Problems for Meaning from Data

R. Nazim Khan

August 8, 2024

Contents

Chapter 1

Data Collection

1. A business analyst is investigating the reasons that owner/operator small businesses fail.

   (a) What data should she collect?

   (b) Classify the data as numeric: discrete or continuous; or categorical: nominal or ordinal.

2. As part of a survey you need the respondent's demographic information. Write survey items (questions) to obtain the following.

   - Age
   - Sex
   - Place of residence
   - Annual income

   Now exchange your questions to another student and complete the answers. When you have both finished, compare answers and discuss. Are the answers what you expected?

3. Identify the type(s) of bias in each of the following situations.

   (a) A radio talk back poll.

   (b) A survey questionnaire sent out by email.

   (c) A random sample of readers of the Business magazine surveyed for business confidence.

Chapter 2

Exploratory Data Analysis: Graphical Presentation and Numerical Summaries of Data, Summation Notation, Transformation of Data

Graphical Presentation and Numerical Summaries of Data, Measures of Central Tendency

From the **Start** menu select **Mathematics and Statistics Software** and then `R`. Select either the 32 or 64 bit version. An `R` console appears. Type `library(Rcmdr)` and hit Return. The `R Commander` window will appear. The window has two parts. The top half lists any `R` commands that are run, and the bottom half contains any output. The top half has two tabs, `R Script` and `R Markdown`. We will be mainly concerned withe the `R Script` tab.

1. The data set *House*, within the file *House.txt*, consists of the following information for properties sold by two competing real estate agents in 2005.

    - Material: What building material was used in the construction of the house? Fibro, single brick or double brick.
    - Bedrooms: The number of bedrooms in the house.
    - Area: Block size of the house (in m^2).
    - YrBuilt: The year in which the house was built.
    - Seller: Having values *Jim* and *Jenny*, the names of the two competing real estate agents.
    - Extras: How many of the following extras were part of the house? Air-conditioning, dishwasher, built-in-robes, gas connection, intercom, skylight(s) and satellite dish.
    - Pool: Did the house have a swimming pool (*Yes* or *No*).
    - Price: Selling price of the house (in thousands of dollars).

(a) Read the data into `R` and obtain the following summaries.

(b) Count the number of houses sold be each agent.
No. sold by Jenny: **No. sold by Jim:**

(c) Count the number of houses that sold for more than $400,000.
Ans:

(d) Now find the number of houses sold by *Jenny* that went for more than $400,000.
Ans:

(e) Obtain the mean selling price for Jim and for Jenny. (Use the `tapply` function.)
Jim's Average Selling Price:
Jenny's Average Selling Price:

(f) Produces a two-way table of counts of material type by seller.

(g) Now prepare histograms of price by seller. You should be able to find the appropriate code by online search.

(h) In light of your analysis above, write a short paragraph describing the house sales for the the two sellers.

2. In this question, we examine the BMI (Body Mass Index) data. The file *BMI.txt* contains the heights and weights of a sample of individuals. The Body Mass Index of a particular person is given by

$$\text{BMI} = \frac{\text{weight}}{\text{height}^2}$$

where 'height' is the person's height in metres and 'weight' is their weight in kilograms.

(a) Read the data into `R` as previously.

(b) Calculate the BMI for an individual who weighs 65kg and is 1.70m tall. Type

```
65/1.7^2
```

in the `R Script` window, select it and submit it. **Ans:**

(c) Create the variable `bmi` that contains the BMIs of the individuals in the data set. **Make sure this is the active data set.**

(d) Produce descriptive statistics for the BMIs and copy them below.

3. The critical factor in washer manufacture is the hole diameter. The specified diameter for a particular type of washer is 50 mm (so the diameter needs to be at least 50 mm). Additionally, washers with diameters greater than 50.5 mm are too big. Your company has been awarded a tender to manufacture these washers. Two machines are available for the manufacture. In order to select the one that best meets the requirements, you first gather some data, consisting of the diameters of a sample of 200 washers, 100 from each of the two machines. This data is contained in the file *Washers.txt*. The aim of this question is to determine which of the two machines is the better choice, by performing some exploratory analysis on the data.

(a) Verify that there are indeed 100 washers from each of Machines A and B.

(b) Generate summary statistics and produce histograms for the diameters by machine.

(c) Which machine would you choose, and why?

(d) Could we suspect the same conclusion as in (c) by looking at the descriptive statistics alone? In particular, how can we interpret the medians, minimum/maximum values, means and standard deviations/variances?

4. (a) For the House data, produce a bar chart of the material type used in the house construction.

(b) For the House data, create a column that contains the Ages of the houses sold. The data is for houses sold in 2005, so we need to subtract the YrBuilt from 2005 to obtain the age of the house. Call the new variable AgeHouse. Obtain descriptive statistics for "AgeHouse".

(c) Produce a scatter plot of Price against AgeHouse. Do you see any relationship between Price and Age of houses?

5. The datafile *Aviation.txt* contains data on the numbers of sectors flown by the major Australian domestic airlines.

(a) Plot a pie chart of the data by typing the following commands in the `R Script` window and submitting it.

```
with(Aviation, pie(NumberOfSectorsFlown, labels=levels(Airline),
 xlab="", ylab="", main="Airline",
  col=rainbow_hcl(length(levels(Airline)))))
```

(b) Plot a bar chart of the data by submitting the command
`with(Aviation, barplot(NumberOfSectorsFlown, xlab="Airline", ylab="Frequency"))`.

