S^2 is a measure of the variation around the regression line and that $\left(Y_i - \hat{\beta}_0 - \beta_1 X_i\right)^2$ is computed for each observed pair $(X_i;\ Y_i)$ as the squared difference of the observed Y_i-value and the estimated Y_i value by using the equation of the regression line. Sometimes S^2 is also indicated as

$$S^2 = \frac{1}{n-2} \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2.$$

Distribution of the estimators

If we want to derive test statistics to test hypotheses about the theoretical parameters of a regression line (ie β_0 and β_1) we need to understand the "behaviour" of the estimators of the parameters. In other words we are interested in the distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$.

From study unit 1 it can be deduced that, if $Y_1,\ ...,\ Y_n$ are independent with $Y_i \sim n\left(\mu_i;\ \sigma^2\right)$ then $U = \Sigma a_i Y_i$ and $V = \Sigma b_i Y_i$ are jointly normally distributed with means $E(U) = \Sigma a_i \mu_i$ and $E(V) = \Sigma b_i \mu_i$, variances $Var(U) = \sigma^2 \Sigma a_i^2$ and $Var(V) = \sigma^2 \Sigma b_i^2$ and covariance $Cov(U,\ V) = \sigma^2 \Sigma a_i b_i$. (See "sums of independent normal variates" just above theorem 1.2.)

Let us apply these powerful (and handy!) results to $\hat{\beta}_0$ and $\hat{\beta}_1$.

We will start with $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{\Sigma Y_i \left(X_i - \overline{X}\right)}{\Sigma \left(X_i - \overline{X}\right)^2}$$

Suppose we use the notation $d^2 = \Sigma \left(X_i - \overline{X}\right)^2 = \Sigma X_i \left(X_i - \overline{X}\right)$.

Then, $\hat{\beta}_1 = \frac{1}{d^2}\Sigma Y_i \left(X_i - \overline{X}\right) = \Sigma a_i Y_i$ where $a_i = \frac{X_i - \overline{X}}{d^2}$.

(a) $E\left(\hat{\beta}_1\right) = \Sigma a_i E\left(Y_i\right)$

$$= \Sigma \frac{X_i - \overline{X}}{d^2} \left(\beta_0 + \beta_1 X_i\right)$$

$$= \frac{1}{d^2}\beta_0 \Sigma \left(X_i - \overline{X}\right) + \frac{1}{d^2}\beta_1 \Sigma X_i \left(X_i - \overline{X}\right)$$

$$= 0 + \frac{1}{d^2}\beta_1 d^2$$

$$= \beta_1$$

Thus $\hat{\beta}_1$ is an unbiased estimator for β_1.

(b) $Var\left(\hat{\beta}_1\right) = \sigma^2 \Sigma a_i^2$

$$= \frac{\sigma^2}{d^4} \Sigma \left(X_i - \overline{X}\right)^2$$

$$= \frac{\sigma^2 d^2}{d^4}$$

$$= \frac{\sigma^2}{d^2}$$

Now we do the same for $\hat{\beta}_0$.

$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$

$$= \frac{1}{n}\Sigma Y_i - \frac{\overline{X}}{d^2}\Sigma Y_i \left(X_i - \overline{X}\right)$$

$$= \Sigma \left[\frac{1}{n} - \frac{\overline{X}}{d^2}\left(X_i - \overline{X}\right)\right] Y_i$$

$$= \Sigma b_i Y_i \text{ where } b_i = \frac{1}{n} - \frac{\overline{X}}{d^2}\left(X_i - \overline{X}\right)$$

(a) $E\left(\hat{\beta}_0\right) = \Sigma\left[\frac{1}{n} - \frac{\overline{X}}{d^2}\left(X_i - \overline{X}\right)\right][\beta_0 + \beta_1 X_i]$

$= \beta_0 \Sigma\left[\frac{1}{n} - \frac{\overline{X}}{d^2}\left(X_i - \overline{X}\right)\right] + \beta_1 \Sigma X_i\left[\frac{1}{n} - \frac{\overline{X}}{d^2}\left(X_i - \overline{X}\right)\right]$

$= \beta_0\left[1 - \frac{\overline{X}}{d^2}\Sigma\left(X_i - \overline{X}\right)\right] + \beta_1\left[\overline{X} - \frac{\overline{X}}{d^2}\Sigma X_i\left(X_i - \overline{X}\right)\right]$

$= \beta_0 - 0 + \beta_1\overline{X} - \beta_1\frac{\overline{X}}{d^2}d^2$

$= \beta_0$

$\therefore \hat{\beta}_0$ is an unbiased estimator for β_0.

(b) $Var\left(\hat{\beta}_0\right) = \sigma^2\Sigma\left[\frac{1}{n} - \frac{\overline{X}}{d^2}\left(X_i - \overline{X}\right)\right]^2$

$= \sigma^2\Sigma\left[\frac{1}{n^2} - \frac{2\overline{X}}{nd^2}\left(X_i - \overline{X}\right) + \frac{\overline{X}^2}{d^4}\left(X_i - \overline{X}\right)^2\right]$

$= \sigma^2\left[\frac{n}{n^2} - 0 + \frac{\overline{X}^2}{d^4}d^2\right]$

$= \sigma^2\left[\frac{1}{n} + \frac{\overline{X}^2}{d^2}\right]$

If you go back to the introduction of this subsection you will notice that all that remains is to simplify

$Cov\left(U,\ V\right) = Cov\left(\hat{\beta}_0;\ \hat{\beta}_1\right) = \sigma^2\Sigma a_i b_i$

$= \sigma^2\Sigma\frac{X_i - \overline{X}}{d^2}\left[\frac{1}{n} - \frac{\overline{X}}{d^2}\left(X_i - \overline{X}\right)\right]$

$= \frac{-\sigma^2\overline{X}}{d^2}$

This long derivation was actually the proof of the following theorem:

Theorem 8.1

$\hat{\beta}_0$ and $\hat{\beta}_1$ are jointly normally distributed with

$$E\left(\hat{\beta}_0\right) = \beta_0; \qquad E\left(\hat{\beta}_1\right) = \beta_1;$$

$$Var\left(\hat{\beta}_0\right) = \sigma^2\left(\frac{1}{n} + \frac{\overline{X}^2}{d^2}\right); \qquad Var\left(\hat{\beta}_1\right) = \frac{\sigma^2}{d^2};$$

$$Cov\left(\hat{\beta}_0;\ \hat{\beta}_1\right) = \frac{-\sigma^2\overline{X}}{d^2} \qquad \text{with } d^2 = \Sigma\left(X_i - \overline{X}\right)^2.$$

The next theorem is also very important and it is assumed without proof here.

Theorem 8.2

$$\frac{\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)^2}{\sigma^2} \sim \chi^2_{n-2}$$

and is independent of $\hat{\beta}_0$ and $\hat{\beta}_1$.

Example 8.2

An experiment is performed to estimate the relationship between the yield (bags per hectare) of a certain variety of maize and the amount of fertilizer (metric tons per hectare) applied, using a new kind of fertilizer. Twelve farms are chosen in a certain district, and divided into four groups of three farms each in a random fashion. Each group receives a certain amount of fertilizer per hectare, and the yields are recorded:

Fertilizer Metric tons/hectare	Yield Bags/hectare
0	15; 12; 17
2	24; 20; 21
4	21; 31; 28
6	36; 32; 31

The problem is to estimate the relationship between yield and amount of fertilizer.

Before proceeding, draw a graph of the data of example 8.2, plotting X (fertilizer) on the horizontal axis and the response Y (yield) on the vertical axis. By inspection of this graph we conclude that a straight line would probably be adequate. Theoretically we should know the form of the regression line even before the data are collected. However, in practice this is very often impossible and we have to be guided by a graph or other means. We shall assume the straight line to be the true model. Our data and computations may be summarised in tabular form as follows:

	X_i	Y_i	$X_i - \overline{X}$	$\left(X_i - \overline{X}\right)^2$	$Y_i\left(X_i - \overline{X}\right)$	$\hat{Y}_i$	$Y_i - \hat{Y}_i$	$\left(Y_i - \hat{Y}_i\right)^2$
	0	15	−3	9	−45	15	0	0
	0	12	−3	9	−36	15	−3	9
	0	17	−3	9	−51	15	2	4
	2	24	−1	1	−24	21	3	9
	2	20	−1	1	−20	21	−1	1
	2	21	−1	1	−21	21	0	0
	4	21	1	1	21	27	−6	36
	4	31	1	1	31	27	4	16
	4	28	1	1	28	27	1	1
	6	36	3	9	108	33	3	9
	6	32	3	9	96	33	−1	1
	6	31	3	9	93	33	−2	4
Total	36	288	0	60	180	288	0	90

This is computed after you have solved the equation of the regression line.

$$\overline{X} = 3; \qquad \overline{Y} = 24; \qquad \hat{\beta}_1 = \frac{180}{60} = 3;$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{X} = 24 - 9 = 15;$$

The estimated regression line is $Y = 15 + 3X$.

We use this line to compute $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ for example if $X = 0 \Rightarrow \hat{Y}_i = 15$
if $X = 2 \Rightarrow \hat{Y}_i = 15 + 3(2) = 21$ et cetera

$$\therefore S^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = \frac{90}{10} = 9$$

We call $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ the *predictions* and $Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$ the residuals, that is the difference between the observations and predictions.

As a final step one usually draws the estimated regression line on the scatter plot. (We repeat this example electronically with JMP in activity 8.10 of the workbook.)

8.4 Inference on the coefficients

We are interested especially in the coefficients β_0 and β_1. We would like to test hypotheses and construct confidence intervals.

The following two theorems follow directly from the definition of a t-variable:

Theorem 8.2

$$T_0 = \frac{\hat{\beta}_0 - \beta_0}{S\sqrt{\frac{1}{n} + \frac{\overline{X}^2}{d^2}}} \text{ is a } t_{n-2} \text{ variate.}$$

Theorem 8.3

$$T_1 = \frac{\hat{\beta}_1 - \beta_1}{\frac{S}{d}} \text{ is a } t_{n-2} \text{ variate.}$$

These two theorems may be used to test the significance of $\hat{\beta}_0$ and $\hat{\beta}_1$ or to construct confidence intervals for β_0 and β_1.

(a) **Inference on $\boldsymbol{\beta_0}$**

Suppose we wish to test whether $E(Y) = c$ if $X = 0$, in other words whether the regression line has a particular **intercept on the Y-axis**. The null hypothesis is $H_0 : \beta_0 = c$.

We compute

$$T_{0;c} = \frac{\left(\hat{\beta}_0 - c\right)}{S\sqrt{\frac{1}{n} + \frac{\overline{X}^2}{d^2}}}$$

and reject H_0 against $H_1 : \beta_0 \neq c$ at the α level of significance if $|T_{0;c}| > t_{\frac{\alpha}{2};n-2}$ similarly for one-sided tests. (The most common null hypothesis is $H_0 : \beta_0 = 0$, that is the regression line passes through the origin.)

A $100(1-\alpha)\%$ confidence interval for β_0 is

$$\left(\hat{\beta}_0 - t_{\frac{\alpha}{2};n-2} S\sqrt{\frac{1}{n} + \frac{\overline{X}^2}{d^2}}; \qquad \hat{\beta}_0 + t_{\frac{\alpha}{2};n-2} S\sqrt{\frac{1}{n} + \frac{\overline{X}^2}{d^2}}\right).$$

(b) **Inference on β_1**

Suppose we wish to test whether the regression line has a particular **slope**. The null hypothesis is $H_0 : \beta_1 = c$ which may be rejected in favour of $H_1 : \beta_1 \neq c$ at the α level if $|T_{1;c}| > t_{\frac{\alpha}{2};n-2}$ where $T_{1;c} = \dfrac{\hat{\beta}_1 - c}{\dfrac{S}{d}}$.

In a similar fashion we will perform a one-sided test. A $100\,(1-\alpha)\,\%$ confidence interval for β_1 is easily seen to be

$$\left(\hat{\beta}_1 - t_{\frac{\alpha}{2};n-2}\frac{S}{d};\qquad \hat{\beta}_1 + t_{\frac{\alpha}{2};n-2}\frac{S}{d}\right).$$

Example 8.2(a) (example 8.2 continued)

Test $H_0 : \beta_0 = 0$ against $H_1 : \beta_0 \neq 0$ and compute a 95% confidence interval for β_1.

Solution

For this example we have already computed

$$\overline{X} = 3;\quad \hat{\beta}_0 = 15;\quad \hat{\beta}_1 = 3;\quad d^2 = 60;\quad S^2 = \frac{90}{10} = 9 \quad \text{and} \quad n = 12.$$

To test $H_0 : \beta_0 = 0$ we compute

$$T_{0;0} = \frac{\hat{\beta}_0 - 0}{S\sqrt{\dfrac{1}{n} + \dfrac{\overline{X}^2}{d^2}}} = \frac{15 - 0}{3\sqrt{\dfrac{1}{12} + \dfrac{9}{60}}} \approx 10.3510.$$

We will reject H_0 at the 5% level of significance in favour of $H_1 : \beta_0 \neq 0$ if

$$|T_{0;0}| > t_{0.025;n-2} = t_{0.025;10} = 2.228 \text{ (table III)}.$$

Since $10.351 > 2.228$ we reject H_0 and conclude that the regression line does not pass through the origin. This could be expected in this example, since we do not expect "no yield" if we do not apply fertilizer.

A 95% confidence interval for β_1 is $\hat{\beta}_1 \pm t_{0.025;10}\dfrac{S}{d} = \left[3 - \dfrac{(2.228)\,(3)}{\sqrt{60}};\ 3 + \dfrac{(2.228)\,(3)}{\sqrt{60}}\right]$. that is $(2.14;\ 3.86)$.

We can use this two-sided interval to test a two-sided alternative, for example:
Test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$.

Since the confidence interval does not include zero, it implies the slope β_1 is significantly different from zero. This means that we could expect an increase of between 2.14 and 3.86 bags/ha for every additional metric ton/ha fertilizer applied. The farmer may now decide whether the price/ton of fertilizer is comparable with the price he or she receives for 2.14 and 3.86 bags of maize. Of course the yield will not increase indefinitely as more and more fertilizer is added. **Strictly speaking, we can only make predictions between 0 and 6 tons/ha.**

8.5 Inference on the regression line

(a) **Confidence limits for the regression line**

Suppose we choose a value X of the independent variable. We predict that the response will be

$$\hat{Y}(X) = \hat{\beta}_0 + \hat{\beta}_1 X.$$

How accurate is this prediction? We note that $\hat{Y}(X)$ is a normal variate, being a linear combination of two normal variates $\hat{\beta}_0$ and $\hat{\beta}_1$. For the same reason $\hat{Y}\,(X)$ is independent of $\hat{\sigma}^2$. Its mean and variance are

$$\begin{aligned} E\left[\hat{Y}(X)\right] &= E\left(\hat{\beta}_0\right) + XE\left(\hat{\beta}_1\right) \\ &= \beta_0 + \beta_1 X \end{aligned}$$

$$\begin{aligned} Var\left[\hat{Y}(X)\right] &= Var\left(\hat{\beta}_0\right) + X^2 Var\left(\hat{\beta}_1\right) + 2Xcov\left(\hat{\beta}_0;\ \hat{\beta}_1\right) \\ &= \sigma^2\left[\left(\frac{1}{n} + \frac{\overline{X}^2}{d^2} + \frac{X^2}{d^2} - \frac{2X\overline{X}}{d^2}\right)\right] \\ &= \sigma^2\left[\frac{1}{n} + \frac{\left(X-\overline{X}\right)^2}{d^2}\right] \end{aligned}$$

A confidence interval for $(\beta_0 + \beta_1 X)$ is seen to be

$$\left(\hat{\beta}_0 + \hat{\beta}_1 X\right) \pm t_{\frac{\alpha}{2};n-2} S\sqrt{\frac{1}{n} + \frac{\left(X-\overline{X}\right)^2}{d^2}}$$

Example 8.2(b) (example 8.2 continued)

We tabulate the calculations for the 95% confidence interval for $\beta_0 + \beta_1 X$ for the different values of X, using $t_{0.025;10} = 2.228$.

X	$\hat{\beta}_0 + \hat{\beta}_1 X$	$S\sqrt{\frac{1}{n} + \frac{(X-\overline{X})^2}{d^2}} = 3\sqrt{\frac{1}{12} + \frac{(X-3)^2}{60}}$	Lower limit	Upper limit
0	15	1.44914	11.77	18.23
1	18	1.1619	15.41	20.59
2	21	0.94868	18.89	23.11
3	24	0.86603	22.07	25.93
4	27	0.94868	24.89	29.11
5	30	1.1619	27.41	32.59
6	33	1.44914	29.77	36.23

For example we are 95% sure that the *mean* yield on farms where 5 tons/ha fertilizer is applied is between 27.41 and 32.59 bags/ha.

Plot these limits on the graph constructed for this example, and connect the points by means of a smooth curve.

In general, the confidence limits have the following form:

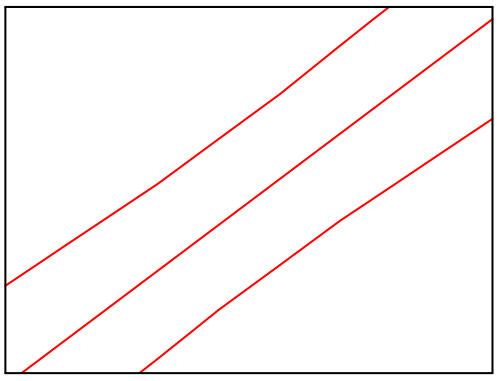


Figure 8.4

The limits form a "confidence band". The band is at its narrowest where $X = \overline{X}$. The limits show how accurately we have estimated the regression line. We can obtain a narrower band by increasing n and $\Sigma \left(X_i - \overline{X}\right)^2$.

(b) **Confidence limits for a future observation**

Suppose we choose a value of X with the intention of obtaining a further observation $Y_0(X)$ independent of $Y_1, \ ..., \ Y_n$. Where can we expect this observation to lie? $Y_0(X)$ is a random variable with mean $\beta_0 + \beta_1 X$ and variance σ^2 according to the assumptions of our model. We predict $Y_0(X)$ to be

$$\hat{Y}_0(X) = \hat{\beta}_0 + \hat{\beta}_1 X.$$

Consider the random variable $Y_0(X) - \hat{Y}_0(X)$. We see that

$$E\left[Y_0(X) - \hat{Y}_0(X)\right] = 0$$

$$Var\left[Y_0(X) - \hat{Y}_0(X)\right] = Var\left(Y_0(X)\right) + Var\left(\hat{Y}_0(X)\right)$$

(if $Y_0(X)$ and $\hat{Y}_0(X)$ are independent)

$$= \sigma^2 + \sigma^2 \left[\frac{1}{n} + \frac{\left(X - \overline{X}\right)^2}{d^2}\right]$$

$$= \sigma^2 \left[1 + \frac{1}{n} + \frac{\left(X - \overline{X}\right)^2}{d^2}\right]$$

It follows that the $100(1-\alpha)\%$ confidence limits for $Y_0(X)$ are

$$\hat{\beta}_0 + \hat{\beta}_1 X \pm t_{\frac{\alpha}{2};n-2} S \sqrt{1 + \frac{1}{n} + \frac{\left(X - \overline{X}\right)^2}{d^2}}$$

which appear rather similar to those of the previous paragraph but which are wider because of the extra term under the square root sign.

Example 8.2(c) (example 8.2 continued)

Similar to example 8.2(b) we can do the calculations in tabular form. We leave this for an activity in the workbook. With $t_{0,025;10} = 2.228$ we find

X	Lower limit	Upper limit
0	7.58	22.42
1	10.83	25.17
2	13.99	28.01
3	17.04	30.96
4	19.99	34.01
5	22.83	37.17
6	25.58	40.42

Suppose a farmer applies 4 tons of fertilizer to 1 ha of land. He or she can be 95% sure that the yield will be between about 20 and 34 bags (not allowing for meteorological variations). Now plot these limits on the graph of the data and join these with smooth curves.

The fact that the variance expressions differ is a rather technical point because the latter variance expression is derived on the assumption that this "future" observation is independent of the observations used and hence $Cov\left(\hat{\beta}_1,\ e_p\right) = 0$. You will learn more about this in some of our honours courses where we deal with the mathematical detail!

8.6 Relationship between tests for correlation and regression

Theorem 5.4 deals with a correlation problem and theorem 8.2 deals with a regression problem. The two t-statistics are computationally the same, however, as will now be shown.

