Introduction to statistics
# Contents

Introduction ................................................................. 1–3
Objectives ................................................................. 1–3
Readings ................................................................. 1–4
What is statistics? ...................................................... 1–4
Some terminology ...................................................... 1–5
Accurate data collection .............................................. 1–6
Sampling methods ...................................................... 1–8
Types of variables ...................................................... 1–10
Questionnaire design ................................................. 1–11
Discussion points ....................................................... 1–15
Additional readings .................................................... 1–15
Summary ................................................................. 1–16
Review questions ....................................................... 1–16
Introduction

This Study Guide serves two key purposes. Firstly, it contains an explanation of the statistical techniques and concepts covered in this course. Secondly, it illustrates those techniques and concepts using applied examples. In many cases, the Study Guide contains more information than is in the textbook, so it is essential that students become familiar with its contents.

In addition to the Study Guide, students should ensure that they complete all the required readings from the textbook each week and that review questions are attempted. At the very least, the recommended problems should be attempted, but it is strongly suggested that students also complete a selection from the additional problems list.

As with many subjects, this course is one where understanding comes from practicing the concepts as much as possible. Note that because this is a postgraduate course, a greater depth of understanding is required than in undergraduate courses. You will be tested not just on your recall of the study material, but also on the application and interpretation of the techniques. It is expected that you will learn the statistical techniques well enough to use them in real life situations, analysing and explaining results, and then making sound business and financial decisions.

There are ‘discussion points’ in each module of this Study Guide. These are designed to get you talking with other students (and the teaching staff) about the concepts covered in this course. Flexible students are expected to take advantage of the flexible student mailing list for this purpose. On-campus students can use classroom situations and study groups for these discussions. Exploring these topics with other students is a very important part of gaining a fuller understanding of course material.

This week we learn some of the basic terminology of statistics and consider issues relating to surveys. For those students who are not familiar with Microsoft Excel, this is also a time to learn some of the basics of this package so that in weeks to come, when it is used to perform statistical analyses, you will feel comfortable working with it. This basic understanding of Excel will be assumed knowledge for the remainder of the course.

Objectives

On completion of this module you should be able to:

- define and explain what is meant by ‘statistics’
- define terms such as sample, sampling frame, population, statistic, parameter, descriptive statistics, and inferential statistics
- explain the importance of accurate data collection
- understand and apply survey sample methods
- define the measurement scales nominal, ordinal, interval and ratio
- evaluate the effectiveness and worth of a survey and survey results
- consider the basics of questionnaire design and
- utilise basic Excel functions.
Readings

Source

Students may use either the 4th (2005) or 5th (2008) editions of Levine et al.

Textbook
Levine et al. 4th edition
Ch. 1 and Excel Primer (pp. 1–48)

Or

Levine et al. 5th edition
Ch. 1 (pp. 1–17), Section 7.1 to 7.2 (pp. 252–261) and Excel Companion (pp. 18–30)

Important note: although the Excel Companion (or Excel Primer) may be revision for many students, it is assumed knowledge for the remainder of the course, so must be studied carefully.

Course website
Visit the course website for links to any supplementary material for this week.

What is statistics?

The term statistics has a variety of interpretations in common use today. When many people think of statistics, they think about quoting figures. A common joke says that ‘80% of statistics are made up on the spot’. Although this may not be strictly true, it does give an indication of what the general public really thinks statistics is! Many people believe that statistics is just about quoting batting averages in cricket or points scored in footy matches. The media often use the word ‘statistic’ to refer to deaths. In that sense, becoming a statistic is definitely not a desirable outcome! Given that there are so many different ways this term is used, it is important to understand what statistics is and why students would need to study a course in statistics!

We’ll begin this module by defining the term statistics as we’ll be using it in this course. Like any field of study, statistics has its own language and terminology, so we’ll examine with some basic definitions. These will be terms we’ll use frequently throughout the term. We’ll also look at sampling and data collection and briefly explore good questionnaire design.

Statistics involves planning, collecting, analysing data, and reporting and interpreting results. It lets us take raw data and derive information that will help in sound analytical decision making. Studying statistics means learning how to use and interpret statistical techniques, always keeping that goal of decision making in mind. Although statistics is often considered a branch of mathematics, studying statistics does not mean simply doing mathematical calculations; it also requires a thoughtful justification of the techniques chosen, careful interpretation and application of results, and consideration of any ethical issues involved. Given that complexity, it is important that statisticians have the ability to translate these complex statistical
analyses into a format that someone without statistical training can understand; they need to be able to do all the calculations and then summarise the result in simple terms. The goal of this course is that students gain a level of competency in all these areas.

Statistics is often divided into two main areas: descriptive statistics and inferential statistics. Descriptive statistics involves collecting, summarising and describing data sets; it is exploratory in nature giving an overview of the main characteristics of the data set. Inferential statistics means estimating characteristics of sample data to discover patterns and make inferences about the population (or the future). We will be exploring both of these areas in this course. In many cases, real world problems require that both descriptive and inferential statistics be used. Often descriptive statistics is used to explore and summarise a data set, and with that deeper understanding inferences are able to be made.

Some terminology

**Element**—an object on which a measurement is made. For example, a registered Australian voter or a student at CQU.

**Population**—the set of all possible elements that could be observed. For example, all the registered voters in Australia or all students enrolled at CQU.

**Sample**—a selected portion of the population. It is a collection of sampling units drawn from a sampling frame (see below for definitions of sampling units and frames). A sample is drawn and examined when it would be too expensive or time consuming to look at every single element in the population.

**Parameter**—a characteristic of the population, for example, the preferred prime minister or the average age of students at CQU.

**Statistic**—a characteristic of the sample that is used to estimate a parameter. For example if we took a random sample of 100 CQU students and found the average age of this sample, we could use that to estimate the average age of all students at CQU. Similarly, pollsters often take a sample of Australian voters, and ask them their preferred prime minister. The results from this sample are used to tell us something about the parameter (the preferred prime minister of all registered Australian voters).

**Sampling unit**—non-overlapping collections of elements from a population. Ideally the sampling units are the same as the elements, but sometimes it is cheaper to sample groups. For example, sampling households instead of individual voters might be more convenient and cheaper. Care needs to be taken that sampling larger units (rather than individual elements) does not lead to bias in the results.

**Frame** (or **sampling frame**)—a list of sampling units. For example, the electoral roll is a list of all registered Australian voters and the student records system contains a list of all students enrolled at CQU. The sample is therefore drawn by selecting sampling units from this sampling frame.

**Census**—the measurement or observation of all possible elements from the population. In other words, the sample contains the entire population. For example, every five years the Australian Bureau of Statistics conducts a census of all people in Australia.

**Variable**—a characteristic of an item or an individual.
Accurate data collection

Good quality data is essential for effective decision making in the business environment. Data is obtained from either:

- **a primary source** where the data collector analyses the data. For example, a company analyses information from their own client database.
- **a secondary source** where the data is collected and then made available for others to use. For example, data is collected by the Australian Bureau of Statistics who then make it available for download (sometimes for a fee).

Some primary sources

An experiment

With an experiment, the effects of various treatments are compared in a controlled setting. For example, if two brands of air-bags were tested in new cars, crash test dummies might be placed in cars which have been fitted with the different brands, and then used to measure the potential damage to car occupants. An experiment normally requires careful experimental design and often advanced statistical techniques. It usually involves a number of repetitions to ensure that results accurately reflect the population. For example, testing the air-bags in only one type of car is not likely to give accurate results—that car may have been special or different in some way which would affect the outcome.

Personal interviews

With personal interviews, an interviewer asks questions of the respondent and notes the responses. Because of the personal interaction, there is usually a good response rate. The interviewer can also get extra (non-verbal) information by noting such things as body language. They can ensure that any misunderstandings of the questions are minimised. To properly conduct a personal interview requires training in interview techniques. An untrained interviewer can easily lead the respondent to a certain response (either intentionally or unintentionally). Questions need to be phrased in a neutral manner and the interviewer must be careful that their body language and voice intonation do not lead the respondent to certain answers. When recording the responses, the interviewer could easily make (minor) errors, but this can be minimised by taping the sessions and reviewing them at a later time.

Telephone interviews

Telephone interviews are similar to personal interviews, but are usually cheaper to conduct (since the interviewer is not required to travel). The sample frame may not accurately reflect the population, however, since not everyone has a telephone and not everyone will be home when phone calls are made. In addition, the respondent’s body language cannot be noted. Telephone interviews have a reasonably high non-response rate. Often people find the phone calls annoying or intrusive and do not wish to be involved. This may lead to a sample which poorly reflects the population.
Self-administered questionnaires (paper or web-based)
Self-administered questionnaires are very cheap to administer and are often the first method used by people who are not trained in statistics. They can be physically handed to respondents, mailed or emailed out, or made available via a website. Usually they have a low response rate and so follow up is often required (for example reminder phone calls, emails or letters asking that people complete and return their questionnaires). Often an incentive (such as a ‘free gift’) can be used to encourage participation, but care needs to be used in selecting the form of this gift so that it does not bias the results.

Self-administered questionnaires tend to attract responses mainly from those who have very strong feelings (either positive or negative) on the issue being surveyed. Web-based surveys are particularly prone to receiving only extreme responses. The sample is self-selecting and so it is highly unlikely to reflect the population accurately. Because of these problems, self-administered questionnaires need to be very carefully designed to encourage participation by everyone in the identified sample group and to avoid leading or ambiguous questions. See the discussion on non-probability samples below for more information.

Direct observation
A person counts events as they occur (for example, counting cars as they cross a bridge to get an idea of traffic flow on the bridge). Electronic equipment is sometimes used to measure the events.

Focus groups
Focus groups are a form of direct observation that are often used in market research. These use open-ended questions and allow time for discussion of the issues raised. A focus group has a moderator who leads the discussion. Other group studies include brainstorming, the Delphi Method and the nominal-group technique (not covered in this course).

Survey errors
Coverage error
It is very important that the sampling frame is constructed in such a way that it accurately reflects the population. If this does not occur, such as if certain groups of elements are excluded, the result is coverage error. Coverage error creates selection bias, since certain kinds of elements will not be included in the sample. The sample would then estimate characteristics of the (faulty) sampling frame, rather than the population.

Nonresponse error
Nonresponse error occurs when the survey fails to collect data from all elements in the sample. This error can be because not everyone is willing to respond to a survey (such as discussed in self-administered questionnaires above). The result of nonresponse error is nonresponse bias in the survey results. Follow up of non-responding elements assists in minimising this error and bias.

Sampling error
When taking a sample, the goal is usually that it be drawn from a population in such a way that all the elements have an equal chance of being selected. Because this selection is random, each time a sample is drawn it is likely to contain different individual elements. Sampling error is due to this chance difference from sample to
sample. Although sampling error can be reduced by increasing the sample size, this obviously comes with an increase in cost (in money and/or time).

**Measurement error**

Measurement errors are problems with the recorded responses. These can be the result of ambiguous wording of questions, errors by the respondent or the ‘halo effect’. The halo effect is when a respondent feels the need to please the interviewer. Proper interviewing technique (as discussed in the section on personal interviews above) can help alleviate this problem.

**Ethical issues**

The various survey errors only become ethical problems if the interviewer or survey designer intentionally causes these to occur. This might mean intentionally excluding a group of individuals from a survey (coverage error/selection bias) where it is believed they would respond in a way contrary to the purposes of the survey. When nonprobability samples (discussed shortly) are used to make inferences about the population, this also creates ethical problems.

**Sampling methods**

**Non-probability samples**

With non-probability samples, elements are chosen without considering the probability of occurrence. In many cases, the elements in these samples are self-selected (such as is the case with web-based surveys). Although non-probability samples have advantages such as being quick and cheap to conduct, most statistical techniques rely on probability samples and so cannot be used on data resulting from a non-probability sample. In addition, non-probability samples introduce selection bias into results. For these reasons, and for the remainder of this course, we will work only with probability samples.

**Probability samples**

When a probability sample is drawn, elements are chosen based on knowledge of the probability of occurrence. Probability samples enable the inference of unbiased generalisations about the population. Although it is often difficult to achieve a true probability sample (i.e. it is difficult to ensure absolute randomness in your selection), this should always be the goal in sampling. We will be considering four kinds of probability samples in this course: simple random sampling, systematic sampling, stratified sampling and cluster sampling.

**Sampling with replacement**—randomly selected elements are returned to the frame after they are selected (and so could potentially be re-selected).

**Sampling without replacement**—randomly selected elements are not returned to the frame after selection.
Simple random sample
A simple random sample is one in which every item in the frame has the same chance of being selected. Also, every sample of a certain size has the same chance of being selected as every other sample of the same size. We use $n$ to represent the number of elements in the sample and $N$ for the number of units in the frame (or population).

Random number tables
Random number tables contain random digits from 0 to 9. Numbers are taken from these tables to select sample items. See Table E.1 in Appendix E of Levine et al. (4th and 5th editions) for a random number table. Elements in the sampling frame are given numbers and then, when random numbers are found using the table, the corresponding coded elements are selected. Often numbering systems may already exist for the elements in the sampling frame. For example, we might use invoice numbers, student ID numbers, etc.

(Pseudo) random number generators
Random number generators are similar to random number tables. Most statistical software and many calculators will generate random numbers which can be used in the same way as those taken from a table. Random number generators are often prefaced with the word pseudo since it is not possible to use computers to generate truly random numbers (although algorithms now exist that do a pretty good job).

Systematic sample
Given $N$ individuals in the frame and $n$ in the sample, the frame is partitioned in to $k$ groups where $k = N/n$. An item is chosen randomly from the first $k$ items and then every $k$th item after this is sampled. This method is frequently used in the production of goods and in street polls (among other places) where a simple random sample would be very difficult to draw (since we often won’t know how many goods are going to be produced until after it has happened and people on the street are constantly moving).

For example, if 200 customers ($N$) are in a sampling frame and a sample of 20 ($n$) is required, $k = 200/20 = 10$. An item is randomly chosen from the first 10 (say item 7) and then every 10th item is chosen after that. The sample would be items 7, 17, 27, etc.

Stratified sample
The frame is divided into strata, from within each of which a simple random sample is drawn. Strata are grouped according to a similar characteristic (e.g., we might separate people into three groups: high, medium and low income earners).

Cluster sample
The frame is divided into clusters so that each cluster is representative of the population. A random sample of clusters is drawn and every item within the chosen clusters is studied (a census is conducted within the selected clusters). Cluster sampling will bring savings in travelling time, but it is difficult to ensure that clusters are truly representative of the population (e.g., the suburbs of a city might be a logical choice for clusters, but they are not normally all representative of the entire city).
Types of variables

Variables are described as either qualitative (categorical) or quantitative (numerical):

- **Qualitative (categorical) random variables**—responses are categories such as: yes or no, male or female, low, medium or high income earner, etc.
- **Quantitative (numerical) random variables**—responses are numerical such as: height, weight, time, distance.

Quantitative variables can be either discrete or continuous:

- **Discrete random variables** are numerical responses resulting from counting (they can take only integer values). Examples include the number of students in a classroom, the number of mobile phones models marketed by a company etc.
- **Continuous random variables** arise from a measuring process. Examples include the speed of a car travelling along a highway, your weight, etc.

**Measurement scales**

Qualitative variables can be:

- Nominal—there is no particular order to the categories (e.g., male/female, yes/no)
- Ordinal—there is an order to the categories (e.g., first year at uni, second year at uni, third year at uni, etc.)

Quantitative variables can be:

- Interval—there is no true zero point (e.g., temperature—degrees Celsius & Fahrenheit have different zero points).
- Ratio—there is a true zero point (e.g., a person’s height, weight, etc.).

**Example 1–1**

For each of the following random variables, determine whether the variable is qualitative or quantitative. If the variable is quantitative, determine whether the variable of interest is discrete or continuous. In addition, determine the level of measurement.

(a) Number of mobile phones per household
(b) Mobile phone service provider
(c) Number of text messages sent per month
(d) Length (in minutes) of longest call made during a month
(e) Colour of mobile phone
(f) Monthly charge (in dollars and cents) for calls made
(g) Ownership of a car charge kit
(h) Number of calls made per month
(i) Whether there is a telephone line connected to a computer modem in the household
(j) Whether there is a fax machine in the household
Solution 1–1

(a) Quantitative, discrete (since we talk about whole mobile phones), ratio (since there is a true zero point).

(b) Qualitative, nominal (there will be no particular order to the service providers).

(c) Quantitative, discrete (counting whole messages), ratio (true zero point).

(d) Quantitative, continuous (time is measured on a continuous scale, although here we are rounding this measurement to the nearest minute and so could argue that it becomes discrete and ordinal), ratio (true zero point since we can’t have a call length of a negative time).

(e) Qualitative (categories will be names of colours), nominal (there is no particular ordering to colour names).

(f) Quantitative, continuous (money is measured on a continuous scale, although here we are rounding to the nearest cent), ratio (true zero point).

(g) Qualitative (effectively only two options: own a car charge kit, don’t own a car charge kit), nominal (there is no particular order to own or don’t own).

(h) Quantitative, discrete (counting whole numbers of calls), ratio (true zero point).

(i) Qualitative (connected or not connected), nominal (no particular order to these options).

(j) Qualitative (fax machine present or not present), nominal (no particular order to these options).

Questionnaire design

Questionnaire design is an area where a large amount of information and guidance is available; however, we will only briefly consider it in this course. The way that survey questions are phrased determines whether the variable is quantitative or qualitative and determines the level of measurement. This choice is based on the kind of information that is required.

Quantitative or numerical questions are relatively easy to design and the results of these are easy to analyse. See Example 1–2 (c) below for some examples. It should always be made clear what the units of measure are (for example, dollars and cents, hours, kilograms, months etc.). Listing categories (for example, age categories of ‘under 20’, ‘20 to 29’, ‘30 to 39’, etc.) transforms a quantitative variable into a qualitative variable. Although this often makes people feel more comfortable in responding (people might not want to reveal their exact age, income, etc.) it also results in poorer quality information and makes analysis of the data much more difficult. Because of this, where possible, categorising quantitative variables should be avoided.

Qualitative or categorical questions require a little more thought. The most common kinds of qualitative questions are dichotomous, multiple choice, response scales and open-ended questions.

Dichotomous questions have only two possible outcomes. Examples are true or false, yes or no, male or female etc. These kinds of questions need to be used carefully, as
in the wrong situation they can oversimplify a problem. For example, ‘Do you believe that overseas earnings should be taxed?’ is a question which requires a yes or no response, but, respondents might want to give a qualified yes or no (yes under certain circumstances or no under certain circumstances).

**Multiple choice** questions are appropriate when a finite list of answers exists. Care needs to be taken that options are mutually exclusive (there is no overlap between categories) and exhaustive (every possible option is covered). An example of a multiple choice question is:

<table>
<thead>
<tr>
<th>Which one of the following best describes your primary field of employment (please circle one)?</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Medical or health profession</td>
</tr>
<tr>
<td>B. Education</td>
</tr>
<tr>
<td>C. Business or government</td>
</tr>
<tr>
<td>D. Information technology</td>
</tr>
<tr>
<td>E. Scientific or technical</td>
</tr>
<tr>
<td>F. Other</td>
</tr>
</tbody>
</table>

**Response scales** allow the respondent to indicate their position from a range of options. A couple of examples are:

<table>
<thead>
<tr>
<th>On the following scale, circle the one value which best describes how you rate the ease of use of PHStat2.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extremely easy</td>
</tr>
<tr>
<td>1  2  3  4  5  6  7</td>
</tr>
<tr>
<td>Extremely difficult</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Please tick the one option below that best summarises your response to the following statement:</th>
</tr>
</thead>
<tbody>
<tr>
<td>All major shopping centres should trade on Sundays.</td>
</tr>
<tr>
<td>Strongly disagree</td>
</tr>
<tr>
<td>Disagree</td>
</tr>
<tr>
<td>No opinion</td>
</tr>
<tr>
<td>Agree</td>
</tr>
<tr>
<td>Strongly agree</td>
</tr>
</tbody>
</table>

**Open-ended** questions produce results that are much more difficult to analyse and so should be used sparingly and only in certain circumstances. They have the benefit of allowing the respondent complete freedom in how they answer. Because of this, they are often used as a final question on a survey to allow the respondent to say anything they have not yet had the opportunity to express. An example of an open-ended question is:
Introduction to statistics

What is your opinion of the current fringe benefit tax laws?

Guidelines for designing questions
Always keep the following guidelines in mind when designing questions:

- keep questions simple
- phrase questions so that their meaning is clear to every respondent (avoid ambiguous questions)
- avoid leading questions
- ensure spelling and grammar is correct
- aim for an easy to read and attractive layout
- for qualitative questions, offer an adequate choice of responses (mutually exclusive and exhaustive)
- keep the questionnaire as short as possible
- order questions carefully (early questions should be simple and build rapport with the respondent, group questions by topics and consider how early questions may impact on the thoughts or feelings of the respondent in answering later questions)
- keep questions pertinent to the objectives of the survey and
- pre-test the questionnaire on a small group of people (a pilot study) to observe any errors or short comings.

Example 1–2
The manager of an electronics company is interested in determining whether customers who purchased a digital camera over the past 12 months were satisfied with their purchase. The manager is planning to survey these customers using the contact information given on warranty cards submitted after the purchases.

(a) Describe the population and frame. What differences are there between the population and the frame? How might these differences affect the results?

(b) Develop three qualitative questions that you feel would be appropriate for this survey.

(c) Develop three quantitative questions that you feel would be appropriate for this survey.

(d) How could a simple random sample of warranty cards be selected?

(e) If the manager wanted to select a sample of warranty cards for each brand of digital camera sold, how should the sample be selected? Explain.
Solution 1–2

(a) The population is all the customers who have purchased a digital camera in the past 12 months.

The sampling frame is the list of customers who have purchased a digital camera in the past 12 months and have returned the warranty card. This would be obtained from the records maintained by the company (presumably they will maintain a database of customers).

Not every customer who purchased a digital camera will return the warranty card. For this reason, the sampling frame is likely to be smaller than the population. It may also be that particular subsets of the population are more likely to return the warranty cards and so this may lead to bias in the results of the survey (the sample may not be representative of the population).

(b) There are many possible answers to this question, but three examples are:

1. What is your gender (please circle one)?
   Male    Female

2. What brand of digital camera did you purchase (please tick one)?
   □ Pentax
   □ Nikon
   □ Canon
   □ Olympus
   □ Sony
   □ Panasonic
   □ HP
   □ Other (please specify): ____________________________

3. How satisfied are you with your digital camera (please tick one)?
   □ Very dissatisfied
   □ Dissatisfied
   □ No opinion
   □ Satisfied
   □ Very satisfied

(c) There are many possible answers to this question, but three examples are:

1. What price (in dollars) did you pay for your digital camera?
   $ ___________ .00

2. How long ago (in months) did you purchase your digital camera?
   ___________ months
3. How many times have you brought your digital camera in for service or repair since you purchased it?

(d) The warranty cards will probably have some numbering system. This means that random numbers can be generated either by computer or from a random number table. The warranty cards that correspond to these random numbers can then be selected. If the cards are not already numbered, they can be allocated numbers (perhaps in the order of date and time the camera was purchased) and then a random sample selected.

(e) This would require stratified sampling, where the strata are the types of digital camera. Since each strata is likely to be a different size, the percentage of customers within each strata should be established (e.g., 20% purchased a Pentax camera) and then that percentage of the total sample be drawn from each stratum (so 20% of the sample size will be taken from those who purchased Pentax cameras).

**Discussion points**

**Discussion point 1–1**

Examine and complete Problem 7.10 from your textbook (p. 261) (Problem 1.21 on p. 21 of the 4th edition). What ethical issues are associated with this scenario? Give an example of how this information could potentially be misused.

**Discussion point 1–2**

Examine and complete Problem 7.56 from your textbook (p. 277) (Problem 1.48 on p. 24 of the 4th edition). Compare your answer with other students. What conclusion can you reach regarding Internet polls? What solutions can you offer?

**Discussion point 1–3**

Examine and complete Problem 1.9 from the textbook (p. 11, in the 5th edition only). Compare your answer with other students. When might it be appropriate to use of each of these two styles? When would it not be appropriate? What conclusion can you reach about surveys and useful information?

**Additional readings**

If you are intending to use Microsoft Excel and PHStat2 on your home computer for this course, then please ensure that you install PHStat2 by following the instructions in Appendix F (or G in the 4th edition) of the textbook.

**Textbook**

Levine et al. 4th edition

Appendix G (PHStat2 User’s Guide)

Or
Introduction to statistics

Levine et al. 5th edition
Appendix F (Microsoft Excel and PHStat2 FAQs)

Students using the 4th edition may also find Appendices F & H to be useful.

Textbook
Levine et al. 4th edition
Appendices F & H

Summary
Now that you have completed this module, turn back to the objectives at the beginning of the module. Have you achieved these objectives?

Ensure that you attempt the recommended problems in the list of review questions below and at least a sample of problems from the optional list. This will help you to identify any areas of difficulty you have in achieving the module’s objectives.

Review questions

Recommended problems
Levine et al. 4th edition: Questions 1.2, 1.4, 1.6, 1.8, 1.14, 1.18, 1.22, 1.24, 1.29 to 1.43, 1.50, 1.52 and 1.58 from the textbook.

Levine et al. 5th edition: Questions 1.6, 1.8, 1.10, 1.12 to 1.18, 1.24, 1.25, 1.28, 7.2, 7.4, 7.6, 7.7, 7.8, 7.12, 7.14, 7.44 to 7.49, 7.58 and 7.60 from the textbook.

Optional problems
Levine et al. 4th edition: Choose a selection of problems from Questions 1.1, 1.3, 1.5, 1.7, 1.9 to 1.13, 1.15 to 1.17, 1.19 to 1.21, 1.23, 1.25 to 1.28, 1.44 to 1.49, 1.51, 1.53 to 1.57 and 1.59 to 1.61.

Levine et al. 5th edition: Choose a selection of problems from Questions 1.1 to 1.5, 1.7, 1.11, 1.21 to 1.23, 1.26, 1.27, 7.1, 7.3, 7.5, 7.9, 7.11, 7.13, 7.15, 7.16, 7.57 and 7.59.
Presenting data
Contents

Introduction ............................................................................................................... 2–3
Objectives .................................................................................................................. 2–3
Readings .................................................................................................................... 2–3
Graphical displays for qualitative data ................................................................. 2–4
Graphical displays for quantitative data ............................................................... 2–12
Scatterplots and times series plots ........................................................................ 2–23
Important graphical considerations ...................................................................... 2–26
Discussion points ..................................................................................................... 2–27
Summary ................................................................................................................. 2–28
Review questions .................................................................................................... 2–28
Introduction

There is an old saying which says ‘a picture is worth a thousand words’. When examining data sets we could just as equally say ‘a picture is worth a thousand numbers’. Most people prefer to see a picture rather than read page after page of numbers. Human beings are generally visual creatures and so it is important that we can present data in appealing and accurate ways. Whether we’re trying to explain to a client the need to cut back their spending or interpret a company’s annual report, it’s important that we can both understand pictorial summaries of data and create our own.

This week we explore ways to present data graphically. Because this is an area with much potential for misuse, it is one we should consider carefully. The media offers many examples of misleading graphs and statistics (as an exercise, after you complete this module, try looking in your local paper, a news website or watch the evening news for some examples of poor graphical displays). We are going to explore ways to produce simple, accurate and reliable graphical summaries of data and critique some poor examples. Remember that you’ll need to know how to do these calculations by hand, as well as using Excel and PHStat2.

Objectives

On completion of this module you should be able to:

- graph a bar chart, pie chart and grouped (side-by-side) bar chart (by hand and using Excel/PHStat2)
- produce (by hand and using Excel/PHStat2) and interpret a Pareto chart
- produce a stem-and-leaf plot (by hand and using Excel/PHStat2)
- construct a frequency distribution (by hand and using Excel/PHStat2)
- plot an ogive and histogram (by hand and using Excel/PHStat2)
- produce a scatterplot and time series plot (by hand and using Excel/PHStat2)
- interpret the data presentations listed above, and apply the results and conclusions in real world examples and
- discover and describe common graphical errors, explore ethical issues related to these and explain how to overcome graphical errors.

Readings

**Source**

Textbook  
Levine et al. 4th edition  
Ch. 2 & Excel Handbook Ch. 2 (pp. 94–101)

Or
Graphical displays for qualitative data

Pie charts
A pie chart allows data to be displayed using a circle divided into slices. The entire circle represents 100% of the data and the slices represent the percentage breakdown of categories. In general, it is more difficult to interpret the relative sizes of the categories in a pie chart than in other graphical displays (such as the bar chart or Pareto chart). A pie chart is preferred when you are most interested in comparing the proportion of the entire data set that is in each category. When the number of categories is large, a pie chart can be difficult to interpret and other graphical types would be preferred.

We will illustrate the construction of a pie chart by hand using a (very) simple example. A group of two hundred members of a particular gym were asked to name their preferred milk type. It was found that 120 preferred no fat milk, 60 preferred low fat milk and 20 preferred full cream milk. This information is summarised in the first two columns of the table below.

Next we need to calculate what percentage of gym members fall into each of the three categories. Since there are 20 who prefer full cream milk, this corresponds to
\[
\frac{20}{200} \times 100\% = 10\%.
\]
Similar for the low fat category we find
\[
\frac{60}{200} \times 100\% = 30\%.
\]
and for the no fat category
\[
\frac{120}{200} \times 100\% = 60\%.\]

Finally, we need to calculate the number of degrees that should be allocated to each category. Remembering that there are 360 degrees in a circle, we find that
\[
10\% \text{ of } 360 = 36^\circ \text{ should be allocated for the full cream category, } 30\% \text{ of } 360 = 108^\circ \text{ for the low fat category and } 60\% \text{ of } 360 = 216^\circ \text{ for the no fat category.}
\]

Clearly to graph this by hand you would need to use a protractor to measure the appropriate degrees. For this reason, in an exam situation, you would only be required to graph a fairly simple pie chart. The pie chart produced using Excel follows.

<table>
<thead>
<tr>
<th>Milk preference</th>
<th>Frequency</th>
<th>Percentage</th>
<th>Degrees</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full cream</td>
<td>20</td>
<td>10</td>
<td>36</td>
</tr>
<tr>
<td>Low fat</td>
<td>60</td>
<td>30</td>
<td>108</td>
</tr>
<tr>
<td>No fat</td>
<td>120</td>
<td>60</td>
<td>216</td>
</tr>
<tr>
<td><strong>Total:</strong></td>
<td><strong>200</strong></td>
<td><strong>100</strong></td>
<td><strong>360</strong></td>
</tr>
</tbody>
</table>
Presenting data

Pie chart of milk preferences for 200 members of a particular gym

[Diagram showing a pie chart with percentages for Full cream (60%), Low fat (30%), and No fat (10%).]

Important note: the fancy three-dimensional and exploded pie charts available in Excel should be avoided. These tend to distort the view of the relative sizes of categories and so violate principles of graphical excellence (see discussion later).

Bar charts and grouped bar charts

In a bar chart, each category is represented by a bar, the length of which corresponds to the frequency (or percentage) for that category. It is different from a histogram (discussed later) in two important ways:

1. a bar chart is used for qualitative (categorical) data whereas a histogram is used for quantitative (numerical) data and
2. the bars on a bar chart must have a gap between them whereas on a histogram the bars must touch.

When the categories are number based (such as, yearly sales figures, response to a survey question using a five-point scale, etc.) the bars are vertical with labels across the x-axis. When categories are word-based (such as, in the milk preference example) the bars are horizontal with labels up the y-axis.

A bar chart might be preferred to a pie chart where you were most interested in comparing categories. Bar charts can be based on raw frequencies or percentages, depending on the situation and on the purpose of the graph.

Returning to the milk preference example, and using the frequencies, the following bar chart was produced (using Excel). Note how the bars are plotted horizontally.
**Presenting data**

Bar chart of milk preferences for 200 members of a particular gym

<table>
<thead>
<tr>
<th>Milk preference</th>
<th>Total frequency</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full cream</td>
<td>20</td>
<td>16</td>
<td>4</td>
</tr>
<tr>
<td>Low fat</td>
<td>60</td>
<td>40</td>
<td>20</td>
</tr>
<tr>
<td>No fat</td>
<td>120</td>
<td>50</td>
<td>70</td>
</tr>
</tbody>
</table>

Note that in this example, the comparison might have been better made using the relative frequencies within each category.
Pareto charts

A Pareto chart is essentially a vertical bar chart with a couple additional features. Firstly, categories are ordered from most to least frequent, and secondly, a cumulative percentage line is included on the same chart.

Pareto charts are commonly used in Total Quality Management, where a key consideration is the search for causes of problems in products and processes. Pareto analysis involves tallying the number and types of defects which occur with a particular product or service. The resulting data is displayed using a Pareto chart.

In practice, it is often the case that only a few categories are responsible for most of the problems. Pareto charts make it easy to distinguish between larger numbers of categories and allow the separation of the ‘vital few’ from the ‘trivial many’ categories. The ‘80/20 rule’ says that about 80 percent of the quality problems can be attributed to only 20 percent of the categories. In other words if you want to improve quality you should identify (and hopefully solve) the ‘vital few’ problems. Solving ‘trivial many’ problems is far less likely to result in a significant improvement to your quality and processes.

We will illustrate Pareto charts using a slightly more complex data set than our milk preference example. A group of students were asked for their views on a particular course. Those who had struggled with the course identified a number of perceived problems with the course structure, material and teaching. The following table summarises these results, by listing the problems identified and the number of times each problem was identified.

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Problem</th>
</tr>
</thead>
<tbody>
<tr>
<td>34</td>
<td>Failure to keep up with reading</td>
</tr>
<tr>
<td>2</td>
<td>Teaching staff poorly prepared</td>
</tr>
<tr>
<td>26</td>
<td>Failure to complete required homework exercises</td>
</tr>
<tr>
<td>3</td>
<td>Course too difficult</td>
</tr>
<tr>
<td>9</td>
<td>Computer software difficult to learn</td>
</tr>
<tr>
<td>8</td>
<td>Course material is not interesting</td>
</tr>
<tr>
<td>5</td>
<td>Teaching staff not available when needed</td>
</tr>
<tr>
<td>11</td>
<td>Study Guide and textbook difficult to understand</td>
</tr>
<tr>
<td>98</td>
<td>Total</td>
</tr>
</tbody>
</table>

The first step in producing a Pareto chart is to order categories from most to least frequent (see table below). Then, to produce the cumulative percentage curve, we need to calculate the relative frequency of each category and then, from this, the cumulative relative frequency. The relative frequency is found by dividing the frequency by the sum of all frequencies. For the first category this means $34/98 \approx 0.346939$ (to 6 decimal places) and then expressing this as a percentage (34.6939%). Similarly for the second category we find $26/98 \times 100\% = 26.5306\%$. See the table below for the remaining values. The tables in the cumulative column are found by adding each new category to the running total. For example, the two most common categories account for $34.6939 + 26.5306 = 61.2245\%$ of the problems.
Presenting data

<table>
<thead>
<tr>
<th>Problem</th>
<th>Frequency</th>
<th>Relative frequency (%)</th>
<th>Cumulative relative frequency (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Failure to keep up with reading</td>
<td>34</td>
<td>34.6939</td>
<td>34.6939</td>
</tr>
<tr>
<td>Failure to complete required homework exercises</td>
<td>26</td>
<td>26.5306</td>
<td>61.2245</td>
</tr>
<tr>
<td>Study Guide and textbook difficult to understand</td>
<td>11</td>
<td>11.2245</td>
<td>72.4490</td>
</tr>
<tr>
<td>Computer software difficult to learn</td>
<td>9</td>
<td>9.1837</td>
<td>81.6327</td>
</tr>
<tr>
<td>Course material is not interesting</td>
<td>8</td>
<td>8.1633</td>
<td>89.7960</td>
</tr>
<tr>
<td>Teaching staff not available when needed</td>
<td>5</td>
<td>5.1020</td>
<td>94.8980</td>
</tr>
<tr>
<td>Course too difficult</td>
<td>3</td>
<td>3.0612</td>
<td>97.9592</td>
</tr>
<tr>
<td>Teaching staff poorly prepared</td>
<td>2</td>
<td>2.0408</td>
<td>100</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>98</strong></td>
<td><strong>100</strong></td>
<td></td>
</tr>
</tbody>
</table>

Next, a vertical bar chart is produced with the frequency on the left-hand y-axis and the cumulative percentage on the right-hand y-axis. The default graph produced by Excel needs a little work to ensure that it meets standards of graphical excellence. A slightly modified graph is given below. Note that category names have been abbreviated in some cases to assist in making the graph more readable. This needs to be done with extreme care, however, so that it does not cause the loss of vital information contained in the category descriptions.

Note that this example is one where the 80/20 rule does not perfectly describe the behaviour of the data. It is clear that 80% is not reached until the fourth category. It is
also clear, however, that most of the problems identified relate to the first two or three categories.

Many texts recommend creating Pareto charts so that the 100% value on the cumulative frequency axis lines up with the total frequency on the frequency axis. This causes the cumulative curve to rest on top of the first bar and proceeds upwards from there. The Pareto chart using this format would look as follows:
Example 2–1

A DVD hire company deals with a number of complaints regarding their rental DVDs. The number of times each complaint occurred is given in the table below.