6. In a large corporation, a very small group of employees have extremely high salaries, whereas the majority of employees receive much lower salaries. In a few sentences describe the shape of such a data set and how the mean and median would compare. When might you want to report the mean, and when might you want to report the median?

Summation Notation, Sample Mean and Sample Variance

[Note that some of these questions need the definitions

$$\text{sample mean} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \ , \ \text{and sample variance} = s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \ ,$$

where $x_1, x_2, \ldots, x_n$ denotes a sample data set of n numbers. Also note that $s_x^2 = \frac{1}{n-1}\left(\sum_{i=1}^n x_i^2 - n\bar{x}^2\right)$]

7. (a) Given that $x_1 = 1, \quad x_2 = -1, \quad x_3 = 2, \quad x_4 = 6, \quad x_5 = 0$, evaluate the following sums.

$$\text{i. } \sum_{i=1}^{5} x_i, \quad \text{ii. } \sum_{i=1}^{5} x_2\ , \quad \text{iii. } \sum_{i=1}^{5} x_i^2\ ,$$

$$\text{iv. } \sum_{i=1}^{5}(x_i - 2)\ , \qquad \text{v. } \bar{x}\ , \qquad \text{vi. } s_x^2.$$

(b) Given that $\sum_{i=1}^{6} y_i = 20$, evaluate

$$\text{i. } \sum_{i=1}^{6}(y_i - 3), \quad \text{ii. } \sum_{k=1}^{6}(-3y_k), \quad \text{iii. } \sum_{j=1}^{6} y_j - 3.$$

8. For five data values x_i, $i = 1, 2, \ldots, 5$, the following calculations were made:

$$\sum_{i=1}^{5} x_i = 10, \qquad \sum_{i=1}^{5} x_i^2 = 29.$$

Use this information to calculate $\bar{x}$, s^2 and s.

9. Find simple expressions for

$$\text{i. } \sum_{i=1}^{10}(x_2 + 2), \quad \text{ii. } \sum_{j=1}^{3} i(j+1), \quad \text{iii. } \frac{\sum_{i=1}^{3}(2ij)}{\sum_{k=1}^{3}(3k+1)}$$

10. Suppose $x_1, x_2, \ldots, x_n$ is a set of n numbers.

(a) Suppose c is some constant. Prove the summation formula

$$\sum_{i=1}^{n}(x_i - c) \quad = \quad \sum_{i=1}^{n} x_i - nc.$$

Hence determine the value of c (expressed as a formula in terms of $x_1, x_2, \ldots, x_n$) such that

$$\sum_{i=1}^{n}(x_i - c) = 0.$$

(b) Prove that $\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$.

Linear Transformations of Data

11. Suppose we have a sample of n data points $x_1, x_2, \ldots, x_n$. If a linear transformation $u_i = ax_i + b$, $i = 1, 2, \ldots, n$ is applied to these data, show that

    (i) $\bar{u} = a\bar{x} + b$ (ii) $s_u^2 = a^2 s_x^2$.

12. A data set contains 100 observations. Which of the following statements are true and which are false? Give reasons for your answers.

    (a) If you add 10 to each observation, the median increases by 10.

    (b) If you add 10 to each observation, the variance increases by 10.

    (c) If you multiply each observation by 3, the mean is multiplied by 3.

    (d) If you multiply each observation by 3, the variance is multiplied by 3.

    (e) If you change the sign of each observation, the sign of the mean is unchanged.

    (f) If you change the sign of each observation, the sign of the standard deviation is changed.

13. The transformation $u_i = \frac{x_i - \bar{x}}{s_x}$ is called *standardisation*. It is the most commonly used transformation in statistics. Noting that the equation

$$\frac{x_i - \bar{x}}{s_x} = ax_i + b$$

    is satisfied for $a = 1/s_x$ and $b = -\bar{x}/s_x$, we see that standardisation is a linear transformation. Hence, using the facts that $\bar{u} = a\bar{x} + b$ and $s_u^2 = a^2 s_x^2$ for a linear transformation $u_i = ax_i + b$, show that $\bar{u} = 0$ and $s_u^2 = 1$.

    Establishing this tells us that standardisation of a data set will always result in the transformed data having mean equal to 0 and sample variance equal to 1. The only thing that really changes as a result of a linear transformation is the "scale" on which the data can be viewed; the main characteristics of the data do not change.

14. Suppose $x_1 = -2$, $x_2 = 5$, $x_3 = 3$, $x_4 = -3$, $x_5 = 0$ and $x_6 = 8$.

    (a) Calculate $\bar{x}$ and s_x.

    (b) Given that

$$u_i = 3 + 4x_i, \qquad i = 1, 2, \ldots, 6,$$

    determine the values of the sample mean $\bar{u}$ and standard deviation s_u of the set of numbers u_1, u_2, u_3, u_4, u_5 and u_6.

    (c) Write down a linear transformation that would convert x_i to v_i, $i = 1, 2, \ldots, 6$, such that $\bar{v} = 0$ and $s_v = 1$.