For the correlation problem let

$$\begin{aligned} S_{11} &= \Sigma\left(Y_{1i} - \overline{Y}_1\right)^2 \\ S_{22} &= \Sigma\left(Y_{2i} - \overline{Y}_2\right)^2 \\ S_{12} &= \Sigma\left(Y_{1i} - \overline{Y}_1\right)\left(Y_{2i} - \overline{Y}_2\right) \end{aligned}$$

then $R = S_{12}/\sqrt{S_{11}S_{22}}$ and the t-statistic of theorem 5.4 for testing $H_0 : \rho = 0$ can be written

$$\begin{aligned} T &= \sqrt{n-2}\frac{S_{12}/\sqrt{S_{11}S_{22}}}{\sqrt{1-S_{12}^2/S_{11}S_{22}}} \\ &= \sqrt{n-2}\frac{S_{12}}{\sqrt{S_{11}S_{22}}} \cdot \frac{\sqrt{S_{11}S_{22}}}{\sqrt{S_{11}S_{22}-S_{12}^2}} \\ &= \sqrt{n-2}\frac{S_{12}}{\sqrt{S_{11}S_{22}-S_{12}^2}}. \end{aligned}$$

Likewise, for the regression model, let

$S_{11} = \Sigma\left(X_1 - \overline{X}\right)^2$
$S_{22} = \Sigma\left(Y_i - \overline{Y}\right)^2$
$S_{12} = \Sigma\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right) = \Sigma Y_i\left(X_i - \overline{X}\right)$ as can be shown easily.

Then

$$\begin{aligned} \hat{\beta}_1 &= \frac{S_{12}}{S_{11}} \\ d^2 &= S_{11} \\ S^2 &= \frac{\Sigma\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)^2}{(n-2)} \\ (n-2)\,S^2 &= \Sigma\left(Y_i - \overline{Y} + \hat{\beta}_1\overline{X} - \hat{\beta}_1 X_i\right)^2 \\ &= \Sigma\left[\left(Y_i - \overline{Y}\right) - \hat{\beta}_1\left(X_i - \overline{X}\right)\right]^2 \\ &= \Sigma\left(Y_i - \overline{Y}\right)^2 - 2\hat{\beta}_1\Sigma\left(Y_i - \overline{Y}\right)\left(X_i - \overline{X}\right) + \hat{\beta}_1^2\Sigma\left(X_i - \overline{X}\right)^2 \\ &= S_{22} - 2\frac{S_{12}}{S_{11}}S_{12} + \frac{S_{12}^2}{S_{11}^2}S_{11} \\ &= S_{22} - \frac{S_{12}^2}{S_{11}} \Rightarrow S^2 = \frac{S_{22} - \frac{S_{12}^2}{S_{11}}}{(n-2)} \end{aligned}$$

This is a very handy alternative formula if you do not like to compute $\hat{Y}_i$ for each different X_i.

The test statistic for testing $H_0 : \beta_1 = 0$ is

$$T_{1;0} = \frac{\hat{\beta}_1}{S/d} = \frac{S_{12}/S_{11}}{\sqrt{\left(S_{22} - S_{12}^2/S_{11}\right) / \left((n-2)\, S_{11}\right)}}$$

$$= \frac{\left(\frac{S_{12}}{S_{11}}\right)\sqrt{(n-2)\, S_{11}}}{\sqrt{\frac{S_{22}S_{11} - S_{12}^2}{S_{11}}}}$$

$$= \sqrt{n-2}\frac{S_{12}}{\sqrt{S_{11}S_{22} - S_{12}^2}}.$$

The two t-statistics are therefore computed in exactly the same way and their distributions under the two null hypotheses ($\rho = 0$ and $\beta_1 = 0$, respectively) are the same. When we draw inference on β_0 or $(\beta_0 + \beta_1 X)$, however, the control variable should not be a random variable, or at least the variance of X must be much smaller than the variance of Y.

The last section of this study unit is optional and you will not be examined on it. It does, however, give a smooth transition from second-year level statistics to third year if you intend to major in Statistics. You will appreciate the matrix approach when you start working with more complicated models.

8.7 Simple linear regression in matrix notation

In matrix notation we may write the simple linear regression model as

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} E_1 \\ E_2 \\ \vdots \\ E_n \end{bmatrix}$$

or $\underline{y} = X\underline{\beta} + \underline{e}$, say.

(Please note that we now use small letters for the independent X-variates because we reserve the capital letter X for the so-called design matrix.)
The least squares criterion states that we should minimise

$$\left(\underline{y} - X\underline{\beta}\right)' \left(\underline{y} - X\underline{\beta}\right)$$

which is the same as $\Sigma \left(Y_i - \beta_0 - \beta_1 x_i\right)^2$.

The first derivative with respect to $\underline{\beta}$ is set equal to zero to obtain

$$\left(X'X\right)\underline{\beta} = X'\underline{y}$$

which is the same as

$$\begin{bmatrix} n & \Sigma x_i \\ \Sigma x_i & \Sigma x_i^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} \Sigma Y_i \\ \Sigma x_i Y_i \end{bmatrix}.$$

The solution yields the least squares estimators:

$$\begin{aligned} \hat{\underline{\beta}} &= (X'X)^{-1} X'\underline{y} \\ &= \frac{1}{n\Sigma x_i^2 - (\Sigma x_i)^2} \begin{bmatrix} \Sigma x_i^2 & -\Sigma x_i \\ -\Sigma x_i & n \end{bmatrix} \begin{bmatrix} \Sigma Y_i \\ \Sigma x_i Y_i \end{bmatrix} \end{aligned}$$

which yields the same estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ as before. To find the covariance matrix of $\hat{\underline{\beta}}$, note that the covariance matrix of $\underline{y}$ is

$$Cov\left(\underline{y},\ \underline{y}'\right) = \sigma^2 I_n$$

while $\hat{\underline{\beta}} = A\underline{y}$ where $A = (X'X)^{-1} X'$.

$$\begin{aligned} \therefore Cov\left(\hat{\underline{\beta}},\ \hat{\underline{\beta}}'\right) &= ACov\left(\underline{y},\ \underline{y}'\right) A' \\ &= \sigma^2 (X'X)^{-1} X' I \left[(X'X)^{-1} X'\right]' \\ &= \sigma^2 (X'X)^{-1} X'X (X'X)^{-1} \\ &= \sigma^2 (X'X)^{-1} \\ &= \frac{\sigma^2}{n\Sigma (x_i - \overline{x})^2} \begin{bmatrix} \Sigma x_i^2 & -\Sigma x_i \\ -\Sigma x_i & n \end{bmatrix} \end{aligned}$$

which is what we obtained before.

Exercise 8.1

1. The amount of lime in a concrete mixture (X) and the hardness of the mixture (Y) were measured and found to be as follows:

X	Y		
5	1.041;	1.047;	1.023
10	1.060;	1.050;	1.069
15	1.046;	1.075;	1.054
20	1.066;	1.075;	1.077
25	1.080;	1.069;	1.073
30	1.095;	1.061;	1.069

   Find a suitable regression model.

2. A study was done to find the effect of temperature (X) on the yield (Y) of a chemical process. The following data were obtained (where $X = (\text{temp} - 300)/20$).

X	-5	-4	-3	-2	-1	0	1	2	3	4	5
Y	12	7	20	13	21	18	22	18	28	32	29

   (a) Plot the data to verify that simple linear regression is a suitable model.

   (b) Estimate β_0, β_1 and σ^2.

   (c) Find a 95% confidence interval for β_1.

   (d) Find a 95% confidence interval for $\beta_0 + 2\beta_1$ (ie the mean yield at 340°C).

   (e) Find a 95% confidence interval for the yield if a further experiment is performed at 360°C.

3. Eleven plots of land were each treated with a certain dosage of fertilizer (X) and the yield (Y) recorded:

X	2	3	4	1	0	2	4	0	1	2	3
Y	30	36	47	28	12	31	54	5	18	29	42

   (a) Plot the points to determine whether a straight line would represent an adequate model.

   (b) Compute the regression line and draw it on the graph.

   (c) Assume normality and find a 95% confidence interval for the slope; interpret the result.

   (d) What is the expected yield at $X = 4$?

   (e) Find a 95% confidence interval for the expected yield at $X = 4$.

   (f) Find a 95% confidence interval for the yield which one may expect to obtain if, in a new experiment, a dosage of $X = 4$ is applied.

4. Let $Y_i \sim n\left(\beta_0 + \beta_1 X_i;\ \sigma^2\right);\ i = 1, ..., n;$ and let $\hat{\beta}_0$ and $\hat{\beta}_1$ be the least squares estimators for β_0 and β_1.

   (a) Write down $Var\left(\hat{\beta}_0 + \hat{\beta}_1 X\right)$ and show that this is a minimum at $X = \overline{X}$.

   (b) Calculate the covariance between $\hat{\beta}_0 + \hat{\beta}_1 X_1$ and $\hat{\beta}_0 + \hat{\beta}_1 X_2$.

   (c) Find k such that $\hat{\beta}_0 + \hat{\beta}_1\left(\overline{X} - k\right)$ and $\hat{\beta}_0 + \hat{\beta}_1\left(\overline{X} + k\right)$ are uncorrelated.

8.8 Learning outcomes

After studying study unit 8, you should **be able to**

- define the concepts *bivariate data analysis* and *regression experiment*
- draw a scatter plot of two numerical variables and describe the nature of the relationship between the two variables
- determine the coefficients of a linear equation using the method of least squares
- compute the estimate of the variance around the line and explain its use
- perform and interpret the hypothesis test $H_0 : \beta_0 = c$
- derive a confidence interval for the population regression *intercept*, β_0
- perform and interpret the hypothesis test $H_0 : \beta_1 = c$
- derive a confidence interval for the population regression *slope*
- derive confidence limits for the regression line
- derive confidence limits for a future observation of a value for a regression experiment
- explain the relationship between tests for *correlation* and tests for the regression *slope*

A. Solutions to exercises

Exercise 1.1

1. $T \sim t_{10}$

   $P(T \geq x) = 0.01$ with $v = 10$ and $p = 0.01$ we therefore find $x = 2.764$

   $\Longrightarrow P(T \geq 2.764) = 0.01$

   If $P(T \geq x) = 0.01 \Longrightarrow P(T \leq x) = 1 - 0.01 = 0.99$

   $\Longrightarrow P(T \leq 2.764) = 0.99$

   Since the t-distribution is symmetric

   $$P(T \geq x) = 0.01 \quad \Longrightarrow \quad P(T \leq -x) = 0.01$$
   $$\Longrightarrow \quad P(T \leq -2.764) = 0.01$$

   Now $P(T \leq -2.764) = 0.01$ and $P(T \geq 2.764) = 0.01$

   $\Longrightarrow P(-2.764 \leq T \leq 2.764) = 1 - (0.01 + 0.01)$

   Thus $P(-2.764 \leq T \leq 2.764) = 0.98$

   $P(T \geq x) = 0.25 \Longrightarrow x = 0.7$

   $P(T \geq 0.7) = 0.25$ by symmetry $P(T \leq -0.7) = 0.25$

   $\Longrightarrow P(-0.7 \leq T \leq 0.7) = 1 - (0.25 + 0.25)$

   $\Longrightarrow P(-0.7 \leq T \leq 0.7) = 0.5$

2. $X \sim F_{5;12}$

   $P(X > x) = 0.05$ with $v_1 = 5$ and $v_2 = 12$

   $\Longrightarrow x = 3.11$ (Note: $F_{0.05;5;12} = 3.11$)

   Thus $P(X > 3.11) = 0.05$

   $P(X > x) = 0.025$ $F_{0.025;5;12} = 3.89$

   $$\Longrightarrow P(X > 3.89) = 0.025$$
   $$P(X < 3.89) = 1 - P(X > 3.89)$$
   $$= 1 - 0.025$$
   $$= 0.975$$

Thus $P(X < 3.89) = 0.975$

$$
\begin{aligned}
P(X > x) = 0.01 \qquad & F_{0.01;5;12} = 5.06 \\
\implies P(X > 5.06) = 0.01 \implies P(X < 5.06) &= 1 - P(X > 5.06) \\
P(X < 3.89) &= 1 - 0.01 \\
&= 0.99
\end{aligned}
$$

Thus, $P(X < 5.06) = 0.99$

* * * * *

Exercise 2.1

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

$$
\begin{aligned}
E(T) &= E\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2\right) \\
&= E\left(\frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2 - 2\overline{X}\sum_{i=1}^{n}X_i + \sum_{i=1}^{n}\overline{X}^2\right)\right) \\
&= E\left(\frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2 - 2n\overline{X}^2 + n\overline{X}^2\right)\right) \\
&= E\left(\frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2 - n\overline{X}^2\right)\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}E\left(X_i^2\right) - nE\left(\overline{X}^2\right)\right)
\end{aligned}
$$

Now

$$
\begin{aligned}
Var(X_i) &= E(X_i^2) - \mu^2 \\
\implies E(X_i^2) &= Var(X_i) + \mu^2 \\
&= \theta + \mu^2
\end{aligned}
$$

$$
\begin{aligned}
Var(\overline{X}) &= E(\overline{X}^2) - \mu^2 \\
\implies E(\overline{X}^2) &= Var(\overline{X}) + \mu^2 \\
&= \frac{\theta}{n} + \mu^2
\end{aligned}
$$

Thus

$$\begin{aligned}
E(T) &= \frac{1}{n-1}\left(\sum_{i=1}^{n} E\left(X_i^2\right) - nE\left(\overline{X}^2\right)\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}\left(\theta+\mu^2\right) - n\left(\frac{\theta}{n}+\mu^2\right)\right) \\
&= \frac{1}{n-1}\left(n\theta + n\mu^2 - \frac{n\theta}{n} - n\mu^2\right) \\
&= \frac{1}{n-1}\left(n\theta - \theta\right) \\
&= \frac{\theta}{n-1}\left(n-1\right) \\
&= \theta
\end{aligned}$$

2. $T_1 = \frac{1}{n}\sum_{i}^{n}\left(X_i-\mu\right)^2$ and $T_2 = \frac{1}{n-1}\sum_{i}^{n}\left(X_i-\overline{X}\right)^2$

$$\begin{aligned}
E(T_1) &= E\left(\frac{1}{n}\sum_{i}^{n}\left(X_i-\mu\right)^2\right) \\
&= E\left(\frac{1}{n}\sum_{i}^{n}\left(X_i-\mu\right)\left(X_i-\mu\right)\right) \\
&= E\left(\frac{1}{n}\left(\sum_{i}^{n}\left(X_i^2 - 2\mu X_i + \mu^2\right)\right)\right) \\
&= E\left(\frac{1}{n}\left(\sum_{i}^{n} X_i^2 - 2\mu\sum_{i}^{n} X_i + \sum_{i}^{n}\mu^2\right)\right) \\
&= \frac{1}{n}\left(\sum_{i}^{n} E\left(X_i^2\right) - 2\mu\sum_{i}^{n} E\left(X_i\right) + \sum_{i}^{n} E\left(\mu^2\right)\right) \\
&= \frac{1}{n}\left(\sum_{i}^{n}\left(\theta+\mu^2\right) - 2\mu\sum_{i}^{n}\mu + \sum_{i}^{n}\mu^2\right) \\
&= \frac{1}{n}\left(n\theta + n\mu^2 - 2\mu n\mu + n\mu^2\right) \\
&= \frac{1}{n}\left(n\theta + n\mu^2 - 2n\mu^2 + n\mu^2\right) \\
&= \frac{1}{n}\left(n\theta\right) \\
&= \theta
\end{aligned}$$

$$\begin{aligned} E(T_2) &= E\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2\right) \\ &= \theta \text{ result from 1} \end{aligned}$$

Note $\sum_{1}^{n}\frac{(X_i-\mu)^2}{\sigma^2}\sim\chi_n^2$ and $\sum\frac{\left(X_i-\overline{X}\right)^2}{\sigma^2}\sim\chi_{n-1}^2$

Now

$$\begin{aligned} T_1 &= \frac{1}{n}\sum_{i}^{n}(X_i-\mu)^2 \\ &= \frac{\sigma^2}{n}\sum_{i}^{n}\frac{(X_i-\mu)^2}{\sigma^2} \text{ multiply numerator and denominator by } \sigma^2 \\ &= \frac{\sigma^2}{n}Y \text{ where } Y=\sum_{i}^{n}\frac{(X_i-\mu)^2}{\sigma^2} \end{aligned}$$

$$\begin{aligned} Var(T_1) &= Var\left(\frac{\sigma^2}{n}Y\right) \\ &= \frac{\sigma^4}{n^2}Var(Y) \text{ now } Var(Y)=2n \text{ since } Y\sim\chi_n^2 \\ &= \frac{\sigma^4}{n^2}\times 2n \\ &= \frac{2\sigma^4}{n} \end{aligned}$$

$$\begin{aligned} T_2 &= \frac{1}{n-1}\sum_{1}^{n}(X_i-\overline{X})^2 \\ &= \frac{\sigma^2}{n-1}\sum_{1}^{n}\frac{\left(X_i-\overline{X}\right)^2}{\sigma^2} \\ &= \frac{\sigma^2}{n-1}\sum_{1}^{n}\left(\frac{X_i-\overline{X}}{\sigma}\right)^2 \\ &= \frac{\sigma^2}{n-1}Y \text{ where } Y=\sum_{i}^{n}\left(\frac{X_i-\overline{X}}{\sigma}\right)^2 \text{. Thus } Y\sim\chi_{n-1}^2 \text{ and } Var(Y)=2(n-1) \end{aligned}$$

$$\begin{aligned} Var\left(T_2\right) &= Var\left(\frac{\sigma^2}{n-1}Y\right) \\ &= \frac{\sigma^4}{\left(n-1\right)^2}Var\left(Y\right) \\ &= \frac{\sigma^4}{\left(n-1\right)^2} \times 2\left(n-1\right) \\ &= \frac{2\sigma^4}{n-1} \end{aligned}$$

Thus $\frac{2\sigma^4}{n} < \frac{2\sigma^4}{n-1} \therefore T_1$ has a smaller variance than T_2.

3 $E(X_i) = c_i\theta_1 + c_i^2\theta_2$

$$\begin{aligned} Q\left(\theta\right) &= \sum_i^n \left(X_i - E(X_i)\right)^2 \\ &= \sum_i^n \left(X_i - \left(c_i\theta_1 + c_i^2\theta_2\right)\right)^2 \\ &= \sum_i^n \left(X_i - c_i\theta_1 - c_i^2\theta_2\right)^2 \end{aligned}$$

$$\begin{aligned} \frac{dQ}{d\theta_1} &= 2\sum_i^n \left(X_i - c_i\theta_1 - c_i^2\theta_2\right) \times -c_i \\ &= -2\sum_i^n \left(c_iX_i - c_i^2\theta_1 - c_i^3\theta_2\right) \end{aligned}$$

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

$$\begin{aligned} \Longrightarrow \quad 0 &= -2\sum_i^n \left(c_iX_i - c_i^2\theta_1 - c_i^3\theta_2\right) \\ &= -2\left(\sum_i^n c_iX_i - \theta_1\sum_i^n c_i^2 - \theta_2\sum_i^n c_i^3\right) \end{aligned}$$

Making θ_1 subject of the formula

$$\begin{aligned}\theta_1 \sum_i^n c_i^2 &= \sum_i^n c_i X_i - \theta_2 \sum_i^n c_i^3 \\ \theta_1 &= \frac{\sum_i^n c_i X_i - \theta_2 \sum_i^n c_i^3}{\sum_i^n c_i^2} \quad \text{.....(1)}\end{aligned}$$

Making θ_2 subject of the formula

$$\begin{aligned}\theta_2 \sum_i^n c_i^3 &= \sum_i^n c_i X_i - \theta_1 \sum_i^n c_i^2 \\ \theta_2 &= \frac{\sum_i^n c_i X_i - \theta_1 \sum_i^n c_i^2}{\sum_i^n c_i^3} \quad \text{.....(2)}\end{aligned}$$

Now

$$\begin{aligned}\frac{dQ}{d\theta_2} &= 2\sum_i^n \left(X_i - c_i\theta_1 - c_i^2\theta_2\right) \times -c_i^2 \\ &= -2\sum_i^n \left(c_i^2 X_i - c_i^3\theta_1 - c_i^4\theta_2\right) \\ &= -2\left(\sum_i^n c_i^2 X_i - \theta_1 \sum_i^n c_i^3 - \theta_2 \sum_i^n c_i^4\right) \\ 0 &= -2\left(\sum_i^n c_i^2 X_i - \theta_1 \sum_i^n c_i^3 - \theta_2 \sum_i^n c_i^4\right)\end{aligned}$$

Making θ_1 subject of the formula

$$\begin{aligned}\theta_1 \sum_i^n c_i^3 &= \sum_i^n c_i^2 X_i - \theta_2 \sum_i^n c_i^4 \\ \theta_1 &= \frac{\sum_i^n c_i^2 X_i - \theta_2 \sum_i^n c_i^4}{\sum_i^n c_i^3} \quad \text{.....(3)}\end{aligned}$$

Making θ_2 subject of the formula

$$\begin{aligned}
\theta_2 \sum_i^n c_i^4 &= \sum_i^n c_i^2 X_i - \theta_1 \sum_i^n c_i^3 \\
\theta_2 &= \frac{\sum_i^n c_i^2 X_i - \theta_1 \sum_i^n c_i^3}{\sum_i^n c_i^4} \quad \text{.....(4)}
\end{aligned}$$