<table>
<thead>
<tr>
<th>Complaint</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scratched disc</td>
<td>125</td>
</tr>
<tr>
<td>Dirty disc</td>
<td>116</td>
</tr>
<tr>
<td>Cracked disc</td>
<td>21</td>
</tr>
<tr>
<td>Wrong DVD</td>
<td>54</td>
</tr>
<tr>
<td>Too expensive</td>
<td>39</td>
</tr>
<tr>
<td>Coarse language</td>
<td>26</td>
</tr>
<tr>
<td>Explicit content</td>
<td>41</td>
</tr>
<tr>
<td>Boring</td>
<td>29</td>
</tr>
<tr>
<td>Too violent</td>
<td>18</td>
</tr>
<tr>
<td>Too soppy</td>
<td>27</td>
</tr>
<tr>
<td>Not funny</td>
<td>33</td>
</tr>
<tr>
<td>Rental period too</td>
<td>14</td>
</tr>
<tr>
<td>Bad movie</td>
<td>12</td>
</tr>
<tr>
<td>Rude staff</td>
<td>9</td>
</tr>
<tr>
<td>Store closed</td>
<td>4</td>
</tr>
</tbody>
</table>

Using this data:

(a) construct a pie chart

(b) construct a bar chart

(c) construct a Pareto diagram

(d) which graphical method do you think is best to portray this data?

(e) based on the results of (a) through (d), what conclusions can you make concerning the most common complaints at the DVD hire company?

Solution 2–1

(a) Below is the default pie chart produced by PHStat2. Clearly with so many categories, the default pie chart is hard to read.
(b) Below is the bar chart produced by PHStat2. Ensure that you are confident in producing a bar chart by hand as well as using PHStat2.

![Bar Chart]

(c) Below is the Pareto chart produced by PHStat2. Ensure that you are confident in producing a Pareto chart by hand as well as using PHStat2.

![Pareto Diagram]
Presenting data

(d) In this case, the pie chart can be eliminated since the labels on this are too close together and so very difficult to read. It is also difficult to distinguish between the similarly sized slices in the pie chart. Although this graph could have been improved with a little work (changing font sizes, re-sizing the pie chart itself, etc.), there are a lot of categories and this makes visual comparisons of the sizes of the slices difficult to make.

With the bar chart, it is also difficult to distinguish between categories with similar frequencies, particularly given that, in this example, the similar sized categories are not next to each other.

The Pareto chart has the advantage over the bar chart that it sorts the complaints from most frequent to least frequent. This makes it easier to see quickly what the most common complaints are in this example. The Pareto chart also includes the cumulative polygon on the same graph.

Therefore the Pareto chart appears to be the best graphical method for this data set.

(e) The two most common complaints received by the DVD hire company are clearly scratched and dirty discs. Together these account for 41% of the complaints. Although these two categories don’t quite make up 80% of the frequencies, they do clearly stand out from the remaining categories. In considering which are the ‘vital few’ problems, these are the most likely candidates. These problems might both be solved by (more) regular checking and cleaning of the discs. The third most common complaint is ‘wrong DVD’.

Graphical displays for quantitative data

With the graphical displays discussed until now, the data was qualitative. Usually it related to frequencies of occurrence of various word (or discrete number) categories. Often, however, data comes to us as just raw numbers, such as daily sales figures, weekly petrol costs, the salaries of the employees of a firm or the cost of an identical basket of goods from various supermarkets. Different graphical displays are used to summarise quantitative data. We will examine a few of these here.

Stem-and-leaf diagrams

Stem-and-leaf plots are a quick way to organise data into groups. They were designed to enable this to be easily done by hand, but Excel/PHStat2 (and other software) also allows this to be done quickly. For smaller data sets, they are particularly useful as an alternative to a histogram. The digits of the data values are separated into a stem and a leaf. For example, if we had a data value of 49, we might make the 4 the stem and the 9 the leaf. When there are more than two digits in the numbers, the decision of how to make this division depends on the type of data and the choice of the person producing the graph. The goal is to find a balance between too many and too few categories, by choosing stems and leaves appropriately. See the discussion below for some pointers on how to make those choices. But first, let’s look at a simple example.
Twenty office workers are randomly selected, and their weekly spending on newspapers is recorded with the following results (expressed in whole dollars):

\[
\begin{array}{cccccccccccc}
0 & 10 & 45 & 30 & 65 & 80 & 14 & 18 & 21 & 39 \\
54 & 8 & 26 & 28 & 36 & 41 & 5 & 12 & 14 & 19 \\
\end{array}
\]

The first step in creating the stem-and-leaf display is to order the data from smallest to largest:

\[
\begin{array}{cccccccccccc}
0 & 5 & 8 & 10 & 12 & 14 & 14 & 18 & 19 & 21 \\
26 & 28 & 30 & 36 & 39 & 41 & 45 & 54 & 65 & 80 \\
\end{array}
\]

This data consists of one or two digit numbers, so we can choose the first digit as the stem and the second as the leaf. For the one digit numbers, the stem would be 0 and the leaf would be the digits we are given (0, 5 and 8). We can see that the smallest stem will be 0 (from data value 0) and the largest will be 8 (from 80).

We would then write these values down the side of the page as follows:

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

Now we can begin to add the leaves one by one. The first data value is 0, so we should add a 0 to the right of the 0 stem. Then for 5, we add a 5 next to the 0 stem etc.

```
0 0 5
1
2
3
4
5
6
7
8
```
If we continue in this fashion for the remainder of the data set, the following stem-and-leaf diagram will result:

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

Choosing stems and leaves for messy data

There are many cases where the data do not lend themselves so neatly to stems and leaves. For example, for data such as 0.0149, 394.235 or 190 653 we cannot simply choose the first digit as the leaf and the next as the stem—we would lose too much information.

Often there are a number of ways we could make the choice and still end up with a nice stem-and-leaf plot. We need to choose stems in such a way that we have enough to see the shape of the data set without spreading the data into so many categories that we lose all the information. Anything from about 5 to 15 stems is appropriate depending on the size of the data set. Generally, small data sets need fewer stems and larger data sets need more.

Let’s demonstrate using some examples.

1. Given the following data values (a subset of a data set): 0.0149, 0.9832, 0.2532, 0.4501, 0.7019, … it might be sensible to round the numbers to two decimal places, use the first digit after the decimal point (0, 9, 2, 4, 7, …) as the stem, and the second digit as the leaves (1, 8, 5, 5, 0, …).
2. Given the following data values: 394.235, 388.583, 392.891, 393.998, 397.852, … it might be sensible to round the number off to one decimal place, use the first three digits as the stems (394, 388, 392, 394, 397, …) and digit after the decimal point as the leaf (2, 6, 9, 0, 9, …). Note how the rounding has affected these values (for example 393.998 rounds to 394.0).
3. Given the following data values 190653, 121987, 154028, 161923, … it might be sensible to round the numbers to three significant figures (so 190653 would become 191000, 121987 would become 122000, 154028 would become 154000 and 161923 would become 162000, …), use the first two digits as the leaves (19, 12, 15, 16, …) and the third digit (the third significant figure) as the leaves (1, 2, 4, 2, …).

Frequency distribution and frequency polygon

A frequency distribution is a way of summarising numerical data in the form of class intervals and frequencies. The graph produced by plotting the frequency distribution is called a frequency polygon. In order to produce a frequency distribution, you must
first decide how many classes are required and then establish the boundaries of each class.

**Classes and boundary points**

Usually between 5 and 15 classes are used, but the exact number depends on the size and type of data you have, and also depends on the statisticians’ judgement. Selecting too few or too many will result in a loss of useful information (particularly when graphed).

Once the number of classes is chosen, the following formula can be a useful in determining the width of each class interval:

\[
\text{Interval width} \approx \frac{\text{Range}}{\text{Number of classes}}
\]

The range is found via: \( \text{Range} = \text{highest data value} - \text{lowest data value} \).

Returning to the example of the amount spent on newspapers, let’s say we wanted to establish a frequency distribution with 8 classes. We find that the range is \( 80 - 0 = 80 \). The interval width would therefore be \( \approx \frac{80}{8} = 10 \). The first interval will have a lower bound equal to the first data point and an upper bound of the first data point plus the interval width: \( 0 + 10 = 10 \). So that the next class does not include the value 10 as well (the classes cannot overlap), we will say that the first class is ‘0 to less than 10’. The second class will therefore be ‘10 to less than 20’ since we add another class width of 10. See the table that follows for the remaining classes. Note that in this case we have been forced to add another class to include our maximum data value, since it would not be included in the eighth class ‘70 to less than 80’ (it is not always necessary to do this!).

Sometimes using the class width and boundaries found with the formula above can end up being quite messy (for example, it may give awkward endpoints). When doing this by hand we should try to use nice round boundaries wherever possible (and leave messy exactness to the computer). We might use the calculated class width and endpoints values as a starting point, rounding them to more workable numbers. So, for example, if the range was 30 and the number of intervals chosen was seven we would get a class width of 4.2857142857… we might choose to round this to one decimal place, 4.3. The rule for calculating boundaries should be viewed as a rule of thumb (a guide), not as absolute law. As always, it’s important to use your common sense.

**Tallying**

Once class intervals are established, the next step is to determine the number of data values that fall within each class. Using tally marks is a simple way of doing this. Normally we mark with a diagonal line up to four times and on the fifth we put a line through the preceding four values like this: ///. We then start a new set of diagonal lines. So, for example, if we had // // we would know that there were 12 (5+5+2) data points in that class. In this way we can quickly look at the category and easily see how many values have fallen in that class. Completing the tally for our newspaper spending example, we find the tally values as given in the table which follows.

**Frequency distribution**

The frequency column in the table below summarises the values in the tally column. The tally column is not a necessary part of a frequency distribution, but it is helpful
when doing this by hand. The frequency distribution consists of just the class and frequency columns of the table.

<table>
<thead>
<tr>
<th>Class</th>
<th>Tally</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 to less than 10</td>
<td>///</td>
<td>3</td>
</tr>
<tr>
<td>10 to less than 20</td>
<td>///</td>
<td>6</td>
</tr>
<tr>
<td>20 to less than 30</td>
<td>///</td>
<td>3</td>
</tr>
<tr>
<td>30 to less than 40</td>
<td>///</td>
<td>3</td>
</tr>
<tr>
<td>40 to less than 50</td>
<td>///</td>
<td>2</td>
</tr>
<tr>
<td>50 to less than 60</td>
<td>/</td>
<td>1</td>
</tr>
<tr>
<td>60 to less than 70</td>
<td>/</td>
<td>1</td>
</tr>
<tr>
<td>70 to less than 80</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>80 to less than 90</td>
<td>/</td>
<td>1</td>
</tr>
</tbody>
</table>

**Class midpoints**

Class midpoints (sometimes called class marks) are the midpoint of each class interval. They are exactly halfway across the class interval and are calculated as the average of the two class endpoints. In our example above, the midpoint of the first class is halfway from 0 to 10, or in other words 5. The remaining midpoints will therefore be 15, 25, 35, 45, 55, 65, 75 and 85.

**Frequency polygon**

A frequency polygon is simply a line graph of the frequency distribution values. For the newspaper spending example, the following frequency polygon is produced.

**Ogive**

An ogive (sometimes called a cumulative percentage polygon or cumulative relative frequency polygon) is a graph of the cumulative frequency distribution. For the newspaper spending example, we find the following:
### Presenting data

<table>
<thead>
<tr>
<th>Class</th>
<th>Frequency</th>
<th>Relative frequency (%)</th>
<th>Cumulative relative frequency (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 to less than 10</td>
<td>3</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>10 to less than 20</td>
<td>6</td>
<td>30</td>
<td>45</td>
</tr>
<tr>
<td>20 to less than 30</td>
<td>3</td>
<td>15</td>
<td>60</td>
</tr>
<tr>
<td>30 to less than 40</td>
<td>3</td>
<td>15</td>
<td>75</td>
</tr>
<tr>
<td>40 to less than 50</td>
<td>2</td>
<td>10</td>
<td>85</td>
</tr>
<tr>
<td>50 to less than 60</td>
<td>1</td>
<td>5</td>
<td>90</td>
</tr>
<tr>
<td>60 to less than 70</td>
<td>1</td>
<td>5</td>
<td>95</td>
</tr>
<tr>
<td>70 to less than 80</td>
<td>0</td>
<td>0</td>
<td>95</td>
</tr>
<tr>
<td>80 to less than 90</td>
<td>1</td>
<td>5</td>
<td>100</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>20</strong></td>
<td></td>
<td><strong>100</strong></td>
</tr>
</tbody>
</table>

The ogive based on this data is as follows.

![Cumulative Percentage Polygon of Amounts Spent on Newspapers](image)

### Histograms

A histogram is a chart which uses vertical bars to display the frequency distribution. Unlike a bar chart, which is used for qualitative data, the histogram is used for quantitative data. This means that the bars on a histogram must touch to indicate the continuous nature of the data. Also, histograms always use vertical bars, unlike bar charts where the bars can be vertical or horizontal. A histogram is preferred to a stem-and-leaf diagram when data sets are larger.

The histogram for the newspaper spending example follows. Notice how the default graph produced by PHStat2 uses the messy bin values on the x-axis. It is possible to use Excel to produce a histogram using the midpoints (see below).
Example 2–2

The following data represent the actual weight of potato chips found in bags labelled 50 grams. The manufacturer aims to overfill the bags by 5 grams to allow for settling and dehydrating of the chips prior to sale. The results of fill weights in a sample of 20 consecutive 50-gram bags are listed below (reading from left to right in the order of being filled):

59.4  56.8  56.0  57.9  59.2  51.7  57.5  54.8  52.6  51.5
51.6  55.7  53.7  54.1  59.6  52.4  55.6  54.4  50.2  56.1

(a) Construct a stem-and-leaf display.

(b) Construct the frequency distribution and the percentage distribution.

(c) Plot the frequency histogram.

(d) Plot the percentage polygon.

(e) Form the cumulative percentage distribution.

(f) Plot the cumulative percentage polygon.
(g) On the basis of the results of (a) through (f), does there appear to be any concentration of the bag weights around specific values?

(h) If you had to make a prediction of the weight of potato chips in the next bag, what would you predict? Why?

**Solution 2–2**

We’ll demonstrate this solution by hand (which is an important skill for the exam) as well as using Excel and PHStat2.

(a) The first step in creating the stem-and-leaf display is to order the data from smallest to largest:

<table>
<thead>
<tr>
<th>50.2</th>
<th>51.5</th>
<th>51.6</th>
<th>51.7</th>
<th>52.4</th>
<th>52.6</th>
<th>53.7</th>
<th>54.1</th>
<th>54.4</th>
<th>54.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>55.6</td>
<td>55.7</td>
<td>56.0</td>
<td>56.1</td>
<td>56.8</td>
<td>57.5</td>
<td>57.9</td>
<td>59.2</td>
<td>59.4</td>
<td>59.6</td>
</tr>
</tbody>
</table>

There are a number of ways we could select stems and leaves for this data. Probably the easiest way would be to use the first two digits as the stem (so 50 through to 59) and use the third digit as the leaves (these would be 2, 5, 6, 7, etc. reading across the rows). Once every data point has been entered into the stem-and-leaf diagram, the following will be the result:

<table>
<thead>
<tr>
<th>Stem unit: 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>50 2</td>
</tr>
<tr>
<td>51 5 6 7</td>
</tr>
<tr>
<td>52 4 6</td>
</tr>
<tr>
<td>53 7</td>
</tr>
<tr>
<td>54 1 4 8</td>
</tr>
<tr>
<td>55 6 7</td>
</tr>
<tr>
<td>56 0 1 8</td>
</tr>
<tr>
<td>57 5 9</td>
</tr>
<tr>
<td>58</td>
</tr>
<tr>
<td>59 2 4 6</td>
</tr>
</tbody>
</table>

When PHStat2 is used to produce the stem-and-leaf diagram, we get the plot given below. To do this we begin by entering the data in a single column in Excel. Next sort the data and then select **PHStat | Descriptive Statistics | Stem-and-Leaf Display**. Enter the cell range for the data and a title. Note that the Summary Statistics box may already be checked; however, this information is not needed here so click on the box to remove this option from your output.
(b) The range of the data is $59.6 - 50.2 = 9.4$. Since this is quite a small data set, using about 6 to 8 classes is sensible; we will work here with 8. Then

$$\text{Width of interval} = \frac{9.4}{8} = 1.175$$

Since the data is given to one decimal place, we could round this number off to 1.2. So the first class will be 50.2 to less than 51.4 (\(= 50.2 + 1.2\)).

Using this method, the following table below can be produced.

<table>
<thead>
<tr>
<th>Weight of bag (grams)</th>
<th>Tally</th>
<th>Number of bags</th>
</tr>
</thead>
<tbody>
<tr>
<td>50.2 to less than 51.4</td>
<td>/</td>
<td>1</td>
</tr>
<tr>
<td>51.4 to less than 52.6</td>
<td>///</td>
<td>4</td>
</tr>
<tr>
<td>52.6 to less than 53.8</td>
<td>//</td>
<td>2</td>
</tr>
<tr>
<td>53.8 to less than 55.0</td>
<td>///</td>
<td>3</td>
</tr>
<tr>
<td>55.0 to less than 56.2</td>
<td>///</td>
<td>4</td>
</tr>
<tr>
<td>56.2 to less than 57.4</td>
<td>/</td>
<td>1</td>
</tr>
<tr>
<td>57.4 to less than 58.6</td>
<td>//</td>
<td>2</td>
</tr>
<tr>
<td>58.6 to less than 59.8</td>
<td>///</td>
<td>3</td>
</tr>
</tbody>
</table>

The percentage distribution can be calculated as demonstrated in the table below:

<table>
<thead>
<tr>
<th>Weight of bag (grams)</th>
<th>Number of bags</th>
<th>Percentage of bags</th>
</tr>
</thead>
<tbody>
<tr>
<td>50.2 to less than 51.4</td>
<td>1</td>
<td>(\frac{1}{20}) \times 100 = 5%</td>
</tr>
<tr>
<td>51.4 to less than 52.6</td>
<td>4</td>
<td>(\frac{4}{20}) \times 100 = 20%</td>
</tr>
<tr>
<td>52.6 to less than 53.8</td>
<td>2</td>
<td>(\frac{2}{20}) \times 100 = 10%</td>
</tr>
<tr>
<td>53.8 to less than 55.0</td>
<td>3</td>
<td>(\frac{3}{20}) \times 100 = 15%</td>
</tr>
<tr>
<td>55.0 to less than 56.2</td>
<td>4</td>
<td>(\frac{4}{20}) \times 100 = 20%</td>
</tr>
<tr>
<td>56.2 to less than 57.4</td>
<td>1</td>
<td>(\frac{1}{20}) \times 100 = 5%</td>
</tr>
<tr>
<td>57.4 to less than 58.6</td>
<td>2</td>
<td>(\frac{2}{20}) \times 100 = 10%</td>
</tr>
<tr>
<td>58.6 to less than 59.8</td>
<td>3</td>
<td>(\frac{3}{20}) \times 100 = 15%</td>
</tr>
</tbody>
</table>

(c) Frequency histogram

Make sure you are confident in producing the frequency histogram by hand. Remember that in the exam room, you will not be able to use Excel and PHStat2 (although you will need to use a calculator).
Type your bins (as calculated above) into column B of the spreadsheet, remembering to reduce the larger value of each bin slightly (as explained in Excel Strategies for Generating Frequency Distributions for Numerical Data on page 59 of the text). For the histogram below, I’ve chosen bin values of 50.2, 51.3999, 52.5999, etc. Type the midpoints of the bins into column C of the spreadsheet. I’ve used midpoints of 50.8, 52, 53.2, etc.

To produce the following frequency histogram, begin by selecting PHStat | Descriptive Statistics | Histograms and Polygons. Enter the data, bin and midpoint cell ranges and ensure Histogram is checked. In the histogram below the legend and cumulative percentage curve have been deleted.

Note that the default labels on the x-axis are the class boundaries, not the midpoints (which Figure 2.2 of the text uses). By using Excel’s chart of the frequencies and midpoints generated by the histogram facility in PHStat2 it is possible to produce the graph below.

(d) Percentage polygon
Presenting data

Make sure that you are confident to produce the percentage polygon by hand. The percentage polygon is produced in PHStat2 using a similar procedure as the histogram above. Ensure that Percentage Polygon option is checked.

(e) Cumulative percentage distribution

<table>
<thead>
<tr>
<th>Weight of bags (grams)</th>
<th>Percentage of bags (%)</th>
<th>Cumulative percentage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>50.2 to less than 51.4</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>51.4 to less than 52.6</td>
<td>20</td>
<td>25</td>
</tr>
<tr>
<td>52.6 to less than 53.8</td>
<td>10</td>
<td>35</td>
</tr>
<tr>
<td>53.8 to less than 55.0</td>
<td>15</td>
<td>50</td>
</tr>
<tr>
<td>55.0 to less than 56.2</td>
<td>20</td>
<td>70</td>
</tr>
<tr>
<td>56.2 to less than 57.4</td>
<td>5</td>
<td>75</td>
</tr>
<tr>
<td>57.4 to less than 58.6</td>
<td>10</td>
<td>85</td>
</tr>
<tr>
<td>58.6 to less than 59.8</td>
<td>15</td>
<td>100</td>
</tr>
</tbody>
</table>
(f) Cumulative percentage polygon

![Cumulative Percentage Polygon](image)

(g) The bag weights seem to be fairly evenly distributed over the interval from about 51 to about 60. There are no obvious outliers. Looking at the data in the order they occurred, there appears to be no obvious pattern.

(h) Because there seems to be no real pattern in the way the data are occurring, our best prediction would be somewhere around the middle of the data we’ve seen so far, so perhaps a prediction of about 55 is best. (Note: we’ll learn about how to make accurate forecasts later in this course.)

**Scatterplots and times series plots**

A *scatterplot* is a graph of two variables, which allows a visual check for any relationship between the variables. For example, you might graph quantity of goods sold on the x-axis and profit earned on the y-axis. This would allow you to study the relationship of profit to number of goods sold. Scatterplots produced by hand would normally use an ‘×’ symbol for each data point, but software (such as Excel) uses a variety of symbols. We will illustrate with a simple example. The following data is the quantity of mobile phones sold and the corresponding daily profit for a small phone sales shop for a random sample of 10 days. The scatterplot of this data follows.
A *time series plot* is used instead of a scatterplot where one of the variables involves regular time intervals. In this case, the researcher is interested in the behaviour of a particular variable over time. For example, ‘are profits increasing?’, ‘has the price of a share decreased over the last year?’ or ‘have daily customer numbers stabilised at a particular restaurant?’. With a time series plot, the time variable is always plotted on the $x$-axis and the variable of interest (profit, share price, customer numbers, etc.) on the $y$-axis. The following example illustrates a time series plot.

<table>
<thead>
<tr>
<th>Quantity of mobiles sold</th>
<th>Daily profit ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>500</td>
</tr>
<tr>
<td>19</td>
<td>690</td>
</tr>
<tr>
<td>8</td>
<td>380</td>
</tr>
<tr>
<td>11</td>
<td>430</td>
</tr>
<tr>
<td>24</td>
<td>900</td>
</tr>
<tr>
<td>16</td>
<td>750</td>
</tr>
<tr>
<td>22</td>
<td>1100</td>
</tr>
<tr>
<td>15</td>
<td>780</td>
</tr>
<tr>
<td>12</td>
<td>560</td>
</tr>
<tr>
<td>9</td>
<td>480</td>
</tr>
</tbody>
</table>
Example 2–3

In recent years, the cost of holiday accommodation on a particular island has been increasing. There was, however, a reduction as a reaction to reduced air travel in the aftermath of the attacks of September 11, 2001. Since then, rising fuel costs have increased the cost of commercial flights and so further discouraged travel to the island, but despite this, the cost of accommodation has continued to increase. The following data represents the cost of a double room for one night’s accommodation on the island for the years 1995 to 2006.

<table>
<thead>
<tr>
<th>Year</th>
<th>Cost of double room ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1995</td>
<td>100</td>
</tr>
<tr>
<td>1996</td>
<td>120</td>
</tr>
<tr>
<td>1997</td>
<td>130</td>
</tr>
<tr>
<td>1998</td>
<td>145</td>
</tr>
<tr>
<td>1999</td>
<td>170</td>
</tr>
<tr>
<td>2000</td>
<td>230</td>
</tr>
<tr>
<td>2001</td>
<td>200</td>
</tr>
<tr>
<td>2002</td>
<td>195</td>
</tr>
<tr>
<td>2003</td>
<td>180</td>
</tr>
<tr>
<td>2004</td>
<td>185</td>
</tr>
<tr>
<td>2005</td>
<td>190</td>
</tr>
<tr>
<td>2006</td>
<td>205</td>
</tr>
</tbody>
</table>

(a) Set up a time series plot (with cost of the double room on the y-axis and year on the x-axis).

(b) Based on the results of (a), is there any pattern to the cost of the double room between 1995 and 2006?

Solution 2–3

(a) The time series plot is given below. Note that this is a time series plot, and not a scatterplot, since the data is recorded at regular intervals (each year). This means that the data values on the graph must be connected with straight lines.
(b) There is a clear upward trend in the cost of the double room over the years 1995 to 2001. At this time (around the September 11 attacks), the cost of the room decreases steadily for three years until it begins to slowly increase again from 2004 onwards. It does not return to the heights that were experienced prior to September 11, presumably due to increasing petrol prices and reduced demand for accommodation.

**Important graphical considerations**

As we said in the introduction of this chapter, producing accurate graphs is something that many people struggle with. It is very easy to produce graphs that look good, but that violate some important principles. We’ll start by offering a list of suggestions on what to do to produce reliable and accurate graphs, then talk about some commonly made errors and finally consider ethical issues.

*Source for many of these guidelines: Levine et al. 2005, *Statistics for managers using Microsoft Excel.*

**Graphical excellence**

A good graph should *always* include the following:

- **A clear, meaningful title**—the title should describe not only the type of graph (e.g., histogram, pie chart, etc.) but also what the graph is showing (daily profit at a restaurant, amount spent on newspapers, etc.).

- **Appropriate axis labels or a legend**—the choice of axis labels or a legend will depend on the type of graph used (and the kind of data) but, a descriptor of the variable on each axis or a legend to distinguish between multiple variables (or sometimes both) is essential.

- **The graph should be able to stand on its own.** A good graph should include enough information so that it can be viewed and understood without the need for further explanation. Usually, in a statistical report, a graph will also be discussed, but it is important that this is not used as a means to explain the graph. Discussion of graphs is normally for the purpose of interpretation of the results presented in the graph, rather than explanation. A meaningful title and axis labels or a legend will obviously assist with producing a self-explanatory graph.

- **Show the data clearly.** It should display complex information in a simple and precise manner.

- **Focus the viewer on the substance of the graph.**

- **Encourage comparisons of data.**

- **Serve a clear purpose.**

- **Tell the truth about the data.** Aside from the (obviously essential) need to present the correct data, it is also important to avoid presenting the data in such a way that the viewer is mislead as to the truth.

- **The graph should not distort the data.** Many fancy graphical displays (such as many of those available in Excel) tend to distort data. For example, using three-dimensional pie charts is likely to overemphasise the important of categories at the front of the graph and underemphasise those at the back. The textbook also contains examples of graphs that distort results.
• **Avoid chartjunk.** Chartjunk is unnecessary information or adornments that clutter the graph. A good rule of thumb is to keep the graph simple.

• **Maximise the data-ink ratio.** The data-ink ratio is the proportion of the graph’s ink that is used for display of useful information.

• **Avoid too much white space.** The actual graph (a pie chart, a bar chart, etc.) should take up the majority of the space allocated for a graph. Too much empty space (white space) around a graph is distracting, unnecessary and lessens the impact of the actual graph.

• **Include zero on the y-axis.** Truncating the graph so that the vertical axis does not begin at zero can distort results, overemphasising differences between data values.

• **Avoid using pictures instead of rectangles in bar charts etc.** Often if a picture is used instead of a bar chart, the reader will interpret the area (or volume) of the picture rather than the height. This can also distort results or overemphasis bigger frequencies. See the text for some examples.

**Common graphical errors and ethical issues**

The text contains a number of examples of poorly presented graphs. Graphs can mislead viewers by being emotionally charged. At times there is a clear intention to mislead the viewer. Violating the guidelines above are some of the most common ways of making errors. A desire to make a graph look attractive must always be balanced with the need to clearly and accurately display the data. Too often, graphics used in the media, focus solely on pretty presentations at the expense of accuracy.

An ethical issue arises when someone intentionally uses a graph to mislead the viewer. In many cases, graphs are poorly produced more from ignorance of the statistical considerations, rather than as an intentional tool to mislead. Packages such as Excel have not assisted in improving graphical excellence, where many of the graphical types available which contain unnecessary adornments or which distort results. In many cases, default graphs require a fairly large expenditure of time to adapt them to an appropriate format and avoid graphical errors. Unfortunately Excel’s graphical features do not appear to have been designed by statisticians. There are, however, many other statistical packages available, which significantly outperform Excel in producing good quality graphics.

**Discussion points**

**Discussion point 2–1**

Examine and complete Problems 2.49 and 2.51 on p. 65 of the 5th edition textbook (Problems 2.47 and 2.49 on pp. 83–84 of the 4th edition). Discuss your answers. Are there any ethical issues associated with these graphical displays? Give an example of how such displays could potentially be misused. What have you learned about graphical excellence? How will this be helpful to you in the workplace?
Summary

Now that you have completed this module, turn back to the objectives at the beginning of the module. Have you achieved these objectives?

Ensure that you attempt the recommended problems in the list of review questions below and at least a sample of problems from the optional list. This will help you to identify any areas of difficulty you have in achieving the module’s objectives.

Review questions

Recommended problems

Levine et al. 4th edition: Questions 2.2, 2.4, 2.8, 2.10, 2.12, 2.18, 2.20, 2.24, 2.28, 2.34, 2.38, 2.40, 2.48, 2.50 to 2.62, 2.70 and 2.72.

Levine et al. 5th edition: Questions 2.2, 2.4, 2.10, 2.12, 2.14, 2.16, 2.20, 2.22, 2.26, 2.28, 2.30, 2.34, 2.36, 2.38, 2.44, 2.50, 2.52, 2.55 to 2.61, 2.80 and 2.82.

Optional problems

Levine et al. 4th edition: Choose a selection of problems from Questions 2.1, 2.3, 2.5 to 2.7, 2.9, 2.11, 2.13 to 2.17, 2.19, 2.21 to 2.23, 2.25 to 2.27, 2.29 to 2.33, 2.35 to 2.37, 2.39, 2.41 to 2.43, 2.47, 2.49, 2.63 to 2.69, 2.71, 2.73 to 2.80, 2.83 and 2.84.

Levine et al. 5th edition: Choose a selection of problems from Questions 2.1, 2.3, 2.5 to 2.9, 2.11, 2.13, 2.15, 2.17 to 2.19, 2.21, 2.23 to 2.25, 2.27, 2.29, 2.31 to 2.33, 2.35, 2.37, 2.39 to 2.43, 2.45, 2.49, 2.51, 2.53, 2.54, 2.62 to 2.79, 2.81 and 2.83 to 2.85.
Numerical descriptive measures
Contents

Introduction ............................................................................................................... 3–3
Objectives .................................................................................................................. 3–3
Readings .................................................................................................................... 3–3
Measures of central tendency .................................................................................... 3–4
Measures of spread .................................................................................................... 3–7
Distribution shape .................................................................................................... 3–12
Empirical rule and Chebyshev’s rule ...................................................................... 3–16
Correlation and covariance ...................................................................................... 3–18
Ethical issues ........................................................................................................... 3–21
Discussion points ..................................................................................................... 3–22
Summary ................................................................................................................. 3–22
Review questions ..................................................................................................... 3–22
Introduction

Although pictures are very important, there are still times when they don’t tell us all we need to know about a data set. We should almost always use both graphical and numerical summaries to explore a data set before we even think about more complex analyses, but often these tell us all we need to know about the data and there is no need to look deeper.

This week we explore various measures of the central tendency (describing where the ‘middle’ of the data set is), spread of data (are the data points similar to each other or spread out?) and measures of correlation and covariance between variables (for example, do high values of one variable correspond to high values in another variable?).

Remember that you’ll need to know how to do these calculations by hand as well as using Excel and PHStat2.

Objectives

On completion of this module you should be able to:

- calculate and interpret measures of central tendency (mean, median & mode) for samples and populations
- calculate and interpret quartiles, range, interquartile range, variance and standard deviation
- calculate and interpret the coefficient of variation
- understand and utilise the empirical rule and Chebyshev’s Rule
- construct a box-and-whisker plot
- calculate covariance and correlation and
- discuss pitfalls and ethical issues relating to descriptive measures.

Readings

Source

Textbook
Levine et al. 4th edition
Ch. 3 (including Section 3.7 on the CD which accompanies the text) & Excel Handbook Ch. 3 (pp. 153–155)

Or

Levine et al. 5th edition
Ch. 3 (excluding the section on Z scores) & Excel Companion Ch. 3 (pp. 143–145)
Numerical descriptive measures

Course website  Visit the course website for links to any supplementary material for this week.

Optional reading  Levine et al. 4th and 5th editions
Appendices A, B and C

Some students may find portions of Appendices A, B and C useful or necessary reading.

Measures of central tendency

Often when we look at a data set, we want to know what a typical result is or get some idea of the middle of the data. We will look at three different ways of measuring the typical value: the mean, median and mode.

The mean

When people talk about the average, they are referring to the mean (sometimes referred to as the arithmetic mean). All data values contribute to this measure, but this also means that the mean is greatly affected by any large data values. When a sample data set is used, the sample mean is found via:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

where the term $\bar{X}$ (read as ‘X bar’) is used to indicate the sample mean, $\sum_{i=1}^{n} X_i$ is the sum of all the data values (the $X_i$ values) and $n$ is the sample size.

If the data is available for the entire population, the population mean is found via:

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

where the term $\mu$ (Greek letter ‘mu’) is used to indicate the population mean, $\sum_{i=1}^{N} X_i$ is the sum of all the data values (the $X_i$ values) and $N$ is the population size.

Note: population parameters are always denoted by Greek letters (for example $\mu$, $\sigma$, $\pi$) and sample values by Roman letters ($X$, $S$ and $p$).

The median

The median is the middle value of the data set where the data has been sorted from smallest to largest. This means that half of the data are less than or equal to the median and half of the data are greater than or equal to the median. Unlike the mean, the median is not affected by extreme data values. The median is found via:

$$\text{Median} = \frac{n + 1}{2} \times \text{ranked value}$$
The mode

The mode is the data value that occurs most often. It is possible that a data set has no mode or that there are several modes.

Example 3-1

A manufacturer of mobile phones has been concerned that the latest model of the battery is not lasting as long as anticipated. They take a random sample of 20 phones and batteries, and record how long they take to go flat (this is done by turning the phones on and leaving them switched on until the battery goes flat). The following data (battery life in hours) are the result:

```
42  42  48  45  51  45  48  44  43  42
46  46  47  48  40  48  42  48  51  50
```

Calculate the mean, median and mode of the battery lives. Which measures of the battery life do you think are best and worst for this data? Why? How will this information be of use to the manufacturer?

Solution 3-1

Note that we will demonstrate these calculations ‘by hand’ and leave it as an exercise for the reader to duplicate the results using PHStat2.

Mean:

\[
\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{42+42+48+45+\ldots+51+50}{20} = \frac{916}{20} = 45.8
\]

Median:

\[
\frac{n+1}{2} = \frac{20+1}{2} = 10.5\th\ observation.\ Ordering\ the\ data\ from\ smallest\ to\ largest\ gives\ us\ (reading\ from\ left\ to\ right)\ the\ following.
\]

```
40  42  42  42  42  43  44  45  45  46
46  47  48  48  48  48  48  50  51  51
```

The 10.5th observation will be between 46 and 46, so the median is 46.

Mode: 48 (48 hours of battery life occurred five times and so is the ‘most typical’).

The mean and median are very similar and probably a more accurate measure of the ‘middle’ of the data. The mode is the most common data value, but in such a small data set, this does not necessarily reflect the ‘middle’ of the data.

The manufacturer will need to have an idea about what is ‘average’ or ‘normal’ in terms of battery life. This is important to maintain quality and for benchmarking purposes.

The geometric mean

The geometric mean measures the rate of change of a variable over time. It is found via:

\[
\bar{X}_g = \left( X_1 \times X_2 \times \ldots \times X_n \right)^{1/n}
\]

Note: whenever the word ‘mean’ is used alone it refers to the arithmetic mean. The geometric mean is always distinguished from the arithmetic mean by including the word ‘geometric’.
The geometric mean rate of return gives the average percentage return of an investment over time. It is found via:

$$\bar{R}_g = \left[ \left(1 + R_1 \right) \times \left(1 + R_2 \right) \times \ldots \times \left(1 + R_n \right) \right]^{1/n} - 1$$

**Example 3–2**

The total rate of return (%) of three blue-chip stocks is given in the table below for the years 2003, 2004 and 2005.

<table>
<thead>
<tr>
<th>Year</th>
<th>Stock A</th>
<th>Stock B</th>
<th>Stock C</th>
</tr>
</thead>
<tbody>
<tr>
<td>2003</td>
<td>3.64</td>
<td>1.12</td>
<td>-0.25</td>
</tr>
<tr>
<td>2004</td>
<td>2.32</td>
<td>1.70</td>
<td>1.03</td>
</tr>
<tr>
<td>2005</td>
<td>0.09</td>
<td>-3.50</td>
<td>2.08</td>
</tr>
</tbody>
</table>

(a) Calculate the geometric mean rate of return for each stock.

(b) Compare these results.