15. (a) The data in the file *Salaries.txt* consists of the salaries of 500 individuals. Compute descriptive statistics for this data, and plot a histogram of the data .

(b) Based on your answer to (a), write a brief summary (one paragraph) of the key features of the data set "Salaries".

Tables

16. An engineer is designing a battery for use in a device that will be subjected to some extreme variations in temperature. He has three possible choices for the plate material. For testing purposes he selects three temperatures. Four batteries are tested at each combination of plate material and temperature and the tests are run in random order. The battery life (hours) under each set of conditions is given in Table below.

(a) What can you infer from this table regarding which is the best material to select for the batteries?

(b) Find the summary statistics for this data. First think about what sort of statistics you should be interested in.

	Temperature (°C)					
Material	-10		20		55	
1	130	155	34	40	20	70
	74	180	80	75	82	58
2	150	188	136	122	25	70
	159	126	106	115	58	45
3	138	110	174	120	96	104
	168	160	150	139	82	60

Chapter 3

Probability

1. Suppose a sample space contains 6 outcomes denoted by $e_1, e_2, \ldots, e_6$. Let A, B and C be the events

$$A = \{e_1, e_2, e_3\},\ B = \{e_3, e_4, e_5\},\ \text{and } C = \{e_5, e_6\},$$

and suppose probabilities are assigned as follows:

$$P(A) = 0.4\ ,\ P(B) = 0.5\ ,\ P(C) = 0.5\ ,\ P(A \cap B) = 0.2\ ,\ P(B \cap C) = 0.2\ .$$

   (a) By using a Venn diagram, or otherwise, verify that such an assignment of probabilities is possible.

   (b) Find the probabilities of the events $A \cup B$, $A \cup B \cup C$ and $C \cap \bar{B}$.

   (c) The event $A \cap B$ can be described in words as "both the event A and the event B occur". Describe similarly in words, the three events of part (b).

2. (a) What is wrong with the following assignment of probabilities?

$$P(A) = 0.6, \qquad P(B) = 0.3, \qquad P(A \cap B) = 0.4.$$

   (b) What is wrong with the following assignment of probabilities?

$$P(A) = 0.6, \qquad P(B) = 0.7, \qquad P(A \cup B) = 0.5.$$

3. Of three events A, B, and C, suppose A and B are independent and B and C are disjoint. Their probabilities are $P(A) = 0.6$, $P(B) = 0.2$ and $P(C) = 0.2$. Calculate the probabilities of the following events:

   (a) Both B and C occur.

   (b) At least one of A and B occurs.

   (c) B does not occur.

   (d) All three events occur.

4. In a regional city, two lawn companies fertilise lawns during the summer. Company A has 72% of the market. Thirty per cent of the lawns fertilised by Company A could be rated as very healthy one month after the service. Company B has the other 28% of the market. Twenty per cent of the lawns fertilised by Company B could be rated as very healthy one month after the service. A lawn that has been fertilised by one of these companies within the last month is selected randomly and is rated as very healthy.

   (a) what is the revised probability that Company A fertilised the lawn?

   (b) What is the corresponding probability that Company B fertilised the lawn?

5. Of 250 employees of a company, a total of 130 are full-time employees. The remainder are part-time employees. There are 150 males working for this company, 85 of whom are full-time employees.

   (a) What is the probability that an employee chosen at random

      i. is a part-time employee?

      ii. is female and a full-time employee?

      iii. is a full-time employee, given that the employee is female?

      iv. is a female, given that the employee is full-time?

   (b) Are the events "employee chosen at random is female" and "employee chosen at random is full-time" statistically independent?

6. Let $P(A \cup B) = 0.9$, $P(A) = 0.5$ and $P(B \mid A) = 0.4$. Calculate $P(B)$ and $P(A \mid B)$. Are A and B independent? Are A and B disjoint? Justify your answers.

7. Let $P(A) = 0.2$, $P(A \cup B) = 0.6$, and suppose the events A and B are independent. Calculate $P(B)$.

8. Let $P(A) = 0.6$, $P(A \mid B) = 0.6$ and $P(A \cup B) = 0.8$. Calculate $P(B)$ and $P(A \cap B)$. Are A and B independent? Justify your answer.

9. A toy manufacturer buys pre-assembled robotic arms from three different suppliers – 50% of the total order comes from Supplier 1, 30% of the total order comes from Supplier 2, and the remaining 20% from Supplier 3. Past data shows that the quality control standards of the three suppliers are different. Two percent of the arms produced by Supplier 1 are defective, while Suppliers 2 and 3 produce defective arms at the rates of 3% and 4% respectively.

   Let S_i be the event that a given arm comes from Supplier i, $i = 1, 2, 3$ and let D be the event that a given arm is defective.

   (a) Draw a tree diagram that models this situation.

   (b) What proportion of the arms in the manufacturer's inventory are non-defective?

   (c) If an arm is found to be defective, what is the probability that it came from Supplier 1? Give your answer to 4 decimal places.

(d) The company wants to increase the proportion of non-defective inventory to 98%. They decide to target Supplier 3 and require them to improve their quality control and thus reduce the proportion of defective arms that they deliver. What must Supplier 3 reduce its proportion of defective production to in order to satisfy the client's demand?