Finding θ_1 by equating equations 2 and 4

$$\begin{aligned}
\frac{\sum_i^n c_i X_i - \theta_1 \sum_i^n c_i^2}{\sum_i^n c_i^3} &= \frac{\sum_i^n c_i^2 X_i - \theta_1 \sum_i^n c_i^3}{\sum_i^n c_i^4} \\
\sum_i^n c_i^4 \left(\sum_i^n c_i X_i - \theta_1 \sum_i^n c_i^2 \right) &= \sum_i^n c_i^3 \left(\sum_i^n c_i^2 X_i - \theta_1 \sum_i^n c_i^3 \right) \\
\left(\sum_i^n c_i^4 \right) \left(\sum_i^n c_i X_i \right) - \theta_1 \left(\sum_i^n c_i^2 \right) \left(\sum_i^n c_i^4 \right) &= \left(\sum_i^n c_i^2 X_i \right) \left(\sum_i^n c_i^3 \right) - \theta_1 \left(\sum_i^n c_i^3 \right)^2 \\
\left(\sum_i^n c_i^4 \right) \left(\sum_i^n c_i X_i \right) - \left(\sum_i^n c_i^2 X_i \right) \left(\sum_i^n c_i^3 \right) &= \theta_1 \left(\sum_i^n c_i^2 \right) \left(\sum_i^n c_i^4 \right) - \theta_1 \left(\sum_i^n c_i^3 \right)^2 \\
\left(\sum_i^n c_i^4 \right) \left(\sum_i^n c_i X_i \right) - \left(\sum_i^n c_i^2 X_i \right) \left(\sum_i^n c_i^3 \right) &= \theta_1 \left(\left(\sum_i^n c_i^2 \right) \left(\sum_i^n c_i^4 \right) - \left(\sum_i^n c_i^3 \right)^2 \right) \\
\widehat{\theta}_1 &= \frac{\left(\sum_i^n c_i X_i \right) \left(\sum_i^n c_i^4 \right) - \left(\sum_i^n c_i^2 X_i \right) \left(\sum_i^n c_i^3 \right)}{\left(\left(\sum_i^n c_i^2 \right) \left(\sum_i^n c_i^4 \right) - \left(\sum_i^n c_i^3 \right)^2 \right)}
\end{aligned}$$

Finding θ_2 by equating equations 1 and 3

$$\begin{aligned}
\frac{\sum_i^n c_i X_i - \theta_2 \sum_i^n c_i^3}{\sum_i^n c_i^2} &= \frac{\sum_i^n c_i^2 X_i - \theta_2 \sum_i^n c_i^4}{\sum_i^n c_i^3} \\
\sum_i^n c_i^3 \left(\sum_i^n c_i X_i - \theta_2 \sum_i^n c_i^3 \right) &= \sum_i^n c_i^2 \left(\sum_i^n c_i^2 X_i - \theta_2 \sum_i^n c_i^4 \right) \\
\left(\sum_i^n c_i X_i \right) \left(\sum_i^n c_i^3 \right) - \theta_2 \left(\sum_i^n c_i^3 \right)^2 &= \left(\sum_i^n c_i^2 X_i \right) \left(\sum_i^n c_i^2 \right) - \theta_2 \left(\sum_i^n c_i^4 \right) \left(\sum_i^n c_i^2 \right) \\
\left(\sum_i^n c_i X_i \right) \left(\sum_i^n c_i^3 \right) - \left(\sum_i^n c_i^2 X_i \right) \left(\sum_i^n c_i^2 \right) &= \theta_2 \left(\sum_i^n c_i^3 \right)^2 - \theta_2 \left(\sum_i^n c_i^4 \right) \left(\sum_i^n c_i^2 \right) \\
\left(\sum_i^n c_i X_i \right) \left(\sum_i^n c_i^3 \right) - \left(\sum_i^n c_i^2 X_i \right) \left(\sum_i^n c_i^2 \right) &= \theta_2 \left(\left(\sum_i^n c_i^3 \right)^2 - \left(\sum_i^n c_i^4 \right) \left(\sum_i^n c_i^2 \right) \right) \\
\widehat{\theta}_2 &= \frac{\left(\sum_i^n c_i X_i \right) \left(\sum_i^n c_i^3 \right) - \left(\sum_i^n c_i^2 X_i \right) \left(\sum_i^n c_i^2 \right)}{\left(\left(\sum_i^n c_i^3 \right)^2 - \left(\sum_i^n c_i^4 \right) \left(\sum_i^n c_i^2 \right) \right)}
\end{aligned}$$

4.

$$\begin{aligned}
Q(\theta) &= \sum_{i=1}^{n} (X_i - E(X_i))^2 \\
&= \sum_{i=1}^{n-1} (X_i - E(X_i))^2 + (X_n - E(X_n))^2 \\
&= \sum_{i=1}^{n-1} (X_i - \theta_1)^2 + (X_n - (\theta_1 + \theta_2))^2 \\
&= \sum_{i=1}^{n-1} (X_i - \theta_1)^2 + (X_n - (\theta_1 + \theta_2))^2 \\
&= \sum_{i=1}^{n-1} (X_i - \theta_1)^2 + (X_n - \theta_1 - \theta_2)^2 \\
\frac{dQ}{d\theta_1} &= 2 \sum_{i=1}^{n-1} (X_i - \theta_1) \times -1 + 2 (X_n - \theta_1 - \theta_2) \times -1 \\
&= -2 \left(\sum_{i=1}^{n-1} (X_i - \theta_1) + X_n - \theta_1 - \theta_2 \right)
\end{aligned}$$

$$
\begin{aligned}
0 &= -2\left(\sum_{i=1}^{n-1}(X_i-\theta_1)+X_n-\theta_1-\theta_2\right)\\
0 &= \sum_{i=1}^{n-1}(X_i-\theta_1)+X_n-\theta_1-\theta_2\\
0 &= \sum_{i=1}^{n-1}X_i-\sum_{i=1}^{n-1}\theta_1+X_n-\theta_1-\theta_2\\
0 &= \sum_{i=1}^{n-1}X_i-(n-1)\,\theta_1-\theta_1+X_n-\theta_2\\
0 &= \sum_{i=1}^{n-1}X_i-n\theta_1+\theta_1-\theta_1+X_n-\theta_2\\
0 &= \sum_{i=1}^{n-1}X_i-n\theta_1+X_n-\theta_2\\
0 &= \sum_{i=1}^{n}X_i-n\theta_1-\theta_2\\
\Longrightarrow\ \widehat{\theta}_1 &= \frac{\sum_{i=1}^{n}X_i-\theta_2}{n}\text{.....(1)}\\
\widehat{\theta}_2 &= \sum_{i=1}^{n}X_i-n\theta_1\text{.....(2)}
\end{aligned}
$$

$$
\begin{aligned}
\frac{dQ}{d\theta_2} &= 2\left(X_n-\theta_1-\theta_2\right)\times -1\\
0 &= -2\left(X_n-\theta_1-\theta_2\right)\\
0 &= X_n-\theta_1-\theta_2\\
\Longrightarrow\ \theta_1 &= X_n-\theta_2\text{.....(3)}\\
\Longrightarrow\ \theta_2 &= X_n-\theta_1\text{.....(4)}
\end{aligned}
$$

Equating 2 and 4

$$\begin{aligned}
X_n - \theta_1 &= \sum_{i=1}^{n} X_i - n\theta_1 \\
n\theta_1 - \theta_1 &= \sum_{i=1}^{n} X_i - X_n \\
\theta_1 (n-1) &= \sum_{i=1}^{n-1} X_i + X_n - X_n \\
\theta_1 (n-1) &= \sum_{i=1}^{n-1} X_i \\
\theta_1 &= \frac{\sum_{i=1}^{n-1} X_i}{n-1} \\
\theta_1 &= \overline{X}
\end{aligned}$$

Equating 1 and 3

$$\begin{aligned}
\frac{\sum_{i=1}^{n} X_i - \theta_2}{n} &= X_n - \theta_2 \\
\sum_{i=1}^{n} X_i - \theta_2 &= nX_n - n\theta_2 \\
\sum_{i=1}^{n} X_i - nX_n &= \theta_2 - n\theta_2 \\
\sum_{i=1}^{n} X_i - nX_n &= \theta_2 (1-n) \\
nX_n - \sum_{i=1}^{n} X_i &= \theta_2 (n-1) \\
nX_n - X_n - \sum_{i=1}^{n-1} X_i &= \theta_2 (n-1) \\
X_n (n-1) - \sum_{i=1}^{n-1} X_i &= \theta_2 (n-1) \\
X_n - \frac{\sum_{i=1}^{n-1} X_i}{n-1} &= \theta_2 \\
X_n - \overline{X} &= \theta_2 \\
&\Longrightarrow \ \widehat{\theta}_2 = X_n - \widehat{\theta}_1
\end{aligned}$$

5. $f_X(x_i,\theta)=\frac{1}{\sqrt{2\pi\theta}}e^{\frac{-\frac{1}{2}(x_i-\mu)^2}{\theta}}$

The maximum likelihood is

$$\begin{aligned}
L(\theta) &= \prod_{i=1}^{n} f_X(x_i,\theta) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\theta}} e^{\frac{-\frac{1}{2}(x_i-\mu)^2}{\theta}} \\
&= \frac{1}{(2\pi\theta)^{n/2}} e^{\frac{-\frac{1}{2}\sum(x_i-\mu)^2}{\theta}} \\
Log\,L(\theta) &= \frac{-n}{2}\log 2\pi - \frac{-n}{2}\log\theta - \frac{1}{2}\frac{\sum(X_i-\mu)^2}{\theta} \\
\frac{dLog\,L(\theta)}{d\theta} &= \frac{-n}{2\theta} - \frac{1}{2}\times -1\left(\frac{\sum(X_i-\mu)^2}{\theta^2}\right) \\
0 &= \frac{-n}{2\theta} + \frac{1}{2}\left(\frac{\sum(X_i-\mu)^2}{\theta^2}\right) \\
\frac{n}{2\theta} &= \frac{1}{2}\left(\frac{\sum(X_i-\mu)^2}{\theta^2}\right) \\
\frac{n}{\theta} &= \frac{\sum(X_i-\mu)^2}{\theta^2} \\
n\theta &= \sum(X_i-\mu)^2 \\
\widehat{\theta} &= \frac{1}{n}\sum(X_i-\mu)^2
\end{aligned}$$

6. $f_X(x,\theta)=\theta(1-\theta)^{x-1}$ $\qquad x>1$

The maximum likelihood is

$$\begin{aligned}
L(\theta) &= \prod_{i=1}^{n} f_X(x_i,\theta) \\
&= \prod_{i=1}^{n} \theta(1-\theta)^{X_i-1} \\
&= \theta^n(1-\theta)^{\sum X_i-n} \\
Log\,L(\theta) &= n\log\theta + \left(\sum X_i - n\right)\log(1-\theta) \\
\frac{dLog\,L(\theta)}{d\theta} &= \frac{n}{\theta} + \frac{1}{1-\theta}\times\left(\sum X_i - n\right)\times -1 \\
0 &= \frac{n}{\theta} - \frac{1}{1-\theta}\left(\sum X_i - n\right) \\
\frac{1}{1-\theta}\left(\sum X_i - n\right) &= \frac{n}{\theta} \\
\theta\left(\sum X_i - n\right) &= n(1-\theta) \\
\theta\sum X_i - n\theta &= n - n\theta \\
\theta\sum X_i - n\theta + n\theta &= n
\end{aligned}$$

$$\begin{aligned}\theta\sum X_i &= n\\ \theta &= \frac{n}{\sum X_i}\\ \widehat{\theta} &= \frac{1}{\overline{X}}\end{aligned}$$

7. $f_X(x,\theta)=\frac{1}{\theta}e^{-\frac{x}{\theta}}\qquad x>0$

The maximum likelihood is

$$\begin{aligned}L(\theta) &= \prod_{i=1}^{n} f_X(x_i,\theta)\\ &= \prod_{i=1}^{n}\frac{1}{\theta}e^{-\frac{x_i}{\theta}}\\ &= \frac{1}{\theta^n}e^{-\frac{\sum X_i}{\theta}}\\ Log\,L(\theta) &= -n\log\theta-\frac{\sum X_i}{\theta}\\ \frac{dLog\,L(\theta)}{d\theta} &= \frac{-n}{\theta}+\frac{\sum X_i}{\theta^2}\\ 0 &= \frac{-n}{\theta}+\frac{\sum X_i}{\theta^2}\\ 0 &= -n\theta+\sum X_i\\ n\theta &= \sum X_i\\ \theta &= \frac{\sum X_i}{n}\\ \widehat{\theta} &= \overline{X}\end{aligned}$$

8. $U(\theta)$ follows a normal distribution

$$\begin{aligned}1-\alpha &= P(-z\le U(\theta)\le z)\\ 0.95 &= P(-1.96\le U(\theta)\le 1.96)\\ &= P(-1.96\le\frac{\overline{X}-\theta}{\frac{\sigma}{\sqrt{n}}}\le 1.96)\\ &= P\left(-1.96\times\frac{\sigma}{\sqrt{n}}\le\overline{X}-\theta\le 1.96\times\frac{\sigma}{\sqrt{n}}\right)\\ &= P\left(-\overline{X}-1.96\times\frac{\sigma}{\sqrt{n}}\le\theta\le-\overline{X}+1.96\times\frac{\sigma}{\sqrt{n}}\right)\\ &= P\left(\overline{X}-1.96\times\frac{\sigma}{\sqrt{n}}\le\theta\le\overline{X}+1.96\times\frac{\sigma}{\sqrt{n}}\right)\end{aligned}$$

∴The interval is $\left(\overline{X}-1.96\times\frac{\sigma}{\sqrt{n}};\overline{X}-1.96\times\frac{\sigma}{\sqrt{n}}\right)$

* * * * *

Exercise 4.1

1. We want to test

   $H_0 : \pi_1 = 0.10; \pi_2 = 0.20; \pi_3 = 0.20;$ and $\pi_4 = 0.5.$

   H_1 : At least one of the proportions is different from the one specified above.

   Now $N_1 = 125; N_2 = 185; N_3 = 230; N_4 = 460.$

   Thus,

$$\begin{aligned} n &= N_1 + N_2 + N_3 + N_4 \\ &= 125 + 185 + 230 + 460 \\ &= 1\,000. \end{aligned}$$

   The expected values are

$$n\pi_1 = 1\,000 \times 0.1 = 100; \qquad n\pi_2 = 1\,000 \times 0.2 = 200$$
$$n\pi_3 = 1\,000 \times 0.2 = 200; \qquad n\pi_4 = 1\,000 \times 0.5 = 500.$$

   The test statistic

$$\begin{aligned} Y^2 &= \sum_{i=1}^{4} \frac{(N_i - n\pi_i)^2}{n\pi_i} \\ &= \frac{(125-100)^2}{100} + \frac{(185-200)^2}{200} + \frac{(230-200)^2}{200}\frac{(460-500)^2}{500} \\ &= 6.25 + 1.125 + 4.5 + 3.2 \\ &= 15.075. \end{aligned}$$

   From table IV, we see that $\chi^2_{\alpha;k-1} = \chi^2_{0.01;3} = 11.3449$. We will reject H_0 if $Y^2 \geq 11.3449$.

   Since $15.075 > 11.3449$, we reject H_0 at the 1% level and conclude that at least one of the proportions is different from the one specified.

2. We want to test H_0 : The probabilities for the four classes will be in the ratio $\theta : 2\theta : 4\theta : 1 - 7\theta$.

   Let N_0, N_1, N_2 and N_3 denote the plants bearing white flowers, yellow flowers, orange flowers and those that fail to germinate, respectively, where $N = N_0 + N_1 + N_2 + N_3$.

Estimate θ according to the maximum likelihood method gives

$$\begin{aligned}
L(\theta) &= \prod_{i=1}^{n} P(X_i = r_i) \\
&= \underbrace{\theta\theta\theta...\theta}_{N_0 \text{ times}} \times \underbrace{2\theta 2\theta....2\theta}_{N_1 \text{ times}} \times \underbrace{4\theta 4\theta....4\theta}_{N_2 \text{ times}} \times \underbrace{(1-7\theta).....(1-7\theta)}_{N_3 \text{ times}} \\
&= (\theta)^{N_0}(2\theta)^{N_1}(4\theta)^{N_2}(1-7\theta)^{N_3} \\
Log\ L(\theta) &= N_0 \log\ \theta + N_1\ \log\ 2\theta + N_2\ \log\ 4\theta + N_3\ \log\ (1-7\theta) \\
\frac{dLog\ L(\theta)}{d\theta} &= \frac{N_0}{\theta} + \frac{2N_1}{2\theta} + \frac{4N_2}{4\theta} + \frac{-7N_3}{1-7\theta}.
\end{aligned}$$

Setting $\frac{dLog\ L(\theta)}{d\theta} = 0$

$$\begin{aligned}
0 &= \frac{N_0}{\theta} + \frac{2N_1}{2\theta} + \frac{4N_2}{4\theta} - \frac{7N_3}{1-7\theta} \\
0 &= \frac{N_0}{\theta} + \frac{N_1}{\theta} + \frac{N_2}{\theta} - \frac{7N_3}{1-7\theta} \\
\frac{7N_3}{1-7\theta} &= \frac{N_0}{\theta} + \frac{N_1}{\theta} + \frac{N_2}{\theta} \\
7N_3 &= (1-7\theta)\left(\frac{N_0+N_1+N_2}{\theta}\right) \\
7\theta N_3 &= (1-7\theta)(N_0+N_1+N_2) \\
7\theta N_3 &= N_0+N_1+N_2-7N_0\theta-7N_1\theta-7N_2\theta \\
7N_0\theta+7N_1\theta+7N_2\theta+7\theta N_3 &= N_0+N_1+N_2 \\
7\theta(N_0+N_1+N_2+N_3) &= N_0+N_1+N_2 \\
\widehat{\theta} &= \frac{N_0+N_1+N_2}{7(N_0+N_1+N_2+N_3)}
\end{aligned}$$

In this case

$$\begin{aligned}
\widehat{\theta} &= \frac{16+28+40}{7(100)} \\
&= \frac{84}{700} \\
&= 0.12.
\end{aligned}$$

The estimated probabilities are therefore $\widehat{\theta} = 0.12;\ 2\widehat{\theta} = 0.24;\ 4\widehat{\theta} = 0.48; 1-7\widehat{\theta} = 0.16$

The expected frequencies are $n\widehat{\pi}_i$

Class	Observed frequencies	Expected frequencies
White	16	12
Yellow	28	24
Orange	40	48
Fail to germinate	16	16

Therefore

$$\begin{aligned} Y^2 &= \sum_{i=1}^{4} \frac{(N_i - n\widehat{\pi}_i)^2}{n\widehat{\pi}_i} \\ &= \frac{(16-12)^2}{12} + \frac{(28-24)^2}{24} + \frac{(40-48)^2}{48} + \frac{(16-16)^2}{16} \\ &= 1.3333 + 0.6667 + 1.3333 + 0 \\ &= 3.3333. \end{aligned}$$

We have $k - r - 1 = 4 - 1 - 1 = 3$ degrees of freedom (one parameter estimated) and $\chi^2_{0.05;2} = 5.99147$. We reject H_0 if $Y^2 \geq 5.991$.

Since $3.3333 < 5.99147$, H_0 cannot be rejected at the 5% level, thus the seed man's claim may be true, that is, there is no sufficient evidence to refute the seed man's claim.

3.

H_0 : The sample comes from a $n(\mu, 100)$ distribution.
H_1 : The sample does not come from a $n(\mu, 100)$ distribution.

X-interval	Z-interval	Expected probability (π_i)	Observed frequency	Expected frequency
$X < 3.255$	$Z < -0.67$	0.2514	7	10.056
$3.255 \leq X < 10$	$-0.67 \leq Z \leq 0$	0.2486	6	9.944
$10 \leq X < 16.745$	$0 \leq Z < 0.67$	0.2486	15	9.944
$X \geq 16.745$	$Z \geq 0.67$	0.2514	12	10.056

$$\begin{aligned} Y^2 &= \sum_{i=1}^{4} \frac{(N_i - \widehat{e}_i)^2}{\widehat{e}_i} \\ &= \frac{(7-10.056)^2}{10.056} + \frac{(6-9.944)^2}{9.944} + \frac{(15-9.944)^2}{9.944} + \frac{(12-10.056)^2}{10.056} \\ &= 0.9287 + 1.5643 + 2.5707 + 0.3758 \\ &= 5.4395 \end{aligned}$$

(a) $\chi^2_{\alpha;k-1} = \chi^2_{0.10;3} = 6.25139$

$\chi^2_{\alpha;k-r-1} = \chi^2_{0.10;2} = 4.60517$

Since $4.60517 < 5.4395 < 6.25139$, the decision is uncertain at the 10% level of significance.

(b) $\chi^2_{\alpha;k-1} = \chi^2_{0.05;3} = 7.81473$

$\chi^2_{\alpha;k-r-1} = \chi^2_{0.05;2} = 5.99147$

Since $5.4395 < 5.99147$, we do not reject H_0 at the 5% and conclude that the sample is from a normal distribution with $\sigma^2 = 100$ in other words, $n(\mu, 100)$.

4. (a)

H_0 : The sample comes from a Poisson distribution with $\theta = 2$.
H_1 : The sample does not come from a Poisson distribution with $\theta = 2$.