**Solution 3–2**

(a) \( \bar{R}_g = \left[ \left(1 + R_1 \right) \times \left(1 + R_2 \right) \times \ldots \times \left(1 + R_n \right) \right]^{1/n} - 1 \)

Note that in these calculations we have converted the rates of return from percentages to decimals (for example, 3.64% became 0.0364).

Stock A: \( \bar{R}_g = \left[ \left(1 + 0.0364 \right) \times \left(1 + 0.0232 \right) \times \left(1 + 0.0009 \right) \right]^{1/3} - 1 = 0.0201 = 2.01\% \) (rounded to two decimal places)

Stock B: \( \bar{R}_g = \left[ \left(1 + 0.0112 \right) \times \left(1 + 0.0170 \right) \times \left(1 - 0.0350 \right) \right]^{1/3} - 1 = -0.0025 = -0.25\% \)

Stock C: \( \bar{R}_g = \left[ \left(1 - 0.0025 \right) \times \left(1 + 0.0103 \right) \times \left(1 + 0.0208 \right) \right]^{1/3} - 1 = 0.0095 = 0.95\% \)

Stock B has the worst rate of return, whilst Stock A has the best rate of return. Stock B’s poor result is mainly due to its negative rate of return in 2005. Although stock A also dropped considerably in 2005, its rate of return has stayed positive. Stock C has behaved very differently from the other two, with its rate of return actually increasing over the three years. This may make it a better choice than Stock A. Stock C may continue this trend and increase again in 2006, whereas Stock A seems to show a downward trend in returns.

**Approximating the mean for a frequency distribution**

If instead of raw data, you have a frequency distribution, it is not always possible to calculate the arithmetic mean. It is, however, possible to approximate the mean using the following formula:

$$\bar{X} = \frac{\sum_{j=1}^{c} m_j f_j}{n}$$

where \( c \) is the number of classes in the frequency distribution, \( m_j \) is the midpoint of the \( j \)th class and \( f_j \) is the frequency of the \( j \)th class. Example 3–4 (see later) demonstrates the calculation of the approximate mean.
Quartiles
Quartiles divide the (ordered) data into four subgroups. The three quartiles are often referred to as $Q_1$, $Q_2$, and $Q_3$. The second quartile, $Q_2$, is the median of the data. It separates the smallest half of the data from the largest half. The first quartile, $Q_1$, is often called the lower quartile (abbreviated to LQ). It separates the smallest quarter of the data from the largest three quarters. The third quartile, $Q_3$, is often called the upper quartile (UQ). It separates the smallest three quarters of the data from the largest quarter. When data is ordered from smallest to largest, the lower and upper quartiles are found via the following:

Lower quartile $= \frac{n+1}{4}$ ranked value

Upper quartile $= \frac{3(n+1)}{4}$ ranked value

Depending on the sample size ($n$), the ranked value found using these formulae may not be a nice whole number. For example, if $n = 14$, then we would be looking for the $\frac{14+1}{4} = 3.75th$ and $\frac{3(14+1)}{4} = 11.25th$ values. There are three methods commonly used for dealing with this problem:

1. Round the ranked value to the nearest whole number. For $n = 14$, we would then look for the 4th and 11th values for the lower and upper quartiles respectively.
2. Find the value half way between two data points. For $n = 14$, we would take the lower quartile as the value 0.5 of the way from the 3rd ranked value to the 4th ranked value and the upper quartile as the value 0.5 of the way from the 11th to 12th ranked values.
3. Find the value the appropriate fraction of the way between two data points. For $n = 14$, we would take the lower quartile as the value 0.75 of the way from the 3rd ranked value to the 4th ranked value and the upper quartile as the value 0.25 of the way from the 11th to 12th ranked values.

Of these three methods, the third is preferred. We will demonstrate how to find these quartiles in Example 3–3.

Quartiles are also referred to as percentiles. The lower quartile is the 25th percentile (since 25% of the data values are less than or equal to this value), the median is the 50th percentile and the upper quartile is the 75th percentile. Using this terminology it also makes sense to discuss other percentile values (e.g., the 90th percentile, the 5th percentile, etc.).

Measures of spread

Range
The range is the difference between the largest and the smallest data values:

$$\text{Range} = X_{\text{largest}} - X_{\text{smallest}}$$
**Interquartile range**

The interquartile range is the difference between the upper and lower quartiles:

\[ IQR = UQ - LQ = Q_3 - Q_1 \]

**Variance and standard deviation**

The variance and the standard deviation are measures of the variability in the data. Unlike the range and interquartile range, they take account of all the data values. Variance and standard deviation measure ‘average’ dispersion of data around the mean. The variance finds the average of squared differences of the data values from the mean. Because of this, its value is always nonnegative. If the value was zero, this would indicate that there is no variation in the data (all the data values would be identical).

The sample variance is given by

\[ S^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1} \]

and the population variance by

\[ \sigma^2 = \frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N} \]

The standard deviation is the square root of the variance and so is given by

\[ S = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}} \]

and

\[ \sigma = \sqrt{\frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}} \]

for sample and population respectively.

**A better computational formula for the variance and standard deviation**

Although the text uses the formulae above, these are computationally more intensive and for that reason are often avoided. For hand calculations the formulae below are an easier way to find the variance and standard deviation. Statistics mode on a calculator will quickly offer the \( \sum X \) and \( \sum X^2 \) values, and these can then be substituted into the formula to give the variance.

The **sample variance** is found via:

\[ S^2 = \frac{\sum X^2 - (\sum X)^2}{n-1} \]
and the **sample standard deviation** is found via:

\[ S = \sqrt{\frac{\sum X^2 - \left(\frac{\sum X}{n}\right)^2}{n-1}} \]

The **population variance** is found via:

\[ \sigma^2 = \frac{\sum X^2 - \left(\frac{\sum X}{N}\right)^2}{N} \]

and the **population standard deviation** is found via:

\[ \sigma = \sqrt{\frac{\sum X^2 - \left(\frac{\sum X}{N}\right)^2}{N}} \]

The variance is expressed in terms of squared units of measurement. For example, with the data in Example 3–1, the variance is expressed in terms of squared hours of battery life. This makes it very difficult to interpret. The standard deviation, on the other hand, is expressed in the same units as the original data. In Example 3–1, this means it is expressed in terms of battery hours. When we discuss the empirical rule and Chebyshev’s rule (see below) we will talk more about interpreting standard deviations.

**Coefficient of variation**

The standard deviation’s size is dependent on the kind of variable being measured. This makes comparisons difficult. We couldn’t say, for example, that for the data measuring battery life (in Example 3–1), a standard deviation of 8 hours, indicates greater variability than a standard deviation of 4 kilograms for body weights in a sample of 20 people. One way of comparing variability between data sets which use different measurement units, or that have different magnitudes, is via the coefficient of variation.

The **coefficient of variation** measures the ratio of the standard deviation to the mean, expressing this as a percentage. Since this is a relative comparison of the standard deviation to its mean, it therefore takes account of the type and relative size of values in the data being compared. It is always expressed as a percentage, rather than in the units of the data.

The **sample coefficient of variation** is found via:

\[ CV = \left(\frac{S}{\bar{X}}\right) \times 100\% \]

and the **population coefficient of variation** is found via:

\[ CV = \left(\frac{\sigma}{\mu}\right) \times 100\% \]
Using the coefficient of variation, we could say that a value of $CV = 50$ indicates much more variability than a value of $CV = 25$.

**Example 3–3**

Returning to the data set from Example 3–1, compute the variance, standard deviation, coefficient of variation, lower and upper quartiles, interquartile range and range.

**Solution 3–3**

We find that $\sum X = 42 + 42 + 48 + \ldots + 50 = 916$ and $\sum X^2 = 42^2 + 42^2 + 48^2 + \ldots + 50^2 = 42154$ and so the variance is:

$$S^2 = \frac{\sum X^2 - \left( \frac{\sum X}{n} \right)^2}{n} = \frac{42154 - (916)^2}{19} = 10.5895 \text{ (to 4 decimal places).}$$

The standard deviation is:

$$S = \sqrt{S^2} = \sqrt{\frac{42154 - (916)^2}{19}} = 3.2541 \text{ (to four decimal places).}$$

Note: in calculating the standard deviation we have avoided a rounding error by taking the square root of the variance before it was rounded (the fraction) rather than after it was rounded (the 10.5895).

The coefficient of variation is: $CV = \left( \frac{S}{X} \right) \times 100\% = \left( \frac{3.2541}{45.8} \right) \times 100\% = 7.11\% \text{ (to 2 decimal places).}$

The lower quartile (or first quartile) is $Q_1 = \frac{n+1}{4}$ th value = $\frac{20+1}{4} = 5.25$ th value. Since there is no 5.25th data value, we go 25% of the way from the 5th value (42) to the 6th value (43). This gives a lower quartile of 42.25.

The upper quartile is $Q_3 = \frac{3(n+1)}{4}$ th value = $\frac{3(20+1)}{4} = 15.75$ th value. Since the 15th value is 48 and the 16th value is also 48, 75% of the way from the 15th to the 16th value is 48 and so the upper quartile is 48.

The interquartile range is: $Q_3 - Q_1 = 48 - 42.25 = 5.75$.

The range is: $X_{\text{largest}} - X_{\text{smallest}} = 51 - 40 = 11$.

**Variance and standard deviation for a frequency distribution**

The approximate standard deviation of a frequency distribution is

$$S = \sqrt{\frac{\sum_{j=1}^{c} (m_j - \bar{X})^2 f_j}{n-1}}$$

**Alternative computational formula for variance of frequency data**

As we saw for the earlier variance formulae, it is possible to use an easier computational formula for the sample variance of frequency distributions:
Numerical descriptive measures

\[ S^2 = \frac{\sum m^2f - \left(\frac{\sum mf}{n}\right)^2}{n-1} \]

As before, the sample standard deviation is the square root of the sample variance:

\[ S = \sqrt{S^2} = \sqrt{\frac{\sum m^2f - \left(\frac{\sum mf}{n}\right)^2}{n-1}} \]

**Example 3–4**

Participants at a recent accounting for small businesses workshop were asked to complete an anonymous survey. The table below contains data taken from this survey: a frequency distribution of the number of staff employed by each of the 50 small businesses in attendance at the workshop. Note that fractional (part-time) staff were recorded in this survey, so for example 2.75 staff could mean two full time and one staff member employed for \(\frac{3}{4}\) of the hours in a working week.

Approximate the arithmetic mean and standard deviation of the number of attendees.

<table>
<thead>
<tr>
<th>Class</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 to less than</td>
<td>16</td>
</tr>
<tr>
<td>5 to less than</td>
<td>19</td>
</tr>
<tr>
<td>10 to less than</td>
<td>5</td>
</tr>
<tr>
<td>15 to less than</td>
<td>7</td>
</tr>
<tr>
<td>20 to less than</td>
<td>2</td>
</tr>
<tr>
<td>25 to less than</td>
<td>1</td>
</tr>
</tbody>
</table>

**Solution 3–4**

We begin by using a table to discover the necessary information to calculate the mean:

<table>
<thead>
<tr>
<th>Class</th>
<th>Frequency</th>
<th>Midpoint</th>
<th>(m_i f_i)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 to less than</td>
<td>16</td>
<td>2.5</td>
<td>16 × 2.5 = 40</td>
</tr>
<tr>
<td>5 to less than</td>
<td>19</td>
<td>7.5</td>
<td>19 × 7.5 = 142.5</td>
</tr>
<tr>
<td>10 to less than</td>
<td>5</td>
<td>12.5</td>
<td>62.5</td>
</tr>
<tr>
<td>15 to less than</td>
<td>7</td>
<td>17.5</td>
<td>122.5</td>
</tr>
<tr>
<td>20 to less than</td>
<td>2</td>
<td>22.5</td>
<td>45</td>
</tr>
<tr>
<td>25 to less than</td>
<td>1</td>
<td>27.5</td>
<td>27.5</td>
</tr>
<tr>
<td>Total</td>
<td>50</td>
<td></td>
<td>440</td>
</tr>
</tbody>
</table>

\[ \bar{X} = \frac{\sum m_i f_i}{n} = \frac{440}{50} = 8.8 \]

Next we need to add some columns to our table (note that normally you would not rewrite the whole table—we’re doing this here only to demonstrate how the solution evolves):
Numerical descriptive measures

<table>
<thead>
<tr>
<th>Class</th>
<th>( f )</th>
<th>( m )</th>
<th>( mf )</th>
<th>( m^2f )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 to less than 5</td>
<td>16</td>
<td>2.5</td>
<td>40</td>
<td>( (2.5)^2 \times 16 = 100 )</td>
</tr>
<tr>
<td>5 to less than 10</td>
<td>19</td>
<td>7.5</td>
<td>142.5</td>
<td>( (7.5)^2 \times 19 = 1068.75 )</td>
</tr>
<tr>
<td>10 to less than 15</td>
<td>5</td>
<td>12.5</td>
<td>62.5</td>
<td>781.25</td>
</tr>
<tr>
<td>15 to less than 20</td>
<td>7</td>
<td>17.5</td>
<td>122.5</td>
<td>2143.75</td>
</tr>
<tr>
<td>20 to less than 25</td>
<td>2</td>
<td>22.5</td>
<td>45</td>
<td>1012.5</td>
</tr>
<tr>
<td>25 to less than 30</td>
<td>1</td>
<td>27.5</td>
<td>27.5</td>
<td>756.25</td>
</tr>
<tr>
<td>Total</td>
<td>50</td>
<td>440</td>
<td>5862.5</td>
<td>50</td>
</tr>
</tbody>
</table>

So we have \( \sum mf = 440 \), \( \sum m^2f = 5862.5 \) and \( S^2 = \frac{5862.5 - \frac{(440)^2}{50}}{49} \), and so \( S = \sqrt{\frac{5862.5 - \frac{(440)^2}{50}}{49}} = 6.3736 \) (to 4 decimal places).

Distribution shape

Box-and-whisker plots

A five-number summary is a list of the following values:

- minimum
- lower quartile
- median
- upper quartile
- maximum

These values give useful information about the way the data is distributed. A box-and-whisker plot is a visual presentation of the five-number summary (see the example that follows). A box is used to enclose the median. The box extends down to the lower quartile and up to the upper quartile. Therefore, we know that 50% of the data falls within the central box. Tails then extend from the edges of this box out to the minimum and maximum values.

For example, the five-number summary below gives customer numbers per day at a small financial advice business. The box-and-whisker plot follows. The central box extends from 20 to 40 (the lower and upper quartiles) with a vertical line at 30 (the median). Tails (or whiskers) extend from the box to the minimum and maximum values.
Box-and-whisker plot of customer numbers per day

Five-number Summary

Minimum  10
First Quartile  20
Median  30
Third Quartile  40
Maximum  50

Excel/PHStat2 will also produce box-and-whisker plots of several data sets on the same axes. This allows easy comparisons of the shape of the distributions.

Shape of a distribution

The shape of a distribution tells us something about the characteristics of the data. **Skewness** occurs when a distribution lacks symmetry. We describe the distribution as either positively skewed (skewed to the right) or negatively skewed (skewed to the left) or as symmetrical.

The following diagrams illustrate the kind of distribution that might be expected in each of these three cases. When a distribution is **negatively skewed**, the mean is less than the median. The skewed portion is indicated by the long, thin part of the curve to the left. When a distribution is **positively skewed**, the mean is greater than the median. When the distribution is **symmetrical**, the mean will equal (or be very close to) the median.
Numerical descriptive measures

The box-and-whisker plot can also indicate where skewness exists. The example given above was symmetrical and so revealed no skewness. The upper and lower quartiles were equal distances away from the median, and the minimum and maximum values were also equal distances away from the median. A box-and-whisker plot with negative skew would have a longer whisker towards the minimum than towards the maximum (indicating the outer values are skewed), and the median would be closer to the right side of the box (indicating the middle 50% of values are skewed). The reverse would be true for a positively skewed distribution.

**Example 3–5**

Returning to the data set in Example 3–1 again,

(a) Construct a box-and-whisker plot.

(b) If the manufacturer intended to issue a statement saying that their batteries will last more than 50 hours, what would you advise them? Why?

(c) Suppose the first value was 142 instead of 42. Repeat (a) above and comment on the differences.

(d) Repeat the analysis of Examples 3–1, 3–2 and Example 3–3 (parts (a) and (b)) with the new data value. Comment on the differences.

(e) How would you describe the shape of the original data set? The revised data set?

**Solution 3–5**

(a) The box-and-whisker plot (produced using PHStat2) is:
Numerical descriptive measures

(b) Considering that the mean, median and mode are all less than 50 hours, it would not be wise to make such a claim about the battery lives. We can see that only three of the 20 observations (or 15%) are over 50 hours and so again this indicates there is little likelihood that the manufacturer’s batteries will be able to match the claim.

(c) Mean: $\bar{X} = \frac{\sum X_i}{n} = \frac{142 + 42 + 48 + 45 + \ldots + 51 + 50}{20} = \frac{1016}{20} = 50.8$

Median: 46.5 (since the order of the data will change as below).

40 42 42 42 43 44 45 45 46 46
47 48 48 48 48 48 50 51 51 142

Mode: 48 (as before).

The mode is clearly unchanged from the earlier calculations. The median has changed only slightly. The mean, however, has been most obviously impacted by this one value. Since the new data value (142 hours) is so different from the others, it is described as an outlier. Outliers can have a huge impact on the value of the mean and so when data sets have outliers, often the median is a more reliable measure of the ‘middle’ of the distribution of data.

(d) Given that now $\sum X = 142 + 42 + 48 + \ldots + 50 = 1016$ and $\sum X^2 = 142^2 + 42^2 + 48^2 + \ldots + 50^2 = 60554$, the variance is:

$$S^2 = \frac{\sum X^2 - (\frac{\sum X}{n})^2}{n-1} = \frac{60554 - (\frac{1016}{20})^2}{19} = 470.5895 \text{ (to 4 decimal places)}$$

The standard deviation is: $S = \sqrt{S^2} = 21.6931 \text{ (to four decimal places)}$.

The coefficient of variation is:

$$CV = \left( \frac{S}{\bar{X}} \right) 100\% = \left( \frac{21.693074}{50.8} \right) 100\% = 42.70\% \text{ (to 2 decimal places)}.$$

The interquartile range is: $Q_3 - Q_1 = 48 - 43 = 5$.

The range is: $X_{\text{largest}} - X_{\text{smallest}} = 142 - 40 = 102$.

All measures except the interquartile range have increased dramatically with the change of one data value. If we did not know that this was the result of a change in only one data value, and we compared these measures of variability with those found in part (d), we might think that the data was much more spread out. We know, however, that here this is due to one outlying value.

(e) The original data are fairly symmetrical since the mean and the median are very similar. The box-and-whisker plot revealed that the median was not quite centred in the box extending from the lower to the upper quartile, so there is a very small amount of skew evident in the central 50% of the data values. But, the whisker extending to the maximum value is longer than that extending to the minimum value, so skewness is in the opposite direction in the tails of the distribution (as compared to the central 50% of the distribution).
The data with the first value adjusted, however, reveal a slight right skew since the mean is greater than the median (this cannot accurately be described as a skew, however, since the new value is clearly an outlier).

**Empirical rule and Chebyshev’s rule**

**The empirical rule**

The empirical rule is a rule of thumb that states (when a data set is normally distributed) approximately the percentage of data values which lie within a certain number of standard deviations of the mean. The normal distribution (which will be discussed in more detail in Week 6) is often referred to as the bell-shaped curve. The empirical rule says that for normally distributed data:

- 68% of values will be within one standard deviation either side of the mean: $\mu \pm 1\sigma$
- 95% of values will be within two standard deviations either side of the mean: $\mu \pm 2\sigma$
- 99.7% of values will be within three standard deviations either side of the mean: $\mu \pm 3\sigma$

**Chebyshev’s Rule**

Chebyshev’s Rule (or the Bienaymé-Chebyshev Rule) applies to any data distribution, regardless of shape. It can therefore be used whether the distribution of the data is known or unknown. When appropriate (i.e., when data is known to be normally distributed), however, the empirical rule is preferred since it gives more accurate information for normally distributed data.

Chebyshev’s Rule states:

A proportion of at least $1 - \frac{1}{k^2}$ data values are within $k$ standard deviations of the mean $(\mu \pm k\sigma)$.

**Example 3–6**

Returning to the original data set in Example 3–1 above, answer the following questions.

(a) According to Chebyshev’s Rule, what percentage of these battery lives are expected to be within ±1 standard deviation of the mean? Within ±2 standard deviations of the mean? Within ±3 standard deviations of the mean?

(b) Assume that the manufacturer knows that the mean life of the population of batteries is 48.2 hours and the standard deviation of the population of batteries is 3.1 hours. What percentage of data values are actually within ±1 standard deviation of the mean? Within ±2 standard deviations of the mean? Within ±3 standard deviations of the mean?

(c) Discuss the differences in your answers to (a) and (b).
Solution 3–6

(a) Using \( \left(1 - \frac{1}{k^2}\right) \times 100\% \), where \( k \) is the number of standard deviation. So within one standard deviation indicates \( k = 1 \) and so

\[
\left(1 - \frac{1}{1^2}\right) \times 100\% = 0\% \text{ (not very helpful!!!)}
\]

so at least 0% of observations are expected to be within 1 standard deviation of the mean. Using \( k = 2 \) gives

\[
\left(1 - \frac{1}{2^2}\right) \times 100\% = 75\%
\]

so at least 75% of observations are expected to be within 2 standard deviations of the mean. Finally, using \( k = 3 \) gives

\[
\left(1 - \frac{1}{3^2}\right) \times 100\% = 88.89\%
\]

so at least 88.89% of observations are expected to be within 3 standard deviations of the mean.

(b) We are given \( \mu = 48.2 \) and \( \sigma = 3.1 \). We need to determine the intervals \( \mu \pm k\sigma \).

For \( k = 1 \),

\[
\mu \pm 1\sigma = 48.2 \pm 3.1 = 45.1 \text{ to } 51.3
\]

Examining the data we see that 11 observations are between 45.1 and 51.3 (46 to 51) or \( 11/20 \times 100\% = 55\% \). So 55% of our data is within one standard deviation of the mean.

For \( k = 2 \),

\[
\mu \pm 2\sigma = 48.2 \pm 2(3.1) = 42 \text{ to } 54.4
\]

All but one of the observations is between 42 and 54.4 (42 to 54) or \( 19/20 \times 100\% = 95\% \). So 95% of the observations are within two standard deviations of the mean.

For \( k = 3 \),

\[
\mu \pm 3\sigma = 48.2 \pm 3(3.1) = 38.9 \text{ to } 57.5
\]

All (100%) of the data is within 3 standard deviations of the mean.

(c) Chebyshev’s Rule applies to any distribution and so in some sense it indicates a worst case. It says that at the very least, 0% of data will be within one standard deviation of the mean, at least 75% within two and at least 88.89% will be within three. Our data are less spread out than the worst case of the rule, but it is consistent with this rule since it is covered by the ‘at least’ in the rule.
Correlation and covariance

Correlation and covariance are two measures of the strength of a relationship between two numerical variables. The first step in determining whether any relationship exists, however, is always to produce a scatterplot which will give a visual indication of any relationship.

Covariance

Covariance measures the strength of the linear relationship between two numerical variables. The sample covariance is defined by:

$$\text{cov}(X,Y) = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$$

The size of the covariance is related to the variable being measured rather than the relative strength of relationship. A covariance of 10, for example, tells us only that the relationship between the two variables is positive. It does not tell us how strong the relationship is, nor does necessarily it enable us to say that this is reveals a stronger relationship than a covariance of 5. The correlation coefficient overcomes this problem.

Correlation

The correlation coefficient measures the relative strength of a linear relationship. A value of -1 indicates perfect negative correlation, a value of 0 indicates no relationship exists and a value of +1 indicates a perfect positive relationship. The symbol $\rho$ (Greek letter ‘rho’) is used for population correlation and $r$ is used to represent sample correlation.
Numerical descriptive measures

Perfect negative correlation

\[ \rho = -1 \]

The sample correlation is found via

\[ r = \frac{\text{cov}(X,Y)}{S_X S_Y} \]

where \( \text{cov}(X,Y) \) is the covariance of the two variables,

\[ S_X = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}} \quad \text{and} \quad S_Y = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}{n-1}}. \]

Given that calculating these values is fairly involved and messy, the correlation coefficient is usually calculated using the following formula:

\[ r = \frac{n \sum X Y - (\sum X)(\sum Y)}{\sqrt{(n \sum X^2 - (\sum X)^2)(n \sum Y^2 - (\sum Y)^2)}} \]

When working by hand, a calculator will find the values of \( \sum X \), \( \sum X^2 \), \( \sum Y \), \( \sum Y^2 \) and \( \sum XY \) values, and these can then be substituted into this formula.

Care needs to be taken when interpreting correlation. Just because two variables have strong correlation does not imply that one causes the other. Correlation could be caused by chance, by a third (perhaps unknown) variable, or by cause-and-effect. Further investigation would be necessary to determine which of these is appropriate in any given scenario. But, if causation does exist, this does imply that there will be correlation (but not necessarily vice versa).

**Example 3–7**

A real estate agency is worried that many of their agents are using poor sales techniques and that this is having a negative impact on sales. They believe this is because many of their agents received very low scores on their compulsory training course exam (an exam which is sat prior to beginning employment with the agency). They randomly select 10 of their agents, recording their exam score (out of 200) and the number of sales they made in the year 2005. This data is recorded in the following table:
Numerical descriptive measures

<table>
<thead>
<tr>
<th>Score</th>
<th>Sales</th>
</tr>
</thead>
<tbody>
<tr>
<td>185</td>
<td>212</td>
</tr>
<tr>
<td>122</td>
<td>143</td>
</tr>
<tr>
<td>157</td>
<td>184</td>
</tr>
<tr>
<td>165</td>
<td>182</td>
</tr>
<tr>
<td>183</td>
<td>201</td>
</tr>
<tr>
<td>191</td>
<td>235</td>
</tr>
<tr>
<td>121</td>
<td>154</td>
</tr>
<tr>
<td>158</td>
<td>187</td>
</tr>
<tr>
<td>166</td>
<td>178</td>
</tr>
<tr>
<td>102</td>
<td>146</td>
</tr>
</tbody>
</table>

(a) Produce a scatterplot of the data. Does there appear to be any correlation between exam score and sales? Explain.

(b) Compute the correlation coefficient. Comment on this value and its meaning for the real estate agency.

Solution 3–7

(a) If you were to produce a graph by hand, it would look something like the one below. The scatterplot below does appear to indicate there is some correlation between exam score and sales figures. Generally, the higher the exam score received, the greater the sales figures.
(b) Using the CORREL function in Excel, we can quickly calculate the correlation. We will demonstrate how to do this calculation by hand here and leave it as an exercise for the reader to check the result using Excel.

Using the computational formula we find that \( \sum X = 1550 \), \( \sum Y = 1822 \), \( \sum X^2 = 248518 \), \( \sum Y^2 = 339684 \) and \( \sum XY = 289872 \) and so

\[
r = \frac{n \sum XY - (\sum X)(\sum Y)}{\sqrt{n \sum X^2 - (\sum X)^2}(n \sum Y^2 - (\sum Y)^2)}
\]

\[
= \frac{10(289872) - (1550)(1822)}{\sqrt{10(248518) - (1550)^2}(10(339684) - (1822)^2)}
\]

\[
= 0.934265 \text{ (to 6 decimal places).}
\]

Since this value is quite close to one, there is strong positive correlation between exam score and sales figures.

**Ethical issues**

There is a fair degree of scepticism evident in the community when it comes to statistics. Many times this is due to experiences with someone misquoting or misusing the statistical results. It is therefore very important to analyse and report data in a fair and ethical manner. The following are some guidelines and comments that may be helpful in doing so:

- **Report results accurately but in a neutral and objective manner.** As well as doing and reporting the statistical calculations, this *always* requires prose to explain/discuss/interpret results in the context of the problem.

- **Choose the most appropriate numerical descriptive measures for a particular data set.** Knowing the shape of the distribution can also influence the choice of descriptive measures that you use. For example, the centre of a highly skewed data set might be better described by the median rather than the mean.

- **Interpretation of numerical values is subjective (although the actual calculations are objective).** Care must be taken to distinguish interpretive comments (even those made by experts) from factual information.

- **Poor presentation is not necessarily the same as unethical presentation of results.** Unethical behaviour occurs when:
  - an inappropriate summary method is chosen wilfully or
  - when some results or analyses are not reported because it would not support a particular position or favoured theory. It is therefore particularly important to report both good and bad results.

Source for many of these guidelines: Levine et al. 2005, *Statistics for managers using Microsoft Excel.*
Discussion points

Discussion point 3–1
Examine and complete Problem 3.60 on p. 137 of your textbook (Problem 3.61 on pp. 147–148 of the 4th edition). Discuss your answers to parts (c) and (d) (parts (e), (f) and (g) in the 4th edition). Why is it important to know whether these particular data sets are skewed? What implications are there for you as a decision maker if the data contained errors? What possible strategies could you use to avoid these problems?

Discussion point 3–2
Examine and complete Problem 3.74 on p. 140 of your textbook (Problem 3.73 on p. 150 of the 4th edition). Compare and discuss your answer with those of other students. What themes do you notice in your responses? How might understanding this issue be helpful to you in the workplace?

Summary
Now that you have completed this module, turn back to the objectives at the beginning of the module. Have you achieved these objectives?
Ensure that you attempt the recommended problems in the list of review questions below and at least a sample of problems from the optional list. This will help you to identify any areas of difficulty you have in achieving the module’s objectives.

Review questions

Recommended problems


5th edition text users please note: it is not necessary to attempt any parts of any questions relating to Z scores. In this course, outliers can be identified using visual checks of the data or using graphical methods such as stem-and-leaf plots or histograms.

Problem 3–1: Additional frequency distribution problem
A real estate specialises in selling small businesses. The following frequency distribution is taken from a random sample of 50 businesses which have been listed with the agency recently. Approximate the arithmetic mean and standard deviation of the value of the businesses listed with the agency.
<table>
<thead>
<tr>
<th>Value of business</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 to less than $50,000</td>
<td>4</td>
</tr>
<tr>
<td>$50,000 to less than $100,000</td>
<td>8</td>
</tr>
<tr>
<td>$100,000 to less than $150,000</td>
<td>10</td>
</tr>
<tr>
<td>$150,000 to less than $200,000</td>
<td>18</td>
</tr>
<tr>
<td>$200,000 to less than $250,000</td>
<td>7</td>
</tr>
<tr>
<td>$250,000 to less than $300,000</td>
<td>3</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>50</strong></td>
</tr>
</tbody>
</table>

Solution 3–1

$\overline{x} = $150,000, \ S = 165,137.5339 = $165,137.53.

Optional problems

Levine et al. 4th edition: Choose a selection of problems from Questions 3.1, 3.3 to 3.5, 3.7, 3.9, 3.11 to 3.13, 3.15 to 3.17, 3.19, 3.21, 3.23 to 3.25, 3.27 to 3.29, 3.31 to 3.33, 3.35 to 3.37, 3.41 to 3.43, 3.45, 3.55, 3.57 to 3.61, 3.63 to 3.67, 3.69 to 3.71, 3.74, 3.75 and from the CD 3.77, 3.79 and 3.80.

Levine et al. 5th edition: Choose a selection of problems from Questions 3.1, 3.3 to 3.5, 3.7 to 3.15, 3.17 to 3.19, 3.21, 3.23, 3.25 to 3.27, 3.29 to 3.35, 3.37, 3.39, 3.41 to 3.43, 3.55, 3.57 to 3.65 and 3.67 to 3.72.

5th edition text users please note: it is not necessary to attempt any parts of any questions relating to Z scores. In this course, outliers can be identified using visual checks of the data or using graphical methods such as stem-and-leaf plots or histograms.
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>4–3</td>
</tr>
<tr>
<td>Objectives</td>
<td>4–3</td>
</tr>
<tr>
<td>Readings</td>
<td>4–3</td>
</tr>
<tr>
<td>Basic Probability</td>
<td>4–3</td>
</tr>
<tr>
<td>Conditional probability</td>
<td>4–7</td>
</tr>
<tr>
<td>Bayes’ theorem</td>
<td>4–10</td>
</tr>
<tr>
<td>Counting rules</td>
<td>4–11</td>
</tr>
<tr>
<td>Ethical issues</td>
<td>4–13</td>
</tr>
<tr>
<td>Discussion points</td>
<td>4–14</td>
</tr>
<tr>
<td>Summary</td>
<td>4–14</td>
</tr>
<tr>
<td>Review questions</td>
<td>4–14</td>
</tr>
</tbody>
</table>
Introduction
This week we explore the topic of **probability**. This material is a foundation for much of the remainder of the course and so it is important that it is learned well. Probability is a topic that is used often in everyday life, having applications everywhere from the court room to medical drug trials to weather forecasting to the stock exchange. Because of the widespread use (and misuse) of probability, it is important for us to consider how to ethically and professionally employ these techniques in the workplace.

Objectives
On completion of this module you should be able to:
- demonstrate understanding of basic probability concepts including sample spaces, events, contingency tables, marginal and joint probability
- use and compute conditional probabilities
- use the addition and multiplication rules
- explain and demonstrate statistical independence and
- apply Bayes’ theorem and counting rules.

Readings

**Source**

**Textbook**
Levine et al. 4th & 5th editions
Ch. 4 (including Section 4.5 on the CD which accompanies the text) & Excel Handbook Ch. 4 (4th edition: p. 185, 5th edition: p. 177)

**Course website**
Visit the course website for links to any supplementary material for this week.

Basic Probability

**What is probability?**
Probability is the likelihood or chance that a particular event will occur. An event that is certain has a probability of 1. An impossible event has a probability of 0. All probabilities are values from 0 to 1 (inclusive).

The event we are interested in is described by a capital letter, such as $E$ (for event), or sometimes some more meaningful letter. For example, if we are interested in whether
it rains today, we might define \( R = \) the event that it rains today. The probability of an event is denoted by \( P(E) \) (or \( P(R) \) for the rain example).

**Three approaches to probability**

*A priori classical probability*—the probability of success is based on prior knowledge of the process involved. This kind of approach to probability is sometimes referred to as the relative frequency of occurrence. The probability of occurrence of an event, \( E \), is found via:

\[
P(E) = \frac{\text{Number of possible outcomes in which event occurs}}{\text{Total number of possible outcomes}}
\]

**Examples**

- A six-sided die has six possible outcomes. Therefore, the probability of rolling a six is 1/6.
- If two dice are rolled, there are \( 6 \times 6 = 36 \) possible outcomes. The probability of rolling a total of six (adding the two values together) is 5/36 since you could roll 1 & 5, 2 & 4, 3 & 3, 4 & 2 or 5 & 1.

**Empirical classical probability**—probabilities are based on observed data, not on prior knowledge of the process. This method involves an experiment to produce outcomes. The probability of occurrence of an event, \( E \), is found via:

\[
P(E) = \frac{\text{Number of outcomes in which event occurs}}{\text{Total number of outcomes}}
\]

**Example**

If we observed that one out of five cars in a car park is white, we could say that the probability that a car in the car park is white is 1/5.

**Subjective probability**—the chance of an event is assigned by a particular individual. This means that different people will assign different probabilities. Probabilities are based on the feelings or insights of the person determining the probability. It ranges from a guess as a worst case to accurate probabilities (particularly where the person has knowledge, understanding and experience of the event). Subjective probability is useful when probabilities of events can’t be determined empirically. An example might be that a doctor could assign a fairly accurate probability on the life expectancy of a cancer patient.

**Some definitions**

*An event*—a possible type of occurrence or the outcome of an experiment. For example, rolling a six on a single die or there is rain on Tuesday.

*A simple event*—can be described by a single characteristic.

*Sample space*—the collection or list of all possible events.

*Complement*—the complement of an event \( A \), written \( A' \), is all the events which are not part of the event \( A \). Since the event must either occur or not occur, we can say that \( P(A) + P(A') = 1 \).

*Joint event*—an event that has two or more characteristics.
Contingency table—a table of cross-classifications.

Simple (marginal) probability—the probability of occurrence of a simple event.

Joint probability—a probability referring to two or more events.

Important note: to be statistically correct, and because probability is always a number between 0 and 1 (inclusive), answers to probability questions should also be expressed as values between 0 and 1 and not as percentages.

Addition rules

If two events \( A \) and \( B \) are being consider, the probability of either of the events occurring is given using the following general addition rule:

\[
P(\text{or } A \text{ and } B) = P(A) + P(B) - P(A \text{ and } B)
\]

Mutually exclusive events are events where the occurrence of one means that the other(s) cannot occur. For example, the variable of gender results in mutually exclusive outcomes. A person is either male or female, not both. In manufacturing, whether a product is good or defective are mutually exclusive outcomes.

If two events, \( A \) and \( B \), are mutually exclusive, then the probability of either of the events occurring is found using the following addition rule for mutually exclusive events:

\[
P(A \text{ or } B) = P(A) + P(B)
\]

since \( P(A \text{ and } B) = 0 \).

Example 4–1

A store manager is interested in whether repeat customers spend more money in his store than first-time customers do. He takes a random sample of 200 customers over a one-month period with the following results:

<table>
<thead>
<tr>
<th>New Customer</th>
<th>Spent $200 or less</th>
<th>Spent over $200</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>67</td>
<td>54</td>
<td>121</td>
</tr>
<tr>
<td>No</td>
<td>29</td>
<td>50</td>
<td>79</td>
</tr>
<tr>
<td>Total</td>
<td>96</td>
<td>104</td>
<td>200</td>
</tr>
</tbody>
</table>

(a) Give an example of a simple event.

(b) Give an example of a joint event.