10. You are given the following probabilities involving events A and B:

$$P(B^c \mid A) = 0.4, P(B) = 2 \times P(A), P(A \cup B) = 0.6.$$

(a) Show that $P(B \mid A) = 0.6$. **(1 mark)**

(b) Show that $P(A \cap B) = 0.6\ P(A)$.

(c) Calculate $P(A)$.

(d) Calculate $P(A \mid B)$.

(e) Are the events A and B independent? Justify your answer.

11. You are given the following probabilities involving events A and B:

$$P(\overline{B} \mid \overline{A}) = 0.2, P(A) = 0.3, \text{ and } \mathrm{P(B \mid A)} = 0.4.$$

(a) Using a Venn diagram or otherwise, show that $P(A \cup B) = 0.86$.

(b) Show that $P(A \cap B) = 0.12$.

(c) Hence find $P(B)$.

(d) Calculate $P(A \mid B)$.

(e) Are the events A and B independent? Justify your answer.

12. A method used to select offshore regions for oil deposits is to sample surface waters from lakes, streams and swamps for the presence of hydrocarbons. The probability of finding oil deposits in a region where surface waters show the presence of hydrocarbons is 0.8; otherwise this probability is only 0.1. In a particular region, initial drilling indicates that the probability of oil deposits in this region is 0.6.

Let O denote the event that oil deposit is present in this area and let H denote the event that surface water in the region shows the presence of hydrocarbons.

(a) Let $P(H) = x$. Draw a well labelled tree diagram to represent the information provided.

(b) Obtain an equation for $P(O)$ in terms of x and solve it to obtain x.

(c) If oil is found in the region, what is the probability that the surface water in the region will show the presence of hydrocarbons?

Chapter 4

Random Variables

Discrete random variables: probability mass functions, cumulative distribution functions, expectation and variance

1. Decide which of the following functions could represent probability mass functions for a random variable X. For those which could, write out the table for the probability mass function and determine $\mathrm{E}(X)$. Also draw the graph of the cumulative distribution function corresponding to the first valid pmf.

   (a) $p(x) = \dfrac{x}{10}$, where $x = 1,\ 2,\ 3,\ 4$;

   (b) $p(x) = 0.7x$, where $x = -1,\ 0,\ 1,\ 2$;

   (c) $p(x) = \dfrac{x^2}{14}$, where $x = 0,\ 1,\ 2,\ 3$;

   (d) $p(x) = \dfrac{1}{x}$, where $x = 1,\ 2,\ 3$;

   (e) $p(x) = (10 - x)/40$, where $x = 0,\ 1,\ 2,\ 3,\ 4$.

2. A discrete random variable X takes the values 1, 4 and 6, with probabilities 0.4, 0.5 and 0.1, respectively. Calculate $E(X)$ and $E(X^2 - 2X)$.

3. * Suppose that X is a discrete random variable with pmf $p_X(x)$.

   (a) Prove, for constants a and b, that $E(aX + b) = aE(X) + b$. (Recall that for any function g, $E[g(X)] = \sum_x g(x)p_X(x)$).

   (b) The *variance* of X is defined as $\mathrm{Var}(X) = E[(X - E[X])^2]$. Prove that $E[(X - E[X])^2] = E(X^2) - (E[X])^2$ (and hence, that the right-hand side may also be used as a representation of the variance of X).

4. Determine whether the following are true or false. Provide reasons for your answers.

   (a) The mean of a discrete random variable is always positive.

   (b) If the mean of a random variable is positive then the random variable only takes on non-negative values.

5. The table below shows the probability mass function of a random variable X, where c is a constant.

x	0	1	2	3
$P_X(x)$	c	c^2	$c^2 + c$	$3c^2 + 2c$

   (a) What is the value of c?

   (b) Using this value of c, find $E(X)$.

6. An economist determines the following table for next year's WA unemployment rate (rounded to nearest 0.5%) with the corresponding probabilities of occurrence:

Rate (%)	2	2.5	3	3.5	4	4.5
Probability	0.13	0.25	c	0.21	0.09	0.02

   (a) Define the random variable, say X, of interest to the economist with this data.

   (b) Find the value of c.

   (c) Tabulate, using appropriate symbols, the p.m.f. of X.

   (d) Sketch a graph of the cumulative distribution function of X.

   (e) Calculate

      i. $P(X = 1)$

      ii. $P(X \geq 3)$

      iii. $P(X < 4)$.

   (f) Find $E(X)$ and describe in words what this tells you.

   (g) Find $Var(X)$ and describe in words what this tells you.

7. *Based on past sales data, an appliance store stocks five window air conditioner units for the coming week. The weekly consumer demand D for this type of unit has the probability distribution given below.

d	0	1	2	3	4	5	6	7	8
$p_D(d)$	0.05	0.05	0.08	0.16	0.30	0.16	0.10	0.05	0.05

   (a) Let the random variable X denote the number of air conditioner units left at the end of the week. Find the probability distribution of X.

   (b) Find the cdf of X and plot it.

   (c) If an order is received when the store is out of stock, then a special stockout order is required. Let the random variable Y denote the number of special stockout orders. Find the probability distribution of Y.

   (d) Find the expected values of X and Y and interpret them.

   (e) The store makes a profit of \$150 on each air conditioner sold from the weekly available stock, but only \$100 for each unit sold from the special stockout orders. Find the expected weekly profit for the company from the sales of the air conditioners.