Numbers	Observed frequency	Expected probability, (π_i)	Expected frequency
0	21	0.1353	14
1	30	0.2707	27
2	27	0.2707	27
3	16	0.1804	18
4	3	0.0902	9
5	2	0.0361	4
6	1	0.0121	1

Because of small expected frequencies, pool the classes "$X = 5$" and "$X = 6$".

$$\begin{aligned} Y^2 &= \sum_{i=1}^{k} \frac{(N_i - n\pi_i)^2}{n\pi_i} \\ &= \frac{(21-14)^2}{14} + \frac{(30-27)^2}{27} + \frac{(27-27)^2}{27} + \frac{(16-18)^2}{18} + \frac{(3-9)^2}{9} + \frac{(3-5)^2}{5} \\ &= 3.5 + 0.3333 + 0 + 0.2222 + 4 + 0.8 \\ &= 8.8555 \end{aligned}$$

Reject H_0 if $Y^2 \geq \chi^2_{0.05;5} = 11.0705$. Since $Y^2 = 8.8555 < 11.0705$, we do not reject H_0 at the 5% level of significance and conclude that the data come from a Poisson distribution with $\theta = 2$.

(b)

H_0 : The data come from a Poisson distribution.
H_1 : The data do not come from a Poisson distribution.

The maximum likelihood estimator is $\widehat{\theta} = \overline{X}$.

$$\begin{aligned} \Longrightarrow \quad \widehat{\theta} &= \frac{(0 \times 21) + (1 \times 30) + (2 \times 27) + (3 \times 16) + (4 \times 3) + (5 \times 2) + (6 \times 1)}{100} \\ &= \frac{160}{100} \\ &= 1.6 \end{aligned}$$

Numbers	Observed frequency	Expected probability, (π_i)	Expected frequency
0	21	0.2019	20
1	30	0.323	32
2	27	0.2584	26
3	16	0.1378	14
4	3	0.0551	6
5	2	0.0176	2
6	1	0.0047	0

Pool "$X = 4$", "$X = 5$" and "$X = 6$".

$$\begin{aligned} Y^2 &= \sum_{i=1}^{k} \frac{(N_i - n\widehat{\pi}_i)^2}{n\widehat{\pi}_i} \\ &= \frac{(21-20)^2}{20} + \frac{(30-32)^2}{32} + \frac{(27-26)^2}{26} + \frac{(16-14)^2}{14} + \frac{(6-8)^2}{8} \\ &= 0.05 + 0.125 + 0.0385 + 0.2857 + 0.5 \\ &= 0.9992 \end{aligned}$$

Reject H_0 if $Y^2 \geq \chi^2_{0.05;5-1-1} = Y^2 \geq \chi^2_{0.05;3} = 7.81473$. Since $Y^2 = 0.9992 < 7.81473$, we cannot reject H_0 at the 5% level of significance and conclude that the data come from a Poisson distribution. Thus, $\hat{\theta} = 1.6$.

5.

H_0 : The sample comes from a $n(32, 64)$ distribution.
H_1 : The sample does not come from a $n(32, 64)$ distribution

Lifetime X-interval	Z-interval	Expected probability, (π_i)	Observed frequency	Expected frequency
less than 16	$Z < -2$	0.0228	6	2.28
16 to 20	$-2 < Z < -1.5$	0.044	9	4.4
20 to 24	$-1.5 < Z < -1$	0.0919	12	9.19
24 to 28	$-1 < Z < -0.5$	0.1498	16	$14.98 \longrightarrow 15$
28 to 32	$-0.5 < Z < 0$	0.1915	20	$19.15 \longrightarrow 19$
32 to 36	$0 < Z < 0.5$	0.1915	22	$19.15 \longrightarrow 19$
36 to 40	$0.5 < Z < 1$	0.1498	10	$14.98 \longrightarrow 15$
above 40	$Z > 1$	0.1587	5	$15.87 \longrightarrow 16$

Pool first three classes.

Expected frequencies: $16, 15, 19, 19, 15, 16$

$$\begin{aligned} Y^2 &= \sum_{i=1}^{k} \frac{(N_i - n\widehat{\pi}_i)^2}{n\widehat{\pi}_i} \\ &= \frac{(27-16)^2}{16} + \frac{(16-15)^2}{15} + \frac{(20-19)^2}{19} + \frac{(22-19)^2}{19} + \frac{(10-15)^2}{15} + \frac{(5-16)^2}{16} \\ &= 7.5625 + 0.0667 + 0.0526 + 0.4737 + 1.6667 + 7.5625 \\ &= 17.3847 \end{aligned}$$

Reject H_0 if $Y^2 \geq \chi^2_{0.05;5} = 11.0705$. Since $Y^2 = 17.3847 > 11.0705$, we reject H_0 at the 5% level of significance and conclude that the sample does not come from a $n(32, 64)$ distribution.

6.

H_0 : The sample comes from a $n(25, 12^2)$ distribution.
H_1 : The sample does not come from a $n(25, 12^2)$ distribution.

The probability of each interval is $\frac{1}{5} = 0.2$

We know that

$$\begin{aligned} P(Z \leq -0.842) &= 0.2 \\ P(Z \leq -0.253) &= 0.4 \\ P(Z \leq 0.253) &= 0.6 \\ P(Z \leq 0.842) &= 0.8 \end{aligned}$$

Z-interval	Equal probability intervals	Expected frequency, e_i	Tally	Observed frequency, N_i	$(N_i - \widehat{e}_i)$
$Z \leq -0.842$	$-\infty < X \leq 14.896$	6	𝍸 |	8	2
$-0.842 \leq Z \leq -0.253$	$14.896 < X \leq 21.964$	6	𝍸	5	-1
$-0.253 \leq Z \leq 0.253$	$21.964 < X \leq 28.036$	6	|||	3	-3
$0.253 \leq Z \leq 0.842$	$28.036 < X \leq 35.104$	6	𝍸	5	-1
$Z \geq 0.842$	$35.104 < X \leq \infty$	6	𝍸 ||||	9	3
	Total	30		30	

$$\begin{aligned} Y^2 &= \sum_{i=1}^{k} \frac{(N_i - e_i)^2}{e_i} \\ &= \frac{4}{6} + \frac{1}{6} + \frac{9}{6} + \frac{1}{6} + \frac{9}{6} \\ &= \frac{24}{6} \\ &= 4 \end{aligned}$$

We reject H_0 at the 10% level if $Y^2 \geq \chi^2_{\alpha;k-1} = \chi^2_{0.10;4} = 7.77944$. Since $Y^2 = 4 < 7.77944$, we do not reject H_0 at the 10% level of significance and conclude that the data come from a $n(25, 12^2)$ distribution.

7. We have to test $H_0 : \beta_2 = 3$ against $H_1 : \beta_2 \neq 3$.

We will reject H_0 at the 10% level of significance (two-sided) if $A > 0.9073$ or if $A < 0.7153$.

Since $n = 11 < 50$, we will use the test statistic A, that is, $A = \dfrac{\frac{1}{n}\sum_{i=1}^{n}\left|X_i - \overline{X}\right|}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}}$

where

$$\begin{aligned} \overline{X} &= \tfrac{220}{11} &&= 20 \\ \sum_{i=1}^{n}\left|X_i - \overline{X}\right| &= |-3| + |2| + \cdots + |4| &&= 40 \\ \sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 &= \sum X_i^2 - n\overline{X}^2 &&= 4\,576 - 11(20)^2 = 176 \end{aligned}$$

$$A = \frac{\frac{1}{n}\sum_{i=1}^{n}\left|X_i - \overline{X}\right|}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}} = \frac{\frac{1}{11}(40)}{\sqrt{\frac{1}{11}(176)}} = \frac{3.636363636}{4} \approx 0.9091.$$

Since $0.9091 > 0.9073$, we reject H_0 at the 10% level and conclude that the kurtosis of the sample is significantly different from the kurtosis of the normal distribution.

8.

$$\overline{X} = 25 \qquad \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 = 200$$

$$\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^3 = -320 \qquad \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^4 = 4\,000$$

We have to test for skewness and kurtosis.

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 $|\beta_1| > 0.534$ (in other words if $\beta_1 < -0.534$ or if $\beta_1 > 0.534$).

$$\begin{aligned}
\beta_1 &= \frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^3}{\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2\right]^{\frac{3}{2}}} \\
&= \frac{\frac{1}{50}(-320)}{\left[\frac{1}{50}(200)\right]^{\frac{3}{2}}} \\
&= \frac{-6.4}{[4]^{\frac{3}{2}}} \\
&= \frac{-6.4}{8} \\
&= -0.8
\end{aligned}$$

Since $-0.8 < -0.534$, we reject H_0 at the 10% level.

Test for kurtosis:

We have to test $H_0 : \beta_2 = 3$ against $H_1 : \beta_2 \neq 3$.

We will reject H_0 if $\beta_2 > 3.99$ or $\beta_2 < 2.15$.

$$\begin{aligned}
\beta_2 &= \frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^4}{\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2\right]^{2}} \\
&= \frac{\frac{1}{50}(4\,000)}{\left[\frac{1}{50}(200)\right]^{2}} \\
&= \frac{80}{16} \\
&= 5
\end{aligned}$$

Since $5 > 3.99$, we reject H_0 at the 10% level.

The sample failed both tests and hence we conclude that the sample is not from a normal distribution.

9.

$$\overline{X} = 50 \qquad \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = 16$$

$$\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^3 = 6.4 \qquad \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^4 = 819.2$$

We have to test for skewness and kurtosis.

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 $\beta_1 < -0.127$ or if $\beta_1 > 0.127$ in other words if $|\beta_1| > 0.127$.

$$\begin{aligned}
\beta_1 &= \frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^3}{\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2\right]^{\frac{3}{2}}} \\
&= \frac{6.4}{[16)]^{\frac{3}{2}}} \\
&= \frac{6.4}{64} \\
&= 0.1
\end{aligned}$$

Since $0.1 < 0.127$, we do not reject H_0 at the 10% level.

Test for kurtosis:

We have to test $H_0 : \beta_2 = 3$ against $H_1 : \beta_2 \neq 3$.

We will reject H_0 if $\beta_2 > 3.26$ or $\beta_2 < 2.76$.

$$\begin{aligned}
\beta_2 &= \frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^4}{\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2\right]^{2}} \\
&= \frac{819.2}{[16]^2} \\
&= 3.2
\end{aligned}$$

Since $2.76 < 3.2 < 3.26$, we do not reject H_0 at the 10% level.

The sample shows that the data are from a normal distribution.

* * * * *

Exercise 5.1

1.

H_0 : There is no relationship between gender and smoking.
H_1 : There is a relationship between gender and smoking.

Observed frequencies are:

	Gender		
Smoked	Male	Female	Total
Yes	26	14	40
No	24	36	60
Total	50	50	100

Expected values are $e_{ij} = \dfrac{N_{i.} \times N_{.j}}{N}$.

The expected frequencies are:

	Gender		
Smoked	Male	Female	Total
Yes	20	20	40
No	30	30	60
Total	50	50	100

$$\begin{aligned} Y^2 &= \sum_{i=1}^{h}\sum_{j=1}^{k} \frac{(N_{ij} - e_{ij})^2}{e_{ij}} \\ &= \frac{(26-20)^2}{20} + \frac{(14-20)^2}{20} + \frac{(24-30)^2}{30} + \frac{(36-30)^2}{30} \\ &= 1.8 + 1.8 + 1.2 + 1.2 \\ &= 6 \end{aligned}$$

We reject H_0 if $Y^2 \geq \chi^2_{\alpha;(r-1)(c-1)} = \chi^2_{0.025;1} = 5.02389$. Since $Y^2 = 6 > 5.02389,$ we reject H_0 at the $2\frac{1}{2}\%$ level and conclude that there is a relationship between gender and smoking.

2.

H_0 : There is no association between appearance and intelligence.
H_1 : There is an association between appearance and intelligence.

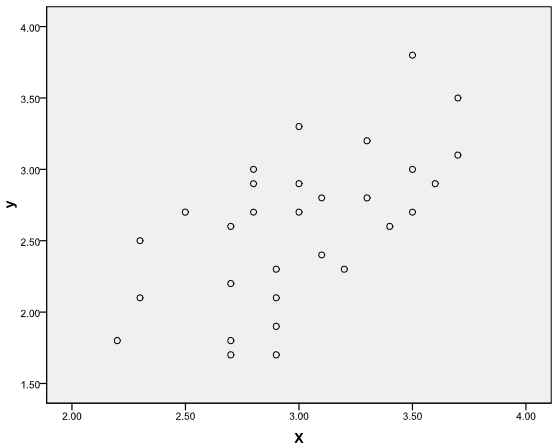
The expected values are:

	VH	H	A	L	Total
A	6	9	9	6	30
O	8	12	12	8	40
U	6	9	9	6	30
Total	20	30	30	20	100

$$\begin{aligned} Y^2 &= \sum_{i=1}^{h}\sum_{j=1}^{k}\frac{(N_{ij}-e_{ij})^2}{e_{ij}} \\ &= \frac{(9-6)^2}{6}+\frac{(12-9)^2}{9}+\cdots+\frac{(9-9)^2}{9}+\frac{(11-6)^2}{6} \\ &= 1.5+1+0.4444+2.6667+0+0.0833+0.3333+0.125+1.5+0.4444+0+4.1667 \\ &= 12.2638 \end{aligned}$$

We reject H_0 if $Y^2 \geq \chi^2_{\alpha;(r-1)(c-1)} = \chi^2_{0.1;6} = 10.6446$. Since $Y^2 = 12.2638 > 10.6446$, H_0 is rejected at the 10% level and we conclude that there is an association between appearance and gender. High IQ tends to go with attractiveness.

3.



A scatter plot of Y against X

The scatter plot shows that there is a linear relationship between X and Y.

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 0$$

$n = 30 \qquad \sum x = 90 \quad \sum x^2 = 274.88$

$\sum xy = 237.91 \quad \sum y = 78 \quad \sum y^2 = 211.14$

$\overline{x} = 3 \qquad \overline{y} = 2.6$

$$\begin{aligned} \sum (x-\overline{x})^2 &= \sum x^2 - \frac{(\sum x)^2}{n} \\ &= 274.88 - \frac{(90)^2}{30} \\ &= 4.88 \end{aligned} \qquad \begin{aligned} \sum (y-\overline{y})^2 &= \sum y^2 - \frac{(\sum y)^2}{n} \\ &= 211.14 - \frac{(78)^2}{30} \\ &= 8.34 \end{aligned}$$

$$\begin{aligned} \sum (x-\overline{x})(y-\overline{y}) &= \sum xy - \frac{(\sum x)(\sum y)}{n} \\ &= 237.91 - \frac{(90)(78)}{30} \\ &= 3.91 \end{aligned}$$

$R = \dfrac{3.91}{\sqrt{4.88}\sqrt{8.34}} = \dfrac{3.91}{6.379592463} \approx 0.6129$

Using table IX, the critical value $= 0.4226$. Since $0.6129 > 0.4226$, we reject H_0 at the 1% level of significance and conclude that there is a significant positive correlation.

4. $H_0 : \rho = 0$ against $H_1 : \rho \neq 0$

$n = 39 \qquad R = -0.35$

$U = \dfrac{1}{2}\log_e \dfrac{1-0.35}{1+0.35} \approx -0.3654$

$\eta = \dfrac{1}{2}\log_e \dfrac{1-\rho}{1+\rho} = \dfrac{1}{2}\log_e \dfrac{1-0.2}{1+0.2} \approx -0.2027$

The test statistic is

$$\begin{aligned} z &= \sqrt{n-3}(U-\eta) \\ &= \sqrt{39-3}(-0.3654+0.2027) \\ &= \sqrt{36} \times -0.1627 \\ &= -0.9762. \end{aligned}$$

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.

5. $H_0 : \rho_1 = \rho_2$ against $H_1 : \rho_1 \neq \rho_2$

$R_1 = -0.6 \qquad n_1 = 33$
$R_2 = -0.8 \qquad n_2 = 153$

$U_1 = \dfrac{1}{2}\log_e \dfrac{1-0.6}{1+0.6} \approx -0.6931$

$U_2 = \frac{1}{2}\log_e \frac{1-0.8}{1+0.8} \approx -1.0986$

$\eta_i = \frac{1}{2}\log_e \frac{1-\rho_i}{1+\rho_i}, i = 1, 2$

The test statistic is

$$\begin{aligned} z &= \frac{U_1 - U_2}{\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}} \\ &= \frac{-0.6931 + 1.0986}{\sqrt{\frac{1}{33-3} + \frac{1}{153-3}}} \\ &= \frac{0.4055}{\sqrt{0.04}} \\ &= 2.0275. \end{aligned}$$

$\alpha = 0.05$, $\alpha/2 = 0.025$, and $Z_{0.025} = 1.96$. We reject H_0 if $Z > 1.96$ or $Z < -1.96$.

Since $2.0275 > 1.96$, we reject H_0 at the 5% level and conclude that $\rho_1 \neq \rho_2$.

6. $R = 0.7 \qquad n = 10$

$U = \frac{1}{2}\log_e \frac{1+0.7}{1-0.7} \approx 0.8673$

The 95% confidence interval is

$$P\left(U - \frac{1.96}{\sqrt{n-3}} < \eta < U + \frac{1.96}{\sqrt{n-3}}\right)$$
$$P\left(0.8673 - \frac{1.96}{\sqrt{7}} < \eta < 0.8673 + \frac{1.96}{\sqrt{7}}\right)$$
$$P\left(0.1265 < \eta < 1.6081\right).$$

Now $\frac{e^{0.1265} - e^{-0.1265}}{e^{0.1265} + e^{-0.1265}} = \frac{1.1348 - 0.8812}{1.1348 + 0.8812} \approx 0.1258$

and $\frac{e^{1.6081} - e^{-1.6081}}{e^{1.6081} + e^{-1.6081}} = \frac{4.9933 - 0.2003}{4.9933 + 0.2003} \approx 0.9229$

in other words, 95% confidence interval for ρ is $(0.1258; 0.9229)$.

7. The contingency table is

	A_1	A_2	Total
B_1	0	5	5
B_2	6	1	7
Total	6	6	12

Note $x = 0$ and $P(X \leq 0) = 0.008$.

8. The contingency table is

Contracted influenza	Brown	White	Total
Yes	1	5	6
No	5	1	6
Total	6	6	12

H_0 : The two strains are equally susceptible
H_1 : The white mice are more susceptible

$N = 12 \quad k = 6 \quad n = 6 \quad x = 1$

$P(X \leq 1) = 0.04 \Rightarrow$ Since $0.04 < 0.05$, we reject H_0 and conclude that white mice are more susceptible than brown mice.

9. H_0 : The number of children is independent of father's level of training.
H_1 : The number of children is not independent of father's level of training.

The expected frequencies are:

	Number of children				
Training	0	1	2	more than 2	Total
Primary school	15	25	30	30	100
Secondary school	9	15	18	18	60
College	4.5	7.5	9	9	30
University	1.5	2.5	3	3	10
Total	30	50	60	60	200

Since the expected frequency for university is less than five we combine the categories college and university to produce the following table:

	Number of children				
Training	0	1	2	more than 2	Total
Primary school	15	25	30	30	100
Secondary school	9	15	18	18	60
College & university	6	10	12	12	40
Total	30	50	60	60	200

The test statistic is

$$\begin{aligned} Y^2 &= \sum_{i=1}^{h}\sum_{j=1}^{k}\frac{(N_{ij}-e_{ij})^2}{e_{ij}} \\ &= \frac{(18-15)^2}{15}+\frac{(22-25)^2}{25}+\cdots+\frac{(15-12)^2}{12}+\frac{(15-12)^2}{12} \\ &= 0.6+0.36+0+0+1+5.4+0.5+0.5+0+3.6+0.75+0.75 \\ &= 13.46. \end{aligned}$$

We reject H_0 if $Y^2 \geq \chi^2_{\alpha;(r-1)(c-1)} = \chi^2_{0.05;6} = 12.5916$.

Since $Y^2 = 13.46 > 12.5916$, we reject H_0 at the 5% level and conclude that the number of children depends on the father's level of training.

10. (a) $H_0 : \rho = 0.2$ against $H_1 : \rho < 0.2$

$n = 19 \qquad R = 0.5$

$$U = \frac{1}{2}\log_e \frac{1+0.5}{1-0.5} \approx 0.5493$$

$$\eta = \frac{1}{2}\log_e \frac{1-\rho}{1+\rho} = \frac{1}{2}\log_e \frac{1+0.2}{1-0.2} \approx 0.2027$$

The test statistic is

$$\begin{aligned} z &= \sqrt{n-3}(U-\eta) \\ &= \sqrt{19-3}(0.5493-0.2027) \\ &= \sqrt{16} \times 0.3466 \\ &= 1.3864. \end{aligned}$$

$\alpha = 0.05$ and $Z_{0.05} = 1.645$. Reject H_0 if $Z < -1.645$.

Since $1.3864 > -1.645$, we do not reject H_0 and conclude that $\rho = 0.2$ at the 5% level.

(b) $\alpha = 0.05, \alpha/2 = 0.025$ and $Z_{0.025} = 1.96$.