(c) What is the complement of ‘new customer’?

(d) Why is ‘new customer who spent over $200’ a joint event.

If a customer is selected at random, what is the probability that he or she:

(e) is a new customer?

(f) spent $200 or less?

(g) is not a new customer and spent over $200?
(h) is not a new customer or spent over $200?

(i) Explain the difference between your answers to (g) and (h).

Solution 4–1

(a) Recall that a simple event is something that can be described by a single characteristic. There are a number of simple events in this example; however, a couple of examples are:

- ‘spent $200 or less’
- ‘is a new customer’.

(b) A joint event is an event that has two or more characteristics. A couple of examples are:

- ‘a new customer who spent more than $200’
- ‘a new customer who spent $200 or less’.

(c) ‘Is not a new customer’—note that you could word this many ways.

(d) Since joint events have two or more characteristics, and ‘new customer’ and ‘spent over $200’ are two characteristics, ‘new customer’ and ‘spent over $200’ is a joint event.

(e) $P(\text{new customer}) = \frac{121}{200} = 0.605$

(f) $P(\text{spent $200 or less}) = \frac{96}{200} = 0.48$

(g) $P(\text{not a new customer and spent over $200}) = \frac{50}{200} = 0.25$

(h) $P(\text{not a new customer or spent over $200}) = \frac{79}{200} + \frac{104}{200} - \frac{50}{200} = \frac{133}{200} = 0.665$

(i) The answer to part (g) is an ‘and’ problem, meaning the answer has two parts: ‘not a new customer’ and ‘spent over $200’. This means we look at only one space (cell) in the contingency table where both events occur.

Part (h) has three parts since it is the probability of ‘not a new customer’ plus ‘spent $200 or less’ minus the probability of the overlap of the two events (‘not a new customer and spent over $200’). Therefore we are adding a row total and column total and subtracting the overlapping cell. If we did not do the subtraction, we would be counting those customers in the ‘not a new customer and spent over $200’ cell twice. Another way to think of this probability is by adding up the three cells in the contingency table that relate to the options:

$\frac{29}{200} + \frac{50}{200} + \frac{54}{200} = \frac{133}{200} = 0.665$. 

4–6
Conditional probability

Conditional probability

Conditional probability is the probability that an event will occur, given that another event has already happened. It is expressed as \( P(A|B) \) which is read as ‘the probability of \( A \) given \( B \)’. This probability is found via:

\[
P(A|B) = \frac{P(A \text{ and } B)}{P(B)}
\]

Note this rule applies only if \( P(B) \neq 0 \).

Statistical independence

Independent events are events when the occurrence of one event has no effect on the probability that another will occur. Dependent events means that when one event occurs, this changes the probability that another will occur. In order to demonstrate statistically that two events are independent, we use the following rule.

Two events are statistically independent if and only if \( P(A|B) = P(A) \).

Multiplication rules

When events \( A \) and \( B \) are independent, the probability \( P(A \text{ and } B) \) can be found via:

\[
P(A \text{ and } B) = P(A)P(B).
\]

When events \( A \) and \( B \) are not independent, the probability \( P(A \text{ and } B) \) can be found via:

\[
P(A \text{ and } B) = P(A|B)P(B).
\]

Example 4–2

Using the table in Example 4–1 above, answer the following questions:

(a) Given that a customer spent over $200, what is the probability that he or she is a new customer?

(b) Given that a customer is new, what is the probability that he or she spent over $200?

(c) Explain the difference in your answers for (a) and (b).

(d) Are the two events, new customer and customer spent over $200 statistically independent?
Solution 4–2

(a) \( P(\text{new customer|spent over $200}) = \frac{54}{104} = 0.5192 \) (to 4 decimal places)

(b) \( P(\text{spent over $200|new customer}) = \frac{54}{121} = 0.4463 \) (to 4 decimal places)

(c) The conditional events have been reversed. Part (a) explores only those customers who have spent over $200 and so of these, what is the probability that a customer is a new customer. Part (b) explores only the new customers and so of these, what is the probability that a customer spent over $200. Clearly since there are differing numbers of customers who spent over $200 (104 customers in the table) and who are new customers (121 customers), the results of the two conditional probabilities will differ.

(d) The test for statistical independence of two events (A and B), says that they are independent if and only if \( P(A \mid B) = P(A) \).

Since \( P(\text{new customer|spent over $200}) = 0.5192 \) and \( P(\text{new customer}) = 0.605 \) are different, the two events are not statistically independent.

Note we could also examine this by reversing the events. So the events would be independent if and only if \( P(B \mid A) = P(B) \). So since

\[
P(\text{spent over $200|new customer}) = \frac{54}{121} = 0.4463 \quad \text{and} \quad P(\text{spent over $200}) = \frac{104}{200} = 0.52
\]

are not equal, we see again that the two events are not statistically independent.

Example 4–3

Chief Executive Officers (CEOs) of companies are demonstrating increasing mobility between companies. A concerned employer at a particular company worries that constantly changing the CEO in his company has left an unstable work environment for staff, and customers confused by ever changing products. He conducts a survey of 150 large companies, looking at whether the companies’ profit levels have increased or decreased and whether the company has had a new CEO within the last five years. Based on his survey he discovers that of the 150 companies, 67 have reported decreased profits. A total of 31 of the companies have not replaced their CEO within the last five years and of these, 29 have increased their profits.

Produce a contingency table summarising this information and use it to find the probability that a company has increased its profit given that it has replaced its CEO in the last five years.
**Solution 4–3**

We start by determining what the two main events are in this problem. Firstly we are interested in whether a given company has a new CEO in the last five years (we’ll call this event $N$, for new CEO). Secondly we are interested in whether a given company has increased or decreased its profit margin (we’ll call this event $I$, for increased profit). Therefore we are looking to evaluate the probability $P(I \mid N)$.

We will use these two events to set up the table (see below) with new CEO down the side and profit across the top. For each event, we list the two options (yes and no to new CEO, and increased or decreased for profit margin).

\[
\begin{array}{c|c|c}
\text{Increased profit} & \text{Yes} & \text{No} \\
\hline
\text{New CEO in last five years} & & \\
\text{Yes} & & \\
\text{No} & & \\
\text{Total} & & \\
\end{array}
\]

Now we can start filling in the information we’re given. We know that there are a total of 150 companies, so this should be inserted in the grand total cell. We’re told that 67 of the total companies have decreased their profits, so we put the 67 in the total for the decreased profit row. We’re told that 31 companies have not replaced their CEO in the last five years, so this value goes in the total for that row. Finally we’re told that of the 31 companies that have not replaced their CEO, 29 have increased their profits, so we place this in the increased profit and not replaced CEO cell. The table should now look like this:

\[
\begin{array}{c|c|c}
\text{Increased profit} & \text{Yes} & \text{No} \\
\hline
\text{New CEO in last five years} & & \\
\text{Yes} & & \\
\text{No} & 29 & 31 \\
\text{Total} & 67 & 150 \\
\end{array}
\]

Now we can start filling in the blanks using some basic logic. If 67 of the total have decreased their profits, then $150 - 67 = 83$ companies must have increased their profits. If 31 have not replaced their CEO then $150 - 31 = 119$ must have replaced their CEO.

\[
\begin{array}{c|c|c}
\text{Increased profit} & \text{Yes} & \text{No} \\
\hline
\text{New CEO in last five years} & & \\
\text{Yes} & & 119 \\
\text{No} & 29 & 31 \\
\text{Total} & 83 & 67 \\
\end{array}
\]

If, of those companies who have not replaced their CEO, 29 have increased their profit, then $31 - 29 = 2$ have decreased profit. If, of those companies who have increased their profits, 29 have not replaced their CEO, then $83 - 29 = 54$ have replaced their CEO. If of those companies who have decreased their profits, 2 have not replaced their CEO, then $67 - 2 = 65$ have replaced their CEO. The completed table should now look like this:
Probability

<table>
<thead>
<tr>
<th>New CEO in last five years</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>54</td>
<td>65</td>
<td>119</td>
</tr>
<tr>
<td>No</td>
<td>29</td>
<td>2</td>
<td>31</td>
</tr>
<tr>
<td>Total</td>
<td>83</td>
<td>67</td>
<td>150</td>
</tr>
</tbody>
</table>

Now, the probability that a company has increased its profit given that it has replaced its CEO in the last five years is:

\[ P(I \mid N) = \frac{54}{119} = 0.4538 \text{ (to 4 decimal places).} \]

Bayes’ theorem

Bayes’ theorem was developed by Thomas Bayes in the 1700s. It allows information about a second event to be used to revise the probability that a first event has occurred. It extends the conditional probability ideas that we have explored so far.

Bayes’ theorem is:

\[ P(B_i \mid A) = \frac{P(A \mid B_i) P(B_i)}{P(A \mid B_1) P(B_1) + P(A \mid B_2) P(B_2) + \ldots + P(A \mid B_k) P(B_k)} \]

where \( B_i \) is the \( i \)th event out of \( k \) mutually exclusive and collectively exhaustive events. In situations where \( k = 2 \), Bayes’ theorem can be expressed more simply as:

\[ P(B_i \mid A) = \frac{P(A \mid B_i) P(B_i)}{P(A \mid B_1) P(B_1) + P(A \mid B') P(B')} \]

In this course, we are mostly interested in using Bayes’ theorem for two events (i.e., for \( k = 2 \)).

Example 4–4

Two large supermarket chains, ‘Value Mart’ and ‘Goldies’, have been fiercely competitive with their range of products. Data are collected by an external watchdog which indicates that when a new product is introduced to the market, Goldies will begin to retail that product 90% of the time. If Goldies has already begun to market a product, Value Mart will take up the product 85% of the time. When Goldies does not begin to retail a product, Value Mart begins to retail it 25% of the time.

(a) Determine the probability that Goldies will market a product given that Value Mart is already retailing it.

(b) Find the probability that Value Mart retails a given product, \( P(V) \).

(c) Why might this information be useful?
Solution 4–4

Let $V$ be the event that Value Mart retails a product and $G$ the event that Goldies retails a product. Then we know that

\[ P(G) = 0.9 \left(= \frac{90}{100}\right), \quad P(V|G) = 0.85 \left(= \frac{85}{100}\right), \quad P(V|G') = 0.25 \left(= \frac{25}{100}\right). \]

Remember that $G'$ means the complement of $G$ (or ‘not $G$’) and so $P(G') = 1 - P(G) = 1 - 0.9 = 0.1$.

(a) Given that we are looking for $P(G|V)$, Bayes’ Theorem tells us that

\[
P(G|V) = \frac{P(V|G)P(G)}{P(V)} = \frac{P(V|G)P(G)}{P(V|G)P(G) + P(V|G')P(G')} = \frac{0.85(0.9)}{0.85(0.9) + 0.25(0.1)} = 0.9684 \text{ (to 4 decimal places)}.\]

(b) $P(V) = P(V|G)P(G) + P(V|G')P(G')$

\[
= 0.85(0.9) + 0.25(0.1) = 0.79.
\]

(c) There are many possible ways this information might be useful. We’ll discuss it briefly here and leave it as an exercise to the reader to explore other possibilities. The external watchdog may be interested in this information in order to determine whether there is any collusion between the two supermarket chains. Do they choose to only take up certain products or brands, do they choose not to market particular products or brands and does the decision of one supermarket chain impact the decision of the other chain? The information also lets us see whether one supermarket is following the other or whether the decisions appear to be autonomous. For example, we note that when Goldie has begun to retail a product, Value Mart then begins to retail the product 85% of the time, but when Value Mart has begun to retail a product, Goldies begins to retail 96.84% (taken from 0.9684) of the time. Perhaps this indicates that Goldies is more likely to follow Value Mart than the other way round. Further investigation is needed to establish this, but the watchdog may use this as a clue to look further into the possibility of collusion.

Also note that the two events are not statistically independent. You can check this by using the rule that events $A$ and $B$ are statistically independent if and only if $P(A|B) = P(A)$.

Counting rules


Counting rule 1

For $k$ events (mutually exclusive and collectively exhaustive) which can occur in each of $n$ trials, the number of possible outcomes is $k^n$. 
Probability

Example
A standard dice has six possible outcomes \((k = 6)\). These are

- **mutually exclusive**—there is no overlap between outcomes (for example, you cannot roll both a 1 and a 2 on the same dice at the same time!)
- **collectively exhaustive**—there are no other possible outcomes (the only possible outcomes are 1 to 6)

If a dice is rolled three times, then \(n = 3\) there are \(6^3 = 216\) possible outcomes.

Counting rule 2
For \(k_1\) events on the first trial, \(k_2\) on the second trial, \(\ldots\) and \(k_n\) on the \(n\)th trial, then the number of possible outcomes is

\[ k_1 \times k_2 \times \ldots \times k_n. \]

Example
A courier must travel between three drop-offs in the following order: A, B and C. From A, there are two ways to reach B \((k_1 = 2)\). From B, there are four ways to C \((k_2 = 4)\).

There are \(k_1 \times k_2 = 2 \times 4 = 8\) ways of travelling from A to C.

Counting rule 3
The number of ways that \(n\) objects can be arranged in order is

\[ n! = n \times (n-1) \times (n-2) \times \ldots \times 2 \times 1 \]

\(n!\) is read ‘\(n\) factorial’. We define \(0! = 1\) and \(1! = 1\).

Example
The number of ways you can arrange five cars in five parking spaces is

\[ 5! = 5 \times 4 \times 3 \times 2 \times 1 = 120. \]

This means there are five ways you can allocated a car to the first parking space, and then four are left to be allocated to the second space, three for the third, two for the fourth and finally only one is left to be allocated to the fifth parking space.

Counting rule 4
**Permutations**—the number of ways of arranging \(X\) objects selected from \(n\) objects *in order* is

\[ {}^nP_X = \frac{n!}{(n-X)!} \]

Example
The number of ways you can seat three people in six chairs is:

\[ {}^6P_3 = \frac{6!}{(6-3)!} = \frac{6 \times 5 \times 4 \times 3 \times 2 \times 1}{3 \times 2 \times 1} = \frac{720}{6} = 120 \text{ ways.} \]
Counting rule 5

**Combinations**—the number of ways of arranging $X$ objects selected from $n$ objects irrespective of order is

$$n \binom{X}{n} = \binom{n}{X} = \frac{n!}{X!(n-X)!}$$

**Example**

The number of ways you can choose three people from a group of six to be members on a committee irrespective of order is

$$\binom{6}{3} = \frac{6!}{3!(6-3)!} = \frac{720}{6 \times 6} = 20 \text{ ways.}$$

**Example 4–5**

The CEO of a large accounting firm must choose a team of 6 accountants to work on a new project. He has a pool of 14 accountants to choose from. How many ways can the CEO select the team?

**Solution 4–5**

We are looking for the number of ways the CEO can select 6 accountants from the pool of 14 (order will not matter here). So we find that there are

$$\binom{14}{6} = \frac{14!}{6!(14-6)!} = \frac{14!}{6!8!} = \frac{(14)(13)(12)(11)(10)(9)}{(6)(5)(4)(3)(2)(1)} = 3003 \text{ ways.}$$

**Ethical issues**

Because many people are not familiar with concepts of probability, care needs to be taken in how probability results are presented. Presentation of probabilities would only become an ethical issue if results are intentionally presented in a misleading way. Ethical problems might also include excluding important background information which is relevant to the presentation of the probability information.

Subjective probability is open to misuse because it is often based on a person’s feelings, intuition or experience. Although everyone has an opinion, and it is not usually wrong to express these, professionals need to be careful that a good guess is not presented as fact. Without a reasonable level of confidence in a subjective probability, presenting this information can become an ethical problem or at very least misleading. Subjective probability also makes it easy for dishonest people to manipulate information to serve their own purpose.

Priori classical probabilities are based on historical data. To avoid ethical problems, care must be taken that this data is reliable and complete. It should also be remembered that probabilities relate to expectations in the ‘long run’. For example, we know that the probability of rolling a six on a single die is 1/6, but this does not guarantee that if we roll a die six times then we will get exactly one six. But, if we were to roll a fair die multiple times, in the long run, we would average about 1/6 of rolls as sixes.
Discussion points

Discussion point 4–1

Discussion point 4–2
Examine and complete Problem 4.55 on p. 175 of your textbook (Problem 4.52 on p. 184 of the 4th edition). What ethical issues are there relating to this issue? Discuss. Give an example of how this information might potentially be misused. What have you learned about expressing probability in reporting findings?

Summary
Now that you have completed this module, turn back to the objectives at the beginning of the module. Have you achieved these objectives?

Ensure that you attempt the recommended problems in the list of review questions below and at least a sample of problems from the optional list. This will help you to identify any areas of difficulty you have in achieving the module’s objectives.

Review questions

Recommended problems
Levine et al. 4th edition: Questions 4.2, 4.4, 4.6, 4.8, 4.10, 4.14, 4.16, 4.18, 4.20, 4.28, 4.30, 4.32, 4.36, 4.38 to 4.44, 4.46 and 4.50 from the textbook and from the CD 4.54, 4.56, 4.64 and 4.66.

Levine et al. 5th edition: Questions 4.2, 4.4, 4.6, 4.10, 4.11, 4.14, 4.16, 4.18, 4.20, 4.24, 4.30, 4.32, 4.34, 4.38, 4.40 to 4.48 and 4.50 from the textbook and from the CD 4.64, 4.66, 4.72 and 4.74.

Optional problems
Levine et al. 4th edition: Choose a selection of problems from Questions 4.1, 4.3, 4.5, 4.7, 4.9, 4.11 to 4.13, 4.15, 4.17, 4.19, 4.21 to 4.27, 4.29, 4.31, 4.33 to 4.35, 4.37, 4.45, 4.47 to 4.49 and 4.51 to 4.53 from the textbook and from the CD 4.55, 4.57 to 4.63 and 4.65.

Levine et al. 5th edition: Choose a selection of problems from Questions 4.1, 4.3, 4.5, 4.7 to 4.9, 4.12, 4.13, 4.15, 4.17, 4.19, 4.21 to 4.23, 4.25 to 4.29, 4.31, 4.33, 4.35 to 4.37, 4.39, 4.49, 4.51 to 4.55 from the textbook and from the CD 4.61 to 4.63, 4.65, 4.67 to 4.71 and 4.73.
Discrete distributions
Contents

Introduction ............................................................................................................... 5–3
Objectives .................................................................................................................. 5–3
Readings .................................................................................................................... 5–3
Probability distributions for a discrete random variable ........................................... 5–4
Describing a discrete distribution.............................................................................. 5–5
Covariance and probability in finance....................................................................... 5–8
Notes on rounding ................................................................................................... 5–11
The binomial distribution ......................................................................................... 5–11
The Poisson distribution ......................................................................................... 5–14
Discussion points ..................................................................................................... 5–16
Summary ................................................................................................................. 5–17
Review questions..................................................................................................... 5–17
Introduction

Simple probability often explores events that have two outcomes: they either occur or they don’t. We then try to quantify how likely they are to occur by assigning a probability as a number between zero and one. Many real life problems, however, have many possible outcomes. For example, a call centre can receive 0, 1, 2, 3, … complaints in a day or a company may invest money for a certain number of (whole) months. These variables are called ‘discrete’ because they only take discrete, non-negative whole number values (hence our topic for the week is discrete distributions). We will be exploring the expected values (or what might occur ‘on average’) for these kinds of variables and determining variance and standard deviations.

Sometimes we are interested in an event which might be repeated many times. For example, on an assembly line in a factory a product might be measured as either acceptable (often referred to as ‘good’)) or defective. This outcome will be measured for each of the hundreds of items that are produced each day. We might then be interested in the probability that a certain number of the items will be good. The binomial distribution is useful in helping us describe this kind of situation.

The final topic we will explore this week relates to situations where we are interested in the rate at which some event is occurring. For example, if we know that, on average, complaints come in to a consumer helpline at three per hour, what is the probability that we might receive more than four complaints in an hour? More than six complaints in an hour? The Poisson distribution (pronounced ‘pwa sohn’) can be helpful in answering these sorts of questions.

Objectives

On completion of this module you should be able to:

- find expected value, variance and standard deviation for a discrete random variable
- calculate and interpret covariance
- determine the appropriateness of the binomial distribution for certain situations
- calculate probabilities of events for the binomial distribution
- determine the appropriateness of the Poisson distribution for certain situations and
- calculate probabilities of events for the Poisson distribution.

Readings

Source

Textbook Levine et al. 4th edition
Sections 5.1 to 5.3 and 5.5 & Excel Handbook Sections EH5.1, 5.2, 5.4 & 5.5 (pp. 218–221)
Discrete distributions

Or

Levine et al. 5th edition
Sections 5.1 to 5.4 & Excel Companion E5.1 to E5.4 (pp. 211–214)

Note: the Hypergeometric distribution is not covered in this course.

Course website Visit the course website for links to any supplementary material for this week.

Probability distributions for a discrete random variable

A random variable is a variable that can take on different values according to the outcome of a chance experiment. Random variables are either described as discrete or continuous.

Discrete random variables can take on non-negative whole numbers and are the result of a counting process. They are used in situations where fractional numbers don’t make sense.

An upper case letter such as $X$ is used to denote the random variable and the corresponding lower case letter, $x$, to denote the particular value assumed by the random variable. For example, if the random variable is the number of people who arrive at an ATM per minute, we obviously could not have 3.15 people arrive in a given minute. The number of people arriving could be 0, 1, 2, 3, 4, 5, etc. So if $X$ denotes the number of people who arrive at an ATM per minute, then $x$ is the value 0 or 1 or 2 or 3… etc.

Continuous random variables can take on fractional numbers and are the result of measurement rather than counting. For example, a continuous random variable might involve measuring the amount of petrol (in litres) in a storage container. We will explore continuous random variables in more detail next week.

A probability distribution describes a random variable. It is the relative frequency distribution that should theoretically occur for observations from a population. A discrete probability distribution is a list of all possible outcomes of an experiment with their probability of occurrence.

Given the random variable, $X$, let $x_i$, where $i=1,2,\ldots,k$ denote the $k$ distinct values that $X$ may assume. Then, the probability that the random variable $X$ will assume value $x_i$ is denoted by $P(X = x_i)$. For example, in the ATM example, the $x_i$ values would be $x_1 = 0$, $x_2 = 1$, $x_3 = 2$, etc. and then the probability of two people arriving in a given minute would be denoted $P(X = 2)$. The probability distribution for the random variable $X$, associates a probability $P(X = x_i)$ for each of the distinct outcomes $x_i$. 
Example 5–1
A quality assurance check is carried out by a manufacturer prior to shipping a particular model of calculator to the retailers. Each calculator is checked for four key problems. If \( X \) is the number of problems identified on each calculator, then it can take on the values \( x = 0, 1, 2, 3 \) or \( 4 \). The probability distribution then associates a probability with each of these outcome values as follows:

<table>
<thead>
<tr>
<th>( x )</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P(X = x) )</td>
<td>0.8</td>
<td>0.1</td>
<td>0.05</td>
<td>0.03</td>
<td>0.02</td>
</tr>
</tbody>
</table>

We know, therefore, that the probability that a calculator has no problems is 0.8, has one problem is 0.1, has two problems is 0.05, etc.

Important notes:

- The probabilities add to 1 (\( 0.8 + 0.1 + 0.05 + 0.03 + 0.02 = 1 \)) since one of the possible \( x \) value outcomes must occur (there must be either 0, 1, 2, 3 or 4 problems with each calculator).
- All the probabilities associated with each possible outcome are values between 0 and 1 (inclusive).
- Often \( P(X = x_i) \) is abbreviated to \( P(x_i) \). For example \( P(0) \) denotes \( P(X = 0) \), \( P(1) \) denotes \( P(X = 1) \), etc.

We can illustrate the probability distribution graphically as follows:

![Probability distribution graph](image)

**Describing a discrete distribution**

Often we would like to know the mean value of a random variable. Since this is a measure of what is expected in the long run it is known as the *expected value of a discrete random variable* and is defined as:

\[
\mu = E(X) = \sum_{i=1}^{k} x_i P(X = x_i)
\]

The term \( E(X) \) is read as ‘the expected value of \( X \)."
The **variance of a discrete random variable** is defined as:

\[
\sigma^2 = \sum_{i=1}^{k} [x_i - E(X)]^2 P(X = x_i)
\]

and the standard deviation of a discrete random variable is defined as:

\[
\sigma = \sqrt{\sigma^2} = \sqrt{\sum_{i=1}^{k} [x_i - E(X)]^2 P(X = x_i)}
\]

We will illustrate these concepts with an example.

**Example 5–2**

The CEO of a large corporation is concerned that the corporation’s intranet has been suffering with a number of faults and is unavailable much of the time. He arranges for the collection of data which details the number of times where the intranet has been down, and for how long, for a random sample of three hundred business days. The results are given in the following table:

<table>
<thead>
<tr>
<th>Hours unavailable (per day)</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>210</td>
</tr>
<tr>
<td>1</td>
<td>24</td>
</tr>
<tr>
<td>2</td>
<td>30</td>
</tr>
<tr>
<td>3</td>
<td>24</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
</tr>
</tbody>
</table>

(a) Form the probability distribution for the number of hours per day the intranet was down.

(b) Compute the mean or expected number of hours that the intranet will be down per day.

(c) Compute the standard deviation.

What is the probability that on any given day:

(d) the intranet will be down for fewer than two hours?

(e) the intranet will not be down at all?

(f) The intranet will be down for at least three hours?
Solution 5–2

(a) The probability distribution requires only that we find the relative frequencies for each of the number of hours per day, by dividing the frequency by the total days \((210 + 24 + \ldots + 0 + 6 = 300)\). This gives us probabilities for each outcome as follows:

<table>
<thead>
<tr>
<th>Hours unavailable per day ((x))</th>
<th>(P(X = x))</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>(\frac{210}{300} = 0.7)</td>
</tr>
<tr>
<td>1</td>
<td>(\frac{24}{300} = 0.08)</td>
</tr>
<tr>
<td>2</td>
<td>(\frac{30}{300} = 0.1)</td>
</tr>
<tr>
<td>3</td>
<td>(\frac{24}{300} = 0.08)</td>
</tr>
<tr>
<td>4</td>
<td>(\frac{6}{300} = 0.02)</td>
</tr>
<tr>
<td>5</td>
<td>(\frac{0}{300} = 0)</td>
</tr>
<tr>
<td>6</td>
<td>(\frac{6}{300} = 0.02)</td>
</tr>
</tbody>
</table>

(b) \(\mu = E(X) = \sum_{i=1}^{4} x_i P(X = x_i) = (0 \times 0.7) + (1 \times 0.08) + \ldots + (6 \times 0.02) = 0.72\). This means that we can expect the intranet to be down for 0.72 of an hour each day, or about 43 minutes per day.

(c) \(\sigma^2 = \sum_{i=1}^{4} [x_i - E(X)]^2 P(X = x_i)\)

\[
\begin{align*}
&= (0 - 0.72)^2 (0.7) + (1 - 0.72)^2 (0.08) + \ldots + (6 - 0.72)^2 (0.02) \\
&= 1.7216 \\
\end{align*}
\]

\(\sigma = \sqrt{\sigma^2} = \sqrt{1.7216} = 1.3121\) (to 4 decimal places).

(d) The probability that the intranet is down for fewer than two hours in a given day is:

\[P(\text{fewer than 2 hours}) = P(X < 2) = P(X = 0) + P(X = 1) = 0.7 + 0.08 = 0.78.\]

The probability that the intranet will be down for fewer than two hours is 0.78.

(e) The probability that the intranet will not be down at all in a given day is:

\[P(0 \text{ hours}) = P(X = 0) = 0.7.\]

The probability that the intranet will not be down at all is 0.7.
The probability that the intranet is down for at least three hours in a given day is:

\[
P(\text{at least 3 hours}) = P(X \geq 3) = P(X = 3) + P(X = 4) + P(X = 5) + P(X = 6) \\
= 0.08 + 0.02 + 0 + 0.02 = 0.12
\]

The probability that the intranet will be down for at least three hours is 0.12.

Clearly the CEO needs to introduce measures to establish a more reliable intranet!!

**Covariance and probability in finance**

Covariance is a measure of the strength of the relationship between variables. The covariance of two discrete random variables, \( X \) and \( Y \), is defined:

\[
\sigma_{XY} = \sum_{i=1}^{k} \left( x_i - E(X) \right) \left( y_i - E(Y) \right) P(x_i, y_i)
\]

A positive value indicates a positive relationship between the two variables (as one variable increases, so does the other), a value of zero indicates no relationship, and a negative value indicates a negative relationship (as one variable increases, the other decreases).

The *expected value of the sum of two random variables*, \( X \) and \( Y \), is equal to the sum of the expected values:

\[
E(X + Y) = E(X) + E(Y)
\]

The *variance of the sum of two random variables*, \( X \) and \( Y \), is equal to the sum of the variances plus twice the covariance:

\[
\text{var}(X + Y) = \sigma_{X,Y}^2 = \sigma_X^2 + \sigma_Y^2 + 2\sigma_{XY}
\]

The *standard deviation of the sum of two variables* is \( \sigma_{X,Y} = \sqrt{\sigma_{X,Y}^2} \).

In finance, we might be interested in the expected value and standard deviation of a sum of two investments. Because we may want to divide an investment into two unequal portions, we will use \( w \) to represent the portion assigned to one asset and \( (1 - w) \) the portion assigned to the other. Using the expected value and standard deviation formulae above, we can derive the following.

The *portfolio expected return* for a two asset investment is defined:

\[
E(P) = wE(X) + (1 - w)E(Y)
\]

where \( E(P) \) is the expected portfolio return, \( E(X) \) is the expected return of asset \( X \), \( E(Y) \) is the expected return of asset \( Y \), \( w \) is the portion of the portfolio value assigned to asset \( X \) and \( (1 - w) \) is the portion of the portfolio value assigned to asset \( Y \).

The *portfolio risk* is defined:

\[
\sigma_p = \sqrt{w^2\sigma_X^2 + (1 - w)^2\sigma_Y^2 + 2w(1 - w)\sigma_{XY}}
\]
Example 5–3

One of your clients is deciding how they should invest a sum of money. They have obtained a report giving the predicted annual return for a $1000 investment in two different stocks. The following probability distribution was included in the report.

<table>
<thead>
<tr>
<th>Probability</th>
<th>Stock J</th>
<th>Stock K</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>-$50</td>
<td>-$105</td>
</tr>
<tr>
<td>0.2</td>
<td>$10</td>
<td>$2</td>
</tr>
<tr>
<td>0.4</td>
<td>$30</td>
<td>$25</td>
</tr>
<tr>
<td>0.2</td>
<td>$100</td>
<td>$200</td>
</tr>
</tbody>
</table>

Determine the following values:

(a) The expected return of stock J
(b) The expected return for stock K
(c) The standard deviation for stock J
(d) The standard deviation for stock K
(e) The covariance of stock J and stock K
(f) Based on your answers above, what recommendations might you make to the client. Explain.

Solution 5–3

Let $X =$ stock J and $Y =$ stock K.

(a)  $E(X) = \mu_x = \sum_{i=1}^{k} x_i P(X = x_i) = -50(0.2) + 10(0.2) + 30(0.4) + 100(0.2) = $24.

(b)  $E(Y) = \mu_y = \sum_{i=1}^{k} y_i P(Y = y_i) = -105(0.2) + 2(0.2) + 25(0.4) + 200(0.2) = $29.40.

(c)  $\text{Var}(X) = \sigma_x^2 = \sum_{i=1}^{k} [x_i - E(X)]^2 P(X = x_i)$

\[= (-50 - 24)^2 (0.2) + (10 - 24)^2 (0.2) + (30 - 24)^2 (0.4) + (100 - 24)^2 (0.2)\]

\[= 2304\]

$\sigma_x = \sqrt{\sigma_x^2} = \sqrt{2304} = $48

(d)  $\text{Var}(Y) = \sigma_y^2 = \sum_{i=1}^{k} [y_i - E(Y)]^2 P(Y = y_i)$

\[= (-105 - 29.4)^2 (0.2) + (2 - 29.4)^2 (0.2) + (25 - 29.4)^2 (0.4) + (200 - 29.4)^2 (0.2)\]

\[= 9591.44\]

$\sigma_y = \sqrt{\sigma_y^2} = \sqrt{9591.44} = $97.94

Based on your answers above, what recommendations might you make to the client. Explain.
(e) \[ \sigma_{xy} = \sum_{i=1}^{k} [x_i - E(X)][y_i - E(Y)]P(x_i, y_i) \]
\[ = (-50 - 24)(-105 - 29.4)(0.2) + (10 - 24)(2 - 29.4)(0.2) \]
\[ + (30 - 24)(25 - 29.4)(0.4) \]
\[ + (100 - 24)(200 - 29.4)(0.2) \]
\[ = 4648.4 \]

(f) Stock K has a higher expected return than stock J but it also has a higher standard deviation. If the client is more willing to take a risk then they should invest in stock K, but if they are more inclined to look for safer options then they should invest in stock J. Because the covariance is positive, this implies that there is a positive relationship between the two investments. In other words, when one stock increases, so too does the other and when one stock decreases so too does the other.

Example 5–4

Returning to the scenario in Example 5–3, compute the portfolio expected return and the portfolio risk if 90% of the stock is invested in stock J. Repeat when 50% of the stock is invested in stock J.

Solution 5–4

When 90% of the stock is invested in stock J, \( w = 0.9 \). Then

\[ E(P) = wE(X) + (1 - w)E(Y) \]
\[ = 0.9(24) + 0.1(29.4) \]
\[ = $24.54 \]

and

\[ \sigma_p = \sqrt{w^2\sigma_x^2 + (1-w)^2\sigma_y^2 + 2w(1-w)\sigma_{xy}} \]
\[ = \sqrt{(0.9)^2(2304) + (0.1)^2(9591.44) + 2(0.9)(0.1)(4648.4)} \]
\[ = $52.90 \]

When 50% of the stock is invested in stock J, \( w = 0.5 \). Then

\[ E(P) = wE(X) + (1 - w)E(Y) \]
\[ = 0.5(24) + 0.5(29.4) \]
\[ = $26.70 \]

and

\[ \sigma_p = \sqrt{w^2\sigma_x^2 + (1-w)^2\sigma_y^2 + 2w(1-w)\sigma_{xy}} \]
\[ = \sqrt{(0.5)^2(2304) + (0.5)^2(9591.44) + 2(0.5)(0.5)(4648.4)} \]
\[ = $72.79 \]

Although both investments are fairly risky, of these two portfolio options, investing 50% in stock J offers a better return. There is, however, a much larger risk associated with this option.
Notes on rounding

All final answers that are money amounts must be rounded to exactly 2 decimal places (for dollars and cents) and must include the dollar sign. The only exception is if the value is whole dollars, when 0 decimal places are acceptable (but clearly a dollar sign is still required). But, don’t round intermediate steps of working or you will introduce a rounding error (and be marked wrong in assessment items). Some examples:

Correct: $4.57 $100.95 $5,984.56 $495 ✓
Incorrect: $4.6 $100.9543 ✗

When rounding other numbers, use your common sense. In general, larger numbers need fewer decimal places and smaller number needs more. A good rule of thumb is to keep at least two to four decimal places in your final answers. But always keep all the decimal places in intermediate steps of working—using the memory on your calculator will help you do this.

The binomial distribution

A Bernoulli random trial is a trial that has two possible outcomes. These are assigned either the value 0 or 1. A random variable, $X$, that is associated with a Bernoulli random trial is called a Bernoulli random variable. Some examples are:

- Students sitting a particular exam either receive a pass grade (assigned the value $X=1$) or a fail grade (assigned the value $X=0$). Here $X$ is a Bernoulli random variable.
- Last year, a company forecast expected profit for the financial year. At the end of that year they will have either met the forecasted profit ($X=1$) or failed to meet it ($X=0$). Again, $X$ is a Bernoulli random variable.

In many situations, we are more interested in the outcome of a sequence of Bernoulli trials, rather than a single trial. For example, we might be interested in the number of students in the class who received pass grades in the exam or the number of years where the company met their forecast profits.

We denote the series of $n$ Bernoulli trials as $X_1, X_2, \ldots, X_n$. The sum of $n$ independent and identically distributed Bernoulli random variables is denoted by $X$:

$$ X = X_1 + X_2 + \ldots + X_n. $$

Then $X$ is called a binomial random variable.

Notice that we have assumed the $n$ trials are independent and identically distributed. Independent means that the two events (or trials) are statistically independent. In other words, the outcome of one trial does not affect the outcome of any other trials. Identically distributed means that the probability of $X_i=1$ and $X_i=0$ are the same for all Bernoulli trials in the sequence. Usually we denote this as $P(X_i=1) = p$ and $P(X_i=0) = 1-p$ for $i=1,2,\ldots,n$. So in the exam example above, $p$ is the probability that a particular student receives a passing grade in the exam. Therefore, the
probability that they don’t receive a passing grade (i.e., they receive a failing grade) is $1 - p$.

The probability of a particular outcome of a discrete random variable, $X$, is expressed as $P(X = x)$ or just $P(x)$. For a binomial random variable, this can be expressed using the following mathematical formula:

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

Notice that this includes the binomial coefficient $\binom{n}{x}$ which is the combinations formula we first encountered in Week 4.

For the binomial distribution to be appropriate, the random variable must have the following characteristics:

- there are a fixed number of trials (called $n$),
- each trial has two mutually exclusive and collectively exhausted outcomes (called ‘success’ and ‘failure’),
- the probability of success, $p$, is constant across all trials and the probability of failure, $1 - p$, is also constant across all trials and
- the outcome of any trial is independent of the outcome of any other trial. If sampling is done without replacement from an infinite population or if sampling is with replacement from a finite population this helps to ensure independence.