8. A mail-order company conducts a survey to examine the effectiveness of its four annual advertising promotions. A questionnaire is sent out to each of its customers asking how many of the previous year's promotions prompted purchases that would not otherwise have been made. Let the random variable X denote the number of promotions that prompted a purchase. The results of the survey are summarised below:

x	0	1	2	3	4
$p_X(x)$	0.10	0.25	0.40	0.20	0.05

Assume that the survey response is an accurate evaluation of the effectiveness of the advertising, and that consumer behaviour will not change in the coming year.

(a) Calculate the mean and variance of the number of promotions that prompted a purchase.

$\mathbf{E}(\mathbf{X}) =$ $\mathbf{Var}(\mathbf{X}) =$

(b) It is known from historical data that the average value of orders for promotional goods is \$12.50, with the company earning a gross profit of 20% per order. The fixed cost of conducting the four promotions next year is \$15,000, with a variable cost of \$3.00 per customer for postage and handling. Assuming that the survey results can be used as an accurate predictor of behaviour for existing and potential customers, how large a customer base must the company have in order to make a profit of \$100,000 per annum on average? (Hint: first show that the expected income from sales, from a randomly chosen customer, is $E(X) \times 0.20 \times 12.50$.)

9. A motor vehicle inspection unit collected the following data during the past year. Assuming the observed pattern continues to hold, what would be the unit's expected daily revenue this year if it charges \$25 per car?

Number of cars inspected per day (x_i)	20	25	32	40	47	51
Number of days (f_i)	79	121	19	25	39	30

10. The number of sightings N in a evening of an endangered nocturnal rodent has probability distribution given by

$$P(N = n) = \frac{9 - 2n}{25} \ , \quad n = 0, 1, 2, 3, 4.$$

Write down the pmf table for the random variable N, and use it to find

i. $P(N > 2)$, ii. $P(N \leq 1)$, iii. $E(N)$.

Bernoulli and Binomial Distributions

11. On Pingelap Island, 10% of the population is colour blind. A researcher selects 50 people at random from the island. Let the random variable X denote the number of people, out of the 50, who are colour blind.

(a) State the distribution of the random variable X.

(b) Determine the probability that of these 50 people, exactly 7 are colour blind.
Ans:

(c) Determine also the following:

i. $\mathrm{P}(X = 4)$; **Ans:**

ii. $\mathrm{P}(X \leq 6)$. **Ans:**

(d) What are the mean and variance of the random variable X?

12. Let X be the number of heads from 10 tosses of a fair coin. Evaluate the following probabilities:

(a) $\mathrm{P}(X = 5)$;

(b) $\mathrm{P}(X > 7)$;

(c) $\mathrm{P}(3 \leq X \leq 8)$.

13. A botanist researching flower bulbs knows that 90% of its bulbs will flower. They are sold in packets of 12 randomly selected bulbs with a guarantee that the packet will be replaced if 100% flowering is not achieved.

(a) What is the probability that it will be necessary to replace a given packet under this guarantee? Interpret this probability.

(b) What would be the probability of replacing a packet if the guarantee covered only at least 10 out of 12 bulbs flowering? Comment on your findings in parts (a) and (b).

14. Often "bad news" hits the Australian stock market and drags the stock prices down. Suppose the probability of such "bad news" occurring at least once in a month is 0.95.

(a) i. What is the probability distribution of the number of months in a given year in which there is "bad news"?

ii. What is the probability distribution of the number of months in a given year in which there is no "bad news"?

(b) What is the probability that there is at least one month of the year in which there is no "bad news"?

15. For each random variable described below state whether a Binomial random variable would provide a satisfactory model. Give reasons to support your decision and where appropriate state values for n and p.

(a) The number of shares in the ASX Top 500 that will fall in value today.

(b) The number of earthquakes in New Zealand this year.

(c) Of the first 10 earthquakes in New Zealand this year, the number that will be above 6.5 on the Richter scale.

(d) The number of days that the maximum temperature will rise above 35°C in Perth next December.

The Poisson Distributions

16. While John is in his office, he receives 4 phone calls per hour on average. Assume that the number of calls within any interval of time is distributed as Poisson.

    (a) What is the probability that the phone rings at least 4 times between 10am and 11am?

    (b) If John takes a 30 minute lunch break, what is the probability that the phone does not ring during that time?

    (c) What is the expected number of times that the phone will ring during John's lunch break? What is the variance?

    (d) If John arrives at work at 9 am and leaves at 5 pm, what is the expected number of times that the phone will ring during the day? What is the variance?

17. Hummingbirds arrive at a flower at a rate λ per hour.

    (a) How many visits are expected in x hours of observation?

    (b) What is the variance of the number of visits in an hour?

    (c) If significantly more variance is observed than expected, what might this tell you about hummingbird visits?

18. Let $Y \sim \text{Poi}(6)$. Without using `R`, find:

    (a) $\text{P}(Y \geq 3)$;

    (b) $\text{P}(Y \leq 15)$;

    (c) $\text{P}(3 \leq Y \leq 15)$.

19. Bacteria are spread across a plate at an average density of 1000 per square cm. What, therefore, is the probability of seeing at least one cell?

    (a) What is the chance of seeing no bacteria in the viewing field of a microscope if this viewing field is 4×10^{-4} square cm?