The 95% confidence interval for ρ is

$$\begin{aligned} U - \frac{1.96}{\sqrt{n-3}} &< \eta < U + \frac{1.96}{\sqrt{n-3}} \\ 0.5493 - \frac{1.96}{\sqrt{16}} &< \eta < 0.5493 + \frac{1.96}{\sqrt{16}} \\ 0.5493 - \frac{1.96}{4} &< \eta < 0.5493 + \frac{1.96}{4} \\ 0.0593 &< \eta < 1.0393 \end{aligned}$$

Now $\dfrac{e^{0.0593} - e^{-0.0593}}{e^{0.0593} + e^{-0.0593}} = \dfrac{1.0611 - 0.9424}{1.0611 + 0.9424} \approx 0.0592$

and $\dfrac{e^{1.0393} - e^{-1.0393}}{e^{1.0393} + e^{-1.0393}} = \dfrac{2.8272 - 0.3537}{2.8272 + 0.3537} \approx 0.7776.$

Thus, the 95% confidence interval for ρ is $0.0592 < \rho < 0.7776$.

11. $H_0 : \rho_1 = \rho_2$ against $H_1 : \rho_1 < \rho_2$

$R_1 = 0.6 \quad n_1 = 53$
$R_2 = 0.9 \quad n_2 = 53$

$$U_1 = \frac{1}{2}\log_e \frac{1+0.6}{1-0.6} \approx 0.6931$$

$$U_2 = \frac{1}{2}\log_e \frac{1+0.9}{1-0.9} \approx 1.4722$$

The test statistic is

$$\begin{aligned} z &= \frac{U_1 - U_2}{\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}} \\ &= \frac{0.6931 - 1.4722}{\sqrt{\frac{1}{53-3} + \frac{1}{53-3}}} \\ &= \frac{-0.7791}{\sqrt{0.04}} \\ &= -3.8955. \end{aligned}$$

$\alpha = 0.05$ and $Z_{0.05} = 1.645$. We reject H_0 if $Z < -1.645$.

Since $-3.8955 < -1.645$, we reject H_0 at the 5% level and conclude that $\rho_1 < \rho_2$.

12. $n = 10 \quad \sum x = 500 \quad \sum x^2 = 26\,600$
$\sum xy = 10\,090 \quad \sum y = 200 \quad \sum y^2 = 4\,400$
$\overline{x} = 50 \quad \overline{y} = 20$

$$\begin{aligned} \sum (x-\overline{x})^2 &= \sum x^2 - \frac{\sum x}{n} \\ &= 26\,600 - \frac{(500)^2}{10} \\ &= 1\,600 \end{aligned} \qquad \begin{aligned} \sum (y-\overline{y})^2 &= \sum y^2 - \frac{\sum y}{n} \\ &= 4\,400 - \frac{(200)^2}{10} \\ &= 400 \end{aligned}$$

$$\begin{aligned} \sum (x-\overline{x})(y-\overline{y}) &= \sum xy - \frac{\sum x \sum x}{n} \\ &= 10\,090 - \frac{(500)(400)}{10} \\ &= 90 \end{aligned}$$

$$R = \frac{90}{\sqrt{400}\sqrt{1\,600}} = \frac{90}{800} \approx 0.1125$$

$H_0 : \rho = 0$ against $H_1 : \rho \neq 0$

The critical value is 0.6319.

Since $0.1125 < 0.6319$, we do not reject H_0 and conclude that the correlation is not significant.

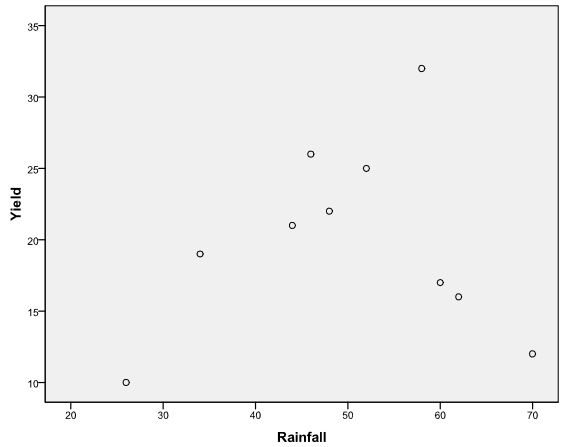
OR

$$\begin{aligned} T &= \frac{R\sqrt{n-2}}{\sqrt{1-R^2}} \\ &= \frac{0.1125\sqrt{8}}{\sqrt{1-0.1125^2}} \\ &= \frac{0.318198051}{\sqrt{0.98734375}} \\ &\approx 0.3202 \end{aligned}$$

The critical value is $t_{\alpha/2;(n-2)} = t_{0.025;8} = 2.306$.

Since $0.3202 < 2.306,$ we do not reject H_0 at the 5% level of significance and conclude that the correlation is not significant.

This is also evidenced by the scatter plot since the scatter diagram shows that the relationship might not be linear. Thus, there is no significant linear relationship between rainfall and yield.



A scatter plot of yield versus rainfall

* * * * *

Exercise 6.1

1. Let $U = \frac{\Sigma\left(X_i - \overline{X}\right)^2}{\sigma^2}$ then $U \sim \chi^2_{n-1}$. From this we may derive the probability expression

$$\begin{aligned} 1-\alpha &= P\left[\chi^2_{\alpha;n-1} > U\right] \\ &= P\left[\chi^2_{\alpha;n-1} > \frac{\Sigma\left(X_i - \overline{X}\right)^2}{\sigma^2}\right] \\ &= P\left[\frac{1}{\chi^2_{\alpha;n-1}} < \frac{\sigma^2}{\Sigma\left(X_i - \overline{X}\right)^2}\right] \\ &= P\left[\frac{\Sigma\left(X_i - \overline{X}\right)^2}{\chi^2_{\alpha;n-1}} < \sigma^2\right]. \end{aligned}$$

Therefore $\left[\frac{\Sigma\left(X_i - \overline{X}\right)^2}{\chi^2_{\alpha;n-1}};\quad \infty\right]$ is a $100\,(1-\alpha)\,\%$ one-sided confidence interval for σ^2 which tests the hypothesis $H_0 : \sigma^2 = c$ against $H_1 : \sigma^2 > c$ where μ is unknown.

Let $U = \frac{\Sigma\left(X_i - \overline{X}\right)^2}{\sigma^2}$ then $U \sim \chi^2_{n-1}$. From this we may derive the probability expression

$$\begin{aligned} 1-\alpha &= P\left[\chi^2_{1-\alpha;n-1} < U\right] \\ &= P\left[\chi^2_{1-\alpha;n-1} < \frac{\Sigma\left(X_i - \overline{X}\right)^2}{\sigma^2}\right] \\ &= P\left[\frac{1}{\chi^2_{1-\alpha;n-1}} > \frac{\sigma^2}{\Sigma\left(X_i - \overline{X}\right)^2}\right] \\ &= P\left[\frac{\Sigma\left(X_i - \overline{X}\right)^2}{\chi^2_{1-\alpha;n-1}} > \sigma^2\right] \\ &= P\left[\sigma^2 < \frac{\Sigma\left(X_i - \overline{X}\right)^2}{\chi^2_{1-\alpha;n-1}}\right]. \end{aligned}$$

Therefore $\left[0; \frac{\Sigma\left(X_i - \overline{X}\right)^2}{\chi^2_{1-\alpha;n-1}}\right]$ is a $100\,(1-\alpha)\,\%$ one-sided confidence interval for σ^2 which tests the hypothesis $H_0 : \sigma^2 = c$ against $H_1 : \sigma^2 < c$ where μ is known.

Let $U = \frac{\Sigma\left(X_i - \mu\right)^2}{\sigma^2}$ then $U \sim \chi^2_n$.

$$\begin{aligned}
1-\alpha &= P\left(\chi^2_{1-\frac{1}{2}\alpha;n} < U < \chi^2_{\frac{1}{2}\alpha;n}\right) \\
&= P\left[\chi^2_{1-\frac{1}{2}\alpha;n} < \frac{\Sigma\left(X_i-\mu\right)^2}{\sigma^2} < \chi^2_{\frac{1}{2}\alpha;n}\right] \\
&= P\left[\frac{1}{\chi^2_{1-\frac{1}{2}\alpha;n}} > \frac{\sigma^2}{\Sigma\left(X_i-\mu\right)^2} > \frac{1}{\chi^2_{\frac{1}{2}\alpha;n}}\right] \\
&= P\left[\frac{\Sigma\left(X_i-\mu\right)^2}{\chi^2_{\frac{1}{2}\alpha;n}} < \sigma^2 < \frac{\Sigma\left(X_i-\mu\right)^2}{\chi^2_{1-\frac{1}{2}\alpha;n}}\right]
\end{aligned}$$

Therefore $\left[\frac{\Sigma\left(X_i-\mu\right)^2}{\chi^2_{\frac{1}{2}\alpha;n}};\quad \frac{\Sigma\left(X_i-\mu\right)^2}{\chi^2_{1-\frac{1}{2}\alpha;n}}\right]$ is a $100\left(1-\alpha\right)\%$ two-sided confidence interval for σ^2 which tests the hypothesis $H_0: \sigma^2 = c$ against $H_1: \sigma^2 \neq c$ where μ is known.

Let $U = \frac{\Sigma\left(X_i-\mu\right)^2}{\sigma^2}$ then $U \sim \chi^2_n$.

$$\begin{aligned}
1-\alpha &= P\left[\chi^2_{\alpha;n} > U\right] \\
&= P\left[\chi^2_{\alpha;n} > \frac{\Sigma\left(X_i-\mu\right)^2}{\sigma^2}\right] \\
&= P\left[\frac{1}{\chi^2_{\alpha;n}} < \frac{\sigma^2}{\Sigma\left(X_i-\mu\right)^2}\right] \\
&= P\left[\frac{\Sigma\left(X_i-\mu\right)^2}{\chi^2_{\alpha;n}} < \sigma^2\right]
\end{aligned}$$

Therefore $\left[\frac{\Sigma\left(X_i-\mu\right)^2}{\chi^2_{\alpha;n}};\quad \infty\right]$ is a $100\left(1-\alpha\right)\%$ one-sided confidence interval for σ^2 which tests the hypothesis $H_0: \sigma^2 = c$ against $H_1: \sigma^2 > c$ where μ is known.

2. $n = 11 \qquad \sum X_i = 110 \qquad \sum X_i^2 = 1\,220$

$$\begin{aligned}
\Sigma\left(X_i-\overline{X}\right)^2 &= \Sigma X_i^2 - \frac{\left(\Sigma X_i\right)^2}{n} \\
&= 1\,220 - \frac{(110)^2}{11} \\
&= 120
\end{aligned}$$

$$\begin{aligned}\Sigma\left(X_i-\mu\right)^2 &= (6-9)^2+(10-9)^2+\cdots+(13-9)^2\\ &= 131\end{aligned}$$

(a) $H_0:\sigma=5$ against $H_1:\sigma<5$

(i) μ is unknown, then the test statistic is

$$\begin{aligned}U &= \frac{\Sigma\left(X_i-\overline{X}\right)^2}{\sigma^2}\\ &= \frac{120}{25}\\ &= 4.8.\end{aligned}$$

The critical value is $\chi^2_{1-\alpha;n-1}=\chi^2_{0.95;10}=3.9403$. Reject H_0 if $U<3.9403$.

Since $4.8>3.9403$, we do not reject H_0 at the 5% level and conclude that $\sigma=5$.

(ii) $\mu=9$, then the test statistic is

$$\begin{aligned}U &= \frac{\Sigma\left(X_i-\mu\right)^2}{\sigma^2}\\ &= \frac{131}{25}\\ &= 5.24.\end{aligned}$$

The critical value is $\chi^2_{1-\alpha;n}=\chi^2_{0.95;11}=4.57481$. Reject H_0 if $U<4.57481$.

Since $5.24>4.57481$, we do not reject H_0 at the 5% level and conclude that $\sigma=5$.

(b) (i) If μ is unknown, a 95% one-sided confidence interval for σ^2 is

$$\begin{aligned}&\left[0;\frac{\Sigma\left(X_i-\overline{X}\right)^2}{\chi^2_{1-\alpha;n-1}}\right]\\ &\left[0;\frac{120}{\chi^2_{0.95;10}}\right]\\ &\left[0;\frac{120}{3.9403}\right]\\ &[0;30.4545]\,.\end{aligned}$$

Thus, the 95% one-sided confidence interval for σ is

$$\begin{aligned}&\left[\sqrt{0};\sqrt{30.4545}\right]\\ &[0;5.5186]\,.\end{aligned}$$

(ii) If $\mu = 9$, a 95% one-sided confidence interval for σ^2 is

$$\left[0; \frac{\Sigma (X_i - \mu)^2}{\chi^2_{1-\alpha;n}}\right]$$
$$\left[0; \frac{131}{\chi^2_{0.95;11}}\right]$$
$$\left[0; \frac{131}{4.57481}\right]$$
$$0; 28.6351.$$

Thus, the 95% one-sided confidence interval for σ is

$$\left[\sqrt{0}; \sqrt{28.6351}\right]$$
$$[0; 5.3512]\,.$$

3. (a) $n = 10 \qquad \alpha = 0.10 \qquad \alpha/2 = 0.05$

$$\chi^2_{\frac{1}{2}\alpha;n} = \chi^2_{0.05;10} = 18.307$$
$$\chi^2_{1-\frac{1}{2}\alpha;n} = \chi^2_{0.95;10} = 3.9403$$

The 90% confidence interval for σ^2, μ is known and $n = 10$ is

$$\left[\frac{\Sigma (X_i - \mu)^2}{\chi^2_{\frac{1}{2}\alpha;n}} < \sigma^2 < \frac{\Sigma (X_i - \mu)^2}{\chi^2_{1-\frac{1}{2}\alpha;n}}\right]$$
$$\left[\frac{\Sigma (X_i - \mu)^2}{18.307} < \sigma^2 < \frac{\Sigma (X_i - \mu)^2}{3.9403}\right].$$

The length of the interval is

$$\Sigma (X_i - \mu)^2 = \left(\frac{1}{3.9403} - \frac{1}{18.307}\right)$$
$$\Sigma (X_i - \mu)^2 = 0.1992.$$

The expected length of the interval is

$$= 0.1992E\left(\Sigma (X_i - \mu)^2\right)$$
$$= 0.1992 \times \sigma^2(n) \quad \text{since } E\left(\Sigma (X_i - \mu)^2\right) = \sigma^2 n$$
$$= 0.1992 \times \sigma^2 \times 10$$
$$= 1.992\sigma^2.$$

(b) $\chi^2_{\frac{1}{2}\alpha;n-1} = \chi^2_{0.05;9} = 16.919$
$\chi^2_{1-\frac{1}{2}\alpha;n-1} = \chi^2_{0.95;9} = 3.32511$

The 90% confidence interval for σ^2, μ is unknown and $n = 10$ is

$$\left[\frac{\Sigma\left(X_i - \overline{X}\right)^2}{\chi^2_{\frac{1}{2}\alpha;n-1}} < \sigma^2 < \frac{\Sigma\left(X_i - \overline{X}\right)^2}{\chi^2_{1-\frac{1}{2}\alpha;n-1}}\right]$$
$$\left[\frac{\Sigma\left(X_i - \overline{X}\right)^2}{16.919} < \sigma^2 < \frac{\Sigma\left(X_i - \overline{X}\right)^2}{3.32511}\right].$$

The length of the interval is

$$\begin{aligned}\Sigma\left(X_i - \overline{X}\right)^2 &= \left(\frac{1}{3.325} - \frac{1}{16.919}\right)\\ \Sigma\left(X_i - \overline{X}\right)^2 &= 0.2416.\end{aligned}$$

The expected length of the interval is

$$\begin{aligned}&= 0.2416E\left(\Sigma\left(X_i - \overline{X}\right)^2\right)\\ &= 0.2416 \times \sigma^2(n-1) \quad \text{since } E\left(\Sigma\left(X_i - \overline{X}\right)^2\right) = \sigma^2(n-1)\\ &= 0.2416 \times \sigma^2 \times 9\\ &= 2.1744\sigma^2.\end{aligned}$$

4. The 95% confidence interval for σ^2, μ is unknown is

$$\left[\frac{\Sigma\left(X_i - \overline{X}\right)^2}{\chi^2_{\frac{1}{2}\alpha;n-1}} < \sigma^2 < \frac{\Sigma\left(X_i - \overline{X}\right)^2}{\chi^2_{1-\frac{1}{2}\alpha;n-1}}\right].$$

The expected length of the interval is

$$\begin{aligned}&= E\left(\Sigma\left(X_i - \overline{X}\right)^2\right)\left[\frac{1}{\chi^2_{\frac{1}{2}\alpha;n-1}} - \frac{1}{\chi^2_{1-\frac{1}{2}\alpha;n-1}}\right]\\ &= \sigma^2(n-1)\left[\frac{1}{\chi^2_{\frac{1}{2}\alpha;n-1}} - \frac{1}{\chi^2_{1-\frac{1}{2}\alpha;n-1}}\right].\end{aligned}$$

Thus,

$$\begin{aligned}\sigma^2(n-1)\left[\frac{1}{\chi^2_{\frac{1}{2}\alpha;n-1}} - \frac{1}{\chi^2_{1-\frac{1}{2}\alpha;n-1}}\right] &\leq 2.5\sigma^2\\ (n-1)\left[\frac{1}{\chi^2_{\frac{1}{2}\alpha;n-1}} - \frac{1}{\chi^2_{1-\frac{1}{2}\alpha;n-1}}\right] &\leq 2.5.\end{aligned}$$

By trial and error, taking $n = 11$:

$$\begin{aligned} &\Longrightarrow \quad 10\left[\frac{1}{3.24697} - \frac{1}{20.4831}\right] \\ &\approx \quad 2.5916 \not\leq 2.5 \end{aligned}$$

taking $n = 12$:

$$\begin{aligned} &\Longrightarrow \quad 11\left[\frac{1}{3.81575} - \frac{1}{21.92}\right] \\ &\approx \quad 2.381 < 2.5 \end{aligned}$$

taking $n = 13$:

$$\begin{aligned} &\Longrightarrow \quad 12\left[\frac{1}{4.40379} - \frac{1}{23.3367}\right] \\ &\approx \quad 2.2107 < 2.5. \end{aligned}$$

Thus, the largest value of n satisfying the condition is $n = 12$.

5. Let sample 1 be unmodified and sample 2 be modified.

$n_1 = 6 \qquad \sum X_{1i} = 180 \qquad \sum X_{1i}^2 = 5\,470$

$$\begin{aligned} \Sigma\left(X_{1i} - \overline{X}\right)^2 &= \Sigma X_{1i}^2 - \frac{\left(\Sigma X_{1i}\right)^2}{n} \\ &= 5\,470 - \frac{(180)^2}{6} \\ &= 70 \end{aligned}$$

$n_2 = 6 \qquad \sum X_{2i} = 192 \qquad \sum X_{2i}^2 = 6\,194$

$$\begin{aligned} \Sigma\left(X_{2i} - \overline{X}\right)^2 &= \Sigma X_{2i}^2 - \frac{\left(\Sigma X_{2i}\right)^2}{n} \\ &= 6\ 194 - \frac{(192)^2}{6} \\ &= 50 \end{aligned}$$

$H_0 : \sigma_1^2 = \sigma_2^2$ against $H_1 : \sigma_1^2 \neq \sigma_2^2$

The test statistic is

$$\begin{aligned} F &= \frac{\sigma_2^2}{\sigma_1^2} \cdot \frac{\sum\limits_{i=1}^{n_1} \left(X_{1i} - \mu_1\right)^2 / n_1}{\sum\limits_{i=1}^{n_2} \left(X_{2i} - \mu_2\right)^2 / n_2} \\ &= 1 \cdot \frac{70/5}{50/5} \\ &= \frac{14}{10} \\ &= 1.4. \end{aligned}$$

The critical values are $F_{0.05;5,5} = 5.05$ and $F_{0.95;5,5} = \dfrac{1}{F_{0.05;5,5}} = \dfrac{1}{5.05} \approx 0.198$

Since $0.198 < F < 5.05$, we cannot reject H_0. The two processes do not differ with respect to precision.