The value of $p$ determines the skewness of the binomial distribution. For $p = 0.5$, the binomial distribution will be symmetrical, for $p < 0.5$ the distribution will be skewed to the left and for $p > 0.5$ the distribution will be skewed to the right.

The mean, or expected value, of the binomial distribution is given by:

$$
\mu = E(X) = np
$$

where $n$ is the number of trials and $p$ is the probability of success. The standard deviation of the binomial distribution is given by:

$$
\sigma = \sqrt{\sigma^2} = \sqrt{\text{var}(X)} = \sqrt{np(1 - p)}
$$

**Example 5–5**

An accountant works mainly with personal income tax cases. He knows from past experience that the probability of a customer being satisfied with the service he offers is 0.8. Given that he sees 8 clients today, determine the probability that:

(a) all 8 customers are satisfied

(b) at least 6 customers are satisfied

(c) fewer than 5 customers are satisfied

(d) What assumptions are necessary in (a) to (c)?

(e) What are the mean and standard deviation of the probability distribution?
Solution 5–5

We have \( n = 8 \) and \( p = 0.8 \). We can calculate probabilities by using:

- the binomial formula \( P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x} \),
- binomial tables (Table E.6 in the textbook) or
- Excel and PHStat2.

Here we’ll demonstrate using the formula and tables and leave it as an exercise for the reader to find the answers using Excel. Remember that you’ll need to be able to do these calculations using both the binomial formula and tables for the exam.

(a) \( P(X = 8) = \frac{8!}{8!} (0.8)^8 (1 - 0.8)^{8-8} = 1 (0.8)^8 (0.2)^0 = 0.1678 \) (to 4 decimal places).

So the probability that the accountant will have exactly 8 satisfied customers in the day is 0.1678.

Note in doing this calculation, \( \frac{8!}{8!} \) can be simplified to a value of one by cancelling the 8! on top and bottom lines and remembering that \( 0! = 1 \).

(b) To find \( P(X \geq 6) \) we will use tables since we would have to do several calculations using the binomial formula. Looking across the bottom of the binomial probability table until we find 0.8 and looking down the rows at the right to find the value of 8 for \( n \), we then find that:

\[
P(X \geq 6) = P(X = 6) + P(X = 7) + P(X = 8)
\]
\[
= 0.2936 + 0.3355 + 0.1678
\]
\[
= 0.7969
\]

So the probability that the accountant will have at least 6 satisfied customers in the day is 0.7969.

(c) To find \( P(X < 5) \) we will again use tables. Since we know that

\[
P(X < 5) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4)
\]
Discrete distributions

\[ P(X < 5) = 1 - P(X = 8) - P(X = 7) - P(X = 6) - P(X = 5) \]

It will be easier to find the second of these two probabilities (it requires less values to be calculated or looked up in the table).

\[ P(X < 5) = 1 - P(X = 8) - P(X = 7) - P(X = 6) - P(X = 5) \]
\[ = 1 - 0.1678 - 0.3355 - 0.2936 - 0.1468 \]
\[ = 0.0563 \]

The probability that the account will have less than 5 satisfied customers is 0.0563.

(d) There are four assumptions made:

1. there are a fixed number of outcomes. In this case, there are \( n = 8 \) trials, so this assumption is met.
2. there are only two mutually exclusive and exhaustives outcomes: satisfied or not satisfied. (Although it is conceivable that some customers might not feel they fit in to either of these categories, we are assuming that this is not the case and that there are only two outcomes.)
3. The probability of success is constant over all customers: \( p = 0.8 \) is the same for all customers.
4. the outcome for each customer is independent of all other customers. (Note: this might not strictly be true, particularly if customers come on the recommendations of others. Often we are forced to make assumptions which are probably true in order to be able to use particular statistical methods. This is one such case, but it is one we hope will lead to small errors in the probabilities calculated.)

Note: we would normally check these assumptions before calculating the probabilities to ensure the distribution is appropriate.

(e) The mean is

\[ \mu = np = 8(0.8) = 6.4 \]

and the standard deviation is

\[ \sigma = \sqrt{np(1-p)} = \sqrt{8(0.8)(0.2)} = 1.1314 \] (to 4 decimal places).

The Poisson distribution

The Poisson distribution describes situations where we are interested in the number of occurrences of a particular event in a certain period, for example, the number of mobile phones sold per week or the number of people calling a helpline per day.

The probability of a particular event for a Poisson random variable can be expressed using the following mathematical formula:
Discrete distributions

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

where \( \lambda \) (Greek letter ‘lambda’) is the expected number occurrences of the event per time period.

The Poisson distribution is always skewed to the right, however, the larger the value of \( \lambda \), the more symmetrical it becomes.

The mean and variance of the Poisson distribution are:

\[ \mu = E(X) = \lambda \]
\[ \sigma^2 = \text{var}(X) = \lambda \]

For the Poisson distribution to be appropriate, the random variable must have the following characteristics:

- interest is in counting the number of occurrences of an event within an interval (which could be time, distance, area, etc.)
- the probability of an event in a given interval is the same for all intervals
- the number of occurrences of an event in an interval is independent of the number of occurrences in any other interval and
- the probability of two or more occurrences of an event in an interval approaches zero as the size of the interval approaches zero.

**Example 5–6**

A new call centre is being set up for a large telephone company. Based on experience at their other similar centres, they expect that calls will come in at an average of 10 per minute. In order to plan their staffing requirements, the company would like to answer the following questions:

(a) What is the probability that less than five calls will be received in any given minute?

(b) What is the probability that at least twenty calls will be received in any given minute?

(c) What is the probability that between seven and thirteen calls will be received in any given minute?

**Solution 5–6**

We have \( \lambda = 10 \) and the time period being considered is minutes. We can calculate probabilities by using:

- the Poisson formula \( P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!} \),
- Poisson tables (Table E.7 in the textbook) or
- Excel and PHStat2.

Here we’ll demonstrate using the formula and tables and leave it as an exercise for the reader to find the answers using Excel.
(a) \[ P(X < 5) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) \]

Since there are so many values in this expression, it is more sensible to use tables than to calculate each one individually using the Poisson formula (but we demonstrate a couple of the calculations below). Using Table E.7, we find \( \lambda = 10 \) and then the appropriate rows for \( X = 0, 1, 2, \ldots \)

\[ P(X < 5) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) \]
\[ = 0.0000 + 0.0005 + 0.0023 + 0.0076 + 0.0189 \]
\[ = 0.0293 \]

The probability that the company will receive less than five calls in a given minute is 0.0293.

If we were to use the Poisson formula for say \( X = 2 \), the following working would result:

\[ P(X = 2) = \frac{e^{-10}10^2}{2!} = 0.0023 \text{ (to 4 decimal places)} \]

and for \( X = 4 \) would be:

\[ P(X = 4) = \frac{e^{-10}10^4}{4!} = 0.0189 \text{ (to 4 decimal places).} \]

(b) \[ P(X \geq 20) = P(X = 20) + P(X = 21) + P(X = 22) + P(X = 23) + P(X = 24) + \ldots \]
\[ = 0.0019 + 0.0009 + 0.0004 + 0.0002 + 0.0001 \]
\[ = 0.0035 \]

The probability that the company will receive at least twenty calls in a given minute is 0.0035. Note that \( P(X = 25) \), \( P(X = 26) \), etc. do exist, although the values are so small that they are zero to four decimal places (and so aren’t tabulated).

(c) \[ P(7 < X < 13) = P(X = 8) + P(X = 9) + P(X = 10) + P(X = 11) + P(X = 12) \]
\[ = 0.1126 + 0.1251 + 0.1251 + 0.1137 + 0.0948 \]
\[ = 0.5713 \]

The probability that the company will receive between seven and thirteen calls in a given minute is 0.5713.

### Discussion points

#### Discussion point 5–1

With other students, discuss and compare your answer to Problems 5.52 & 5.53 from p. 206 of your textbook (Problems 5.39 & 5.40 on p. 213 of the 4th edition). How should you go about determining whether these properties are appropriate for a particular example? What impact will your decision have on the statistical analysis you perform?
Discussion point 5–2

Examine and complete Problem 5.55 on p. 206 of your textbook (Problem 5.43 on p. 213 of the 4th edition). Compare your answers with other students, discuss and justify your decisions.

Summary

Now that you have completed this module, turn back to the objectives at the beginning of the module. Have you achieved these objectives?

Ensure that you attempt the recommended problems in the list of review questions below and at least a sample of problems from the optional list. This will help you to identify any areas of difficulty you have in achieving the module’s objectives.

Review questions

Recommended problems


Levine et al. 5th edition: Questions 5.2, 5.4, 5.6, 5.8, 5.14, 5.15, 5.18, 5.22, 5.28, 5.30, 5.34, 5.40, 5.51 to 5.53, 5.56, 5.70.

Optional problems

Levine et al. 4th edition: Choose a selection of problems from Questions 5.1, 5.3, 5.5, 5.7, 5.9, 5.11, 5.13, 5.15, 5.17 to 5.21, 5.29, 5.31, 5.33 to 5.35, 5.37, 5.43, 5.45 to 5.49 and 5.51 to 5.53.

Levine et al. 5th edition: Choose a selection of problems from Questions 5.1, 5.3, 5.5, 5.7, 5.9 to 5.13, 5.16, 5.17, 5.19 to 5.21, 5.23 to 5.27, 5.29, 5.31 to 5.33, 5.35 to 5.39, 5.41 to 5.43, 5.55 to 5.67, 5.70 and 5.71.
Continuous distributions
# Contents

- Introduction ............................................................................................................... 6–3
- Objectives .................................................................................................................. 6–3
- Readings .................................................................................................................... 6–3
- Continuous random variables .................................................................................... 6–4
- The uniform distribution ........................................................................................... 6–5
- The normal distribution ............................................................................................. 6–7
- Checking the assumption of normality ....................................................................... 6–12
- The exponential distribution ..................................................................................... 6–15
- Discussion points ...................................................................................................... 6–16
- Summary .................................................................................................................... 6–16
- Review questions ...................................................................................................... 6–17
Introduction

This week we will look at three distributions where the random variable can take on any value on a continuum. These are called continuous distributions. This allows the exploration of the distribution of objects or processes that involve continuous measurements, (such as height, weight, time, distance). Examples are the number of kilometres a car can travel per litre of petrol, the money a consumer spends on groceries per month, the amount of money a corporation invests in charity events, the value of GST a company must account for each year, the number of megabytes of hard-disk space employees of a company use to store computer files etc.

The first distribution we will examine this week is the normal distribution. This is often referred to as the ‘bell-shaped curve’ and describes a surprising number of variables quite well. Often it is used to model the distribution of student’s grades in an exam (hence the American expression ‘grading on the curve’), the heights of men or the heights of women (which follow separate normal distributions) and fill weights of grocery packaging (for example, cereal boxes), etc.

The second distribution is the uniform distribution. As its names suggests, this distribution takes on uniform values across its domain. In other words, every outcome is equally likely.

The final distribution we will look at this week is the exponential distribution. This is often used to describe such things as arrival times (for example, the arrival of people at an ATM).

Objectives

On completion of this module you should be able to:

- calculate areas under the standard normal curve
- solve and interpret problems involving the normal distribution
- check assumptions of normality
- calculate probabilities using the uniform distribution,
- solve and interpret problems involving the uniform distribution
- calculate probabilities using the exponential distribution and
- solve and interpret problems involving the exponential distribution.

Readings

Source

Textbook Levine et al. 4th edition Sections 6.1 to 6.4 & Excel Handbook Sections EH6.1 to 6.3 (pp. 276–278)
Continuous random variables

Random variables that can take on any value on a continuum are called continuous random variables. The height of people, for example, is regarded as a continuous random variable. Although measuring equipment (such as a tape measure in this case) only gives a certain degree of accuracy, the more accurate the measuring equipment, the more accurate the measure of a person’s height will be. Other examples include the current outdoor temperature, income, IQ (intelligence quotient) score, monthly amount a company spends on advertising, etc.

With discrete random variables, it made sense to associate a certain probability with each possible outcome of the random variable. This was called a discrete probability distribution. With continuous random variables, however, since there are an infinite number of possible values that the random variable might taken on, it makes no sense to attempt to associate a probability with each one! Instead, probabilities are associated with intervals on the continuum.

For example, imagine staff at a particular company are required to complete a short online survey about working conditions. The time taken for each staff member to complete the survey is recorded in seconds (including a number of decimal places based on parts of seconds). The graph below displays the probability density function of the times to complete the questionnaire. From this graph we can see that most employees took between about 60 and 180 seconds to complete the questionnaire. The probability that an employee takes between 140 and 160 seconds is shaded on the graph. This corresponds to a probability of 0.1359.

Given sufficient information about the distribution itself (i.e., the mathematical formula), we could also calculate the probabilities for any other intervals. The graph is actually a normal distribution with a mean of 120 seconds and a standard deviation of 20 seconds. As an exercise, when you have covered the material on the normal distribution that follows, return to this example and verify the probability displayed on the graph.
A mathematical function of the type pictured above is called a probability density function. A probability density function for a continuous random variable, \( X \), is a mathematical function for which the area under the curve corresponding to any interval is equal to the probability that \( X \) will take on a value in the interval. The probability density function is denoted by \( f(x) \), which indicates the probability density at \( x \).

This week we will examine three key continuous probability density functions: the uniform distribution, the normal distribution and the exponential distribution. Graphs of these distribution follow:

![Graphs of distributions](image)

**The uniform distribution**

The uniform distribution, or rectangular distribution, is so called because of its rectangular shape. It is described by two parameters, \( a \) and \( b \), where \( a \) is the smallest value that \( X \) can assume and \( b \) is the largest value that \( X \) can assume. The continuous uniform probability density function is:

\[
f(x) = \begin{cases} 
\frac{1}{b-a} & \text{for } a \leq x \leq b \\
0 & \text{elsewhere}
\end{cases}
\]

The value \( \frac{1}{b-a} \) gives the height of the rectangle. The mean, or expected value, of the uniform distribution is:

\[
\mu = E(X) = \frac{a+b}{2}
\]

and the standard deviation and variance are (respectively):

\[
\sigma(X) = \sqrt{\frac{(b-a)^2}{12}}
\]

\[
\sigma^2(X) = \frac{(b-a)^2}{12}
\]

**Example 6–1**

A surfer knows that the time between wipe-outs (falling off his surfboard) is uniformly distributed between two minutes and nine minutes in particularly large surf. What is the probability that the time between wipe-outs is:

(a) less than five minutes

(b) between three and four minutes

(c) more than six minutes
(d) What is the expected value of the time between wipe-outs?

(e) What is the standard deviation of the time between wipe-outs?

**Solution 6–1**

As always, it is helpful to produce a graph of the distribution when trying to estimate probability plots. Note that the probability of being anywhere under the rectangle (i.e., the probability of wiping out in between 2 and 9 minutes) is 1. The length of the rectangle is \( b - a \). The height must therefore be \( \frac{1}{b-a} = \frac{1}{9-2} = \frac{1}{7} \).

![Graph](image)

We find probabilities with a uniform distribution by finding areas of the appropriate rectangle (base times height).

(a) Note that since probabilities for values less than two minutes are zero, we have

\[
P(X < 5) = P(2 < X < 5)
\]

The shaded area in the graph below indicates the required probability region.

![Graph](image)

Now using the base times height formula, the desired probability is:

\[
P(X < 5) = P(2 < X < 5) = (5-2) \times \left( \frac{1}{9-2} \right) = \frac{3}{7}
\]

(b) \( P(3 < X < 4) = (4-3) \times \left( \frac{1}{9-2} \right) = \frac{1}{7} \)
The normal distribution

The normal distribution is one of the most important distributions in classical statistics, having applications in a variety of fields. In the introduction, we mentioned the following examples of variables that can be described by a normal distribution: student’s grades in an exam, the heights of men or the heights of women and fill weights of grocery packaging.

The normal distribution is described by two parameters: the mean ($\mu$) and standard deviation ($\sigma$). The normal random variable may take on any value between $-\infty$ and $+\infty$. Although most real-world examples have limits on the value the variable can take, the normal model is often still useful in these situations.

The normal probability density function is denoted by:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

where $\pi \approx 3.14159...$ and the exp notation means $\exp[a] = e^a$ where $e \approx 2.71828...$ is the base of natural logarithms. The normal distribution’s distinctive bell shape has its centre (and peak) at the mean, $\mu$. The parameter $\sigma$, the standard deviation, determines the spread of the distribution. Because a normal distribution is uniquely defined by its mean and standard deviation (or variance), the following notation is often used to
Continuous distributions denote a particular normal distribution: \( N(\mu, \sigma^2) \). Some examples of the normal distribution curves are given below.

The **standard normal distribution** is a particular example of the normal distribution with \( \mu = 0 \) and \( \sigma = 1 \). The standard normal random variable is denoted by \( Z \). Using the notation mentioned above, we can say that \( Z \) is distributed as \( N(0,1) \). The standardised form of a normal random variable \( X \) with mean \( \mu \) and standard deviation \( \sigma \) is:

\[
Z = \frac{X - \mu}{\sigma}
\]

This formula is useful to us, since standard normal probabilities tables can be used to find probabilities (see Appendix E.2 in the text). It means we are not forced to evaluate the (rather messy) normal probability density function each time we want to calculate a probability. We use a lower case \( z \) to indicate a particular outcome of the standard normal random variable \( Z \). The tables in the text used in this course give \( P(Z < z) \).

**Example 6–2**

Evaluate the following probabilities for the standard normal variable, \( Z \):

(a) \( P(Z \leq 1) \)

(b) \( P(Z \leq -2.76) \)

(c) \( P(Z \geq 2.49) \)
Solution 6–2

Check for yourself that you can locate the following probabilities in the cumulative standard normal tables in the text.

(a)  \( P(Z \leq 1) = 0.8413 \)

(b)  \( P(Z \leq -2.76) = 0.0029 \)

(c)  \( P(Z \geq 1.49) = 1 - P(Z < 1.49) = 1 - 0.9319 = 0.0681 \)

Remember that the total probability under the probability density function curve is 1.

Determining probabilities for any normal distribution

The standard normal distribution tables can be used to calculate probabilities for any normal random variable. This is done by transforming the normal random variable, \( X \), into the standard normal variable, \( Z \) using the linear transformation \( Z = \frac{X - \mu}{\sigma} \). We’ll demonstrate with an example.

Example 6–3

A final exam for a particular accountancy course is known by students to be a difficult one. In the past, the mean mark was 62% and the standard deviation was 11%. What proportion of students have received a mark of:

(a) at least 65%
(b) at least 50%
(c) less than 40%
(d) between 70% and 100%
(e) Between what two marks symmetrically distributed around the mean will 95% of the students’ marks fall?
Solution 6–3

We have $\mu = 62$ and $\sigma = 11$.

(a) The proportion of students who received a mark of at least 65% can be illustrated using the graph to the right.

We need to convert this to a standard normal distribution so that normal tables can be used to determine proportions (probabilities). We do this by:

$$Z = \frac{X - \mu}{\sigma} = \frac{65 - 62}{11} = 0.27 \text{ (to 2 dec. pl.)}.$$ 

This indicates that using the standard normal distribution we are looking for the shaded area on the graph (at right) of a standard normal distribution (which has $\mu = 0$ and $\sigma = 1$). Note that we have rounded the $Z$ value to two decimal places since the standard normal tables are tabulated for two decimal places.

Now using Table E.2 we have 

$$P(X > 65) = P(Z > 0.27) = 1 - P(Z < 0.27)$$

$$= 1 - 0.6064 = 0.3936$$

So 39.36% of the students will receive more than 65% on the exam.

Note that whenever we’re solving problems involving the normal distribution it is helpful to draw a quick sketch of the required probability area.

(b) $Z = \frac{X - \mu}{\sigma} = \frac{50 - 62}{11} = -1.09$

$P(X > 50) = P(Z > -1.09) = 1 - P(Z < -1.09)$

$$= 1 - 0.1379 = 0.8621$$

So 86.21% of the students will receive more than 50% on the exam.

(c) $Z = \frac{X - \mu}{\sigma} = \frac{40 - 62}{11} = -2$

$P(X < 40) = P(Z < -2) = 0.0228$

So 2.28% of the students will receive less than 40% on the exam.
Continuous distributions

(d) \[ Z_1 = \frac{X - \mu}{\sigma} = \frac{70 - 62}{11} = 0.73 \] and
\[ Z_2 = \frac{X - \mu}{\sigma} = \frac{100 - 62}{11} = 3.45 \]

\[ P(70 < X < 100) = P(0.73 < Z < 3.45) = 0.99972 - 0.7673 = 0.23242 \]

So 23.24% of students received between 70% and 100% on the exam.

(e) In finding the values which contain 95% of the distribution, we know that 47.5% or 0.475 of the area must be in each half. This means 0.5 - 0.475 = 0.025 must be in the tail. If we search through the centre of Table E.2, we find that at a Z-value of -1.96 has a probability of 0.025. Therefore for \( P(Z_1 < Z < Z_2) = 95\% \) to be true we must have \( Z_1 = -1.96 \) and \( Z_2 = +1.96 \).

Using \( Z = \frac{X - \mu}{\sigma} \) we discover that

\[ -1.96 = \frac{X - 62}{11} \]
\[ -1.96(11) = X - 62 \]
\[ X = 62 - 1.96(11) = 40.44 \]

and

\[ +1.96 = \frac{X - 62}{11} \]
\[ 1.96(11) = X - 62 \]
\[ X = 62 + 1.96(11) = 83.56 \]

So we know that 95% of students will receive between 40.44% and 83.56%.
Checking the assumption of normality

In order to be able to use the normal distribution in real-world examples, we must first test to see whether data can be approximated by the normal distribution. We do this by checking to see whether the characteristics of the data set match those of a normal distribution and by producing a normal probability plot.

The normal distribution has the following characteristics:

- it is symmetrical
- it is bell-shaped
- the mean, median and mode are equal
- the interquartile range equals 1.33 standard deviations.

We can test visually for symmetry and a bell-shape using a histogram (or stem-and-leaf plot for smaller data sets) and using a box-and-whisker plot. We obviously need to use numerical summaries to test for equality of mean, median and mode and to compare the interquartile range to 1.33 standard deviations. We shall demonstrate these in Example 6–4 below.

For a more thorough investigation, we might also examine other characteristics of a normal distribution. For example, we can show that

- approximately 68.26% of the values of a normal distribution fall within ±1 standard deviation of the mean
- approximately 95.44% of the values of a normal distribution fall within ±2 standard deviation of the mean
- approximately 99.73% of the values of a normal distribution fall within ±3 standard deviation of the mean.

The normal probability plot is another useful tool in determining the appropriateness of the normal distribution. In practice, normal probability plots are most often produced by a computer because they tend to be quite messy to produce by hand. However, we will explain how the graph is produced, since it is essential to understand what the software is doing!

Creating a normal probability plot:

- Rank the data from smallest to largest value, giving the smallest observation rank 1 through to the largest with rank $n$.
- For each rank, $i = 1, \ldots, n$, find the percentile of order $i/(n + 1)$. For example, if the sample size is $n = 50$, these values will be $1/(50 + 1) = 0.0196$, $2/(50 + 1) = 0.0392$, $3/(50 + 1) = 0.0588$, $\ldots$, $50/(50 + 1) = 0.9804$ (all to 4 decimal places).
- Find the standard normal distribution values, $Z_i$, which correspond to the cumulative area of these values. For example, $Z_{0.0196} = -2.06$ (found by looking for 0.0196 in the body of the standard normal table), $Z_{0.0392} = -1.76$ through to $Z_{0.9804} = 2.06$.
- Produce a scatterplot of the ordered data values against the standardised normal scores.
If the data are normally distributed, the normal probability plot should fall along a straight line, with the slope of the line given by $1/\sigma$. The graphs below indicate the approximate shapes we might expect for normally distributed data, left-skewed data, right-skewed data and rectangular-shaped distribution data.

**Example 6–4**

Last term, a group of 21 students enrolled in an accounting course on a particular campus. Their scores on the final exam are recorded below. Determine whether or not these marks are normally distributed by evaluating the actual versus theoretical properties and by constructing a normal probability plot.

<table>
<thead>
<tr>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>59</td>
</tr>
<tr>
<td>48</td>
</tr>
<tr>
<td>75</td>
</tr>
<tr>
<td>51</td>
</tr>
<tr>
<td>51</td>
</tr>
<tr>
<td>48</td>
</tr>
<tr>
<td>67</td>
</tr>
<tr>
<td>43</td>
</tr>
<tr>
<td>43</td>
</tr>
<tr>
<td>63</td>
</tr>
<tr>
<td>64</td>
</tr>
<tr>
<td>64</td>
</tr>
<tr>
<td>49</td>
</tr>
<tr>
<td>76</td>
</tr>
<tr>
<td>74</td>
</tr>
<tr>
<td>53</td>
</tr>
<tr>
<td>58</td>
</tr>
<tr>
<td>71</td>
</tr>
<tr>
<td>44</td>
</tr>
<tr>
<td>72</td>
</tr>
<tr>
<td>62</td>
</tr>
</tbody>
</table>

**Solution 6–4**

Recall that a normal distribution is symmetrical and bell-shaped (the mean, median and mode are all equal). The interquartile range is equal to 1.33 standard deviations. The distribution is continuous and has an infinite range. It is these characteristics that we will look for in the data set to determine whether the actual properties reflect the theoretical properties.

We begin by producing a box-and-whisker plot. This requires firstly that we produce the five number summary of median, upper and lower quartiles and minimum and maximum values:
Continuous distributions

<table>
<thead>
<tr>
<th>Five-number Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum 43</td>
</tr>
<tr>
<td>First Quartile 48.5</td>
</tr>
<tr>
<td>Median 59</td>
</tr>
<tr>
<td>Third Quartile 69</td>
</tr>
<tr>
<td>Maximum 76</td>
</tr>
</tbody>
</table>

Box plot of exam marks

We can see that this data set appears to be symmetric. The upper and lower quartiles may be a little too large and too small respectively to fit the normal distribution model (the interquartile range should be 1.33 standard deviations). We also check that the mean and the median are approximately equal (although since the data set is small, the mode will not necessarily be a good measure of the central tendency):

| Mean 58.80952 | Std. Deviation 11.10234 |

Clearly the mean and the median are very close and so this particular characteristic of the normal distribution is found in the sample data. A normal probability plot (below) reveals that the data might approximately follow a straight line, although it could perhaps be better described as reflecting the rectangular-shaped distribution. The positioning of the upper and lower quartiles mentioned above, might also be evidence of a more rectangular shaped distribution.

Now we will compare the interquartile range to 1.33 standard deviations. The interquartile range is:

$$IQR = UQ - LQ = 69 - 48.5 = 20.5$$
The standard deviation is 11.10234, so 1.33 standard deviations are:

\[ 1.33s = 1.33(11.10234) = 14.7661 \text{ (4 dec. pl.)} \]

Since 20.5 is quite different from 14.7661, this characteristic of the normal distribution is not reflected by this data set.

We can conclude that the data is probably not well described by the normal distribution. Since the data set is small, however, it is difficult to draw a completely reliable conclusion. With a larger data set, it might be possible to make a clear decision about whether this data would be best described by a rectangular or normal distribution.

The exponential distribution

The exponential distribution has a distinct right-skewed shape which is useful in describing duration phenomena. For example, the random variable \( X \) might represent the time until a caller waiting in a queue has their call answered or the time until the arrival of the next customer at a particular ATM. The exponential random variable may take on any positive value: \( 0 < x < \infty \). The exponential distribution has one parameter, \( \lambda \), which represents the mean number of arrivals per unit. This parameter is always positive: \( \lambda > 0 \). The exponential cumulative probability density function is:

\[ P(X < x) = 1 - e^{-\lambda x} \]

As an example, imagine we know that on average three customers call a customer helpline each minute. Therefore we have \( \lambda = 3 \). Then, if we wished to find the probability that the next call occurred within 15 seconds (i.e., 0.25 minutes), we would be looking for \( P(X < 0.25) = 1 - e^{-3(0.25)} = 0.5276 \) (to 4 decimal places).
Example 6–5

People are known to arrive at a particular vending machine at a mean rate of 27 per hour. Assuming that these arrival times follow an exponential distribution, find the probability that the next person will arrive:

(a) within one minute
(b) within five minutes
(c) in more than five minutes

Solution 6–5

We have $\lambda = 27$ and $P(X < x) = 1 - e^{-\lambda x} = 1 - e^{-27x}$. Notice that the arrival rate is expressed per hour (so $X$, the arrival time, is measured in hours), but the questions are posed as per minute, so we must convert the minute amounts to fractions of hours.

(a) $P(\text{arrival time} < 1 \text{ minute}) = 1 - e^{-\frac{27}{60}} = 0.3624$ (to 4 decimal places).

(b) $P(\text{arrival time} < 5 \text{ minutes}) = 1 - e^{-\frac{27}{30}} = 0.8946$

(c) $P(\text{arrival time} > 5 \text{ minutes}) = 1 - P(\text{arrival time} < 5 \text{ minutes})$
   $= 1 - 0.8946 = 0.1054$

Discussion points

Discussion point 6–1

With other students, discuss and compare your answer to Problem 6.38 on p. 244 of your textbook (Problem 6.61 on p. 270 of the 4th edition).

Discussion point 6–2

A friend comes to you and makes the statement that ‘anything we measure in the real world is effectively continuous and can be modelled by a normal distribution’. How would you respond? With other students, discuss, compare and justify your answer.

Summary

Now that you have completed this module, turn back to the objectives at the beginning of the module. Have you achieved these objectives?

Ensure that you attempt the recommended problems in the list of review questions below and at least a sample of problems from the optional list. This will help you to identify any areas of difficulty you have in achieving the module’s objectives.
Review questions

Recommended problems

Levine et al. 4th edition: Do Questions 6.4, 6.6, 6.8, 6.10, 6.12, 6.14, 6.16, 6.18, 6.20, 6.24, 6.28, 6.30, 6.32, 6.36 and 6.61 to 6.67.

Levine et al. 5th edition: Do Questions 6.4, 6.6, 6.8, 6.10, 6.12, 6.14, 6.18, 6.24, 6.26, 6.28, 6.32, 6.34 and 6.38 to 6.44.

Optional problems

Levine et al. 4th edition: Choose a selection of problems from Questions 6.1 to 6.3, 6.5, 6.7, 6.9, 6.11, 6.13, 6.15, 6.17, 6.19, 6.21 to 6.23, 6.25 to 6.27, 6.29, 6.31, 6.33 to 6.35 and 6.37.

Sampling distributions
Contents

Introduction .................................................................................................................................. 7–3
Objectives .................................................................................................................................. 7–3
Readings .................................................................................................................................. 7–4
Sampling distributions ............................................................................................................ 7–4
Sampling distribution of the mean .......................................................................................... 7–4
Sampling distribution of the proportion .................................................................................. 7–9
Normal approximation to the binomial and Poisson distributions ........................................... 7–10
Sampling from finite populations ........................................................................................... 7–14
Discussion points ..................................................................................................................... 7–16
Summary .................................................................................................................................. 7–16
Review questions ....................................................................................................................... 7–16
Introduction

When we draw a sample from a population, we use the sample data to estimate a population parameter. For example, we calculate the sample mean (\( \bar{X} \)), standard deviation (\( S \)) and proportion (\( \rho \)) to estimate the population mean (\( \mu \)), standard deviation (\( \sigma \)) and proportion (\( \pi \)), respectively. Because we’ve taken only a sample and not examined the whole population, the estimated values will not be exactly the same as the parameters; there will be a sample error. This week we begin by exploring the measurement of the error of these estimates using the standard error.

Imagine that several different random samples are drawn from the same population. If the mean of each sample was calculated, these means would be slightly (if not very) different. We would hope that they were accurate estimates of the true population mean, but how accurate would depend on a number of factors. The distribution of the estimates of the parameters is called the sampling distribution. This week we explore the sampling distribution of sample means and proportions. We will see what effect our knowledge of the population distribution and the sample size has on our understanding of these sampling distributions. These concepts are an important stepping stone to next week’s material, so they should be learned well.

Other topics covered this week include a study of how, in certain circumstances, we can use the normal distribution to approximate the binomial and Poisson distributions and an investigation of population size (rather than sample size) issues. This latter topic considers the differences that result when we have finite or infinite populations. Although often our population is large enough that we can assume it is infinite (such as the number of people on the planet—although we can quantify this, our sample sizes are usually very small relative to this number!). At other times, the population is clearly finite (such as the number of students in a classroom or the number of staff on the payroll of a small firm). It turns out that when we have a finite population, we need to make a correction to our standard error estimate.

Objectives

On completion of this module you should be able to:

- calculate the standard error of the mean and explain the effect of sample size on the standard error
- explain the concept of a sampling distribution for samples taken from either normal or non-normal populations and understand the central limit theorem
- calculate the standard error of the proportion
- calculate probabilities relating to sample means and proportions
- use the normal approximation to the binomial and Poisson distributions and
- understand and apply sampling techniques for finite populations.
Readings

Source

Textbook
Levine et al. 4th edition
Sections 6.5 to 6.9 & Excel Handbook Sections EH6.4 to 6.5 (pp. 279–280)

Note that Sections 6.8 and 6.9 are on the CD which accompanies the 4th edition text.

Or

Levine et al. 5th edition
Sections 6.6, 7.3 to 7.6 & Excel Companion E7.2 (pp. 281–282)

Note that Sections 6.6 and 7.6 are on the CD which accompanies the 5th edition text.

Course website
Visit the course website for links to any supplementary material for this week.

Sampling distributions

As we said in the introduction, we use sample statistics to estimate population parameters. There are always small differences in the estimates between one sample and the next. Clearly, the larger the sample size (relative to the population size), the more likely it is that the sample is an accurate reflection of the population. The problem is, large samples are not always practical (because of costs in time, money, etc.). Because we usually can’t afford to conduct a census (i.e., we can’t afford to sample every element in the population), we need a way of measuring how much variation there is between sample estimates of parameters. We can do this using the sampling error.

Sampling error is the difference between the result obtained from a sample (such as an estimate of the mean, standard deviation or proportion) and the result that would be obtained from a census.

Sampling error is different from many of the kinds of errors discussed in Week 1 (such as coverage error, measurement error and nonresponse error) because it is due to the randomness of sampling, not due to faulty procedure or lack of response.

Sampling distribution of the mean

We will demonstrate sampling error of the mean using a simple example. Four components of a coffee machine are tested and the number of faults found with each component is recorded (see the table below). Note that this is actually a population since we are interested in only these four components. Therefore, we know that the population has size $N = 4$. 

<table>
<thead>
<tr>
<th>Component</th>
<th>Faults</th>
</tr>
</thead>
<tbody>
<tr>
<td>Component 1</td>
<td>2</td>
</tr>
<tr>
<td>Component 2</td>
<td>3</td>
</tr>
<tr>
<td>Component 3</td>
<td>1</td>
</tr>
<tr>
<td>Component 4</td>
<td>4</td>
</tr>
</tbody>
</table>


The mean number of faults for the components of the coffee machine is:

\[ \mu = \frac{5 + 3 + 6 + 2}{4} = 4 \]

We can illustrate the shape of the distribution graphically as follows:

![Diagram showing distribution](null)

The mean of this particular distribution sits exactly in the centre. Now imagine that we take samples of size 2 from this population (so \( n = 2 \)). Note that there are \( ^{4}C_2 = 4! / 2! \times 2! = 6 \) possible samples. The table below lists all these samples and their sample means.

<table>
<thead>
<tr>
<th>Possible samples</th>
<th>Sample mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>A, B</td>
<td>( \bar{X} = \frac{5 + 3}{2} = 4 )</td>
</tr>
<tr>
<td>A, C</td>
<td>( \bar{X} = \frac{5 + 6}{2} = 5.5 )</td>
</tr>
<tr>
<td>A, D</td>
<td>( \bar{X} = \frac{5 + 2}{2} = 3.5 )</td>
</tr>
<tr>
<td>B, C</td>
<td>( \bar{X} = \frac{3 + 6}{2} = 4.5 )</td>
</tr>
<tr>
<td>B, D</td>
<td>( \bar{X} = \frac{3 + 2}{2} = 2.5 )</td>
</tr>
<tr>
<td>C, D</td>
<td>( \bar{X} = \frac{6 + 2}{2} = 4 )</td>
</tr>
</tbody>
</table>

The distribution of the sample means is graphed below. Note that it is centred on the population mean, \( \mu \).
The mean of the sample means is \( \mu_\bar{X} = \frac{2.5 + 3.5 + 4 + 4.5 + 5.5}{6} = 4 \) which corresponds exactly to the population mean.

Notice that \( \mu_\bar{X} \), the mean of the sample means, is equal to \( \mu \), the population mean.

The arithmetic mean, \( \bar{X} \), is an unbiased estimator of the population mean. Although we can’t be sure that the mean of a particular sample is identical to the population mean, we can be sure that the average of all possible sample means is equal to the population mean.

The sampling distribution of the sample mean, \( \bar{X} \), is the probability distribution associated with the random variable \( \bar{X} \). For any population and sample size (using simple random sampling), the mean of the sampling distribution of the mean is equal to the population mean. In other words,

\[
\mu_\bar{X} = \mu
\]

The variance of the sampling distribution of the mean is

\[
\sigma^2_\bar{X} = \frac{\sigma^2}{n}
\]

and the standard deviation of the sampling distribution, which is referred to as the standard error of the mean, is

\[
\sigma_\bar{X} = \frac{\sigma}{\sqrt{n}}
\]

Looking at this value, we see that the larger the sample size, the smaller this standard error will be. This makes intuitive sense, since we understand that bigger sample sizes are more likely to give a better estimate of population parameters. Since this standard error decreases proportionately with the square root of the sample size, it becomes increasingly difficult to reduce the standard error by increasing \( n \). For example, if the standard error based on a sample of 100 is to be halved, then the sample size must be quadrupled to 400.