    (b) What is therefore the probability of seeing at least one bacterium cell?

20. A firm collects large quantities of data. Occasionally, typing errors cause data to be incorrectly entered. The number of typing errors per 20 pages of data is a Poisson random variable with mean 3.

    (a) What is the probability of there being 10 or more typing errors in 40 pages of data?

    (b) What is the probability of there being between 5 and 9 (inclusive) typing errors in 40 pages of data?

    (c) What is the probability of there being less than 5 typing errors in 20 pages of data?

    (d) What is the mean number of typing errors in 40 pages of data? What is the variance?

Hypothesis Testing for Binomial Proportions

21. According to the Center for Disease Control (CDC), the percent of overweight adults 20 years of age and over in the United States is 69.0% (see link to the URL). A city council believes the proportion of overweight citizens in their city is less than this known national proportion. They take a random sample of 150 adults 20 years of age or older in their city and find that 98 are classified as overweight. Let's use the five step hypothesis testing procedure to determine if there is evidence that the proportion in this city is different from the known national proportion.

    (a) Define the random variable of interest, and state its distribution.

    (b) Write down null and alternative hypotheses that could be tested in order to answer the question "Is the true proportion of overweight people in the city less than the national proportion?".

    (c) State the distribution of the random variable defined in (a), under H_0.

    (d) Write down an expression for the p–value obtained from the data, and evaluate it using `R`.

    (e) We can perform the hypothesis test in `R` . The code is
    `binom.test(x, n, p, alternative="greater")`
    where

    - x is the number of success observed
    - n is the number of trials (or sample size)
    - p is the value of the proportion in the null hypothesis
    - alternative states the alternative hypothesis, taking values "greater", "less" or blank for two-sided test.

    In the `R` console, type in the code for the testing the hypothesis stated above. Compare the p–value from the output with the value obtained previously.

    (f) What should the council conclude at the 2.5% level of significance?

22. Hypersensitivity of teeth, known as dentin hypersensitivity, is a pathological condition in which teeth are sensitive to thermal, chemical and physical stimuli. Patients with dentin hypersensitivity experience pain from hot/cold and sweet/sour solutions and foods. Pain may also be felt when hot or cold air comes in contact with teeth. Pain varies from mild to excruciating. Dentin hypersensitivity is caused by exposure of dentile tubules from attrition, abrasion, erosion, fracture or chipping of teeth, or a faulty restoration (Kishore, A, Mehrota, K. K. and Saimbi, C. S. (2002) Effectiveness of desensitising agents. *J. of Endodontics*, **28**, 34–35.). Past studies have shown that 5% potassium nitrate solution reduces dentin hypersensitivity in 48% of cases. A researcher is testing the effectiveness of a 40% formalin solution as an alternative to potassium nitrate in reducing hypersensitivity. In a sample of 81 patients suffering from dentin hypersensitivity, 49 reported significant pain relief when using 40% formalin solution. What should the researcher conclude regarding the effectiveness of the two desensitising agents?

    (a) Define the random variable of interest, and state its distribution.

(b) Write down null and alternative hypotheses that could be tested in order to answer the question "Is the true proportion of people who experience pain when using a 40% formalin solution more than that for a 5% potassium nitrate solution?".

(c) State the distribution of the random variable defined in (a), under H_0.

(d) Write down an expression for the p–value obtained from the data, and evaluate it using `R`.

(e) Perform the hypothesis test in `R` as in the previous question. Compare the p-value from the output with the value obtained previously.

(f) What should the researcher conclude at the 2.5% level of significance?

23. **Driving and cell phones** In a survey, 1640 out of 2246 randomly selected adults in the United States admitted using a cell phone while driving (based on data from Zogby International). The claim is that the proportion of adults who use cell phones while driving is more than 72%. Based on this data, what do you conclude about this claim?

Hypothesis Testing for Poisson Means

24. The number of customers arriving at a particular fast food outlet between 12:30pm and 12:35pm on a randomly chosen day is a Poisson random variable. The outlet believes that an average of 4 customers enter the outlet between 12:30pm and 12:35pm each day. One staff member believes that the true mean number of arrivals is greater than 4, and that on each day, more staff should be assigned to work during this time interval. On a randomly chosen day, 9 customers are seen to arrive between 12:30pm and 12:35pm. At the 2.5% level of significance, test whether the staff member's suspicion is justified (and hence whether more staff should be used during this time period).

25. Tree Swallows (TS) lay on average 2 eggs per clutch. A new sub-species of Tree Swallows has been discovered by Onny, the ornithologist, who calls it TS1. Onny believes that the sub-species lays more eggs on average than TS. She collects data on the number of eggs per clutch for TS1, and finds the total number of eggs in three clutches is 10. Let λ denote the mean number of eggs per clutch for TS1. She will decide this by performing a hypothesis test based on the data.

(a) In terms of λ, state the null and alternative hypotheses that Onny needs to test.

(b) Let the random variable X denote the number of eggs in three clutches. What is the distribution of X under the null hypothesis?

(c) Find the p–value of the test.

(d) State your conclusion at the 2.5% level of significance regarding the mean number of eggs per clutch for TS1.

(e) Is the Poisson model is appropriate for this data?