The 90% confidence interval for $\dfrac{\sigma_1^2}{\sigma_2^2}$ is

$$P\left(F_{1-\frac{\alpha}{2};n_2-1;n_1-1} < \frac{\sigma_2^2}{\sigma_1^2}\frac{S_1^2}{S_2^2} < F_{\frac{\alpha}{2};n_2-1;n_1-1}\right) = 1-\alpha$$

$$\left[\frac{F_{1-\frac{\alpha}{2};n_2-1;n_1-1}}{S_2^2/S_1^2}; \frac{F_{\frac{\alpha}{2};n_2-1;n_1-1}}{S_2^2/S_1^2}\right]$$

$\alpha = 0.10, \alpha/2 = 0.05$

$F_{1-\frac{\alpha}{2};n_2-1;n_1-1} = F_{0.95;5;5} = \dfrac{1}{F_{0.05;5;5}} = \dfrac{1}{5.05} \approx 0.198.$

$F_{\frac{\alpha}{2};n_2-1;n_1-1} = F_{0.05;5;5} = 5.05$

$S_1^2 = \dfrac{1}{n_1-1}\Sigma\left(X_{1i} - \overline{X}\right)^2 = \dfrac{1}{5}(70) = 14$

$S_2^2 = \dfrac{1}{n_2-1}\Sigma\left(X_{2i} - \overline{X}\right)^2 = \dfrac{1}{5}(50) = 10.$

$\therefore$The 90% confidence interval is

$$\left[\frac{0.198}{10/14}; \frac{5.05}{10/14}\right]$$
$$[0.2772; 7.07]\,.$$

Note the confidence interval for $\dfrac{\sigma_1}{\sigma_2}$ will be

$$\left[\sqrt{0.2772}; \sqrt{7.07}\right]$$
$$[0.5265; 2.6589]\,.$$

6. $n_1 = 10 \qquad \sum X_{1i} = 20 \qquad \sum X_{1i}^2 = 148$

$$\begin{aligned}\Sigma\left(X_{1i} - \overline{X}\right)^2 &= \Sigma X_{1i}^2 - \frac{\left(\Sigma X_{1i}\right)^2}{n} \\ &= 148 - \frac{(20)^2}{10} \\ &= 108\end{aligned}$$

$n_2 = 12 \qquad \sum X_{2i} = 36 \qquad \sum X_{2i}^2 = 152$

$$\begin{aligned}\Sigma\left(X_{1i} - \overline{X}\right)^2 &= \Sigma X_{1i}^2 - \frac{\left(\Sigma X_{1i}\right)^2}{n} \\ &= 152 - \frac{(36)^2}{12} \\ &= 44\end{aligned}$$

$$S_1^2 = \frac{1}{n_1 - 1}\Sigma\left(X_{1i} - \overline{X}\right)^2 = \frac{1}{9}(108) = 12$$
$$S_2^2 = \frac{1}{n_2 - 1}\Sigma\left(X_{2i} - \overline{X}\right)^2 = \frac{1}{11}(44) = 4$$

(a) We need to test: $H_0 : \dfrac{\sigma_2}{\sigma_1} = \dfrac{1}{2}$

$H_0 : \dfrac{\sigma_2^2}{\sigma_1^2} = \dfrac{1}{4}$ against $H_1 : \dfrac{\sigma_2^2}{\sigma_1^2} < \dfrac{1}{4} \Longrightarrow H_0 : \sigma_1^2 = 4\sigma_2^2$ against $H_1 : \sigma_1^2 > 4\sigma_2^2$.

The test statistic is

$$\begin{aligned}F &= \frac{\sigma_2^2}{\sigma_1^2} \cdot \frac{S_1^2}{S_2^2} \\ &= \frac{1}{4} \times \frac{12}{4} \\ &= 0.75.\end{aligned}$$

The critical value is $F_{\alpha;n_1-1;n_2-1} = F_{0.05;9,11} = 2.9$.

Since $0.75 < 2.9$, we do not reject H_0 and conclude that $\sigma_1^2 = 4\sigma_2^2$.

(b) The 95% one-sided confidence interval for $\dfrac{\sigma_2^2}{\sigma_1^2}$ is

$$\left[0; \frac{F_{\alpha;n_1-1;n_2-1}}{S_1^2/S_2^2}\right]$$
$$\left[0; \frac{F_{0.05;9;11}}{12/4}\right]$$
$$\left[0; \frac{2.9}{3}\right]$$
$$[0; 0.9667]\,.$$

Now the 95% one-sided confidence interval for $\dfrac{\sigma_1^2}{\sigma_2^2}$ is

$$\left[\frac{1}{0.9667}; \infty\right]$$
$$[1.0344; \infty]\,.$$

Note $\frac{1}{0.000000001} \approx 1\,000\,000\,000$, thus values close to 0 goes to ∞.

Thus the 95% one-sided confidence interval for $\dfrac{\sigma_1}{\sigma_2}$ is

$$\left[\sqrt{1.0344}; \infty\right]$$
$$[1.0171; \infty]\,.$$

7. $n = 11$ $\quad \sum X_1 = 330$ $\quad \sum X_1^2 = 10\,802$
$\sum X_1 X_2 = 10\,230$ $\quad \sum X_2 = 330$ $\quad \sum X_2^2 = 10\,098$
$\overline{X}_1 = 30$ $\quad \overline{X}_2 = 30$

$$\begin{aligned}\sum \left(X_1 - \overline{X}_1\right)^2 &= \sum X_1^2 - \frac{\left(\sum X_1\right)^2}{n} \\ &= 10\,802 - \frac{(330)^2}{11} \\ &= 902\end{aligned}$$

$$\begin{aligned}\sum \left(X_2 - \overline{X}_2\right)^2 &= \sum X_2^2 - \frac{\left(\sum X_2\right)^2}{n} \\ &= 10\,098 - \frac{(330)^2}{11} \\ &= 198\end{aligned}$$

$$\begin{aligned}\sum \left(X_1 - \overline{X}_1\right)\left(X_2 - \overline{X}_2\right) &= \sum X_1 X_2 - \frac{\left(\sum X_1\right)\left(\sum X_2\right)}{n} \\ &= 10\,230 - \frac{(330)(330)}{11} \\ &= 330\end{aligned}$$

$H_0 : \sigma_1^2 = \sigma_2^2$ against $H_1 : \sigma_1^2 > \sigma_2^2$

$U_{11} = 902$ $\quad U_{12} = 330$ $\quad U_{22} = 198$

$$\begin{aligned}T &= \frac{\sqrt{n-2}\,(U_{11} - U_{22})}{2\sqrt{U_{11}U_{22} - U_{12}^2}} \\ &= \frac{\sqrt{11-2}\,(902 - 198)}{2\sqrt{902 \times 198 - 330^2}} \\ &= \frac{3\,(704)}{2\sqrt{69\,696}} \\ &= \frac{2\,112}{528} \\ &= 4\end{aligned}$$

$\alpha = 0.10;$ $\quad t_{\alpha;(n-2)} = t_{0.1;9} = 1.383$. Reject H_0 if $T > 1.383$.

Since $4 > 1.383$, H_0 is rejected at the 10% level. We conclude that $\sigma_1^2 > \sigma_2^2$.

8. We want to test $H_0 : \sigma_1^2 = \sigma_2^2$ against $H_1 : \sigma_1^2 > \sigma_2^2$.

$\overline{x} = 30 \quad \sum x = 300 \quad \sum x^2 = 9\,090$

$\overline{y} = 50.2 \quad \sum y = 502 \quad \sum y^2 = 25\,238$

$n = 10 \quad \sum xy = 15\,088$

$$\begin{aligned} \sum (x-\overline{x})^2 &= \sum x^2 - \frac{(\sum x)^2}{n} \\ &= 9\,090 - \frac{(300)^2}{10} \\ &= 90 \end{aligned} \qquad \begin{aligned} \sum (y-\overline{y})^2 &= \sum y^2 - \frac{(\sum y)^2}{n} \\ &= 25\,238 - \frac{(502)^2}{10} \\ &= 37.6 \end{aligned}$$

$$\begin{aligned} \sum (x-\overline{x})(y-\overline{y}) &= \sum xy - \frac{(\sum x)(\sum y)}{n} \\ &= 15\,088 - \frac{(300)(502)}{10} \\ &= 28 \end{aligned}$$

$U_{11} = 90 \qquad U_{12} = 28 \qquad U_{22} = 37.6$

$$\begin{aligned} T &= \frac{\sqrt{n-2}\,(U_{11} - U_{22})}{2\sqrt{U_{11}U_{22} - U_{12}^2}} \\ &= \frac{\sqrt{10-2}\,(90 - 37.6)}{2\sqrt{90 \times 37.6 - 28^2}} \\ &= \frac{\sqrt{8}\,(52.4)}{2\sqrt{2\,600}} \\ &= \frac{148.2095813}{101.9803903} \\ &\approx 1.4533 \end{aligned}$$

$\alpha = 0.10; \qquad t_{\alpha;(n-2)} = t_{0.1;8} = 1.397$. Reject H_0 if $T > 1.397$.

Since $1.4533 > 1.397$, we reject H_0 at the 10% level and conclude that the students were more uniform after the remedial training than before that is, $\sigma_1^2 > \sigma_2^2$.

9.

$\overline{X}_1 = 20.1 \quad \sum X_{1i} = 120.6 \quad \sum X_{1i}^2 = 2\,425.08$

$\overline{X}_2 = 19.9 \quad \sum X_{2i} = 119.4 \quad \sum X_{2i}^2 = 2\,377.54$

$\overline{X}_3 = 20.2 \quad \sum X_{3i} = 121.2 \quad \sum X_{3i}^2 = 2\,453.34$

$\overline{X}_4 = 20 \quad \sum X_{4i} = 120 \quad \sum X_{4i}^2 = 2\,400.88$

$n = 6$

$$\begin{aligned} S_1^2 &= \frac{1}{n-1}\left(\sum X_{1i}^2 - \frac{\left(\sum X_{1i}\right)^2}{n}\right) \\ &= \frac{1}{6-1}\left(2\,425.08 - \frac{(120.6)^2}{6}\right) \\ &= \frac{1}{5}(1.02) \\ &= 0.204 \end{aligned} \qquad \begin{aligned} S_2^2 &= \frac{1}{n-1}\left(\sum X_{2i}^2 - \frac{\left(\sum X_{2i}\right)^2}{n}\right) \\ &= \frac{1}{6-1}\left(2\,377.54 - \frac{(119.4)^2}{6}\right) \\ &= \frac{1}{5}(1.48) \\ &= 0.296 \end{aligned}$$

$$\begin{aligned} S_3^2 &= \frac{1}{n-1}\left(\sum X_{3i}^2 - \frac{\left(\sum X_{3i}\right)^2}{n}\right) \\ &= \frac{1}{6-1}\left(2\,453.34 - \frac{(121.2)^2}{6}\right) \\ &= \frac{1}{5}(5.1) \\ &= 1.02 \end{aligned} \qquad \begin{aligned} S_4^2 &= \frac{1}{n-1}\left(\sum X_{4i}^2 - \frac{\left(\sum X_{4i}\right)^2}{n}\right) \\ &= \frac{1}{6-1}\left(2\,400.88 - \frac{(120)^2}{6}\right) \\ &= \frac{1}{5}(0.88) \\ &= 0.176 \end{aligned}$$

$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$ against $H_1 : \sigma_p^2 \neq \sigma_q^2$ for at least one $p \neq q$

The test statistic is

$$\begin{aligned} U &= \frac{\max\limits_i S_i^2}{\min\limits_i S_i^2} \\ &= \frac{1.02}{0.176} \\ &\approx 5.7955. \end{aligned}$$

The critical value is 13.7. H_0 is rejected if $U > 13.7$.

Since $5.7955 < 13.7$, we do not reject H_0 at the 5% level and conclude that the variances of the four populations are equal.

10.

$$\begin{array}{lll} \mu_1 = 5 & \sum X_{1i} = 60 & \sum X_{1i}^2 = 390 \\ \mu_2 = 7 & \sum X_{2i} = 80 & \sum X_{2i}^2 = 730 \\ \mu_3 = 8 & \sum X_{3i} = 85 & \sum X_{3i}^2 = 740 \end{array}$$

$$\begin{aligned} S_1^2 &= \frac{1}{n_1}\sum (X_{1i}-\mu_1)^2 \\ &= \frac{1}{n_1}\left(\sum X_{1i}^2 - 2\mu_1 \sum X_{1i} + n\mu_1^2\right) \\ &= \frac{1}{10}\left(390 - 2\times 5(60) + 10\times 5^2\right) \\ &= \frac{1}{10}\left(390 - 600 + 250\right) \\ &= \frac{1}{10}\left(40\right) \\ &= 4 \end{aligned}$$

$$\begin{aligned} S_2^2 &= \frac{1}{n_2}\sum (X_{2i}-\mu_2)^2 \\ &= \frac{1}{n_2}\left(\sum X_{2i}^2 - 2\mu_2 \sum X_{2i} + n\mu_2^2\right) \\ &= \frac{1}{10}\left(730 - 2\times 7(80) + 10\times 7^2\right) \\ &= \frac{1}{10}\left(730 - 1\,120 + 490\right) \\ &= \frac{1}{10}\left(100\right) \\ &= 10 \end{aligned}$$

$$\begin{aligned} S_3^2 &= \frac{1}{n_3}\sum (X_{3i}-\mu_3)^2 \\ &= \frac{1}{n_3}\left(\sum X_{3i}^2 - 2\mu_3 \sum X_{3i} + n\mu_3^2\right) \\ &= \frac{1}{10}\left(740 - 2\times 8(85) + 10\times 8^2\right) \\ &= \frac{1}{10}\left(740 - 1\,360 + 640\right) \\ &= \frac{1}{10}\left(20\right) \\ &= 2 \end{aligned}$$

Testing $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2$ against $H_1 : \sigma_p^2 \neq \sigma_q^2$ for at least one $p \neq q$

Test statistic is

$$\begin{aligned} U &= \frac{\max\limits_i S_i^2}{\min\limits_i S_i^2} \\ &= \frac{10}{2} \\ &= 5. \end{aligned}$$

The critical value is 4.85. H_0 is rejected if $U > 4.85$.

Since $5 > 4.85$, we reject H_0 at the 5% level and conclude that the variances of the three populations are not the same.

Note that $S_j^2 = \frac{1}{n}\Sigma X_{ji}^2 - 2\mu_j\mu_j + \mu_j^2$ has n degrees of freedom.

* * * * *

Exercise 7.1

1. (a) $H_0 : \mu = 0.5$ against $H_1 : \mu \neq 0.5$

$n = 11 \qquad \sum X_i = 5.83 \qquad \sum X_i^2 = 3.1339$

Since σ^2 is unknown, we have

$$\begin{aligned} S^2 &= \frac{1}{n-1}\left(\Sigma\left(X_i - \overline{X}\right)^2\right) \\ &= \frac{1}{n-1}\left(\Sigma X_i^2 - \frac{(\Sigma X_i)^2}{n}\right) \\ &= \frac{1}{11-1}\left(3.1339 - \frac{(5.83)^2}{11}\right) \\ &= \frac{1}{10}(0.044) \\ &= 0.0044. \end{aligned}$$

Thus, $S = \sqrt{0.0044} \approx 0.0663$.

$\overline{X} = \frac{1}{n}\Sigma X_i = \frac{1}{11}(5.83) = 0.53$

The test statistic is

$$\begin{aligned} T &= \frac{\sqrt{n}\left(\overline{X} - \mu\right)}{S} \\ &= \frac{\sqrt{11}(0.53 - 0.5)}{0.0663} \\ &= \frac{\sqrt{11}(0.03)}{0.0663} \\ &\approx 1.5007. \end{aligned}$$

The critical value is $t_{\alpha/2;(n-1)} = t_{0.05;10} = 1.812$. Reject H_0 if $T > 1.812$ or if $T < -1.812$.

Since $-1.812 < 1.5007 < 1.812$, we do not reject H_0 at the 10% level and conclude that $\mu = 0.5$.

(b) The 90% confidence interval for μ is

$$\begin{aligned} \overline{X} - t_{\alpha/2;(n-1)} \times \frac{S}{\sqrt{n}} &\leq \mu \leq \overline{X} + t_{\alpha/2;(n-1)} \times \frac{S}{\sqrt{n}} \\ 0.53 - t_{0.05;10} \times \frac{0.0663}{\sqrt{11}} &\leq \mu \leq 0.53 + t_{0.05;10} \times \frac{0.0663}{\sqrt{11}} \\ 0.53 - 1.812 \times \frac{0.0663}{\sqrt{11}} &\leq \mu \leq 0.53 + 1.812 \times \frac{0.0663}{\sqrt{11}} \\ 0.53 - 0.0362 &\leq \mu \leq 0.53 + 0.0362 \\ 0.4938 &\leq \mu \leq 0.5662. \end{aligned}$$

2. (a) $H_0 : \mu = 1.6$ against $H_1 : \mu < 1.6$

$n = 6 \qquad \sum X_i = 9 \qquad \sum X_i^2 = 13.548$

Since σ^2 is unknown, we have

$$\begin{aligned} S^2 &= \frac{1}{n-1}\left(\Sigma\left(X_i - \overline{X}\right)^2\right) \\ &= \frac{1}{n-1}\left(\Sigma X_i^2 - \frac{(\Sigma X_i)^2}{n}\right) \\ &= \frac{1}{6-1}\left(13.548 - \frac{(9)^2}{6}\right) \\ &= \frac{1}{5}(0.048) \\ &= 0.0096. \end{aligned}$$

Thus, $S = \sqrt{0.0096} \approx 0.098$.

$\overline{X} = \frac{1}{n}\Sigma X_i = \frac{1}{6}(9) = 1.5$

The test statistic is

$$\begin{aligned} T &= \frac{\sqrt{n}\left(\overline{X} - \mu\right)}{S} \\ &= \frac{\sqrt{6}(1.5 - 1.6)}{0.098} \\ &= \frac{\sqrt{6}(-0.1)}{0.098} \\ &\approx -2.4995. \end{aligned}$$

The critical value is $t_{\alpha;(n-1)} = t_{0.05;5} = 2.015$. Reject H_0 if $T < -2.015$.

Since $-2.4995 < -2.015$, we reject H_0 at the 5% level and conclude that $\mu < 1.6$.

(b) The 95% upper confidence limit for μ is

$$\left(-\infty; \overline{X} + t_{\alpha;(n-1)} \times \frac{S}{\sqrt{n}}\right)$$
$$\left(-\infty; \overline{X} + t_{0.05;5} \times \frac{S}{\sqrt{n}}\right)$$
$$\left(-\infty; 1.5 + 2.015 \times \frac{0.098}{\sqrt{6}}\right)$$
$$(-\infty; 1.5 + 0.0806)$$
$$(-\infty; 1.5806).$$

3. $H_0 : \mu = 100$ against $H_1 : \mu \neq 100$

$n = 16 \qquad \mu = 100 - \sqrt{0.72\sigma^2} \qquad T_0 = \frac{\sqrt{n}\left(\overline{X} - \mu_0\right)}{S} \sim t_{n-1}$

Now

$$\begin{aligned}
\delta &= \frac{\sqrt{n}\,(\mu-\mu_0)}{\sigma} \\
&= \frac{\sqrt{n}\left(100-\sqrt{0.72\sigma^2}-100\right)}{\sigma} \\
&= \frac{\sqrt{n}\left(-\sigma\sqrt{0.72}\right)}{\sigma} \\
&= -\sqrt{n}\sqrt{0.72} \\
&= -\sqrt{16}\sqrt{0.72} \\
&\approx -3.3941.
\end{aligned}$$

Since test is two-sided $\phi = \frac{|\delta|}{\sqrt{2}} = \frac{|-3.3941|}{\sqrt{2}} \approx 2.4.$

Thus, at $\alpha = 0.05,$ the power of the test is approximately $89\%.$

Also at $\alpha = 0.01,$ the power of the test is approximately $67\%.$

4. (a) $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 < \mu_2$

$$\begin{array}{lll}
n_1 = 10 & \sum X_{1i} = 1\,070 & \sum X_{1i}^2 = 115\,990 \\
n_2 = 12 & \sum X_{2i} = 1\,344 & \sum X_{2i}^2 = 152\,328
\end{array}$$

$$\overline{X}_1 = \frac{1}{n_1}\sum X_{1i} = \frac{1}{10} \times 1\,070 = 107 \qquad \overline{X}_2 = \frac{1}{n_2}\sum X_{2i} = \frac{1}{12} \times 1\,344 = 112$$

$$\begin{aligned}
S_1^2 &= \frac{1}{n_1-1}\left(\sum X_{1i}^2 - \frac{\left(\sum X_{1i}\right)^2}{n}\right) & S_2^2 &= \frac{1}{n_2-1}\left(\sum X_{2i}^2 - \frac{\left(\sum X_{2i}\right)^2}{n}\right) \\
&= \frac{1}{10-1}\left(115\,990 - \frac{(1\,070)^2}{10}\right) & &= \frac{1}{12-1}\left(152\,328 - \frac{(1\,344)^2}{12}\right) \\
&= \frac{1}{9}\,(1\,500) & &= \frac{1}{11}\,(1\,800) \\
&\approx 166.6667 & &\approx 163.6364
\end{aligned}$$

$$\begin{aligned}
S_p^2 &= \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} \\
&= \frac{9(166.6667) + 11(163.6364)}{10+12-2} \\
&= \frac{3\,300.0007}{20} \\
&\approx 165
\end{aligned}$$

$$\Longrightarrow S_p = \sqrt{165} \approx 12.8452$$

The test statistic is

$$\begin{aligned} T &= \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \\ &= \frac{(107 - 112) - 0}{12.8452\sqrt{\frac{1}{10} + \frac{1}{12}}} \\ &= \frac{-5}{5.499986051} \\ &\approx -0.9091. \end{aligned}$$

Test is one-tailed. The critical value is $t_{\alpha;(n_1+n_2-2)} = t_{0.05;20} = 1.725$. Reject H_0 if $T < -1.725$.

Since $-0.9091 > -1.725$, we do not reject H_0 at the 5% level and conclude that the means are equal, that is, $\mu_1 = \mu_2$.