The standard error formula also tells us that for a given sample size, the larger the population variance, the larger the variability of the sampling distribution of the mean. Therefore, when the population variation is high, the sample mean will tend to vary more around the population mean than when it is lower.

**Central Limit Theorem**

For almost any population, the sampling distribution of \( \bar{X} \) is approximately normal when the simple random sample size is sufficiently large.
Just what ‘sufficiently large’ really means depends on the nature of the population and on how close an approximation to the normal distribution is required. In general, if the population is skewed, a larger random sample size is required for the sampling distribution of the mean to be approximately normal. A good rule of thumb for a sufficiently large sample is a sample of size of at least 30. If the population is known to be approximately bell-shaped then a smaller sample size will suffice. If the population is normally distributed, no matter what sample size is used, the sampling distribution will also be normal. If the population is relatively symmetrical, samples as small as 15 may be sufficient for the sampling distribution to be approximately normal.

The central limit theorem can be used to make probability statements about the sample mean when the sample size is sufficiently large. As before, the standard normal table is used and this requires the following standardised variable:

$$Z = \frac{\bar{X} - \mu}{\sigma \sqrt{n}} = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

**Example 7–1**

The distribution of times it takes an office worker to complete a particular task is known to have a mean of eight minutes and a standard deviation of two minutes. If random samples of forty tasks are taken, find:

(a) the probability that the average time spent per task will be more than nine minutes

(b) the proportion of sample means that will be between 7.2 and 8.5 minutes

(c) If the random sample had been of only 20 tasks, what changes would this make to your answers in (a) and (b)? What assumptions would you need to make in order to be able to answer (a) and (b) based on a sample of 20 tasks?

(d) Which of the following is more likely to occur: a sample mean below 7.5 minutes in a sample of 30 tasks, a sample mean below 7.5 minutes in a sample of 50 tasks, an individual task taking less than two minutes?

**Solution 7–1**

We are given $\mu = 8$ and $\sigma = 2$.

(a) We are looking for $P(\bar{X} > 9)$. The sample size is large enough that we can make use of the central limit theorem. Therefore

$$Z = \frac{\bar{X} - \mu}{\sigma \sqrt{n}} = \frac{9 - 8}{2 / \sqrt{40}} = 3.16$$

and so

$$P(\bar{X} > 9) = P(Z > 3.16) = 1 - 0.99921 = 0.00079$$

So the probability that the average time spent per task will be more than nine minutes is 0.00079.
(b) We require \( P(7.2 < \bar{X} < 8.5) \).

\[
Z_1 = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{7.2 - 8}{2/\sqrt{40}} = -2.53 \quad \text{and} \quad Z_2 = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{8.5 - 8}{2/\sqrt{40}} = 1.58
\]

\[
P(7.2 < \bar{X} < 8.5) = P(-2.53 < Z < 1.58) = 0.9429 - 0.0057 = 0.9372
\]

So the proportion of sample means that will be between 7.2 and 8.5 minutes is 0.9372.

(c) Clearly reducing the population to 20 takes us below the value of 30 that the Central Limit Theorem specifies as the required sample size. This means that unless the population is normally distributed, or at least relatively symmetrical, we cannot assume the means are normally distributed. Although many situations which involve human performance result in data that is approximately normally distributed, we are not really given enough information here to assume that this is the case. So because we are not told that the distribution is symmetrical or normal, we must assume that, for a sample size of 20, it would not be valid to use the normal distribution to solve the problems in (a) and (b).

(d) According to the central limit theorem we can calculate the probably of a sample mean being below 7.5 minutes in a sample of 30 tasks via:

\[
Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{7.5 - 8}{2/\sqrt{30}} = -1.37
\]

\[
P(\bar{X} < 7.5) = P(Z < -1.37) = 0.0853
\]

The probably of a sample mean being below 7.5 minutes in a sample of 50 tasks via:

\[
Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{7.5 - 8}{2/\sqrt{50}} = -1.77
\]

\[
P(\bar{X} < 7.5) = P(Z < -1.77) = 0.0384
\]

The probability that an individual task will be less than two minutes can be calculated via:

\[
Z = \frac{X - \mu}{\sigma} = \frac{2 - 8}{2} = -3
\]

\[
P(X < 2) = P(Z < -3) = 0.00135
\]

Therefore, the most likely outcome of the three is a sample mean below 7.5 minutes in a sample of 30.
Sampling distribution of the proportion

Just as we saw when examining the sample mean, when we estimate the proportion of a certain characteristic, this estimate will differ from sample to sample. The sample proportion is defined as:

\[ p = \frac{X}{n} = \frac{\text{number of occurrences}}{\text{sample size}} \]

This is an unbiased estimator of the population proportion, \( \pi \). The sampling distribution of the proportion therefore has expectation \( E(p) = \pi \), variance

\[ \sigma_p^2 = \frac{\pi(1-\pi)}{n} \]

and the standard deviation of the sampling distribution of the proportion, called the standard error of the proportion, is

\[ \sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} \]

The sampling distribution of the proportion is approximately normal when the sample size is sufficiently large. A rule of thumb that indicates an adequate sample size, and hence that the normal approximation is adequate, is:

\[ n\pi \geq 5 \text{ and } n(1-\pi) \geq 5 \]

As before, the standard normal table is used and this requires the following standardised variable:

\[ Z = \frac{p - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \]

Example 7–2

Recent research has indicated a growing number of young children from two parent families are being placed in child care so that both parents can work. Although the families increase their income with two pay packets, the cost of child care is often prohibitively high. A particular study indicated that 40% of families have children in child care facilities. If a random sample of 100 two-parent families is selected, find:

(a) the proportion of samples which will have between 40% and 50% of families using child care facilities

(b) the probability of obtaining a sample percentage of greater than 45%

(c) Within what symmetrical limits of the population percentage will 95% of the sample percentages fall?

Solution 7–2

(a) Since \( n\pi = 100(0.4) = 40 \geq 5 \) and \( n(1-\pi) = 100(0.6) = 60 \geq 5 \) the sample size is large enough to assume the sampling distribution of the proportion is approximately normally distributed. We are looking for \( P(0.4 < p < 0.5) \).
**Sampling distributions**

1. \[ Z_1 = \frac{p - \pi}{\sqrt{\frac{\pi (1 - \pi)}{n}}} = \frac{0.4 - 0.4}{\sqrt{\frac{0.4(1-0.4)}{100}}} = 0 \]

2. \[ Z_2 = \frac{p - \pi}{\sqrt{\frac{\pi (1 - \pi)}{n}}} = \frac{0.5 - 0.4}{\sqrt{\frac{0.4(1-0.4)}{100}}} = 2.04 \]

\[ P(0.4 < p < 0.5) = P(0 < Z < 2.04) = 0.9793 - 0.5 = 0.4793 \]

So the proportion of sample between 40% and 50% will be 0.4793.

(b) \[ Z = \frac{p - \pi}{\sqrt{\frac{\pi (1 - \pi)}{n}}} = \frac{0.45 - 0.4}{\sqrt{\frac{0.4(1-0.4)}{100}}} = 1.02 \]

\[ P(p > 0.45) = P(Z > 1.02) = 1 - 0.8461 = 0.1539 \]

So the probability of obtaining a sample percentage of greater than 45% is 0.1539.

(c) We already know that 95% of the standard normal graph is between \(\pm 1.96\) and so \(P(-1.96 < Z < 1.96) = 0.95\).

Using \[ Z = \frac{p - \pi}{\sqrt{\frac{\pi (1 - \pi)}{n}}} \] and rearranging we find that \[ p = \pi + Z \sqrt{\frac{\pi (1 - \pi)}{n}}. \]

Substituting in the two values of \(Z\), we find that:

\[ 0.4 - 1.96 \left( \frac{0.4(0.6)}{100} \right) = 0.3040 \] (to 4 decimal places) and

\[ 0.4 + 1.96 \left( \frac{0.4(0.6)}{100} \right) = 0.4960 \] (to 4 decimal places).

So 95% of the sample percentages will fall between 30.40% and 49.60%.

**Normal approximation to the binomial and Poisson distributions**

**The continuity correction**

When we use the normal approximation to the binomial or Poisson distributions, we are approximating discrete distributions with a continuous distribution. The graph below indicates how this can create differences. The histogram represents a binomial distribution and the solid line a normal distribution. The binomial distribution can only take on particular values which are represented by the bars. These indicate a certain frequency (given by the bar heights) for any \(x\)-value (at the centre of each bar).
The normal distribution, however, is continuous and so can take on any value (which is why it is drawn as a line). The normal distribution is being used to approximate the binomial, but it will never be exact because there is a difference between the discrete and the continuous. Obviously the more bars in the binomial graph (i.e., the larger the data set) the better the normal distribution will approximate the binomial. A similar situation occurs in approximating the Poisson distribution with a normal curve.

In Example 7–3 below, we will approximate $P(X \geq 3)$ on the binomial distribution using the normal distribution. This probability is indicated by the shaded bars in the graph below. Because the probability we are looking for includes everything greater than or equal to 3, the entire bar centred on the value 3 must be included. It can be seen on the continuous (normal) curve that this requires we evaluate from 2.5 upwards and hence we will find $P(X \geq 2.5)$ on the normal curve.

When we approximate a probability such as $P(X \leq 5)$, we would have the opposite case. Again making a continuity correction to include the entire bar at 5, on the normal curve, we would find $P(X \leq 5.5)$.

We use $X_a$ to represent the adjusted number of successes for the discrete random variable $X$. Depending on the probability area required, $X_a$ will be $X \pm 0.5$.
Normal approximation to the binomial distribution

In Week 5, we noted that the binomial distribution is symmetrical when $p = 0.5$. The closer $p$ is to 0.5, the closer the distribution will be to symmetrical. Similarly, the larger the sample size, the more symmetrical the distribution will become. Clearly with more observations, the calculations required to obtain binomial probabilities become much more involved and often tables do not exist for larger sample sizes. Thus, it may be sensible to use the normal approximation to the binomial under certain conditions. Generally we require that $np \geq 5$ and $n(1-p) \geq 5$ in order to use the normal approximation to the binomial distribution.

As before, when solving problems using the normal distribution, we use the transformation formula

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

For the binomial distribution (see the Week 5 material), the mean is

$$\mu = np$$

and the standard deviation is

$$\sigma = \sqrt{np(1-p)}$$

Substituting these into the transformation formula gives $Z = \frac{X - np}{\sqrt{np(1-p)}}$. We also have to include the continuity correction, so this formula becomes:

$$Z = \frac{X - np}{\sqrt{np(1-p)}}$$

Example 7–3

A company offers its sales staff a choice of three salary packages. Package A includes a base salary of $50,000 per year as well as 1% commission on all sales made by the staff member. Package B includes a base salary of $20,000 plus a 4% sales commission and package C consists solely of a 7% sales commission. The company has designed the packages in such a way as to expect equal numbers of staff to choose each option.

(a) If a random selection of six sales staff is taken, what is the probability that at least three will select package C?

(b) If a random selection of twenty sales staff is taken, what is the approximate probability that at least three will select package C?

Solution 7–3

(a) The situation follows a binomial distribution because: there are a fixed number of observations (six sales staff), each observation has two outcomes (select package C or not—we’re interested only in whether they select package C or not, not whether they select A or B when they didn’t select C), the probability of success is constant (here success is choosing package C with $p = 1/3$) and we assume the outcomes of the six sales staff are independent of each other.
So, given that we know this is a binomial distribution situation, we have $n = 6$ and $p = \frac{1}{3}$ and want to quantify $P(X \geq 3)$.

Given that $np = 6\left(\frac{1}{3}\right) = 2 \not\geq 5$ and $n(1 - p) = 6\left(\frac{2}{3}\right) = 4 \geq 5$, we cannot use the normal approximation (and so must use the binomial distribution).

Using the binomial formula (because the $p$ value is not included in the tables in the textbook), $P(X) = \frac{n!}{X!(n-X)!} p^X (1-p)^{n-X}$, we find:

$$P(X \geq 3) = 1 - P(X = 2) - P(X = 1) - P(X = 0) = 1 - \frac{6!}{2!4!} \left(\frac{1}{3}\right)^2 \left(\frac{2}{3}\right)^4 - \frac{6!}{1!5!} \left(\frac{1}{3}\right)^3 \left(\frac{2}{3}\right)^5 - \frac{6!}{0!6!} \left(\frac{1}{3}\right)^6 \left(\frac{2}{3}\right)^6$$

$$= 1 - 0.3292 - 0.2634 - 0.0878$$

$$= 0.3196$$

Given a random selection of six sales staff, the probability that at least three will select salary package C is 0.3196.

(b) Since $np = 20\left(\frac{1}{3}\right) = 6 \geq 5$ and $n(1 - p) = 20\left(\frac{2}{3}\right) = 13 \geq 5$ we can use the normal approximation to the binomial distribution. We find that

$$Z = \frac{X - np}{\sqrt{np(1-p)}} = \frac{2.5 - 20\left(\frac{1}{3}\right)}{\sqrt{20\left(\frac{1}{3}\right)\left(\frac{2}{3}\right)}} = -1.98$$

$$P(Z \geq -1.98) = 1 - 0.0239 = 0.9761$$

Given a random selection of twenty employees, the approximate probability that at least three will select salary package C is 0.9761.

**Normal approximation to the Poisson distribution**

With the Poisson distribution, we found (in Week 5) that the mean was

$$\mu = \lambda$$

and the variance was $\sigma^2 = \lambda$. The standard deviation is therefore

$$\sigma = \sqrt{\lambda}.$$

In order for the normal approximation to the Poisson distribution to be appropriate, we require that the expected number of successes be greater than or equal to five:

$$\lambda \geq 5$$

Substituting the mean and standard deviation into the normal transformation formula

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

and including the continuity correction, we have
\[ Z = \frac{X_a - \lambda}{\sqrt{\lambda}} \]

**Example 7–4**

Customers arrive at a busy takeaway coffee counter at the rate of five per minute.

(a) What is the probability that in any given minute three or fewer customers arrive?

(b) What is the approximate probability that in any given minute three or fewer customers arrive?

(c) Compare your answers from (a) and (b).

**Solution 7–4**

(a) This data can be described by a Poisson distribution. Given \( \lambda = 5 \),

\[
P(X \leq 3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)
\]

\[
= 0.0067 + 0.0337 + 0.0842 + 0.1404
\]

\[
= 0.2650
\]

(Using Poisson probability tables).

(b) Since \( \lambda \geq 5 \), we can use the normal approximation to the Poisson distribution.

\[
Z = \frac{X_a - \lambda}{\sqrt{\lambda}} = \frac{3.5 - 5}{\sqrt{5}} = -0.67
\]

\[
P(X \leq 3) = P(Z \leq -0.67) = 0.2514
\]

(c) There is a difference of 0.0136 between the two values. We can conclude that the normal distribution has been reasonably accurate in approximating the Poisson distribution.

**Sampling from finite populations**

Until now we have assumed that the sample size is small relative to the size of the population (when the population is finite). In practice, this is not always the case. We measure the relative size of the sample to the population via the fraction \( n/N \).

Generally when \( n/N \) is not small, or when \( n/N > 0.05 \), we consider the sample size to be larger and must account for this in our analysis. In other words, we can treat a finite population as infinite when \( n/N \leq 0.05 \), but must make an adjustment to our standard error for larger samples (finite populations) when \( n/N > 0.05 \).

The finite population correction factor (fpc) corrects the standard error to give the proper estimate when the population is finite. It is given by

\[
fpc = \sqrt{\frac{N-n}{N-1}}
\]
We can make a couple of comments about this finite population correction factor:

- When the population size, \( N \), is large relative to the sample size, \( n \), the \( fpc \) is close to 1. This is why we can treat a finite population as infinite when the sample is small relative to the population and is why the correction factor has not been necessary in the material covered until now.

- As \( n \) becomes large and close to \( N \), the \( fpc \) approaches zero. This makes intuitive sense, since the closer the sample size is to the population size, the less sampling error we expect to be present.

Using the finite population correction factor, the **standard error of the mean for a finite population** is

\[
\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}
\]

and the **standard error of the proportion for a finite population** is

\[
\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} \sqrt{\frac{N-n}{N-1}}
\]

**Example 7–5**

The management team of a large company has been investigating the work habits of the employees of the company. They have become concerned that some of the employees are spending large portions of the working day outside having smoking breaks. It is known that 500 employees are regular smokers. It is expected that the time these employees spend smoking per day is normally distributed with a mean of 25 minutes and a standard deviation of eight minutes. If a random sample of 50 of the smokers is selected without replacement, what proportion of the sample means would be greater than 26 minutes?

**Solution 7–5**

We have \( \mu = 25 \) and \( \sigma = 8 \). Since \( n/N = 50/500 = 0.1 > 0.05 \) and the sample has been selected without replacement, we need to use the finite population correction in our calculations.

\[
\mu_{\bar{X}} = \mu = 25
\]

\[
\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} = \frac{8}{\sqrt{50}} \sqrt{\frac{500-50}{500-1}} = 1.0744 \text{ (to 4 decimal places)}
\]

\[
Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{26 - 25}{1.0744} = 0.93
\]

\[
P(\bar{X} > 26) = P(Z > 0.93) = 1 - 0.8238 = 0.1762
\]

So 0.1762 (or 17.62%) of the sample means can be expected to be greater than 26 minutes.
Discussion points

Discussion point 7–1

With other students, discuss and compare your answers to Problems 7.41, 7.42 and 7.43 on p. 276 of the textbook (Problems 6.70, 6.71, 6.72 and 6.73 on p. 271 of the 4th edition).

Discussion point 7–2

Discuss the reason the continuity correction is required. In what situations can the normal distribution be used to approximate the binomial or Poisson distributions? When might this information be useful in practice? Give a couple of examples for each and discuss the implications in terms of time required to do calculations, simplicity of calculations, value for labour costs, accuracy of results, etc.

Summary

Now that you have completed this module, turn back to the objectives at the beginning of the module. Have you achieved these objectives?

Ensure that you attempt the recommended problems in the list of review questions below and at least a sample of problems from the optional list. This will help you to identify any areas of difficulty you have in achieving the module’s objectives.

Review questions

Recommended problems

Levine et al. 4th edition: Do Questions 6.38, 6.40, 6.42, 6.44, 6.50, 6.52, 6.54, 6.60, 6.68 to 6.74 and 6.80 from the textbook and from the CD 6.89 to 6.91, 6.94, 6.96, 6.100 to 6.102 and 6.104.

Levine et al. 5th edition: Do Questions 7.18, 7.20, 7.24, 7.28, 7.30, 7.39 to 7.43 and 7.52 from the textbook and from the CD 6.60, 6.64, 7.66, 7.68 and 7.70.
Problem 7–1: Normal approximation to Poisson problem


The number of cars arriving per minute at a toll booth on a particular bridge is Poisson distributed with a mean of 2.5. What is the probability that in any given minute

(a) no cars arrive?
(b) not more than two cars arrive?

If the expected number of cars arriving at the toll booth per ten-minute interval is 25.0, what is the approximate probability that in any given ten-minute period

(c) not more than 20 cars arrive?
(d) between 20 and 30 cars arrive?

Solution 7–1

(a) 0.0821  (b) 0.5438  (c) 0.1841  (d) 0.6803

Optional problems


Levine et al. 5th edition: Choose a selection of problems from Questions 7.17, 7.19, 7.21 to 7.23, 7.25 to 7.27, 7.29, 7.31 to 7.38, 7.50, 7.51, 7.53 to 7.55 from the textbook and from the CD 6.61 to 6.63, 7.67, 7.69, 7.71 and 7.72.
Confidence interval estimation
Contents

Introduction ............................................................................................................... 8–3
Objectives .................................................................................................................. 8–3
Readings .................................................................................................................... 8–3
Confidence intervals for the mean ............................................................................. 8–4
Confidence intervals for the proportion ................................................................. 8–9
Determining sample size .......................................................................................... 8–11
Confidence intervals for finite populations ............................................................ 8–13
Confidence intervals and auditing ........................................................................... 8–13
Discussion points ..................................................................................................... 8–14
Summary ................................................................................................................. 8–14
Review questions ..................................................................................................... 8–14
Introduction

Until now, when we’ve estimated a population parameter we have used only point estimates. These are single value estimates, and as such will quite often miss the mark. This week we will look at estimation using a **confidence interval**. This is an interval within which we are confident (at some given level) that the true population parameter value will fall. We will call on our knowledge of the standard error of our estimates (for means and proportions) in order to define the endpoints of the interval.

A 95% confidence interval for the mean indicates that we are 95% confident that the population mean will be within this interval. In other words, if we selected 100 samples and calculated the confidence interval based on the sample data in each case, then 95 would result in an interval which contained the population mean, and 5 would produce intervals that did not. Similarly, a 99% confidence interval for the proportion indicates that we are 99% confident that the population proportion will be within the interval. This means that out of 100 samples, 99 would result in an interval which contained the population proportion and 1 would produce an interval that did not.

Note that we typically select confidence levels of 95% and 99% (these are high because we usually want to be reasonably confident of the accuracy of our interval) although less often other values (for example, 90%) are selected.

Objectives

On completion of this module you should be able to:

- calculate and interpret confidence interval estimates for the mean and proportion
- determine sample size for means and proportions
- understand the application of confidence intervals, particularly in auditing and
- consider ethical issues relating to confidence interval estimation.

Readings

Source

<table>
<thead>
<tr>
<th>Textbook</th>
<th>Levine et al. 4th edition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sections 7.1 to 7.7</td>
<td>&amp; Excel Handbook Sections EH7.1 to 7.7 (pp. 324–329)</td>
</tr>
</tbody>
</table>

Or

<table>
<thead>
<tr>
<th>Textbook</th>
<th>Levine et al. 5th edition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Section 8.1 to 8.7</td>
<td>&amp; Excel Companion E8.1 to 8.8 (pp. 322–326)</td>
</tr>
</tbody>
</table>
Confidence intervals for the mean

As we’ve said, we have been estimating parameters using point estimates. We use $\bar{X}$ to estimate $\mu$, $S$ to estimate $\sigma$ and $p$ to estimate $\pi$. But we know that there are a number of reasons why this sample estimate may not be exactly equal to the population parameter. In particular, sampling error causes small differences in the estimates found using different samples. We discovered last week that because of the randomness of the sample selection process, each time we take a sample, we are likely to get different data values and so a different estimate of the parameter. This week we will use our understanding of the sampling distribution to find interval estimates of parameters. This accounts for the chance differences between samples by giving an interval within which we can say with a level of confidence that the true population parameter will be contained.

Imagine we desire a 95% confidence interval for the mean. The 95% is equivalent to the decimal of 0.95. This latter value is known as the confidence coefficient and is denoted by $1 - \alpha$. The value $100(1 - \alpha)\%$ (in this example, the 95%) is called the confidence level.

Recall that the standard error of the estimate of the mean is given by $\sigma_x = \frac{\sigma}{\sqrt{n}}$. When a sample size is large, the Central Limit Theorem tells us that the sampling distribution of the mean will be approximately normally distributed. We saw last week that $\frac{\bar{X} - \mu}{\sigma_x} = \frac{X - \mu}{\sigma} \sqrt{n}$ follows a normal distribution. Therefore we can draw on our knowledge of the normal distribution to get bounds which give us a desired level of confidence. For example, we know that 95% of the area under the standard normal curve is between $Z = -1.96$ and $Z = +1.96$ (make sure you know why this is by looking back at Example 6–3 (e) in Module 6). We found this by looking for $\alpha/2 = 0.05/2 = 0.025$ in the centre of the standard normal tables, which gave us $Z = -1.96$ (see the graph below for the general case). By symmetry, we could therefore say that $Z = +1.96$ created the upper bound.
Confidence interval estimation

The value \( Z = \frac{X - \mu}{\sigma / \sqrt{n}} \) could fall anywhere on the horizontal axis on the graph above.

We know that the probability that \( Z \) falls between \(-Z_{a/2}\) to \(+Z_{a/2}\) is \( 1 - \alpha \). Therefore we can say that

\[
P\left(-Z_{a/2} \leq \frac{X - \mu}{\sigma / \sqrt{n}} \leq +Z_{a/2}\right) = 1 - \alpha
\]

Rearranging this gives us

\[
P\left(\frac{X - Z_{a/2} \sigma}{\sqrt{n}} \leq \mu \leq \frac{X + Z_{a/2} \sigma}{\sqrt{n}}\right) = 1 - \alpha
\]

This tells us that if we repeatedly sample from the population, the proportion of values of \( X \) for which \( \mu \) falls in the interval from \( X - Z_{a/2} \sigma / \sqrt{n} \) to \( X + Z_{a/2} \sigma / \sqrt{n} \) is \( 1 - \alpha \).

Therefore, when our sample size is large (at least thirty) and the population standard deviation is known, the following formula gives the confidence interval for the mean:

\[
X \pm Z_{a/2} \frac{\sigma}{\sqrt{n}}
\]

or

\[
X - Z_{a/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq X + Z_{a/2} \frac{\sigma}{\sqrt{n}}
\]

The choice of the confidence coefficient value depends on a number of factors. The closer the value is to one, the more confident you can be that the interval contains the true population mean. The intervals, however, become increasingly wide as the confidence interval approaches one. Ideally we would like the confidence interval to be very narrow (precise) and to have a high level of confidence that the interval contains the population mean. One way of assisting in achieving this goal is by increasing the sample size, since this will reduce the standard error and so will reduce the width of the confidence interval. Unfortunately increasing sample size is not always possible (because of costs in terms of time, money, etc.) and so a compromise between accuracy and practicality must be made. Commonly used confidence levels are 90%, 95% and 99%. The choice depends on how much risk of obtaining an incorrect interval someone is willing to take.

Note that confidence intervals give an idea of the magnitude of the sampling error. They do not account for any biases in the data from other sources (such as understatements of income by respondents, etc.).

The \( t \) distribution

Often the population standard deviation, \( \sigma \), is unknown and so we are forced to use the sample standard deviation, \( S \), to approximate \( \sigma \) in the confidence interval formula. If the sample size is small (less than thirty), then the Student’s \( t \) distribution (usually just referred to as the \( t \) distribution) must be used. As we shall see below, either the \( t \) distribution or the normal distribution may be used when \( \sigma \) is unknown.
but the sample is large. The \( t \) distribution is similar to the standard normal distribution except that it has more area in the tails of the distribution and is less in the middle. It has one parameter, the degrees of freedom, which in this situation is \( n - 1 \) where \( n \) is the sample size. (Recall in contrast that the normal distribution has two parameters: the mean and standard deviation.) The text goes into more detail about the concept of degrees of freedom, so students should read this carefully.

If the random variable, \( X \), is normally distributed, then
\[
\frac{X - \mu}{S/\sqrt{n}} \sim t
\]
with \( n - 1 \) degrees of freedom. Table E.3 in the textbook gives \( t \) values based on degrees of freedom and area in the upper tail of the distribution. The notation \( t_{n-1} \) is used to denote the \( t \) value with \( n - 1 \) degrees of freedom. The confidence interval formula for the mean with a small sample and \( \sigma \) unknown is:
\[
\bar{X} \pm t_{n-1} \frac{S}{\sqrt{n}}
\]

Finding \( t \) values

For a 95% confidence interval, the area in the upper tail will be
\[
\frac{\alpha}{2} = \frac{(1 - 0.95)/2}{2} = 0.025
\]
If we had a sample size of \( n = 20 \), we would have \( n - 1 = 20 - 1 = 19 \) degrees of freedom and the \( t \) value would be \( t_{19} = 2.0930 \) (check for yourself that you can find this value in the table).

Choosing \( Z \) or \( t \) in the confidence interval formula

Knowing when to use the normal or \( t \) distributions in confidence interval formulae is not just dependent on whether the population standard deviation is known or unknown (the textbook tends to oversimplify this choice). Using the Central Limit Theorem (CLT) we know that for large samples (thirty or more) the sampling distribution of the mean is approximately normal. This allows us to choose between the normal or \( t \) distributions when the sample is large, even if the population standard deviation is unknown. The tables below summarise the choice for a variable, \( X \), based on whether it is, or is not, normally distributed.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>( n &lt; 30 )</th>
<th>( n \geq 30 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \sigma ) Known</td>
<td>( Z )</td>
<td>( Z )</td>
</tr>
<tr>
<td>Unknown</td>
<td>( t )</td>
<td>( t ) (or ( Z ) by CLT)</td>
</tr>
</tbody>
</table>

If the variable, \( X \), is not normally distributed:

<table>
<thead>
<tr>
<th>Sample size</th>
<th>( n &lt; 30 )</th>
<th>( n \geq 30 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \sigma ) Known</td>
<td>-</td>
<td>( Z )</td>
</tr>
<tr>
<td>Unknown</td>
<td>-</td>
<td>( t ) (or ( Z ) by CLT)</td>
</tr>
</tbody>
</table>
Using this additional knowledge, we can see that there are a number of variations on the formula for the confidence interval based on whether $\sigma$ is known or unknown and on whether the normal or $t$ distribution is used. These possibilities are found by using $S$ instead of $\sigma$ or $t$ instead of $Z$ in the confidence interval formula (or both).

**Example 8–1**

A lecturer is interested in the amount of time it takes students to complete a particular assignment. It is known that the standard deviation is 24 minutes. The lecturer takes a random sample of 40 students and discovers that in this sample, the mean time to complete the assignment is 150 minutes.

(a) Set up a 95% confidence interval estimate for the population mean time taken to complete the assignment.

(b) Recalculate your answer to (a) based on a 90% confidence interval.

(c) It was the lecturer’s intention that this assignment take students no more than two hours to complete. Comment on how well the sampled students reflect this goal. Based on your results, what would you recommend to the lecturer?

(d) Suppose an extra class was offered to the students which enabled them to better understand the goal of the assignment. This resulted in a change of the standard deviation for all students to only 12 minutes. What effect would this have on your answer to (a)?

**Solution 8–1**

We are told that $\sigma = 24$, $\bar{X} = 150$ and $n = 40$.

(a) Since the sample size is greater than 30, we will use the normal distribution. For a 95% confidence interval, $Z = 1.96$ and so the confidence interval is

$$\bar{X} \pm Z \frac{\sigma}{\sqrt{n}} = 150 \pm (1.96) \frac{24}{\sqrt{40}}$$

$$= 150 \pm 7.4377 \text{ (to 4 decimal places)}$$

$$142.56 \leq \mu \leq 157.44 \text{ (rounded to 2 decimal places)}.$$

So the lecturer can be 95% confident that the true mean time taken to complete the assignment is between 142.56 and 157.44 minutes.

(b) For a 90% confidence interval, we have the following situation:

![Confidence Interval Diagram](image)
Confidence interval estimation

So $Z = 1.64$ and the 90% confidence interval is

$$\bar{X} \pm Z \frac{\sigma}{\sqrt{n}} = 150 \pm (1.64) \frac{24}{\sqrt{40}}$$

$$= 150 \pm 6.2234$$

$$143.78 \leq \mu \leq 156.22 \text{ (to 2 decimal places)}$$

So the lecturer can be 90% confident that the true mean time taken to complete the assignment is between 143.78 and 156.22 minutes.

Note that it would be equally correct to use either $Z = 1.645$ or $Z = 1.65$ since the value 0.05 does not appear exactly in the standard normal table.

(c) The widest of the two confidence intervals calculated (the 95% confidence interval from 142.56 to 157.44 minutes) does not contain 120 minutes (2 hours). It implies that the average time to complete the assignment is actually much longer than 120 minutes, and so it seems unlikely that the students are going to be able to complete the assignment in 120 minutes. We would expect that only 5% of possible samples would result in an interval which does not contain the population mean. This means that it is possible that this is one of these 5%, however, given the difference in size (120 is much smaller than the values in the interval), this seems unlikely.

Note: also that the mean being discussed here is the average time taken to complete the assignment for a sample of students. It is very likely that there are some individual students who are able to complete the assignment within two hours, however, on average, the students appear unable to do so.

(d) Given $\sigma = 12$, the 95% confidence interval becomes

$$\bar{X} \pm Z \frac{\sigma}{\sqrt{n}} = 150 \pm (1.96) \frac{12}{\sqrt{40}}$$

$$= 150 \pm 3.7188$$

$$146.28 \leq \mu \leq 153.72 \text{ (rounded to 2 decimal places)}.$$ 

So the lecturer can be 95% confident that the true mean time taken to complete the assignment is between 146.38 and 153.72 minutes.

A reduction in the standard deviation has resulted in a smaller interval. In other words less variation in the students times means a smaller interval estimate for the population mean time.

Example 8–2

A group of students are concerned that bags of a particular brand of potato chips weigh less than the 50 grams that the packaging claims. They take a random sample of 20 bags and discover that for this sample, the mean weight is 49.2 grams and the standard deviation is 1.1 grams. Calculate the 95% confidence interval based on their sample data. Do you think the students are justified in their belief? What potential problems are there with the way the students have conducted this experiment?
Solution 8–2

We are told that $\bar{X} = 49.2$ and $S = 1.1$. Since the population standard deviation is unknown (we are given a sample standard deviation) and the sample size is less than 30, the following formula is used to calculate the confidence interval:

$$\bar{X} \pm t_{n-1} \frac{S}{\sqrt{n}}.$$ 

Given our sample size was twenty, $t_{n-1} = t_{20-1} = 2.0930$.

$$\bar{X} \pm t_{n-1} \frac{S}{\sqrt{n}} = 49.2 \pm (2.0930) \frac{1.1}{\sqrt{20}} = 49.2 \pm 0.5148$$

$48.69 \leq \mu \leq 49.71$ (to 2 decimal places)

So the students can be 95% confident that the true population mean weight of the bags of chips is between 48.69 and 49.71 grams.

Based on this confidence interval, it appears that the mean weight of the bags is less than 50 grams and so we can assume that the students are probably correct in their claim. We could not be sure of this, however, unless we knew that the sample from which this confidence interval is calculated is representative of the population. For example we would need to know:

• that the data was approximately normally distributed
• that the sample was truly randomly selected (i.e., the bags were randomly selected from stores, location in stores, etc. and probably randomly selected from town or city also)
• that a sample of size twenty is sufficiently large (which seems doubtful)
• that the scale used to weigh the bags was accurate
• that in the time since packing the bags there were no unavoidable factors (such as settling, moisture content changes, etc.) that may have resulted in a reduction (or increase) in bags weights between the factory and the students hands.

There are many other factors that may also have contributed to this experiments result. As an exercise, consider what other factors you can think of.

Confidence intervals for the proportion

Our point estimate of the population proportion is $p = \frac{X}{n}$. As with the mean, a confidence interval estimate of the population proportion will give us an interval within which we can say with a level of confidence that the true population proportion will be contained. Last week we saw that $Z = \frac{p - \pi}{\pi (1-\pi) \frac{n}{n}}$ was approximately normally distributed. Therefore we know that
Confidence interval estimation

\[ P \left( p - Z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}} \leq \pi \leq p + Z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}} \right) = 1 - \alpha . \]

This tells us that if we repeatedly sample from the population, the proportion of values of \( p \) for which \( \pi \) falls in the interval from \( p - Z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}} \) to \( p + Z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}} \) is \( 1 - \alpha \).

The **confidence interval formula for the proportion** is:

\[ p \pm Z \sqrt{\frac{p(1-p)}{n}} \]

or

\[ p - Z \sqrt{\frac{p(1-p)}{n}} \leq \pi \leq p + Z \sqrt{\frac{p(1-p)}{n}} \]

This requires that the sample size is large enough. Normally we require that both \( X \) and \( n - X \) are greater than 5.

**Example 8–3**

A financial advice firm has been recommending a particular investment opportunity to a large number of its clients. They surveyed a random sample of 500 clients who took advantage of the investment and discover that 408 of them are glad they made the investment. Construct both a 95% and 99% confidence interval for the population proportion of clients who are glad they made the investment.

**Solution 8–3**

We are given \( p = \frac{408}{500} = 0.816 \) and \( n = 500 \). For a 95% confidence interval, \( Z = 1.96 \) and so

\[ p \pm Z \sqrt{\frac{p(1-p)}{n}} = 0.816 \pm (1.96) \sqrt{\frac{0.816(1-0.816)}{500}} \]

\[ = 0.816 \pm 0.033964 \] (to 6 decimal places)

\[ 0.782 \leq \pi \leq 0.850 \] (to 3 decimal places)

So we can say with 95% confidence that the true population proportion of clients who were happy with the investment is between 0.782 and 0.850.

For a 99% confidence interval, \( Z = 2.58 \) and

\[ p \pm Z \sqrt{\frac{p(1-p)}{n}} = 0.816 \pm (2.58) \sqrt{\frac{0.816(1-0.816)}{500}} \]

\[ = 0.816 \pm 0.044708 \]

\[ 0.771 \leq \pi \leq 0.861 \] (to 3 decimal places)
So we can say with 99% confidence that the true population proportion of clients who were happy with the investment is between 0.771 and 0.861.

**Determining sample size**

Normally we would determine the required sample size prior to collecting data. This decision is based on what sampling error is acceptable, as well as constraints such as available time and cost. When estimating the population mean, the sampling error is the amount that is added or subtracted from the sample mean in the confidence interval formula: \( e = Z \frac{\sigma}{\sqrt{n}} \). By rearranging this formula, we find the sample size required when estimating the mean:

\[
 n = \frac{Z^2 \sigma^2}{e^2}
\]

To estimate the required sample size, we therefore need to know the required confidence level (which will determine the value of \( Z \)), the population standard deviation \( \sigma \) and the acceptable sampling error \( e \).