Chapter 5

Continuous Random Variables

1. Suppose X is a continuous random variable whose probability density function is given by
$$f_X(x) = \begin{cases} kx, & \text{for } 0 \leq x \leq 5, \\ 0, & \text{otherwise,} \end{cases}$$
where k is a constant.
   (a) Evaluate k.
   (b) Find $\mathrm{P}(1 \leq X \leq 3)$, $\mathrm{P}(2 \leq X \leq 4)$, $\mathrm{P}(X = 4)$ and $\mathrm{P}(X \leq 3)$.

2. Suppose Y is a continuous random variable with probability density function given by
$$f_Y(y) = \begin{cases} 1/2 - |y|/4, & -2 \leq y \leq 2 \\ 0, & \text{elsewhere.} \end{cases}$$
   (a) Sketch the probability density function of Y.
   (b) From your sketch, or otherwise, verify that this is indeed a valid pdf.
   (c) Evaluate the following probabilities:
      i. $\mathrm{P}(0 \leq Y \leq 0.4)$;
      ii. $\mathrm{P}(Y > 0.5)$;
      iii. $\mathrm{P}(Y = 0)$;
      iv. $\mathrm{P}(Y \leq 0.5)$.
   (d) If X_1 and X_2 are independent $U(-1, 1)$ random variables, then it can be shown that $X_1 + X_2$ has the same distribution as Y Using this fact, and the facts that $E(X_1) = (a + b)/2$ and $\mathrm{Var}(X) = (b - a)^2/12$, find the following:
      i. $\mathrm{P}(X_1 + X_2 \geq 0.5)$;
      ii. $\mathbb{E}[Y]$ and $\mathrm{var}(Y)$.

3. Suppose X is a continuous random variable whose probability density function is given by
$$f_X(x) = \begin{cases} k & \text{for } 1 \leq x \leq 5, \\ 0 & \text{otherwise,} \end{cases}$$
where k is a constant.

(a) Evaluate k.

(b) Find $\mathrm{P}(2.2 \leq X \leq 3.4)$, $\mathrm{P}(0 \leq X \leq 3.4)$, $\mathrm{P}(X = 3)$ and $\mathrm{P}(X \geq 3.4)$.

Chapter 6

Joint Distributions

Joint pmfs, Independent random variables

1. Suppose X and Y are discrete random variables with the joint probability distribution given in the table:

y \ x	3	4	5
-1	0	0.4	0
0	0.3	0	0.1
2	0	0.2	0

   (a) Calculate $E(X), E(Y)$,Var(X), Var(Y) and Cov(X, Y).

   (b) Are X and Y independent? Justify your answer.

   (c) Calculate $P(X \leq Y^2 + 1)$.

   (d) Calculate (i) Var$(X - Y)$, (ii) Var$(3Y + 4X + 2)$, (iii) Cov$(X + Y, Y)$, (iv) Cov$(X + 10, Y - 10)$.

2. The random variables X and Y respectively denote the daily sales of two types of products X and Y at a Delicatessen, where X is highly perishable, and needs to be sold on the day it arrives in the store. Their joint probability distribution is given in the table below.

y \ x	0	1
0	0.15	0.1
1	0.2	0.2
2	0.05	0.3

   Assume that one of product X and two of product Y are stocked by the Delicatessen.

   (a) Calculate $E(X), E(Y)$, Var(X), Var(Y) and Cov(X, Y).

   (b) Are X and Y independent? Justify your answer.

   (c) What is the probability that on a given day the number of sales of X is less than that of Y?

(d) If each product X costs \$3.00 and sells for \$6.80, what is the expected daily profit from the sale of this product?

3. X and Y are independent discrete random variables where X has possible values 1 and 2 with probabilities 0.3 and 0.7 respectively, and Y has possible values 0, 1, 3 and 4 with probabilities 0.2, 0.3, 0.3 and 0.2 respectively.

   (a) Set up a two-way table giving the values of the joint p.m.f. $p_{X,Y}(x, y)$.

   (b) Given that $\text{Var}(X) = \sigma_X^2 = 0.21$ and $Var(Y) = \sigma_Y^2 = 2.2$, calculate

   i. $\text{Var}(X - Y)$ iii. $E\left(\dfrac{X - \mu_X}{\sigma_X}\right)$

   ii. $\text{Var}(X + Y)$ iv. $\text{Var}\left(\dfrac{X - \mu_X}{\sigma_X}\right)$.

4. The marginal distributions for two discrete random variables X and Y as well as the nature of their joint p.m.f. are given in the table below:

x \ y	1	3	5	$p_X(x)$
-1	a	b	?	0.5
1	?	?	?	0.5
$p_Y(y)$	0.2	0.3	0.5	1.0

   (a) Show that if $a = 0.05$ and $b = 0.25$, then $\text{Cov}(X, Y) = 0$.

   (b) Show that there is at least one other pair of values for a and b, different from those in part (a), such that $\text{Cov}(X, Y) = 0$.

Chapter 7

Contingency Tables: The Chi-Squared Test for Independence

1. Four hotels took part in a survey on hotel guest satisfaction. A follow up question was asked of all respondents who were dissatisfied with the service. These guests were asked to indicate the main reason for their dissatisfaction. You are asked to investigate whether the choice of hotel has any bearing on the main reason for dissatisfaction.