(b) The 90% two-sided confidence interval for $\mu_1 - \mu_2$ is

$$\begin{aligned} \left(\overline{X}_1 - \overline{X}_2\right) - t_{\alpha/2;(n_1+n_2-2)} \times S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} &\leq \mu_1 - \mu_2 \leq \left(\overline{X}_1 - \overline{X}_2\right) - t_{\alpha/2;(n_1+n_2-2)} \times S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \\ (107 - 112) - t_{0.05;20} \times 12.8452\sqrt{\frac{1}{10} + \frac{1}{12}} &\leq \mu_1 - \mu_2 \leq (107 - 112) + t_{0.05;20} \times 12.8452\sqrt{\frac{1}{10} + \frac{1}{12}} \\ -5 - 1.725 \times 12.8452\sqrt{\frac{11}{60}} &\leq \mu_1 - \mu_2 \leq -5 + 1.725 \times 12.8452\sqrt{\frac{11}{60}} \\ -5 - 9.4875 &\leq \mu_1 - \mu_2 \leq -5 + 9.4875 \\ -14.4875 &\leq \mu_1 - \mu_2 \leq 4.4875. \end{aligned}$$

5. A large value of ϕ will increase the power $\Longrightarrow$ $|\delta|$ should be the largest, in other words, the maximum.

   Now

   $$\delta = \frac{\mu_1 - \mu_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

   This can only be maximised if $\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ is the smallest. This is achieved if $n_1 = n_2 = 10$ for all $\mu_1 - \mu_2$ and σ.

   δ is maximised if $n_1 = n_2 = 10$.

6. $n_1 = 3 \qquad n_2 = 9 \qquad \mu_1 = \mu_2 + 1.8\sigma\sqrt{2}$

Now

$$\begin{aligned}\delta &= \frac{\mu_1 - \mu_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \\ &= \frac{\mu_2 + 1.8\sigma\sqrt{2} - \mu_2}{\sigma\sqrt{\frac{1}{3} + \frac{1}{9}}} \\ &= \frac{1.8\sigma\sqrt{2}}{\sigma\sqrt{\frac{4}{9}}} \\ &= \frac{1.8\sqrt{2}}{\frac{2}{3}} \\ &\approx 3.8184.\end{aligned}$$

Now $\phi = \dfrac{|\delta|}{\sqrt{2}} = \dfrac{|3.8184|}{\sqrt{2}} \approx 2.7.$

$v = n_1 + n_2 - 2 = 3 + 9 - 2 = 10$

(a) Thus, at $\alpha = 0.05$, the power of the test is approximately 93%.

(b) Also at $\alpha = 0.01$, the power of the test is approximately 73% $\left(\frac{69+77}{2}\right)$.

7. $H_0 : \mu_1 = 2\mu_2$ against $H_1 : \mu_1 > 2\mu_2$

$\Longrightarrow H_0 : \mu_1 - 2\mu_2 = 0$ against $H_1 : \mu_1 - 2\mu_2 > 0$

Now $2\overline{X}_2 \Longrightarrow E\left(2\overline{X}_2\right) = 2E\left(\overline{X}_2\right) = 2\mu_2 \qquad Var\left(2\overline{X}_2\right) = 4Var\left(\overline{X}_2\right)$

$$\overline{X}_1 \sim n\left(\mu_1, \frac{\sigma^2}{n_1}\right) \qquad \overline{X}_2 \sim n\left(\mu_2, \frac{\sigma^2}{n_2}\right)$$

$$2\overline{X}_2 \sim n\left(2\mu_2, \frac{4\sigma^2}{n_2}\right) \qquad \overline{X}_1 - 2\overline{X}_2 \sim n\left(\mu_1 - 2\mu_2, \frac{\sigma^2}{n_1} + \frac{4\sigma^2}{n_2}\right)$$

$$U = \frac{(\overline{X}_1 - \overline{2X}_2) - (\mu_1 - 2\mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{4}{n_2}}}$$

Now $\dfrac{(n_1-1)S_1^2}{\sigma^2} \sim \chi^2_{n_1-1}$ and $\dfrac{(n_2-1)S_2^2}{\sigma^2} \sim \chi^2_{n_2-1}$

$$W = \left[\frac{(n_1-1)S_1^2}{\sigma^2} + \frac{(n_2-1)S_2^2}{\sigma^2}\right] \sim \chi^2_{n_1+n_2-2}$$

$$T = \frac{U}{\sqrt{\frac{W}{n_1+n_1-2}}} \sim t_{n_1+n_1-2}$$

where

$$\begin{aligned} T &= \frac{\left(\frac{(\overline{X}_1-\overline{2X}_2)-(\mu_1-2\mu_2)}{\sqrt{\frac{1}{n_1}+\frac{4}{n_2}}}\right)}{\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_1-2}}} \\ &= \frac{(\overline{X}_1-\overline{2X}_2)-(\mu_1-2\mu_2)}{S_p\sqrt{\frac{1}{n_1}+\frac{4}{n_2}}} \end{aligned}$$

where $S_p^2 = \dfrac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$

If $H_0 : \mu_1 = 2\mu_2 \Longrightarrow \mu_1 - 2\mu_2 = 0$. Thus,

$$T = \frac{(\overline{X}_1-\overline{2X}_2}{S_p\sqrt{\frac{1}{n_1}+\frac{4}{n_2}}}.$$

8. (a) Let Y_i = blood sugar (after) – blood sugar (before)

Patient i	1	2	3	4	5	6	7	8
Y_i	8	6	7	3	12	7	4	9

$n = 8 \quad \sum y_i = 56 \quad \sum y_i^2 = 448$

$\overline{Y} = \dfrac{1}{n}\sum y_i = \dfrac{56}{8} = 7$

$$\begin{aligned} S_y^2 &= \frac{1}{n-1}\left(\sum y_i^2 - \frac{(\sum y_i)^2}{n}\right) \\ &= \frac{1}{8-1}\left(448 - \frac{(56)^2}{8}\right) \\ &= \frac{1}{7}(56) \\ &= 8 \end{aligned}$$

$S_y = \sqrt{8} \approx 2.8284$

$H_0 : \mu = 5$ against $H_1 : \mu > 5$

The test statistic is

$$\begin{aligned} T &= \frac{\sqrt{n}\,(\overline{Y}-\mu)}{S_y} \\ &= \frac{\sqrt{8}\,(7-5)}{2.8284} \\ &= \frac{\sqrt{8}\,(2)}{2.8284} \\ &\approx 2. \end{aligned}$$

The critical value is $t_{\alpha;(n-1)} = t_{0.05;7} = 1.895$. Reject H_0 if $T > 1.895$.

Since $2 > 1.895$, we reject H_0 at the 5% level and conclude that the blood sugar content increases by more than 5 units.

(b) The 95% lower confidence interval is

$$\left(\overline{X} - t_{\alpha;(n-1)} \times \frac{S_y}{\sqrt{n}}; \infty\right)$$
$$\left(\overline{X} - t_{0.05;7} \times \frac{S_y}{\sqrt{n}}; \infty\right)$$
$$\left(7 - 1.895 \times \frac{2.8284}{\sqrt{8}}; \infty\right)$$
$$(7 - 1.895; \infty)$$
$$(5.105; \infty)\,.$$

9. $n_1 = 9 \qquad \overline{X}_1 = 110 \qquad S_1^2 = 180$
$n_2 = 11 \qquad \overline{X}_2 = 120 \qquad S_2^2 = 56$

$H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$

The test statistic is

$$\begin{aligned} T &= \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \\ &= \frac{(110 - 120) - 0}{\sqrt{\frac{180}{9} + \frac{55}{11}}} \\ &= \frac{-10}{\sqrt{25}} \\ &\approx -2 \end{aligned}$$

$$\begin{aligned} v &= \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{S_1^4}{n_1^2(n_1-1)} + \frac{S_2^4}{n_2^2(n_2-1)}} \\ &= \frac{\left(\frac{180}{9} + \frac{55}{11}\right)^2}{\frac{180^2}{9^2(8)} + \frac{55^2}{11^2(10)}} \\ &= \frac{(25)^2}{50 + 2.50} \\ &\approx 11.90. \end{aligned}$$

Interpolating between $v = 11$ and $v = 12$.

$\alpha = 0.10$, $\alpha/2 = 0.05$

$$\begin{aligned} t_{0.05;11.90} &= 1.796 + 0.90(1.782 - 1.796) \\ &= 1.796 - 0.0126 \\ &\approx 1.783 \end{aligned}$$

We reject H_0 if $T < -1.783$ or $T > 1.783$.

Since $-2 < -1.783$, we reject H_0 at the 10% level and conclude that the means are not the same.

10. $\overline{X}_i \sim n\left(\mu_i, \sigma_i^2\right)$. Now $\overline{X}_1 \sim n\left(\mu_1, \frac{\sigma_1^2}{n_1}\right) \Longrightarrow \overline{X}_1 \sim n\left(\mu_1, \frac{2\sigma_2^2}{n_1}\right)$ and $\overline{X}_2 \sim n\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$

$E(\overline{X}_1 - \overline{X}_2) = \mu_1 - \mu_2$

$$\begin{aligned} Var(\overline{X}_1 - \overline{X}_2) &= Var(\overline{X}_1) + Var(\overline{X}_2) \\ &= \frac{2\sigma_2^2}{n_1} + \frac{\sigma_2^2}{n_2} \\ &= \sigma_2^2\left(\frac{2}{n_1} + \frac{1}{n_2}\right) \end{aligned}$$

The test statistic

$$\begin{aligned} T &= \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{2}{n_1} + \frac{1}{n_2}}} \\ &= \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{S\sqrt{\frac{2}{n_1} + \frac{1}{n_2}}} \end{aligned}$$

where $S^2 = \dfrac{\frac{1}{2}(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$

11. $\overline{X}_1 = 40 \qquad S_1^2 = 400 \quad n_1 = 10$

$\overline{X}_2 = 60 \qquad S_2^2 = 720 \quad n_2 = 12$

$\alpha = 0.05$, $\alpha/2 = 0.025$

The degrees of freedom are

$$\begin{aligned} v &= \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{S_1^4}{n_1^2(n_1-1)} + \frac{S_2^4}{n_2^2(n_2-1)}} \\ &= \frac{\left(\frac{400}{10} + \frac{720}{12}\right)^2}{\frac{400^2}{10^2(9)} + \frac{720^2}{12^2(11)}} \\ &= \frac{(100)^2}{505.0505051} \\ &\approx 19.8. \end{aligned}$$

Interpolating between $v = 19$ and $v = 20 \Longrightarrow$ critical value is

$$\begin{aligned} t_{0.025;19.8} &= 2.093 + 0.8(2.086 - 2.093) \\ &= 2.093 - 0.0056 \\ &\approx 2.087. \end{aligned}$$

The 95% two-sided confidence interval for $\mu_1 - \mu_2$ is

$$\begin{aligned} \left(\overline{X}_1 - \overline{X}_2\right) - t_{\alpha/2;v} \times \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} &\leq \mu_1 - \mu_2 \leq \left(\overline{X}_1 - \overline{X}_2\right) - t_{\alpha/2;v} \times \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \\ (40 - 60) - 2.087\sqrt{\frac{400}{10} + \frac{720}{12}} &\leq \mu_1 - \mu_2 \leq (40 - 60) + 2.087\sqrt{\frac{400}{10} + \frac{720}{12}} \\ -20 - 2.087\sqrt{100} &\leq \mu_1 - \mu_2 \leq -20 + 2.087\sqrt{100} \\ -20 - 20.87 &\leq \mu_1 - \mu_2 \leq -20 + 20.87 \\ -40.87 &\leq \mu_1 - \mu_2 \leq 0.87 \end{aligned}$$

12. Testing $H_0 : \mu_1 = \mu_2 = \mu_3$ against $H_1 : \mu_p \neq \mu_p$ for at least one pair $p \neq q$

$k = 3$ $\quad n = 7$ $\quad kn - k = 18$ $\quad k - 1 = 2$

$\overline{X}_1 = 1.9$ $\quad \sum X_{1i} = 13.3$ $\quad \sum X_{1i}^2 = 26.07$

$$\begin{aligned} SS_1 &= \sum (X_{1i} - \overline{X}_1)^2 \\ &= \sum X_{1i}^2 - \frac{\left(\sum X_{1i}\right)^2}{n} \\ &= 26.07 - \frac{(13.3)^2}{7} \\ &= 0.8 \end{aligned}$$

[Note: After entering data in statistics mode this is equal to $n\sigma^2$ or $(n-1)S^2$.]

$\overline{X}_2 = 1.8$ $\quad \sum X_{2i} = 12.6$ $\quad \sum X_{2i}^2 = 23.32$

$$\begin{aligned} SS_2 &= \sum (X_{2i} - \overline{X}_2)^2 \\ &= \sum X_{2i}^2 - \frac{\left(\sum X_{2i}\right)^2}{n} \\ &= 23.32 - \frac{(12.6)^2}{7} \\ &= 0.64 \end{aligned}$$

$\overline{X}_3 = 2.3 \qquad \sum X_{3i} = 16.1 \qquad \sum X_{3i}^2 = 37.75$

$$\begin{aligned} SS_3 &= \sum (X_{3i} - \overline{X}_3)^2 \\ &= \sum X_{3i}^2 - \frac{\left(\sum X_{3i}\right)^2}{n} \\ &= 37.75 - \frac{(16.1)^2}{7} \\ &= 0.72 \end{aligned}$$

$\overline{X} = 2 \qquad \sum\sum X_{ij} = 42 \qquad \sum\sum X_{ij}^2 = 87.14$

$$\begin{aligned} SST &= \sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij} - \overline{X})^2 \\ &= \sum\sum X_{ij}^2 - \frac{\left(\sum\sum X_{ij}\right)^2}{n} \\ &= 87.14 - \frac{(42)^2}{21} \\ &= 3.14 \end{aligned}$$

$$\begin{aligned} SSE &= SS_1 + SS_2 + SS_3 \\ &= 0.8 + 0.64 + 0.72 \\ &= 2.16 \end{aligned}$$

$$MSE = \frac{SSE}{kn - k} = \frac{2.16}{18} = 0.12.$$

$$\begin{aligned} \sum_{i=1}^{3}(\overline{X}_i - \overline{X})^2 &= (1.9 - 2)^2 + (1.8 - 2)^2 + (2.3 - 2)^2 \\ &= 0.14 \end{aligned}$$

$$\begin{aligned} SSTr &= n\sum_{i=1}^{3}(\overline{X}_i - \overline{X})^2 \\ &= 7(0.14) \\ &= 0.98 \end{aligned}$$

$$MSTr = \frac{n\sum_{i=1}^{3}(\overline{X}_i - \overline{X})^2}{k - 1} = \frac{0.98}{2} = 0.49$$

$$F = \frac{MSTr}{MSE} = \frac{0.49}{0.12} \approx 4.0833$$

The ANOVA table is

Source of variation	Sum of squares	Degrees of freedom	Mean square	F
Treatments	0.98	2	0.49	4.0833
Error	2.16	18	0.12	
Total	3.14	20		

The critical value is $F_{0.05;2;18} = 3.55$. Reject H_0 if $F > 3.55$.

Since $4.0833 > 3.55$, we reject H_0 at the 5% level and conclude that there is sufficient evidence to indicate a difference in means.

Multiple comparisons:

$$\begin{aligned}(k-1)F_{\alpha;k-1;kn-k} &= 2 \times F_{0.05;2;18} \\ &= 2 \times 3.55 \\ &= 7.1\end{aligned}$$

$$\begin{aligned}T_{pq} &= \frac{\sqrt{n}(\overline{X}_p - \overline{X}_q)}{\sqrt{2}S} \\ &= \frac{\sqrt{7}(\overline{X}_p - \overline{X}_q)}{\sqrt{2}\sqrt{0.12}} \\ &= \sqrt{\frac{7}{0.24}}(\overline{X}_p - \overline{X}_q)\end{aligned}$$

Reject $H_0 : \mu_p = \mu_q$ if : Now $|T_{pq}| > \sqrt{7.1} \therefore \left|\overline{X}_p - \overline{X}_q\right| > \dfrac{\sqrt{7.1}}{\sqrt{\frac{7}{0.24}}} \approx 0.4934.$

Testing $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$

$\left|\overline{X}_1 - \overline{X}_2\right| = |1.9 - 1.8| = 0.1$

Since $0.1 < 0.4934$, we do not reject H_0 and conclude that the mean for food *A* is the same as the mean for food *B*.

Testing $H_0 : \mu_1 = \mu_3$ against $H_1 : \mu_1 \neq \mu_3$

$\left|\overline{X}_1 - \overline{X}_3\right| = |1.9 - 2.3| = 0.4$

Since $0.4 < 0.4934$, we do not reject H_0 and conclude that the mean for food *A* is the same as the mean for food *C*.

Testing $H_0 : \mu_2 = \mu_3$ against $H_1 : \mu_2 \neq \mu_3$

$\left|\overline{X}_2 - \overline{X}_3\right| = |1.8 - 2.3| = 0.5$

Since $0.5 > 0.4934$, we reject H_0 and conclude that the mean for food *B* is not the same as the mean for food *C*.

13. (a) Testing $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ against $H_1 : \mu_p \neq \mu_p$ for at least one pair $p \neq q$

$k = 4$ $n = 8$ $kn - k = 28$ $k - 1 = 3$

$\overline{X}_1 = 13 \qquad \sum X_{1i} = 104 \qquad \sum X_{1i}^2 = 1\,370.7$

$$\begin{aligned} SS_1 &= \sum (X_{1i} - \overline{X}_1)^2 \\ &= \sum X_{1i}^2 - \frac{\left(\sum X_{1i}\right)^2}{n} \\ &= 1\,370.7 - \frac{(104)^2}{8} \\ &= 18.7 \end{aligned}$$

$\overline{X}_2 = 10 \qquad \sum X_{2i} = 80 \qquad \sum X_{2i}^2 = 809.96$

$$\begin{aligned} SS_2 &= \sum (X_{2i} - \overline{X}_2)^2 \\ &= \sum X_{2i}^2 - \frac{\left(\sum X_{2i}\right)^2}{n} \\ &= 809.96 - \frac{(80)^2}{8} \\ &= 9.96 \end{aligned}$$

$\overline{X}_3 = 12 \qquad \sum X_{3i} = 96 \qquad \sum X_{3i}^2 = 1\,170.68$

$$\begin{aligned} SS_3 &= \sum (X_{3i} - \overline{X}_3)^2 \\ &= \sum X_{3i}^2 - \frac{\left(\sum X_{3i}\right)^2}{n} \\ &= 1\,170.68 - \frac{(96)^2}{8} \\ &= 18.68 \end{aligned}$$

$\overline{X}_4 = 9 \qquad \sum X_{4i} = 72 \qquad \sum X_{4i}^2 = 652.7$

$$\begin{aligned} SS_4 &= \sum (X_{4i} - \overline{X}_4)^2 \\ &= \sum X_{4i}^2 - \frac{\left(\sum X_{4i}\right)^2}{n} \\ &= 652.7 - \frac{(72)^2}{8} \\ &= 4.7 \end{aligned}$$

$\overline{X} = 11 \qquad \sum\sum X_{ij} = 352 \qquad \sum\sum X_{ij}^2 = 4\,004.04$

$$\begin{aligned} SST &= \sum_{i=1}^{k}\sum_{j=1}^{n} (X_{ij} - \overline{X})^2 \\ &= \sum\sum X_{ij}^2 - \frac{\left(\sum\sum X_{ij}\right)^2}{n} \\ &= 4\,004.04 - \frac{(352)^2}{32} \\ &= 132.04 \end{aligned}$$

$$\begin{aligned} SSE &= SS_1 + SS_2 + SS_3 + SS_4 \\ &= 18.7 + 9.96 + 18.68 + 4.7 \\ &= 52.04 \end{aligned}$$

$$MSE = \frac{SSE}{kn-k} = \frac{52.04}{28} \approx 1.8586$$

$$\begin{aligned} \sum_{i=1}^{4}(\overline{X}_i - \overline{X})^2 &= (13-11)^2 + (10-11)^2 + (12-11)^2 + (9-11)^2 \\ &= 10 \end{aligned}$$

$$\begin{aligned} SSTr &= n\sum_{i=1}^{3}(\overline{X}_i - \overline{X})^2 \\ &= 8(10) \\ &= 80 \end{aligned}$$

$$MSTr = \frac{n\sum_{i=1}^{4}(\overline{X}_i - \overline{X})^2}{k-1} = \frac{80}{3} \approx 26.6667$$

$$F = \frac{MSTr}{MSE} = \frac{26.6667}{1.8586} \approx 14.3477$$

The ANOVA table is

Source of variation	Sum of squares	Degrees of freedom	Mean square	F
Treatments	80	3	26.6667	14.3477
Error	52.04	28	1.8586	
Total	132.04	31		

The critical value is $F_{0.05;3;28} = 2.95$. Reject H_0 if $F > 2.95$.

Since $14.3477 > 2.95$, H_0 is rejected at the 5% level and we conclude that there is sufficient evidence to indicate a significant difference in the means of the four brands of feed.