When estimating the population proportion, the sampling error is the amount that is added or subtracted from the sample proportion in the confidence interval formula:

\[
 e = Z \sqrt{\frac{\pi(1-\pi)}{n}}
\]

By rearranging this formula, we find the sample size required when estimating the proportion:

\[
 n = \frac{Z^2 \pi(1-\pi)}{e^2}
\]

To estimate the required sample size, we therefore need to know the required confidence level (which will determine the value of \( Z \)), the population proportion \( \pi \) and the acceptable sampling error \( e \).

Unfortunately, we are unlikely to know the value of \( \pi \) prior to taking the sample! We can either use past information, or expert opinion to estimate its value, or, we can use the value which always give the largest sample size (since this will never underestimate the sample size). A value of \( \pi = 0.5 \) always gives the largest sample size and because of this is often referred to as the most conservative estimate.

**Example 8–4**

A consumer watchdog organisation is interested in the mean amount charged per hour by accountants for their services. Based on studies in other similar countries, the standard deviation is believed to be $12.75. The organisation wants to estimate the mean amount charged per hour to within ±$4 with 95% confidence. What sample size is needed? If 99% confidence were required, what would the required sample size be?
Solution 8–4

We are given $\sigma = 12.75$, $e = 4$ and based on 95% confidence, we know that $Z = 1.96$. The sample size is then

$$n = \frac{Z^2\sigma^2}{e^2} = \frac{(1.96)^2 (12.75)^2}{(4)^2} = 39.03 \approx 40$$

Note: that we always round up to the next whole integer when determining sample size.

So a sample of 40 accountants should be taken to be 95% confident that the estimate of the mean is within ±$4 of the true average.

Based on 99% confidence, we have $Z = 2.58$ and the sample size is

$$n = \frac{Z^2\sigma^2}{e^2} = \frac{(2.58)^2 (12.75)^2}{(4)^2} = 67.63 \approx 68$$

So a sample of 68 accountants should be taken to be 99% confident that the estimate of the mean is within ±$4 of the true average.

Example 8–5

The same group of students that were discussed in Example 8–2 discovered a flaw in the process of random selection of chip bags. They decide to conduct the experiment again and this time work out what sample size will be needed to accurately estimate the population proportion of bags of chips that are underweight.

If they wish to be 95% confident that their estimated proportion is within ±0.025 of the population proportion, determine the required sample size.

Solution 8–5

Given $Z = 1.96$ and $e = 0.025$, we must determine a value for the ‘success’ (in this case success is the number of underweight bags). Since the students previous results were incorrect, the most conservative estimate of $p$ should be used, which is $\pi = 0.5$. Therefore the sample size should be

$$n = \frac{Z^2\pi(1-\pi)}{e^2} = \frac{(1.96)^2 (0.5)(0.5)}{(0.025)^2} = 1536.64 \approx 1537$$

So the students need to sample 1537 bags of chips so that their estimate of the proportion is within ±0.025 of the population proportion with 95% confidence.

Note that they could reduce the required sample size quite dramatically by allowing a larger sampling error. For example, increasing the error to ±0.05 will reduce the sample size to 385 bags of chips. Another possibility (which we are not given the information to be able to demonstrate here) would be to use the estimate of the proportion of underweight bags from their previous data set.
Confidence interval estimation

Confidence intervals for finite populations

As we discovered last week, when the population is finite, the finite population correction factor, \( \sqrt{\frac{N-n}{N-1}} \), must be included in the standard error. The same is true for the confidence interval formula. The confidence interval formulae for the mean and proportion therefore become:

\[
\bar{X} \pm t_{n-1} \frac{S}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}
\]

\[
p \pm Z \sqrt{\frac{p(1-p)}{n}} \sqrt{\frac{N-n}{N-1}}
\]

Confidence intervals and auditing

In the business world, sampling is common to estimate population information. Because of time and money constraints, auditors will rarely be able to look at every item in the population and so confidence intervals are useful. Unlike the confidence intervals we have examined so far, it makes more sense in auditing to look for an estimate of the population total, rather than the population mean. The point estimate for the population total is given by:

\[
\text{Total} = N\bar{X}
\]

The confidence interval estimate for the total is:

\[
N\bar{X} \pm N \left( t_{n-1} \right) \frac{S}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}
\]

Note that this formula includes the finite population correction factor since when auditing we are most likely to be sampling from a finite population!

Example 8–6

A budget eyewear store (which sells frames for glasses, lens cleaner, cases, cloths, etc.) is conducting the end of quarter inventory of its stock. It was determined that there were 1296 items in stock of which a sample of 100 was randomly selected. An audit was conducted which found that the mean value of the merchandise in the sample was $196 and the standard deviation was $67.50. Based on this information, find the 95% confidence interval estimate of the total estimated value of the merchandise in inventory at the end of the quarter.

Solution 8–6

We know that \( N = 1296 \), \( n = 100 \), \( \bar{X} = 196 \), \( S = 67.50 \) and \( t_{n-1,0.025} = 1.9842 \). The 95% confidence interval will be

\[
N \cdot \bar{X} \pm N \cdot t_{n-1,0.025} \frac{S}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} = 1296(196) \pm 1296(1.9842) \frac{67.50}{100} \sqrt{1296-100} \sqrt{1296-1} = 254016 \pm 16681.1092 \text{ (to 4 decimal places)}
\]
Confidence interval estimation

$237,334.11 \leq \text{Population total} \leq $270,697.11

So the store can be 95% confident that the population total merchandise value will be between $237,334.11 and $270,697.11.

Note that although using a sample to estimate the total value of merchandise is quicker and cheaper than looking at every item in stock, the confidence interval that results in this case is very wide and so does not give a very accurate estimate. The cause of this is the relatively large standard error. This result may actually indicate that the store must add up every item in stock.

Discussion points

Discussion point 8–1
With other students, discuss the answer to Problem 8.60 on p. 315 of the textbook (Problem 7.60 on p. 317 of the 4th edition).

Discussion point 8–2
Discuss and compare point estimation and confidence interval estimation. Why are they different and what are their respective uses? What ethical issues arise from point estimation? What kind of estimation do you believe should be used in the workplace? Explain.

Summary

Now that you have completed this module, turn back to the objectives at the beginning of the module. Have you achieved these objectives?

Ensure that you attempt the recommended problems in the list of review questions below and at least a sample of problems from the optional list. This will help you to identify any areas of difficulty you have in achieving the module’s objectives.

Review questions

Recommended problems

Levine et al. 4th edition: Do Questions 7.2, 7.6, 7.8, 7.10, 7.16, 7.18, 7.22, 7.28, 7.30, 7.32, 7.36, 7.40, 7.46, 7.50, 7.52, 7.54, 7.58, 7.60 to 7.66, 7.68, 7.70 and 7.80.

Levine et al. 5th edition: Do Questions 8.2, 8.6, 8.8, 8.10, 8.16, 8.20, 8.24, 8.26, 8.28, 8.34, 8.38, 8.42, 8.50, 8.52, 8.54, 8.58, 8.60 to 8.66, 8.68, 8.70 and 8.80 from the textbook and from the CD 8.92 and 8.98.
Optional problems

Levine et al. 4th edition: Choose a selection of problems from Questions 7.1, 7.3 to 7.5, 7.7, 7.9, 7.11 to 7.15, 7.17, 7.19 to 7.21, 7.23 to 7.29, 7.31, 7.33 to 7.35, 7.37 to 7.39, 7.41 to 7.45, 7.47 to 7.49, 7.51, 7.53, 7.55 to 7.57, 7.59, 7.67, 7.69, 7.71 to 7.79 and 7.81 to 7.87.

Levine et al. 5th edition: Choose a selection of problems from Questions 8.1, 8.3 to 8.5, 8.7, 8.9, 8.11 to 8.15, 8.17 to 8.19, 8.21 to 8.23, 8.25, 8.27, 8.29 to 8.33, 8.35 to 8.37, 8.39 to 8.41, 8.43 to 8.49, 8.51, 8.53, 8.55 to 8.57, 8.59, 8.67, 8.69, 8.71 to 8.79 and 8.81 to 8.84 and from the CD 8.91, 8.93 to 8.97.
Hypothesis testing
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>9-3</td>
</tr>
<tr>
<td>Objectives</td>
<td>9-3</td>
</tr>
<tr>
<td>Readings</td>
<td>9-3</td>
</tr>
<tr>
<td>Examples</td>
<td>9-4</td>
</tr>
<tr>
<td>Discussion points</td>
<td>9-10</td>
</tr>
<tr>
<td>Summary</td>
<td>9-11</td>
</tr>
<tr>
<td>Review questions</td>
<td>9-11</td>
</tr>
</tbody>
</table>
Introduction

Last week we used confidence intervals to allow us to estimate a range within which we had some level of confidence that our population parameter (a mean, proportion or total) would fall. Sometimes we are more interested in seeing whether a sampled parameter estimate is significantly different from a certain value. Remember Example 8–2 from last week’s material? There, some students were interested in discovering whether the weights of bags of chips were consistently less than was printed on the packaging. The question they were asking was if the sample mean was less than 50 grams. Obviously a confidence interval can give us some idea of this, but hypothesis testing gives us a clearer outcome. With hypothesis testing we ask questions such as ‘Is the mean weight of the bags of chips less than 50 grams?’ or ‘Was the mean mark of students on their statistics exam at least 60%?’ As with confidence intervals, we select an acceptable level of confidence. With hypothesis tests, however, we express this as the level of significance and usually pick values of 0.01, 0.05 and 0.10. So 0.01, for example, indicates that the conclusion of the hypothesis test will be wrong 0.01 (or 1%) of the time and so is similar to our 99% confidence interval.

Objectives

On completion of this module you should be able to:

• understand and demonstrate the process required for hypothesis testing
• explain the difference between one- and two-tailed tests
• perform a hypothesis test for the mean and the proportion
• consider ethical issues relating to hypothesis testing.

Readings

Source

Textbook Levine et al. 4th edition
Sections 8.1 to 8.6 & Excel Handbook Sections EH8.1 to 8.3
(pp. 368–371)

Or

Textbook Levine et al. 5th edition
Sections 9.1 to 9.6 & Excel Companion Sections E9.1 to 9.3
(pp. 364–367)

Course website Visit the course website for links to any supplementary material for this week.
Examples

Example 9–1
In the Australian legal system, the accused is considered innocent until proven guilty. Using this information, state the null and alternative hypotheses and discuss the type I and II errors (which should be the larger value, which should be the smaller value and why?).

Solution 9–1
The alternative hypothesis is normally used to describe the statement that we are attempting to prove. So in this situation the null hypothesis should be that the accused is innocent and the alternative that they are guilty. The hypotheses would therefore be:

\[ H_0: \text{the accused is innocent} \]
\[ H_1: \text{the accused is guilty}. \]

We know that \( \alpha = \) probability of a type I error (rejecting the null hypothesis when it is true) and \( \beta = \) probability of a type II error (accepting the null hypothesis when it is false). If we were to draw a diagram of the various outcomes, it might look like this:

![Diagram](image)

Clearly we want to minimise all errors, so our desire is that both \( \alpha \) and \( \beta \) are as small as possible. Unfortunately, they are inversely related, meaning as one decreases, the other must increase. Normally, Australian society demands that we minimise the type I error since we are less tolerant of convicting innocent people. The consequence is that a (hopefully small) portion of those accused of crimes are found not guilty even though they are guilty of the crime.
Example 9–2

A recent graduate from a business degree is considering the benefits of working for various companies. A particular company, Touccancy Inc., has a reputation for treating its employees well. In fact, the company’s website gives some statistics about starting salaries for recent graduates. It claims (in large print) that the mean starting salary for recent graduates is $45,000. In much smaller print, buried in a report on statistics, the website states that the known standard deviation of starting salaries is $5000 and that in a random sample of fifty recently employed graduates, the mean salary was $39,000.

(a) At the 5% level of significance, determine if there is evidence that the claim given in large print is valid, based on the sample data. Use both the critical value and p-value approaches as part of your answer.

(b) Determine the 95% confidence interval for the population mean starting salary and compare this to your answer in (a).

Solution 9–2

(a) Critical value approach


1. \( H_0: \mu = 45000 \)
   \( H_1: \mu \neq 45000 \)

2. \( \alpha = 0.05 \) and \( n = 50 \) (fifty recently employed graduates were sampled).

3. Because \( \sigma = 5000 \) is known, we use a Z test.

4. Having chosen \( \alpha = 0.05 \), our rejection regions are defined by the values \(+1.96\) and \(-1.96\). Given that \( Z \) will be the value of the test statistic, we express a rejection (decision) rule as:
   
   Reject \( H_0 \) if \( Z > +1.96 \) or if \( Z < -1.96 \), otherwise reject \( H_0 \).

5. Given that \( \bar{X} = 39000 \), the test statistic is:

   \[ Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{39000 - 45000}{5000} = -8.49 \]

6. Since \(-8.49 < -1.96\), this falls in the rejection region, so we reject the null hypothesis. We can conclude that the mean starting salary is significantly different from $45,000.

It appears that the claim made by the company is not valid based on this particular sample. Either this sample is not representative of the population, or, the company’s claim is designed as a (dishonest?) recruiting tool, rather than based on facts.
Hypothesis testing

p-value approach

If we had used the p-value approach, we would utilise the five steps of Exhibit 9.2 on p. 339 of the textbook (a simplified version of the ten step approach in Exhibit 8.3 on p. 341 of the 4th edition). The first three steps are the same as given above. The remaining steps are detailed below:

4. Using the test statistic already calculated (-8.49), the p-value approach requires that we find the probability of being more extreme than our test statistic. This means what is the probability of being more than 8.49 standard deviations either in the positive or negative direction. The p-value is therefore \( P(Z < -8.49) + P(Z > 8.49) \). If we search for these values in normal tables, we find that they are not listed meaning the probability is effectively zero. This means our p-value is 0.

5. Clearly \( 0 < 0.05 \) is true (the p-value is less than \( \alpha \)) and so we reject the null hypothesis. As before, we can conclude that the mean starting salary is significantly different from $45,000.

(b) The 95% confidence interval is given by:

\[
\overline{x} \pm \frac{Z}{\sqrt{n}} = 39000 \pm \left( 1.96 \frac{5000}{\sqrt{50}} \right)
\]

\[
= 39000 \pm 1385.93
\]

\[
$37,614.07 < \mu < 40,385.93$
\]

This interval does not contain the company’s claimed mean starting salary. In fact $45,000 is much greater than the values in the interval. This again indicates that the company’s claim appears to be incorrect based on this sample of starting salary.

Example 9–3

An investment advisory company has been having problems with the printery which is responsible for the printing of the company’s weekly stock reports. The company requires that the printing of the report be completed within twenty-four hours of receipt of the necessary files and documentation and specifies a standard deviation of two hours. They are prepared to employ a different printery if they can establish statistically that the printery is not meeting their requirements. A sample of thirty recent printing times for reports has a mean printing time of twenty-five hours.

(a) If the company tests the hypothesis at the 1% level of significance, what decision would be made using the p-value approach to hypothesis testing? Interpret the meaning of the p-value in this problem.

(b) How would your answer in (a) change if the standard deviation had been three hours?

Solution 9–3

(a) Following the five steps of Exhibit 9.2 we get the following:

1. The null hypothesis is: \( H_0 : \mu \leq 24 \)

The alternative hypothesis is: \( H_1 : \mu > 24 \).
Note that the alternative hypothesis here is the statement we are trying to prove. The company believes the printing is taking too long and are considering changing to another printery. Therefore this is reflected in the alternative hypothesis. The null hypothesis always includes the equals sign and reflects the status quo or what should be happening. The company will be quite prepared to accept printing times that are less than 24 hours, hence the one-tailed test is appropriate; they are only interested in showing whether printing times are too long (or not).

2. We are told that the significance level is 1%, so \( \alpha = 0.01 \). The sample size is \( n = 30 \).

3. Because \( \sigma = 2 \) is known, we use a Z-test.

4. Given that \( \bar{X} = 25 \), the test statistic is:

\[
Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{25 - 24}{2/\sqrt{30}} = 2.74
\]

The \( p \)-value is found via \( P(Z > 2.74) \). Using normal tables we find that the \( p \)-value is \( 1 - 0.9969 = 0.0031 \).

Note that because this is a one-tailed test, the \( p \)-value is found in only one region of the normal distribution (contrast this with the previous example where a two-tailed test was used). The \( p \)-value is found by looking at the alternative hypothesis and choosing the probability using that inequality. Since the alternative has a ‘>’ symbol, so too does our \( p \)-value probability.

5. \( 0.0031 < 0.01 \). Since the \( p \)-value is less than \( \alpha \), we reject the null hypothesis. We can conclude that there is evidence that the printery is taking more than the company’s required twenty-four hours to print the reports. Based on this, the company may be justified in finding another printery to produce their reports.

(b) If \( \sigma = 3 \), the test statistic would be

\[
Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{25 - 24}{3/\sqrt{30}} = 1.83
\]

and the \( p \)-value would have been \( P(Z > 1.83) = 1 - 0.9664 = 0.0336 \). Then, since \( 0.0336 < 0.01 \), we would not have rejected the null hypothesis at the 1% level of significance. This would indicate the company would not be justified in changing to another printery when the standard deviation is 3.
**Important note:**

We could still have rejected this null hypothesis at the 5% level of significance (since $0.0336 < 0.05$), indicating that the mean time to print the reports is significantly greater than twenty-four hours at the 5% level. Often the term *highly significant* is used to describe a situation where the null hypothesis is rejected at the 1% level and *significant* to indicate a rejection at the 5% level of significance. Therefore the situation with the modified standard deviation is significant, but not highly significant.

In terms of the decision the company should make, this introduces a degree of ambiguity. The decision is then based on how confident the company wants to be that they are making the right decision. Clearly in the situation where $\sigma = 2$, they would reject the null hypothesis in either case.

**Example 9–4**

A firm of accountants has been established in a small regional town for twenty years. Part of their service includes regular visits to the clients they service to ensure they keep a customer service focus. The accountants must still be contactable while they are away from their desks, so mobile phones are provided. They have discovered, however, that the useful life of the phones is very much reduced by the shortness of the phone’s battery life. In the past, the batteries gave an average talk time of twenty hours, after which time the phone needed recharging. A sample of forty batteries recently revealed an average talk time of eighteen hours, with a standard deviation of four hours.

(a) At the 0.05 significance level, is there evidence that the mean talk time of the batteries has changed from twenty hours?

(b) What assumptions did you make regarding the population distribution in answering part (a)? Explain how you would test these assumptions if you had the talk time data.

**Solution 9–4**

(a) 1. $H_0 : \mu \geq 20$
   $H_1 : \mu < 20$

2. Select $\alpha = 0.05$. A random sample of $n = 40$ batteries is being used.

3. If we assume the population of talk times of the batteries is normally distributed, then the $t$-test is appropriate here since the population standard deviation of the talk times is unknown (we’re given the sample value).

4. Given a sample size of $n$, the test statistics follows a $t$ distribution with $n-1$ degrees of freedom. At $\alpha = 0.05$, the critical value will be $t_{0.05} = -1.6849$ and so the rejection rule is: Reject $H_0$ if $t < -1.6849$, otherwise do not reject $H_0$. 


5. The test statistic is

\[ t = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} = \frac{18 - 20}{\frac{4}{\sqrt{40}}} = -3.16 \]

**Note:** we are using the critical value approach for this hypothesis test. Check that the \( p \)-value approach gives the same conclusion.

6. Since \(-3.16 < -1.6849\), the test statistic is in the rejection region and we reject the null hypothesis. The data provides sufficient evidence to conclude that the mean talk time provided by the batteries is significantly less than 20 hours.

(b) We assumed that the random sample of talk times of batteries comes from a population that is normally distributed. As long as the sample size is reasonably large and the population is not too skewed, the \( t \) distribution provides a good approximation to the sampling distribution of the mean when the standard deviation is unknown. This assumption of normality can be tested by examining the sample and based on the distribution of the sample, making inferences about the likely distribution of the population. We are therefore making the assumption that the sample is reasonably representative of the population. We could test the normality of the sample data by producing:

- a histogram or stem-and-leaf plot and checking for something similar to a bell-shaped curve
- a normal probability plot and checking for departures from a straight line
- a box-and-whisker plot and checking for symmetry.

**Example 9–5**

For many years, women have been under represented in management positions. A twenty-year-old study revealed that only 18% of companies had at least one woman in their management team. A recent study was conducted to discover whether this situation has changed. A survey was sent out to 500 of the largest companies across Australia. Of these, 23% (115) indicated that they had at least one woman in their management team. At the 5% level of significance, can you state that there has been an increase in the proportion of companies who include women in their management teams?
Solution 9–5

\[ H_0 : \pi \leq 0.18 \]
\[ H_1 : \pi > 0.18 \]

At \( \alpha = 0.05 \), the critical value will be \( Z = 1.64 \) and so the rejection rule is: Reject \( H_0 \) if \( Z > 1.64 \), otherwise do not reject \( H_0 \).

The test statistic is

\[
Z = \frac{p - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} = \frac{0.23 - 0.18}{\sqrt{\frac{0.18(0.82)}{500}}} = 2.91
\]

Note that we are using the critical value approach to this hypothesis test. Check for yourself that the \( p \)-value approach gives the same conclusion.

Since \( 2.91 > 1.64 \), the test statistic is in the rejection region and so we reject the null hypothesis. Therefore, the data provides sufficient evidence to conclude that there has been an increase in the proportion of companies who include women in their management teams.

Discussion points

Discussion point 9–1

Consider a situation where a medical laboratory is testing individuals to determine whether they are HIV positive. Given the null hypothesis that an individual is HIV negative, describe the type I and II errors. With other students, discuss the type I and II errors and their relative sizes. What ethical issues must be considered in this situation?

Discussion point 9–2

Imagine that a colleague, Amin, comes to you for some statistical advice. Amin is employed in the finance division of a large retail company. Part of his job is to compile the relevant statistics to be included in the company’s annual report. He has been looking at the company’s figures all year and has repeatedly told his boss that the company’s average weekly revenue is over four million dollars. A sample of twenty weeks from the last year indicates that the mean is just $3.5 million. Amin tells you that he was able to produce a hypothesis test which states that the 3.5 million is not statistically different from 4 million. However he tells you that initially the test was significantly different so he reduced the population standard deviation he was using and changed his significance level from 0.01 to 0.05. In groups, consider your answers to the following questions.

(a) What would your advice to Amin be? Explain your answer to him using statistical theory learned over the last few weeks (and any other knowledge you have that may be useful).

(b) How would it affect your response if Amin told you his boss was going to fire him if his report didn’t make the company look positive? Don’t just give the obvious answer here. Explore the pros and cons and make a considered response.
Summary

Now that you have completed this module, turn back to the objectives at the beginning of the module. Have you achieved these objectives?

Ensure that you attempt the recommended problems in the list of review questions below and at least a sample of problems from the optional list. This will help you to identify any areas of difficulty you have in achieving the module’s objectives.

Review questions

Recommended problems

Levine et al. 4th edition: Do Questions 8.1 to 8.12, 8.16, 8.18, 8.20, 8.22, 8.24, 8.26, 8.28, 8.34, 8.38, 8.40, 8.42, 8.52, 8.54, 8.62, 8.66, 8.68, 8.84 and 8.86.


Optional problems

Levine et al. 4th edition: Choose a selection of problems from Questions 8.13 to 8.15, 8.17, 8.19, 8.21, 8.23, 8.25, 8.27, 8.29 to 8.33, 8.35 to 8.37, 8.39, 8.41, 8.43 to 8.51, 8.53, 8.55 to 8.61, 8.63 to 8.65, 8.67, 8.69 to 8.83, 8.85 and 8.87 to 8.94.

Chi-square tests
Contents

Introduction ............................................................................................................. 10–3
Objectives ................................................................................................................ 10–3
Readings .................................................................................................................. 10–3
Examples ................................................................................................................. 10–3
Discussion points ..................................................................................................... 10–8
Summary ................................................................................................................. 10–9
Review questions ................................................................................................. 10–9
Introduction

This week we focus on three uses of the chi-square \( \chi^2 \) test. We will use the hypothesis testing techniques developed last week, to answer questions such as: are the proportions of a particular variable equal in two or more samples and are two variables statistically independent?

Objectives

On completion of this module you should be able to:

- perform and interpret a \( \chi^2 \) test for the difference between two or more proportions and
- perform and interpret a \( \chi^2 \) test of independence.

Readings

Source

Textbook Levine et al. 4th edition
Sections 11.1 to 11.3 & Excel Handbook Sections EH11.1, 11.2 and 11.4 (pp. 503–507)

Or

Textbook Levine et al. 5th edition
Sections 12.1 to 12.3 & Excel Companion Sections E12.1 to 12.3 (pp. 506–507)

Course website Visit the course website for links to any supplementary material for this week.

Examples

Example 10–1

A sample of 400 ex-students was taken and the students were asked ‘Did you enjoy your university experience?’ The results are given in the table below. Is there evidence of a significant difference between the proportions of males and females who enjoyed their university experience? (Use the 0.05 level of significance.)
### Solution 10–1

The hypotheses are:

\[ H_0 : \pi_1 = \pi_2 \]
\[ H_1 : \pi_1 \neq \pi_2 \]

For a 2 × 2 contingency table there is one degree of freedom, so \( df = 1 \). Given \( \alpha = 0.05 \) (and using Table E.4 in either text), the critical value is \( \chi^2_{0.05, 1} = 3.841 \) and so the decision rule is:

Reject \( H_0 \) if \( \chi^2 > 3.841 \), otherwise do not reject \( H_0 \).

Note that with \( \chi^2 \) tests, we are only ever interested in rejecting the null hypothesis when there are big (not small) differences between the two groups. This means that the hypothesis tests are always one-tailed.

The average proportion is calculated as

\[ \bar{p} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{214 + 107}{280 + 120} = 0.8025 \]

The working required to calculate the test statistic is given in the table below.

Note that all expected frequencies are at least 5 and so the chi-square test is appropriate.

<table>
<thead>
<tr>
<th>Enjoyed university experience?</th>
<th>Gender</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Male</td>
<td>Femal</td>
<td>Total</td>
</tr>
<tr>
<td>Yes</td>
<td>214</td>
<td>107</td>
<td>321</td>
</tr>
<tr>
<td>No</td>
<td>66</td>
<td>13</td>
<td>79</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>280</td>
<td>120</td>
<td>400</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( f_o )</th>
<th>( f_e )</th>
<th>( f_o - f_e )</th>
<th>( (f_o - f_e)^2 )</th>
<th>( (f_o - f_e)^2 / f_e )</th>
</tr>
</thead>
<tbody>
<tr>
<td>214</td>
<td>( \bar{p} \times n_1 = 0.8025 \times 280 ) = 224.7</td>
<td>214 - 224.7 = -10.7</td>
<td>114.49</td>
<td>114.49/224.7 = 0.5095</td>
</tr>
<tr>
<td>107</td>
<td>( \bar{p} \times n_2 = 0.8025 \times 120 ) = 96.3</td>
<td>10.7</td>
<td>114.49</td>
<td>1.1889</td>
</tr>
<tr>
<td>66</td>
<td>( (1 - \bar{p}) \times n_1 = 0.1975 \times 280 ) = 55.3</td>
<td>10.7</td>
<td>114.49</td>
<td>2.0703</td>
</tr>
<tr>
<td>13</td>
<td>( (1 - \bar{p}) \times n_2 = 0.1975 \times 120 ) = 23.7</td>
<td>-10.7</td>
<td>114.49</td>
<td>4.8308</td>
</tr>
</tbody>
</table>

\[ 8.5995 \]
Chi-square tests

Since $8.5995 > 3.841$, we reject the null hypothesis and conclude that there is sufficient evidence to believe that there is a difference in the proportion of males and females who enjoyed their university experience.

Example 10–2

A city is supported by three major IT companies, BMI, Unitses, and Pear. There have been rumours of price-fixing and collusion between the companies. In order to investigate these accusations, a consumer watchdog organisation conducted a survey of 500 consumers of these IT services. The results of this survey are summarised in the table below. At the $\alpha = 0.05$ level of significance, determine whether there is evidence of a significant difference between the consumer satisfaction of the three IT companies.

<table>
<thead>
<tr>
<th>Satisfied with Service</th>
<th>IT Company</th>
<th></th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BMI</td>
<td>Unitse</td>
<td>Pear</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>115</td>
<td>99</td>
<td>108</td>
<td>322</td>
</tr>
<tr>
<td>No</td>
<td>55</td>
<td>102</td>
<td>21</td>
<td>178</td>
</tr>
<tr>
<td>Total</td>
<td>170</td>
<td>201</td>
<td>129</td>
<td>500</td>
</tr>
</tbody>
</table>

Solution 10–2

The hypotheses are:

\[ H_0 : \pi_1 = \pi_2 = \pi_3 \]
\[ H_1 : \text{Not all are } \pi_j \text{ equal} \]

For a $2 \times 3$ contingency table there are $(r - 1)(c - 1) = (2 - 1)(3 - 1) = 2$ degrees of freedom. Given $\alpha = 0.05$, the critical value is $\chi^2_{2, 0.05} = 5.991$ and so the decision rule is:

Reject $H_0$ if $\chi^2 > 5.991$, otherwise do not reject $H_0$.

The average proportion is calculated as

\[ \bar{p} = \frac{X_1 + X_2 + X_3}{n_1 + n_2 + n_3} = \frac{115 + 99 + 108}{170 + 201 + 129} = 0.644 \]

The table below is then used to calculate the test statistic.

Note that all expected frequencies are at least $5$ and so the chi-square test is appropriate.
Chi-square tests

$$o_foe_f - o_foe_f - \frac{(o_foe_f - o_foe_f)^2}{o_foe_f}$$

<table>
<thead>
<tr>
<th></th>
<th>$f_o$</th>
<th>$f_e$</th>
<th>$f_o - f_e$</th>
<th>$(f_o - f_e)^2$</th>
<th>$\frac{(f_o - f_e)^2}{f_e}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>115</td>
<td>$\bar{p} \times n_1 = 109.48$</td>
<td>5.52</td>
<td>30.4704</td>
<td>0.2783</td>
</tr>
<tr>
<td>99</td>
<td>$\bar{p} \times n_2 = 129.444$</td>
<td>-30.444</td>
<td>926.837136</td>
<td>7.1601</td>
<td></td>
</tr>
<tr>
<td>108</td>
<td>$\bar{p} \times n_1 = 83.076$</td>
<td>24.924</td>
<td>621.205776</td>
<td>7.4776</td>
<td></td>
</tr>
<tr>
<td>129</td>
<td>$(1 - \bar{p}) \times n_1 = 60.52$</td>
<td>-5.52</td>
<td>30.4704</td>
<td>0.5035</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>$(1 - \bar{p}) \times n_2 = 71.556$</td>
<td>30.444</td>
<td>926.837136</td>
<td>12.9526</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>$(1 - \bar{p}) \times n_3 = 45.924$</td>
<td>-24.924</td>
<td>621.205776</td>
<td>13.5268</td>
<td></td>
</tr>
</tbody>
</table>

41.8989

Since $41.8989 > 5.991$, we reject the null hypothesis and conclude that there is sufficient evidence to believe that there is a difference in the proportion of satisfied clients for the three IT companies.

**Example 10–3**

A group of researchers is interested in determining whether students who enrol in a university degree straight from school perform better than those who take a year off before beginning university (sometimes called a ‘gap year’). The following information was gathered from a sample of 400 students.

<table>
<thead>
<tr>
<th>Lowest grade received in first year of study</th>
<th>Enrolment group</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>School leaver</td>
<td>Gap year</td>
</tr>
<tr>
<td>HD</td>
<td>27</td>
<td>13</td>
</tr>
<tr>
<td>D</td>
<td>42</td>
<td>18</td>
</tr>
<tr>
<td>C</td>
<td>85</td>
<td>35</td>
</tr>
<tr>
<td>P</td>
<td>121</td>
<td>19</td>
</tr>
<tr>
<td>F</td>
<td>25</td>
<td>15</td>
</tr>
<tr>
<td>Total</td>
<td>300</td>
<td>100</td>
</tr>
</tbody>
</table>

At the 0.01 level of significance, determine whether there is evidence of a significant relationship between the lowest grade a student receives in their first year of study and whether they have come directly from school or have had a ‘gap year’. Interpret your result.

**Solution 10–3**

The hypotheses will be:

$H_0$: there is no relationship between lowest grade received in first year of study and enrolment group

$H_1$: there is a relationship between lowest grade received in first year of study and enrolment group
There are \( (r-1)(c-1) = (5-1)(2-1) = 4 \) degrees of freedom and so the critical value at a significance level of 0.01 is 13.277 and so the rejection rule is

Reject \( H_0 \) if \( \chi^2 > 13.277 \), otherwise do not reject \( H_0 \).

Note that in the table that follows all expected frequencies are at least 5 and so the chi-square test is appropriate.

\[
\begin{array}{cccccc}
 f_o & f_e & f_o - f_e & (f_o - f_e)^2 & (f_o - f_e)^2 / f_e \\
27 & \frac{40 \times 300}{400} = 30 & -3 & 9 & 0.3 \\
13 & \frac{40 \times 100}{400} = 10 & 3 & 9 & 0.9 \\
42 & \frac{60 \times 300}{400} = 45 & -3 & 9 & 0.2 \\
18 & \frac{60 \times 100}{400} = 15 & 3 & 9 & 0.6 \\
85 & 90 & -5 & 25 & 0.2778 \\
35 & 30 & 5 & 25 & 0.8333 \\
121 & 105 & 16 & 256 & 2.4381 \\
19 & 35 & -16 & 256 & 7.3143 \\
25 & 30 & -5 & 25 & 0.8333 \\
15 & 10 & 5 & 25 & 2.5 \\
\end{array}
\]

Since 16.1968 > 13.277, we reject the null hypothesis and conclude that there is sufficient evidence to indicate that there is a relationship between lowest grade received in first year of study and enrolment group.

Interpreting a conclusion such as this can be difficult. One way of assisting is to produce a table where observed and expected frequencies are included together such as is given in the table that follows.
Chi-square tests

Lowest grade received in first year of study | Enrolment group | Total |
--- | --- | --- |
|  | School leaver | Gap year |  |
| HD Obs | 27 | 13 | 40 |
(Exp) | (30) | (10) |
| D Obs | 42 | 18 | 60 |
(Exp) | (45) | (15) |
| C Obs | 85 | 35 | 120 |
(Exp) | (90) | (30) |
| P Obs | 121 | 19 | 140 |
(Exp) | (105) | (35) |
| F Obs | 25 | 15 | 40 |
(Exp) | (30) | (10) |
| Total | 300 | 100 | 400 |

We can now scan through this table looking for any unusual differences between observed and expected value. The students who started university immediately after leaving school seem to have more P grades than expected and those who enrolled after a gap year have less P grades that expected. The school leavers have slightly lower than expected numbers of HD, D, C and F grades, whilst the gap year students have slightly higher than expected HD, D, C and F grades. Because this is such a mixed result, it is difficult to determine which enrolment group is doing better in their first year at university. Although the chi-square test has revealed a significant difference, there appears to be no easy way of determining with any certainty which group has done better in their first year at university.

Perhaps lowest grade is not a good guide to overall performance. What would a better variable be to measure this? GPA (grade point average)? A more accurate measure of overall performance might allow a better interpretation of any differences between the two groups.

Discussion points

Discussion point 10–1

With other students, discuss the answers to Problems 12.57, 12.58 & 12.59 on p. 499 of your textbook (Problems 11.46, 11.48 & 11.49 on p. 496 of the 4th edition).

Discussion point 10–2

With other students, identify the assumptions that must be met in order to use the chi-square test. How practical do you believe it is to require that these assumptions be met in real world problems? Can you think of an example where the assumptions would not be met but the test might still be appropriate? What effect do you believe this would have on the outcome of the test? (Hint: consider the test statistic and the impact of small expected frequencies on this.)
Summary

Now that you have completed this module, turn back to the objectives at the beginning of the module. Have you achieved these objectives?

Ensure that you attempt the recommended problems in the list of review questions below and at least a sample of problems from the optional list. This will help you to identify any areas of difficulty you have in achieving the module’s objectives.

Review questions

Recommended problems


Levine et al. 5th edition: Do Questions 12.4, 12.6, 12.8 (a & b), 12.16, 12.18, 12.24, 12.26, 12.57 to 12.59 and 12.64.

Optional problems

Levine et al. 4th edition: Choose a selection of problems from Questions 11.1, 11.3 (a, b & c), 11.5 (a, b & c), 11.6(a & b), 11.7, 11.9 to 11.11, 11.13, 11.15 to 11.17, 11.19, 11.21, 11.23, 11.53 and 11.56 to 11.61.

Levine et al. 5th edition: Choose a selection of problems from 12.1 to 12.3, 12.5 (a, b & c), 12.7 (a, b & c), 12.9 (a, b & c), 12.10 (a & b), 12.11 to 12.15, 12.17, 12.19 to 12.23, 12.25 and 12.65 to 12.67.
Linear regression
## Contents

Introduction ................................................................. 11–3  
Objectives ................................................................. 11–3  
Readings ................................................................. 11–4  
Examples ................................................................. 11–4  
Discussion points ......................................................... 11–12  
Summary ................................................................. 11–12  
Review questions ......................................................... 11–13
Introduction

This week we will use a technique called **linear regression** to describe the relationship between two continuous variables. For example, we might expect that a person’s height and weight are linearly related. The taller a person is, the more they can be expected to weigh. Of course this will not be a perfect relationship (some tall people are quite thin and light), but in general we might expect a positive relationship between height and weight. Linear regression allows us to quantify this kind of relationship (using data) and then, given an individual’s height, make a prediction about what we would expect their weight to be. In this example, we assume that weight is dependent on the height of the person. So height would be described as an **independent variable** and weight as a **dependent variable**.