(b) **Multiple comparisons**

$$\begin{aligned} (k-1)F_{\alpha;k-1;kn-k} &= 3 \times F_{0.05;3;28} \\ &= 3 \times 2.95 \\ &= 8.85 \end{aligned}$$

$$\begin{aligned} T_{pq} &= \frac{\sqrt{n}(\overline{X}_p - \overline{X}_q)}{\sqrt{2}S} \\ &= \frac{\sqrt{8}(\overline{X}_p - \overline{X}_q)}{\sqrt{2}\sqrt{1.8586}} \\ &= \sqrt{\frac{8}{3.7172}}(\overline{X}_p - \overline{X}_q) \end{aligned}$$

We reject $H_0 : \mu_p = \mu_q$ if

$$\begin{aligned} |T_{pq}| &> 8.85 \\ \therefore \quad \left|\overline{X}_p - \overline{X}_q\right| &> \frac{\sqrt{8.85}}{\sqrt{\frac{8}{3.7172}}} \\ \left|\overline{X}_p - \overline{X}_q\right| &> 2.0278. \end{aligned}$$

Testing $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$

$\left|\overline{X}_1 - \overline{X}_2\right| = |13 - 10| = 3.$

Since $3 > 2.0278$, we reject H_0 and conclude that there is a significant difference between the mean for brand *A* and the mean for brand *B*.

Testing $H_0 : \mu_1 = \mu_3$ against $H_1 : \mu_1 \neq \mu_3$

$\left|\overline{X}_1 - \overline{X}_3\right| = |13 - 12| = 1.$

Since $1 < 2.0278$, we do not reject H_0 and conclude that the means for brand *A* and brand *C* are the same.

Testing $H_0 : \mu_1 = \mu_4$ against $H_1 : \mu_1 \neq \mu_4$

$\left|\overline{X}_1 - \overline{X}_4\right| = |13 - 9| = 4.$

Since $4 > 2.0278$, we reject H_0 and conclude that there is a significant difference between the means for brand *A* and brand *D*.

Testing $H_0 : \mu_2 = \mu_3$ against $H_1 : \mu_2 \neq \mu_3$

$\left|\overline{X}_2 - \overline{X}_3\right| = |10 - 12| = 2.$

Since $2 < 2.0278$, we do not reject H_0 and conclude that the means for brand *B* and brand *C* are the same.

Testing $H_0 : \mu_2 = \mu_4$ against $H_1 : \mu_2 \neq \mu_4$

$\left|\overline{X}_2 - \overline{X}_4\right| = |10 - 9| = 1.$

Since $1 < 2.0278$, we do not reject H_0 and conclude that there is no significant difference between the means for brand *B* and brand *D*.

Testing $H_0 : \mu_3 = \mu_4$ against $H_1 : \mu_3 \neq \mu_4$

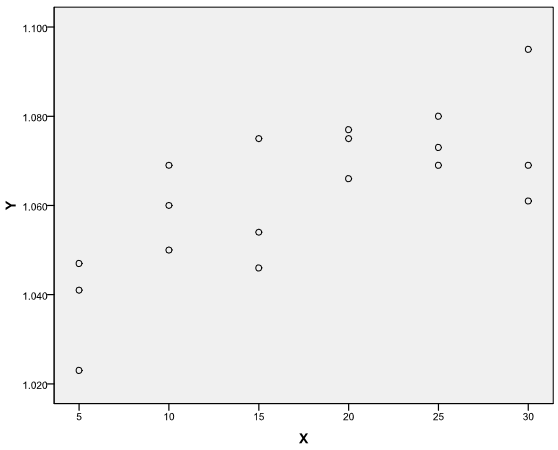
$\left|\overline{X}_3 - \overline{X}_4\right| = |12 - 9| = 3.$

Since $3 > 2.0278$, we reject H_0 and conclude that there is a significant difference between the means for brand *C* and brand *D*.

* * * * *

Exercise 8.1

1.



A scatter plot of Y versus X

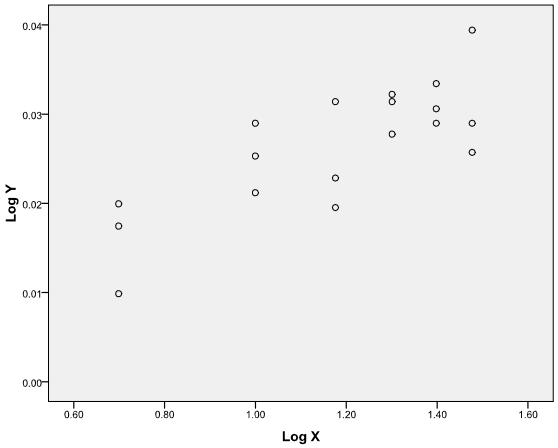
The scatter plot reveals that there is no strong linear relationship. A transformation of the data after taking logs is given below:

$Log_{10}X$	$Log_{10}Y$		
0.7	0.017	0.020	0.010
1	0.025	0.021	0.029
1.2	0.020	0.031	0.023
1.3	0.028	0.031	0.032
1.4	0.033	0.029	0.031
1.5	0.039	0.026	0.029

The transformation makes the points to be more clustered together than the previous original data. Here a suitable transformation is

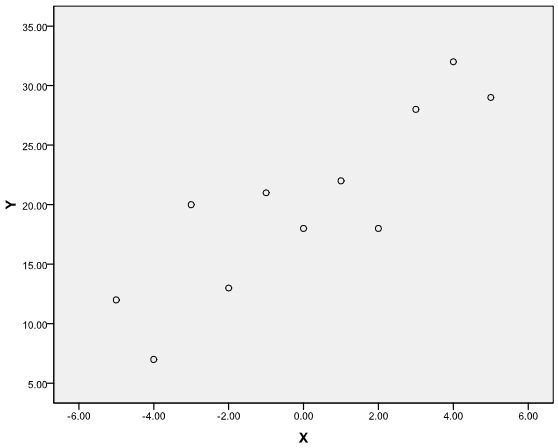
$$Log_{10}Y = \beta_0 + \beta_1 Log_{10}X + \epsilon.$$

This is also evidenced by the scatter plot of the transformed variables shown below:



A scatter plot of $Log_{10}Y$ versus $Log_{10}X$

2. (a)



The points seem to be clustered along a straight line, thus simple linear regression is a suitable model.

(b)

X_i	Y_i	$X_i - \overline{X}$	$\left(X_i - \overline{X}\right)^2$	$Y_i\left(X_i - \overline{X}\right)$	$\widehat{Y}_i$	$Y_i - \widehat{Y}_i$	$\left(Y_i - \widehat{Y}_i\right)^2$
-5	12	-5	25	-60	10	2	4
-4	7	-4	16	-28	12	-5	25
-3	20	-3	9	-60	14	6	36
-2	13	-2	4	-26	16	-3	9
-1	21	-1	1	-21	18	3	9
0	18	0	0	0	20	-2	4
1	22	1	1	22	22	0	0
2	18	2	4	36	24	-6	36
3	28	3	9	84	26	2	4
4	32	4	16	128	28	4	16
5	29	5	25	145	30	-1	1
Total	220	0	110	220			144

$\overline{X}=0$ and $\overline{Y}=0$

$$\begin{array}{rclcrcl}\widehat{\beta}_1 & = & \frac{\sum Y_i\left(X_i-\overline{X}\right)}{\sum\left(X_i-\overline{X}\right)^2} & = & \frac{220}{110} & = & 2\\ \widehat{\beta}_0 & = & \overline{Y}-\widehat{\beta}_1\overline{X} & = & 20-2(0) & = & 20\end{array}$$

The estimated regression line is

$$\widehat{Y}=20+2X$$

$$\begin{aligned}\widehat{\sigma}^2 &= S^2\\ &= \frac{1}{n-2}\sum\left(Y_i-\widehat{Y}_i\right)^2\\ &= \frac{144}{11-2}\\ &= \frac{144}{9}\\ &= 16\end{aligned}$$

(c) The confidence interval for β_1 is

$$\left(\widehat{\beta}_1-t_{\alpha/2;n-2}\times\frac{S}{d};\widehat{\beta}_1+t_{\alpha/2;n-2}\times\frac{S}{d}\right)$$

$$\left(\widehat{\beta}_1-t_{\alpha/2;n-2}\times\frac{S}{\sqrt{\sum\left(X_i-\overline{X}\right)^2}};\widehat{\beta}_1+t_{\alpha/2;n-2}\times\frac{S}{\sqrt{\sum\left(X_i-\overline{X}\right)^2}}\right)$$

$\alpha=0.05$ $\quad\alpha/2=0.025$ $\quad t_{\alpha/2;n-2}=t_{0.025;9}=2.262$
$\widehat{\beta}_1=2$ $\quad S=\sqrt{16}=4$ $\quad d=\sqrt{110}.$

Thus, the 95% confidence interval for $\widehat{\beta}_1$ is

$$\left(\widehat{\beta}_1-t_{\alpha/2;n-2}\times\frac{S}{\sqrt{\sum\left(X_i-\overline{X}\right)^2}};\widehat{\beta}_1+t_{\alpha/2;n-2}\times\frac{S}{\sqrt{\sum\left(X_i-\overline{X}\right)^2}}\right)$$

$$\left(2-2.262\times\frac{4}{\sqrt{110}};2+2.262\times\frac{4}{\sqrt{110}}\right)$$

$$(2-0.8627;2+0.8627)$$

$$(1.1373;2.8627)\,.$$

(d) The mean yield at 340 is

$$\begin{aligned}\widehat{Y}_i &= \beta_0+2\beta_1\\ &= 20+2(2)\\ &= 24.\end{aligned}$$

The confidence interval for $\beta_0 + 2\beta_1$ is

$$\left(\widehat{\beta}_0 + 2\widehat{\beta}_1\right) \pm t_{\alpha/2;n-2} \times S\sqrt{\frac{1}{n} + \frac{\left(X - \overline{X}\right)^2}{d^2}}$$

$$\left(\widehat{\beta}_0 + 2\widehat{\beta}_1\right) \pm t_{\alpha/2;n-2} \times S\sqrt{\frac{1}{n} + \frac{\left(X - \overline{X}\right)^2}{\sum\left(X_i - \overline{X}\right)^2}}$$

$$24 \pm 2.262 \times 4\sqrt{\frac{1}{11} + \frac{(2-0)^2}{110}}$$

$$24 \pm 9.048\sqrt{0.127272727}$$

$$24 \pm 3.2279$$

$$(20.7721; 27.2279)\,.$$

(e) At 360^oC

$$\begin{aligned} X &= (\text{temp} - 300)/20 \\ &= (360 - 300)/20 \\ &= 3. \end{aligned}$$

The 95% confidence interval for the yield if a further experiment is performed at 360^oC is

$$\left(\widehat{\beta}_0 + 2\widehat{\beta}_1\right) \pm t_{\alpha/2;n-2} \times S\sqrt{1 + \frac{1}{n} + \frac{\left(X - \overline{X}\right)^2}{\sum\left(X_i - \overline{X}\right)^2}}$$

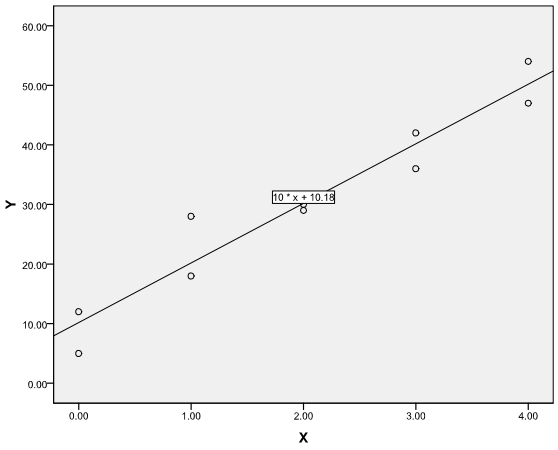
$$(20 + 2(3)) \pm 2.262 \times 4\sqrt{1 + \frac{1}{11} + \frac{(3-0)^2}{110}}$$

$$26 \pm 9.048\sqrt{1.127272727}$$

$$26 \pm 9.7983$$

$$(16.2017; 35.7983)\,.$$

3. (a)



The plot shows that the points are clustered along a straight line. Thus, the simple linear regression would be a suitable model.

(b)

X_i	Y_i	$X_i-\overline{X}$	$\left(X_i-\overline{X}\right)^2$	$Y_i\left(X_i-\overline{X}\right)$	$\widehat{Y}_i$	$Y_i-\widehat{Y}_i$	$\left(Y_i-\widehat{Y}_i\right)^2$
2	30	0	0	0	30.18	−0.18	0.0324
3	36	1	1	36	40.18	−4.18	17.4724
4	47	2	4	94	50.18	−3.18	10.1124
1	28	−1	1	−28	20.18	7.82	61.1524
0	12	−2	4	−24	10.18	1.82	3.3124
2	31	0	0	0	30.18	0.82	0.6724
4	54	2	4	108	50.18	3.82	14.5924
0	5	−2	4	−10	10.18	−5.18	26.8324
1	18	−1	1	−18	20.18	−2.18	4.7524
2	29	0	0	0	30.18	−1.18	1.3924
3	42	1	1	42	40.18	1.82	3.3124
$\sum=22$	332	0	20	200			143.6364

$\overline{X}=\frac{22}{11}=2$ and $\overline{Y}=30.1818$

$$\begin{aligned}\widehat{\beta}_1 &= \frac{\sum Y_i\left(X_i-\overline{X}\right)}{\sum\left(X_i-\overline{X}\right)^2} = \frac{200}{20} = 10\\ \widehat{\beta}_0 &= \overline{Y}-\widehat{\beta}_1\overline{X} = 30.1818-10(2) = 10.1818\end{aligned}$$

∴The estimated regression line is

$$\widehat{Y}=10.1818+10X.$$

(c) The 95% confidence interval for β_1 is

$$\left(\widehat{\beta}_1-t_{\alpha/2;n-2}\times\frac{S}{d};\widehat{\beta}_1+t_{\alpha/2;n-2}\times\frac{S}{d}\right)$$

$t_{\alpha/2;n-2}=t_{0.025;9}=2.262$ $\quad\widehat{\beta}_1=10$ $\quad S^2=\frac{143.6364}{9}=15.9596$
$S=\sqrt{15.9596}\approx 3.9949$ $\quad d=\sqrt{20}.$

Thus, the 95% confidence interval for $\widehat{\beta}_1$ is

$$\begin{aligned}&\left(\widehat{\beta}_1-t_{\alpha/2;n-2}\times\frac{S}{d};\widehat{\beta}_1+t_{\alpha/2;n-2}\times\frac{S}{d}\right)\\ &\left(10-2.262\times\frac{3.9949}{\sqrt{20}};2+2.262\times\frac{3.9949}{\sqrt{20}}\right)\\ &(10-2.0206;10+2.0206)\\ &(7.9794;12.0206).\end{aligned}$$

We are 95% confident that the slope will lie between 7.98 and 12.02. Thus, for every increase of 1 unit in X we expect an increase in Y to range from $(7.98$ to $12.02)$.

(d) The expected yield at $X=4$ is

$$\begin{aligned}Y &= 10.1818+10(4)\\ &= 50.1818.\end{aligned}$$

(e) The 95% confidence interval for the expected yield at $X = 4$ is

$$\left(\widehat{\beta}_0 + \widehat{\beta}_1 X\right) \pm t_{\alpha/2;n-2} \times S\sqrt{\frac{1}{n} + \frac{\left(X - \overline{X}\right)^2}{\sum\left(X_i - \overline{X}\right)^2}}$$

$$50.1818 \pm 2.262 \times 3.9949\sqrt{\frac{1}{11} + \frac{(4-2)^2}{20}}$$

$$50.1818 \pm 9.0364638\sqrt{0.29090909}$$

$$50.1818 \pm 4.8739$$

$$(50.1818 - 4.8739; 50.1818 + 4.8739)$$

$$(45.3079; 55.0557)\,.$$

(f) The 95% confidence interval for the yield which one may expect to obtain, if in a new experiment a dosage of $X = 4$ is applied is

$$\left(\widehat{\beta}_0 + \widehat{\beta}_1 X\right) \pm t_{\alpha/2;n-2} \times S\sqrt{1 + \frac{1}{n} + \frac{\left(X - \overline{X}\right)^2}{\sum\left(X_i - \overline{X}\right)^2}}$$

$$50.1818 \pm 2.262 \times 3.9949\sqrt{1 + \frac{1}{11} + \frac{(4-2)^2}{20}}$$

$$50.1818 \pm 9.0364638\sqrt{1.290909091}$$

$$50.1818 \pm 10.2671$$

$$(50.1818 - 10.2671; 50.1818 + 10.2671)$$

$$(39.9147; 60.4489)\,.$$

4. (a) $Var\left(\widehat{\beta}_0 + \widehat{\beta}_1 X\right) = \sigma^2\left[\frac{1}{n} + \frac{\left(X - \overline{X}\right)^2}{d^2}\right]$

It is a minimum when

$$\begin{aligned}\frac{\left(X - \overline{X}\right)^2}{d^2} &= 0\\ \Longrightarrow \quad \left(X - \overline{X}\right)^2 &= 0\\ \Longrightarrow \quad X - \overline{X} &= 0\\ \Longrightarrow \quad X &= \overline{X}.\end{aligned}$$

(b) The covariance between $\widehat{\beta}_0 + \widehat{\beta}_1 X_1$ and $\widehat{\beta}_0 + \widehat{\beta}_1 X_2$ is

$$\begin{aligned}Cov(\widehat{\beta}_0 + \widehat{\beta}_1 X_1, \widehat{\beta}_0 + \widehat{\beta}_1 X_2) &= Var(\widehat{\beta}_0) + X_1 Cov(\widehat{\beta}_0, \widehat{\beta}_1) + X_2 Cov(\widehat{\beta}_0, \widehat{\beta}_1) + X_1 X_2 Var(\widehat{\beta}_1)\\ &= \sigma^2\left[\frac{1}{n} + \frac{\overline{X}^2}{d^2}\right] + \frac{X_1\left(-\sigma^2\overline{X}\right)}{d^2} + \frac{X_2\left(-\sigma^2\overline{X}\right)}{d^2} + X_2 X_1\frac{\sigma^2}{d^2}\\ &= \frac{\sigma^2}{n} + \frac{\sigma^2\overline{X}^2}{d^2} + \frac{-\sigma^2\overline{X}\left(X_1 + X_2\right)}{d^2} + X_1 X_2\frac{\sigma^2}{d^2}\end{aligned}$$

(c) They are uncorrelated if $Cov = 0$

$$\begin{aligned}
Cov(\widehat{\beta}_0 + \widehat{\beta}_1\left(\overline{X} - k\right), \widehat{\beta}_0 + \widehat{\beta}_1\left(\overline{X} + k\right)) &= Var(\widehat{\beta}_0) + \left(\overline{X} - k\right) Cov(\widehat{\beta}_0, \widehat{\beta}_1) + \left(\overline{X} + k\right) Cov(\widehat{\beta}_0, \widehat{\beta}_1) \\
&\quad + \left(\overline{X} - k\right)\left(\overline{X} + k\right) Var(\widehat{\beta}_1) \\
&= Var(\widehat{\beta}_0) + \left(\overline{X} - k\right) Cov(\widehat{\beta}_0, \widehat{\beta}_1) + \left(\overline{X} + k\right) Cov(\widehat{\beta}_0, \widehat{\beta}_1) \\
&\quad + \left(\overline{X}^2 - k^2\right) Var(\widehat{\beta}_1) \\
&= \sigma^2\left(\frac{1}{n} + \frac{\overline{X}^2}{d^2}\right) + \left(\overline{X} - k\right)\left(\frac{-\sigma^2\overline{X}}{d^2}\right) \\
&\quad + \left(\overline{X} + k\right)\left(\frac{-\sigma^2\overline{X}}{d^2}\right) + \left(\overline{X}^2 - k^2\right)\left(\frac{\sigma^2}{d^2}\right) \\
&= \frac{\sigma^2}{n} + \frac{\sigma^2\overline{X}^2}{d^2} - \frac{\sigma^2\overline{X}^2}{d^2} + \frac{k\sigma^2\overline{X}^2}{d^2} - \frac{\sigma^2\overline{X}^2}{d^2} \\
&\quad - \frac{k\sigma^2\overline{X}^2}{d^2} + \frac{\sigma^2\overline{X}^2}{d^2} - \frac{k^2\sigma^2}{d^2} \\
&= \frac{\sigma^2}{n} + \frac{2\sigma^2\overline{X}^2}{d^2} - \frac{2\sigma^2\overline{X}^2}{d^2} + \frac{k\sigma^2\overline{X}}{d^2} - \frac{k\sigma^2\overline{X}}{d^2} - \frac{k^2\sigma^2}{d^2} \\
&= \frac{\sigma^2}{n} - \frac{k^2\sigma^2}{d^2}
\end{aligned}$$

Now $\Longrightarrow$

$$\begin{aligned}
\frac{\sigma^2}{n} - \frac{k^2\sigma^2}{d^2} &= 0 \\
\frac{\sigma^2}{n} &= \frac{k^2\sigma^2}{d^2} \\
\frac{1}{n} &= \frac{k^2}{d^2} \\
\frac{d^2}{n} &= k^2 \\
\sqrt{\frac{d^2}{n}} &= k \\
\frac{d}{\sqrt{n}} &= k
\end{aligned}$$

Thus, $k = \dfrac{d}{\sqrt{n}}$

If you have worked through all the activities in the workbook and tried to do all the exercises in the study guide, there is a good chance that you will pass STA2601 with a distinction!!