Another example of a linear relationship is exam marks scored by students and the number of hours of study they did before the exam. We expect hours of study to be the independent variable and exam mark the dependent variable (since we hope that the more someone studies the better they do on the exam). There are many potential examples of linear relationships in the business world also. Revenue (the dependent variable) is usually dependent on the quantity of goods sold (the independent variable). In certain circumstances, a linear relationship between these two variables might be appropriate. The income that an accountant earns (dependent variable) may be linearly related to the number of clients he or she has (independent variable). The profit earned by an investor (dependent variable) may be linearly related to the number of shares he has (independent variable).

Of course, not all relationships are going to be linear. It’s often the case that the more units of an item you produce, the less profit per item you’ll make. So in this case trying to describe the relationship with a straight line is not appropriate. Non-linear regression models could be used here. It’s also true that at times there may be a number of independent variables. This situation is called multiple regression. In this course, however, we’ll only be looking at simple linear relationships with one independent and one dependent variable (although interested students can read about some non-linear and multiple regression techniques in the textbook).

Objectives

On completion of this module you should be able to:

- calculate the coefficients of a simple linear regression equation
- interpret the intercept and slope of a simple linear regression equation
- calculate and interpret $r^2$
- perform a residual analysis and
- consider ethical issues and problems relating to linear regression models.
Examples

Example 11–1
An accountant has been considering the method he uses to charge his clients for the services he provides. He has begun to realise that his current method is overly complex and is actually leading to the loss of some clients. In order to improve this situation he collects data from a sample of 9 recent clients (see the table below). He then uses simple linear regression to describe the relationship between the amount the client is charged and the client’s annual income.

(a) Repeat the accountant’s analysis (i.e., set up a scatter diagram and find the linear regressions coefficients).

(b) Interpret the meaning of the coefficients \(b_0\) and \(b_1\).

(c) Predict the amount the accountant should charge a client whose annual income is $450,000.
Solution 11–1

We’ll demonstrate how to do these calculations by hand, using the formulae given in Sections 13.2 and 13.3 of the textbook (Section 12.10 of the 4th edition). Normally a computer is used (for obvious reasons), but remember that in the exam, you may be asked to demonstrate your understanding of the techniques by doing some of the calculations (for small data sets) by hand. We’ll give a quick summary of the results found using PHStat2, but leave it as an exercise for the reader to follow the instructions in the text and obtain these results.

We should always begin a regression problem by producing a scatterplot of the data and looking to see whether it is sensible to fit a straight line to the data. The diagram below reveals what appears to be a fairly clear linear relationship between client income and the amount the client is charged. This indicates it is sensible for us to continue with a linear regression analysis.

<table>
<thead>
<tr>
<th>Client income (in 1000’s of $)</th>
<th>Charge ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td>80</td>
<td>260</td>
</tr>
<tr>
<td>590</td>
<td>704</td>
</tr>
<tr>
<td>476</td>
<td>598</td>
</tr>
<tr>
<td>145</td>
<td>385</td>
</tr>
<tr>
<td>79</td>
<td>340</td>
</tr>
<tr>
<td>298</td>
<td>493</td>
</tr>
<tr>
<td>876</td>
<td>1027</td>
</tr>
<tr>
<td>330</td>
<td>435</td>
</tr>
<tr>
<td>1201</td>
<td>1428</td>
</tr>
</tbody>
</table>

Scatterplot of charge for services versus client income
Now, we can use the table below to calculate all the necessary sums for computing the two coefficients of the linear regression equation.

<table>
<thead>
<tr>
<th>Income</th>
<th>Charge</th>
<th>$X^2$</th>
<th>$Y^2$</th>
<th>$XY$</th>
</tr>
</thead>
<tbody>
<tr>
<td>80</td>
<td>260</td>
<td>6400</td>
<td>67600</td>
<td>20800</td>
</tr>
<tr>
<td>590</td>
<td>704</td>
<td>348100</td>
<td>495616</td>
<td>415360</td>
</tr>
<tr>
<td>476</td>
<td>598</td>
<td>226576</td>
<td>357604</td>
<td>284648</td>
</tr>
<tr>
<td>145</td>
<td>385</td>
<td>21025</td>
<td>148225</td>
<td>55825</td>
</tr>
<tr>
<td>79</td>
<td>340</td>
<td>638401</td>
<td>115600</td>
<td>271660</td>
</tr>
<tr>
<td>298</td>
<td>493</td>
<td>88804</td>
<td>243049</td>
<td>146914</td>
</tr>
<tr>
<td>876</td>
<td>1027</td>
<td>767376</td>
<td>1055729</td>
<td>899652</td>
</tr>
<tr>
<td>330</td>
<td>435</td>
<td>108900</td>
<td>189225</td>
<td>143550</td>
</tr>
<tr>
<td>1201</td>
<td>1428</td>
<td>1442401</td>
<td>2039184</td>
<td>1715028</td>
</tr>
</tbody>
</table>

**Sums:** 4075 5670 3015823 4710832 3708637

(a) Reading from the table, we see that $\sum_{i=1}^{n} X_i^2 = 3015823$, $\sum_{i=1}^{n} Y_i^2 = 4710832$ and $\sum_{i=1}^{n} X_i Y_i = 3708637$. Then

$$SSXY = \sum_{i=1}^{n} X_i Y_i - \frac{\left(\sum_{i=1}^{n} X_i\right) \left(\sum_{i=1}^{n} Y_i\right)}{n}$$

$$= 3708637 - \frac{(4075)(5670)}{9}$$

$$= 1141387$$

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

$$= 3015823 - \frac{(4075)^2}{9}$$

$$= 1170753.55556$$

$$b_1 = \frac{SSXY}{SSX} = \frac{1141387}{1170753.55556} = 0.97491654$$

(Note: 8 decimal places have been used here to avoid a rounding error later.)

$$X = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{4075}{9} = 452.77777778$$
Linear regression

\[
\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} = \frac{5670}{9} = 630
\]

\[
b_0 = \bar{Y} - b_1\bar{X} = 630 - 0.9749\times 452.77777778 = 188.5795 \text{ (to 4 decimal places)}.
\]

So the fitted equation is: \( \hat{Y}_i = 188.58 + 0.97X_i \)

(Note that the two coefficients have now been rounded to 2 decimal places for simplicity in reading the fitted equation.)

Next we should always plot the fitted line next to the data to check the fit. The graph below demonstrates the result. It can be seen that the linear regression equation fits the data set very well.

**Plotting the regression equation with the data**

Once the scatterplot of the data has been produced in Excel, this can be done by selecting the data points on the chart, going to the Chart menu (in Excel), selecting Add Trendline, ensuring that the Linear graph is selected and clicking okay. If you like, you can include the fitted trend equation on the chart by clicking on the Options tab (on the Add Trendline dialog box) and selecting Display equation on chart. See EH12.2 in the text for more information.

Excel and PHStat2 output for the analysis of Example 11–1 is given below. Make sure that you know how to produce this output.
Regression Analysis

<table>
<thead>
<tr>
<th>Regression Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple R</td>
</tr>
<tr>
<td>R Square</td>
</tr>
<tr>
<td>Adjusted R Square</td>
</tr>
<tr>
<td>Standard Error</td>
</tr>
<tr>
<td>Observations</td>
</tr>
</tbody>
</table>

ANOVA

<table>
<thead>
<tr>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>Significance F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>1112757.059</td>
<td>1112757.059</td>
<td>299.877468</td>
</tr>
<tr>
<td>Residual</td>
<td>7</td>
<td>25974.94059</td>
<td>3710.705799</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>8</td>
<td>1138732</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Standard Error</th>
<th>t Stat</th>
<th>P-value</th>
<th>Lower 95%</th>
<th>Upper 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>188.5794577</td>
<td>32.58946182</td>
<td>5.786516473</td>
<td>0.000672749</td>
<td>111.5176259 265.6412894</td>
</tr>
<tr>
<td>Income</td>
<td>0.974916535</td>
<td>0.056298331</td>
<td>17.31697052</td>
<td>5.26366E-07</td>
<td>0.841792137 1.108040934</td>
</tr>
</tbody>
</table>

(b) The value \( b_0 = 188.58 \) gives the portion of the charge which occurs no matter what the value of the client’s annual income. This means that there is effectively a base rate of charge of $188.58. The value \( b_1 = 0.97 \) indicates that for every thousand dollar increase in the client’s annual income, an extra $0.97 is added to the total charge.

(c) Since the fitted equation has been calculated based on an income expressed as thousands of dollars, we substitute \( X = 450 \) into the fitted equation (not \( X = 450,000 \)). Therefore

\[
\hat{y} = 188.5794577 + 0.974916535(450) = 627.29.
\]

So based on this linear regression equation, the accountant should charge $627.29 to a client whose annual income is $450,000. (Notice that we have used all the decimal places in the calculation to avoid a rounding error.)

Example 11–2

Using the information from Example 11–1, calculate the coefficient of determination and interpret its meaning. Determine the standard error of the estimate. How useful is the regression model for predicting the charge?
Solution 11–2

Using the information calculated in Example 11–1 we can find SSR and SST as follows (these values can also be read from the Excel output above):

\[
SSR = b_0 \sum_{i=1}^{n} Y_i + b_1 \sum_{i=1}^{n} X_i Y_i - \frac{\left( \sum_{i=1}^{n} Y_i \right)^2}{n} = 188.5795(5670) + 0.97491654(3708637) - \frac{(5670)^2}{9} = 1112757.0594
\]

\[
SST = \sum_{i=1}^{n} Y_i - \frac{\left( \sum_{i=1}^{n} Y_i \right)^2}{n} = 4710832 - \frac{(5670)^2}{9} = 1138732
\]

Note that we’ve used all the decimal places for \(b_0\) and \(b_1\) to try to avoid rounding errors in the final coefficient of determination answer.

Then

\[
\frac{SSR}{SST} = \frac{1112757.0594}{1138732} = 0.9772 \text{ (to 4 decimal places).}
\]

This indicates that approximately 97.72% of the variation in charges to clients is explained by the variability in the annual incomes.

To calculate the standard error of the estimate we first find

\[
SSE = \sum_{i=1}^{n} Y_i^2 - b_0 \sum_{i=1}^{n} Y_i - b_1 \sum_{i=1}^{n} X_i Y_i
\]

\[
= 4710832 - 188.5795(5670) - 0.97491654(3708637)
\]

\[
= 25974.6828
\]

and then

\[
S_{xx} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{25974.6828}{7}} = 60.9153 \text{ (to 4 decimal places).}
\]

Therefore the standard error of the estimate is $60.92. This is a measure of the variability around the line of regression.

Since the \(r^2\) value is 0.9772 (this is very high), the regression line should be very useful for predicting the charge the accountant should apply to clients. This good \(r^2\) value reflects the close fit of the data to the regression line that we saw earlier.

Note that in real life data sets, it is very rare to obtain such a high \(r^2\) value. Notice also the increasing difficulty of maintaining enough decimal places to avoid rounding errors in calculations. The more steps we take, the more we are forced to round slightly each step, the less reliable our final answers will be. This is an obvious reason to use computing technology, rather than rely on hand calculations which are open to error!
Example 11–3

Using the data and regression equation from Example 11–1, determine the adequacy of the fit of the model. Determine whether the assumptions of the regression model have been seriously violated. Would you recommend that the accountant use this model to set the charge for clients?

Solution 11–3

A table and plots of residual values are given below. These are calculated by finding the difference between the actual charge to each client and the value the regression equation estimates should be charged for the same client. So, for example, the first client (with the income of $80,000) was charged $260. Using the regression model, they would have been charged \( \hat{Y} = 188.5794577 + 0.974916535(80) = 266.57 \) (to 2 decimal places). The difference between these two is \( 260 – 266.57 = -6.57 \). The table below gives the same value including 5 decimal places.

The data do not appear to be completely randomly distributed (the data appear to follow a rough ‘u’ shape). Small and large income values correspond to larger, positive residuals, whilst mid-income values correspond to negative residuals. This suggests that the data might be better described by another model (perhaps a quadratic).

<table>
<thead>
<tr>
<th>Income</th>
<th>Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>80</td>
<td>-6.57278</td>
</tr>
<tr>
<td>590</td>
<td>-59.7802</td>
</tr>
<tr>
<td>476</td>
<td>-54.6397</td>
</tr>
<tr>
<td>145</td>
<td>55.05764</td>
</tr>
<tr>
<td>79</td>
<td>74.40214</td>
</tr>
<tr>
<td>298</td>
<td>13.89541</td>
</tr>
<tr>
<td>876</td>
<td>-15.6063</td>
</tr>
<tr>
<td>330</td>
<td>-75.3019</td>
</tr>
<tr>
<td>1201</td>
<td>68.54578</td>
</tr>
</tbody>
</table>

The residual plot does not reveal any violation of the assumption of homoscedasticity (i.e., the variation around the regression line does not appear to deviate significantly from a constant value for values of income).
A normal probability plot (see below) does not demonstrate any major deviation from the assumption of normality. The s shape in this plot, however, does indicate the possibility of a rectangular shaped distribution.
Plotting the residuals in the order they were observed (see below), does not reveal any particular pattern and so the assumption of independence seems not to have been violated.

We can conclude by saying that a quadratic curve may better model this particular data set, but that otherwise the assumptions of the simple linear regression model have not been clearly violated.

**Discussion points**

**Discussion point 11–1**

Complete Problems 13.23 and 13.24 from p. 533 of the textbook (Problems 12.20 and 12.21 on p. 532 of the 4th edition). With other students, discuss your answers and comment on any violations of assumptions that appear to be occurring in these data sets. What possible solutions can you offer to any violations?

**Discussion point 11–2**

With other students, discuss the answers to Problems 13.68, 13.69 and 13.70 on p. 556 of the textbook (Problems 12.61 and 12.62 on p. 562 of the 4th edition). How practical do you think it is to have to have met these assumptions in solving real world problems. Give examples to justify your response.

**Summary**

Now that you have completed this module, turn back to the objectives at the beginning of the module. Have you achieved these objectives?
Ensure that you attempt the recommended problems in the list of review questions below and at least a sample of problems from the optional list. This will help you to identify any areas of difficulty you have in achieving the module’s objectives.

**Review questions**

**Recommended problems**

Levine et al. 4th edition: Do Questions 12.2, 12.4, 12.6, 12.9 to 12.14, 12.16, 12.20 to 12.22, 12.24, 12.57 to 12.62, 12.66 (a to j & o only), 12.68 (a to g only), 12.70 (a to g, l only) and 12.80.

Levine et al. 5th edition: Do Questions 13.4, 13.6, 13.16, 13.18, 13.23, 13.24, 13.26, 13.28, 13.64 to 13.70, 13.74 (a to f & k only), 13.76 (a to e only), 13.78 (a to e only) and 13.92.

**Optional problems**

Levine et al. 4th edition: Choose a selection of problems from Questions 12.1, 12.3, 12.5, 12.7, 12.8, 12.15, 12.17 to 12.19, 12.23, 12.25 to 12.27, 12.67 (a to j, l, q only), 12.69 (a to g & l only), 12.71 (a to i & k only), 12.72, 12.74 (a to h), 12.75 (a to h only), 12.76 (a to h only), 12.77 (a to g & j to o only), 12.78 (a to g & j to m only), 12.79 and 12.83.

Levine et al. 5th edition: Choose a selection of problems from Questions 13.1 to 13.3, 13.5, 13.7 to 13.15, 13.17, 13.19 to 13.22, 13.25, 13.27, 13.29 to 13.31, 13.73, 13.75 (a to f, m only), 13.77 (a to e only), 13.79 (a to f, h only), 13.80, 13.81 (a to e, k only), 13.82 (a to e only), 13.83 (a to e only), 13.84 (a to f), 13.85 (a to e only), 13.90 (a to e, j only) and 13.93 (a only).
Time series forecasting and index numbers
Contents

Introduction ............................................................................................................. 12–3
Objectives ................................................................................................................ 12–3
Readings .................................................................................................................. 12–3
Examples ................................................................................................................. 12–4
Summary ............................................................................................................... 12–17
Review questions ................................................................................................... 12–17
Introduction

It is very common, in business and financial disciplines, to find that data is collected at regular intervals. Examples of this are the closing price of a particular stock each day, the profit obtained in a store each week or an individual’s income each year. This kind of data is called time series data. Often we want to use such data to make a prediction of what will happen in the future. This is called forecasting and requires an understanding of not only the way the data has behaved until now, but an assumption that the data will continue to behave in the same way. This week we will examine various ways of describing time series data: moving averages, exponential smoothing and linear trends. These are then used to forecast data values.

The second topic we will examine this week is index numbers. Most of us are familiar with one particular index number: the consumer price index (CPI). The CPI allows us to measure the relative change in a particular basket of goods over time. Index numbers have many other applications too, some of which we will refer to this week.

Objectives

On completion of this module you should be able to:

- describe a multiplicative times series model
- use moving averages and exponential smoothing
- use these models to make simple forecasts
- use least squares trend-fitting and forecasting
- choose appropriate forecasting models for data particular sets
- explore common problems that occur with time-series forecasting and
- understand and apply index numbers.

Readings

Source

Textbook
Levine et al. 4th edition
Sections 15.1 to 15.4, 15.6 and 15.8 to 15.9 & Excel Handbook
Sections EH15.1 to 15.3 & 15.5 (pp. 717–719)

Or

Textbook
Levine et al. 5th edition
Sections 16.1 to 16.4, 16.6 and 16.8 to 16.9 & Excel Companion
Sections E16.1 to 16.3 (pp. 702–703)
Examples

Example 12–1

The following data represents the number of new clients obtained by a large company over 15 years.

<table>
<thead>
<tr>
<th>Year</th>
<th>New Clients</th>
<th>Year</th>
<th>New Clients</th>
</tr>
</thead>
<tbody>
<tr>
<td>1991</td>
<td>35</td>
<td>1999</td>
<td>76</td>
</tr>
<tr>
<td>1992</td>
<td>45</td>
<td>2000</td>
<td>60</td>
</tr>
<tr>
<td>1993</td>
<td>41</td>
<td>2001</td>
<td>53</td>
</tr>
<tr>
<td>1994</td>
<td>40</td>
<td>2002</td>
<td>57</td>
</tr>
<tr>
<td>1995</td>
<td>54</td>
<td>2003</td>
<td>70</td>
</tr>
<tr>
<td>1996</td>
<td>72</td>
<td>2004</td>
<td>62</td>
</tr>
<tr>
<td>1997</td>
<td>75</td>
<td>2005</td>
<td>55</td>
</tr>
<tr>
<td>1998</td>
<td>74</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(a) Plot the data on a chart.

(b) Fit a 3-year moving average to the data, plotting the results.

(c) Use a smoothing coefficient of $W = 0.4$ to exponentially smooth the data, again plotting the results. What forecast can you make for 2006?

(d) Use a smoothing coefficient of $W = 0.6$ to exponentially smooth the data, again plotting the results. What forecast can you make for 2006?

(e) Compare the results of (c) and (d).
Solution 12–1

(a) A time-series plot of the data is given below.

(b) The 3-year moving averages are calculated as follows:

\[ MA(3) = \frac{Y_1 + Y_2 + Y_3}{3} = \frac{35 + 45 + 41}{3} = 40.33 \text{ (to 2 decimal places)}. \]

\[ MA(3) = \frac{Y_2 + Y_3 + Y_4}{3} = \frac{45 + 41 + 40}{3} = 42 \]

\[ MA(3) = \frac{Y_3 + Y_4 + Y_5}{3} = \frac{41 + 40 + 54}{3} = 45 \]

etc.
The table below summarises all the values and illustrates how the data is centred.

<table>
<thead>
<tr>
<th>Year</th>
<th>New Clients</th>
<th>MA(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1991</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>1992</td>
<td>45</td>
<td>40.33</td>
</tr>
<tr>
<td>1993</td>
<td>41</td>
<td>42</td>
</tr>
<tr>
<td>1994</td>
<td>40</td>
<td>45</td>
</tr>
<tr>
<td>1995</td>
<td>54</td>
<td>55.33</td>
</tr>
<tr>
<td>1996</td>
<td>72</td>
<td>67</td>
</tr>
<tr>
<td>1997</td>
<td>75</td>
<td>73.67</td>
</tr>
<tr>
<td>1998</td>
<td>74</td>
<td>75</td>
</tr>
<tr>
<td>1999</td>
<td>76</td>
<td>70</td>
</tr>
<tr>
<td>2000</td>
<td>60</td>
<td>63</td>
</tr>
<tr>
<td>2001</td>
<td>53</td>
<td>56.67</td>
</tr>
<tr>
<td>2002</td>
<td>57</td>
<td>60</td>
</tr>
<tr>
<td>2003</td>
<td>70</td>
<td>63</td>
</tr>
<tr>
<td>2004</td>
<td>62</td>
<td>62.33</td>
</tr>
<tr>
<td>2005</td>
<td>55</td>
<td></td>
</tr>
</tbody>
</table>

A graph of the time series alongside the moving averages is also given here.
(c) & (d) The exponentially smoothed values when \( W = 0.4 \) are found as follows:

\[
E_1 = Y_1 = 35
\]

\[
E_2 = WY_1 + (1-W)E_1 = 0.4(45) + 0.6(35) = 39
\]

\[
E_3 = WY_2 + (1-W)E_2 = 0.4(41) + 0.6(39) = 39.8
\]

\[
E_4 = WY_3 + (1-W)E_3 = 0.4(40) + 0.6(39.8) = 39.88
\]

etc.

The table of all these values is given below. Remember that to use the exponential smoothing facility in Excel, you must enter \((1-W)\) as the dampening factor.

<table>
<thead>
<tr>
<th>Year</th>
<th>New Clients</th>
<th>ES(0.4)</th>
<th>ES(0.6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1991</td>
<td>35</td>
<td>35</td>
<td>35</td>
</tr>
<tr>
<td>1992</td>
<td>45</td>
<td>39</td>
<td>41</td>
</tr>
<tr>
<td>1993</td>
<td>41</td>
<td>39.8</td>
<td>41</td>
</tr>
<tr>
<td>1994</td>
<td>40</td>
<td>39.88</td>
<td>40.4</td>
</tr>
<tr>
<td>1995</td>
<td>54</td>
<td>45.528</td>
<td>48.56</td>
</tr>
<tr>
<td>1996</td>
<td>72</td>
<td>56.1168</td>
<td>62.624</td>
</tr>
<tr>
<td>1997</td>
<td>75</td>
<td>63.6701</td>
<td>70.0496</td>
</tr>
<tr>
<td>1998</td>
<td>74</td>
<td>67.8021</td>
<td>72.4198</td>
</tr>
<tr>
<td>1999</td>
<td>76</td>
<td>71.0812</td>
<td>74.5679</td>
</tr>
<tr>
<td>2000</td>
<td>60</td>
<td>66.6487</td>
<td>65.8272</td>
</tr>
<tr>
<td>2001</td>
<td>53</td>
<td>61.1892</td>
<td>58.1309</td>
</tr>
<tr>
<td>2002</td>
<td>57</td>
<td>59.5136</td>
<td>57.4524</td>
</tr>
<tr>
<td>2003</td>
<td>70</td>
<td>63.7081</td>
<td>64.9809</td>
</tr>
<tr>
<td>2004</td>
<td>62</td>
<td>63.0249</td>
<td>63.1924</td>
</tr>
<tr>
<td>2005</td>
<td>55</td>
<td>59.8149</td>
<td>58.2770</td>
</tr>
</tbody>
</table>

A graph of the time series alongside the exponentially smoothed values is also given here.
Remember that we make a forecast using the final smoothed value (which in this case comes from the year 2005).

Using a smoothing coefficient of $W = 0.4$, we can make a forecast of 59.8149 or approximately 60 new clients for the year 2006.

Using a smoothing coefficient of $W = 0.6$ we can make a forecast of 58.2770 or approximately 58 new clients for the year 2006.

(e) The results from the two forecasts are very similar, but with the $W = 0.4$ smoothing coefficient resulting in a slightly larger forecast. This is because this value places more weight on the previous smoothed values, than on the current data. The $W = 0.6$ smoothing coefficient results in a smaller forecast, since it places more weight on the current data value and less on the previous smoothed value. This can be seen in the graph of the two smoothed series above, where we can see that the line with the square symbols ($W = 0.4$) does not hug the data as closely as does the line with the triangle symbols ($W = 0.6$).

Note: although Section 16.4 of the textbook (Section 15.4 in the 4th edition) covers a number of trend models (linear, quadratic and exponential), you will only be expected to use a linear trend in assessment items for this course. The material relating to quadratic and exponential trends, however, can be read for interest and to offer a deeper understanding of ways of modelling trends in data.
Example 12–2

The sales figures of a sailing boat manufacturing company have been increasing fairly steadily over the past twenty years. In 1987 they sold 101 of their Q-type yachts. By 2006 sales had increased to 180 of the same yacht. Plot the data (given in the table below) on a chart and use a linear trend equation to model the sales figures. Plot this line, with the data on another chart. What sales figure would you forecast for 2007? What comments would you make to the company regarding the accuracy of this forecast?

<table>
<thead>
<tr>
<th>Year</th>
<th>Sales of Q-type yachts</th>
</tr>
</thead>
<tbody>
<tr>
<td>1987</td>
<td>101</td>
</tr>
<tr>
<td>1988</td>
<td>98</td>
</tr>
<tr>
<td>1989</td>
<td>105</td>
</tr>
<tr>
<td>1990</td>
<td>123</td>
</tr>
<tr>
<td>1991</td>
<td>121</td>
</tr>
<tr>
<td>1992</td>
<td>129</td>
</tr>
<tr>
<td>1993</td>
<td>123</td>
</tr>
<tr>
<td>1994</td>
<td>134</td>
</tr>
<tr>
<td>1995</td>
<td>116</td>
</tr>
<tr>
<td>1996</td>
<td>142</td>
</tr>
<tr>
<td>1997</td>
<td>159</td>
</tr>
<tr>
<td>1998</td>
<td>145</td>
</tr>
<tr>
<td>1999</td>
<td>147</td>
</tr>
<tr>
<td>2000</td>
<td>145</td>
</tr>
<tr>
<td>2001</td>
<td>164</td>
</tr>
<tr>
<td>2002</td>
<td>150</td>
</tr>
<tr>
<td>2003</td>
<td>173</td>
</tr>
<tr>
<td>2004</td>
<td>185</td>
</tr>
<tr>
<td>2005</td>
<td>170</td>
</tr>
<tr>
<td>2006</td>
<td>180</td>
</tr>
</tbody>
</table>

Solution 12–2

The first step in analysing any data set is always to produce a graph of it. This allows us to check for the appropriateness of the model. This particular data set appears to have a roughly linear increase in sales. There is, of course, a fair amount of random error surrounding this line which we will not attempt to model here.
Using PHStat2 and the simple linear regression procedure, the following output it obtained. Note that we have recoded 1987 as $X = 0$, 1988 as $X = 1$, etc.

We can see that the linear trend is found to be $\hat{Y}_i = 100.9857 + 4.1594X_i$ (rounding all coefficients to four decimal places) and that $r^2 = 0.8897$, indicating that the linear trend is a good fit to the data.

Regression Analysis

Regression Statistics

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple R</td>
<td>0.943246706</td>
</tr>
<tr>
<td>R Square</td>
<td>0.889714349</td>
</tr>
<tr>
<td>Adjusted R Square</td>
<td>0.883587368</td>
</tr>
<tr>
<td>Standard Error</td>
<td>8.901010428</td>
</tr>
<tr>
<td>Observations</td>
<td>20</td>
</tr>
</tbody>
</table>

ANOVA

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>Significance F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>11504.89624</td>
<td>11504.89624</td>
<td>145.2125282</td>
<td>4.717E-10</td>
</tr>
<tr>
<td>Residual</td>
<td>18</td>
<td>1426.103759</td>
<td>79.22798663</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>19</td>
<td>12931</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Coefficients

<table>
<thead>
<tr>
<th></th>
<th>Standard Error</th>
<th>t Stat</th>
<th>P-value</th>
<th>Lower 95%</th>
<th>Upper 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>100.9857143</td>
<td>3.835853092</td>
<td>26.32679403</td>
<td>8.0049E-16</td>
<td>92.92688599</td>
</tr>
<tr>
<td>X</td>
<td>4.159398496</td>
<td>0.345166379</td>
<td>12.0504161</td>
<td>4.717E-10</td>
<td>3.434230843</td>
</tr>
</tbody>
</table>
The fitted trend line alongside the data is given in the chart below. We can see that this model appears to describe the trend in the sales data reasonably well. A graph of the residuals (see below) does not appear to reveal any patterns and so we can conclude that the linear trend line is an adequate model for the sales data.

In order to make a forecast for the year 2007, we simply substitute the appropriate $X$ value in to the fitted trend line. In this case, $X = 20$ and so

$$\hat{Y} = 100.9857 + 4.1594 \times 20 = 180.1737$$

Therefore we can expect sales of about 180 Q-type yachts in 2007.
Given that a number of factors have indicated that this trend line is a reasonably good model of the sales data, we can be reasonably confident that this forecast is accurate. This is, however, a point estimate, and as such is unlikely to pinpoint the exact sales for 2006. A more appropriate method of predicting sales might be to include a prediction interval; however, this is beyond the scope of this course (read Section 13.8 on p. 546 of the textbook [Section 12.8 on p. 547 of the 4th edition] if you are interested in knowing how to do this). In presenting the forecast of 180 yachts sales in 2007, it might be wise to indicate that this is a point estimate, and as such is subject to a small margin of error. By doing so, you are less likely to be sued for inaccurate forecasting!!

**Example 12–3**

Using the data and fitted linear trend line from Example 12–2, calculate the standard error of the estimate ($S_{xy}$) and the mean absolute deviation (MAD). If a quadratic trend line were to be fitted to this same data set, the standard error of the estimate would be $S_{xy} = 9.1584$ and the MAD would be $MAD = 6.7689$. A graph of the residuals for this model is given below. Using your calculations for the linear trend model, compare these with the quadratic trend model and, considering the impact of the principal of parsimony, recommend the better model for this data set.

![Residuals for Quadratic Trend](image)

**Solution 12–3**

We can see that the graphs of residuals for the linear and quadratic models are both very similar (to the naked eye they appear almost identical) and so neither model appears best based solely on residuals graphs.

Reading from the PHStat2 output given in Example 12–2, the standard error of the estimate is $S_{xy} = 8.9010$ (rounded to 4 decimal places). By using residuals stored in Example 2 working, taking the absolute value of these and then the average of the absolute values, we get $MAD = 6.7594$ (rounded to 4 decimal places).
Summarising this information, and the values obtained for a quadratic model in a table, may help in making a comparison.

<table>
<thead>
<tr>
<th>Linear trend</th>
<th>Quadratic trend</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_{xx}$</td>
<td>8.9010</td>
</tr>
<tr>
<td>MAD</td>
<td>6.7594</td>
</tr>
<tr>
<td></td>
<td>9.1584</td>
</tr>
<tr>
<td></td>
<td>6.7689</td>
</tr>
</tbody>
</table>

Examining these values, we can see that the MAD value has actually slightly increased for the quadratic model and so too has the standard error of the estimate. This indicates, that the added complexity of using the quadratic model has not been worthwhile. The principle of parsimony and the MAD value both indicate that the linear trend is the best model to select for this data set.

Note that if we looked at our original plot of the data set, we would have no reason to suspect that these data would be well modelled by a quadratic. If this were the case we would expect to see a curved trend, but these data appears to be linearly increasing.

### Example 12–4

The data given below records the prices of a 20-dose packet of a particular migraine treatment drug over a fifteen year period.

(a) Use this data to calculate a simple price index for the years 1985 to 1999 using 1985 as a base year. Interpret the 1999 value.

(b) Recalculate the simple price index using 1990 as the base year. Interpret the 1999 value.

(c) Describe any trends in the price of the migraine treatment drug from 1985 to 1999.

<table>
<thead>
<tr>
<th>Year</th>
<th>Price ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1985</td>
<td>3.85</td>
</tr>
<tr>
<td>1986</td>
<td>4.15</td>
</tr>
<tr>
<td>1987</td>
<td>4.25</td>
</tr>
<tr>
<td>1988</td>
<td>4.30</td>
</tr>
<tr>
<td>1989</td>
<td>6.20</td>
</tr>
<tr>
<td>1990</td>
<td>5.90</td>
</tr>
<tr>
<td>1991</td>
<td>6.70</td>
</tr>
<tr>
<td>1992</td>
<td>7.00</td>
</tr>
<tr>
<td>1993</td>
<td>7.00</td>
</tr>
<tr>
<td>1994</td>
<td>9.00</td>
</tr>
<tr>
<td>1995</td>
<td>12.00</td>
</tr>
<tr>
<td>1996</td>
<td>13.50</td>
</tr>
<tr>
<td>1997</td>
<td>13.85</td>
</tr>
<tr>
<td>1998</td>
<td>17.20</td>
</tr>
<tr>
<td>1999</td>
<td>19.90</td>
</tr>
</tbody>
</table>
Solution 12–4

(a) The following table illustrates the calculation of the simple price index. Using the formula \( I = \frac{P}{P_{base}} \times 100 \), the value for 1993, for example, was found via

\[ I_{1993} = \frac{7.00}{3.85} \times 100 = 181.8 \]. (All index values have been rounded to one decimal place.)

<table>
<thead>
<tr>
<th>Year</th>
<th>Price</th>
<th>PI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1985</td>
<td>3.85</td>
<td>100</td>
</tr>
<tr>
<td>1986</td>
<td>4.15</td>
<td>107.8</td>
</tr>
<tr>
<td>1987</td>
<td>4.25</td>
<td>110.4</td>
</tr>
<tr>
<td>1988</td>
<td>4.30</td>
<td>111.7</td>
</tr>
<tr>
<td>1989</td>
<td>6.20</td>
<td>161.0</td>
</tr>
<tr>
<td>1990</td>
<td>5.90</td>
<td>153.2</td>
</tr>
<tr>
<td>1991</td>
<td>6.70</td>
<td>174.0</td>
</tr>
<tr>
<td>1992</td>
<td>7.00</td>
<td>181.8</td>
</tr>
<tr>
<td>1993</td>
<td>7.00</td>
<td>181.8</td>
</tr>
<tr>
<td>1994</td>
<td>9.00</td>
<td>233.8</td>
</tr>
<tr>
<td>1995</td>
<td>12.00</td>
<td>311.7</td>
</tr>
<tr>
<td>1996</td>
<td>13.50</td>
<td>350.6</td>
</tr>
<tr>
<td>1997</td>
<td>13.85</td>
<td>359.7</td>
</tr>
<tr>
<td>1998</td>
<td>17.20</td>
<td>446.8</td>
</tr>
<tr>
<td>1999</td>
<td>19.90</td>
<td>516.9</td>
</tr>
</tbody>
</table>

Therefore the price of the migraine treatment drug was 416.9% higher in 1999 than it was in 1985.

(b) We’ll demonstrate here how the hand calculations proceed and leave it as an exercise to verify these using Excel. Using the formula \( I_{\text{new}} = \frac{I_{\text{old}}}{I_{\text{new base}}} \times 100 \), the value for 1993, for example, was found via \( I_{1993} = \frac{181.8}{153.2} \times 100 = 118.6 \).
The price of the migraine treatment drug was 237.3% higher in 1999 than it was in 1990.

(c) In order to describe a trend, the first thing we must do is graph the data. Looking at the graph below, we note there is a very obvious upward trend in the price from 1985 to 1999, and that this trend seems to be becoming more steeply positive as time goes by. This upward turn indicates that a linear trend is unlikely to model the data particularly well (a quadratic trend, for example, may be a better alternative in this case).
As anticipated, the linear trend line illustrated below does not fit the data set particularly well.

Note that we can use the quadratic trend option (by specifying a polynomial fit with order 2) for adding a trendline to a data set with the following results. Because the quadratic models the amplifying increase in price over time, the goodness of fit (and $R^2$ value) are improved. Note that fitting a quadratic trend line is not required knowledge for the exam.
Summary

Now that you have completed this module, turn back to the objectives at the beginning of the module. Have you achieved these objectives?

Ensure that you attempt the recommended problems in the list of review questions below and at least a sample of problems from the optional list. This will help you to identify any areas of difficulty you have in achieving the module’s objectives.

Review questions

Recommended problems

Levine et al. 4th edition: Do Questions 15.2, 15.4, 15.8, 15.10, 15.12, 15.32, 15.50, 15.54, 15.56, 15.60 to 15.63, 15.65, 15.67 to 15.72, 15.74 (a to c, i, j & m).


Optional problems

Levine et al. 4th edition: Choose a selection of problems from Questions 15.1, 15.3, 15.5 to 15.7, 15.9, 15.11, 15.13, 15.4, 15.33, 15.34, 15.51 to 15.53, 15.55, 15.57 to 15.59, 15.73, 15.75 (a, b, c, i, j, k & m), 15.76 (a, b, c, i, j, k & m), 15.77 (a, b, c, i, j, k & m), 15.78 (a, b, i, j, l & m), 15.79 and 15.82.