   Do not use `R` in this question! Write your answers on paper, showing full working.

   (a) State appropriate hypotheses that could be tested to answer the question: "Do the results of the survey provide evidence that the nature of dissatisfaction and the choice of hotel are related?"

   (b) A contingency table, summarising the results of the survey, is given below. The table shows the observed frequencies for each cell, as well as some of the expected frequencies under H_0 (in parentheses). Copy down this table, and without using `R`, calculate the remaining expected frequencies under H_0. Show working!

	Hotel								
	Fijian		Tradeswest		Sheraton		Coral Reef		Totals
Politeness	23	( )	7	( )	37	(33.7410)	67	(62.0192)	134
Knowledge	25	( )	13	( )	25	(30.9712)	60	(56.9281)	123
Responsiveness	13	(11.0024)	5	(6.6906)	13	(15.6115)	31	(28.6954)	62
Other	13	(17.3909)	20	(10.5755)	30	(24.6763)	35	(45.3573)	98
Totals	74		45		105		193		417

   (c) Write down an expression for the relevant test statistic, and state its distribution under H_0 (together with any associated parameters!).

   (d) Without using `R`, calculate the contribution from the upper-left cell to the observed value of the test statistic.

   (e) Given that the observed value of the test statistic is $\chi^2_{\text{obs}} = 20.8059$, carry out the test (without using `R`) at the 5% significance level, and state your conclusion. Is there sufficient evidence to conclude that there is a relationship between the choice of hotel and the nature of dissatisfaction?

2. Enter the data from Question 1 as shown below:

	Fijian	Tradeswest	Sheraton	Coral Reef
Politeness	23	7	37	67
Knowledge	25	13	25	60
Responsiveness	13	5	13	31
Other	13	20	30	35

(a) Use `R` to generate an appropriate output for a test for independence of the two variables of interest, carry out the test, and check that your conclusions are the same as in Question 1.

(b) If there is evidence that the nature of dissatisfaction is related to the choice of hotel, where do the discrepancies lie? Which hotel(s) could be advised to improve their service, and in which area(s)? Do any of the hotels appear to provide significantly better service than the others in a particular area?

Chapter 8

The Normal Distribution

Introduction, Normal Distribution Problems

1. The random variable X has a normal distribution with mean 5 and variance 4. Using probability tables, evaluate the following probabilities.

   (i) $\mathrm{P}(X > 5.7)$ (ii) $\mathrm{P}(|X - 2| \geq 2)$ (iii) $\mathrm{P}(2.8 \leq X \leq 5.1)$.

2. Now evaluate the probabilities in the previous question using `R`.

3. Suppose that $Y \sim \mathrm{N}(3, 2^2)$. Using `R`, find the following probabilities.

   (a) $\mathrm{P}(Y > 0.7)$;

   (b) $\mathrm{P}(2.5 < Y \leq 5.2)$;

   (c) $\mathrm{P}(|Y - 3| \leq 2)$.

4. Suppose $Z \sim N(0, 1)$. Use tables to evaluate the following probabilities.

   (a) $P(Z \leq 1.75)$ (b) $P(|Z| > 2)$ (c) $P(Z^2 < 2)$.

5. Suppose X has a normal distribution with mean 2 and variance 8. Evaluate the following probabilities.

   (i) $\mathrm{P}(X > 2.2)$ (ii) $\mathrm{P}(|X - 2| \geq 3)$ (iii) $\mathrm{P}(1.2 \leq X \leq 3.8)$.

6. Suppose $Z \sim N(0, 1)$. Use tables to find z such that the following probability statements are true.

   (a) $P(Z \leq z) = 0.95$ (b) $P(|Z| \leq z) = 0.95$ (c) $P(-1.8 \leq Z \leq z) = 0.6$.

Sums of Normal Random Variables

7. Suppose X and Y are independent random variables with $X \sim \mathrm{N}(4, 3^2)$ and $Y \sim \mathrm{N}(6, 4^2)$. Suppose $D = Y - X$.

   (i) What is the probability distribution of D?

   (ii) Using your answer to (i), evaluate $\mathrm{P}(Y < X)$.

   (iii) Suppose $U = 4X + 2Y + 3$. What is the probability distribution of U?

8. A pizza company runs a marketing campaign based on their delivery time for pizzas. They claim that they will deliver a pizza in a radius of 5 km within 30 minutes of ordering or it is free. In practice the time it takes to prepare a pizza is normally distributed with mean 15 minutes and standard deviation 2 minutes. The delivery time is also normally distributed with mean 10 minutes and standard deviation 2 minutes.

   (a) Assuming that preparation time is independent of delivery time, what is the distribution of the time it takes for the pizza to be delivered from the moment it is ordered?

   (b) What is the probability that a pizza is delivered free?

   (c) On a busy Saturday evening, a total of 50 pizzas are ordered. What is the probability that more than 3 were delivered free?

   (d) If the company wants to reduce the proportion of pizzas that are delivered free to 1%, what should the delivery time be advertised as?

9. The maintenance department of a city's council finds that it is more cost effective to replace all the streetlight bulbs at once rather than replacing bulbs individually as they burn out. Tests have shown that the lifetime of the bulbs is normally distributed with mean 3,000 hours and standard deviation 200 hours.

   (a) If the department wants no more than 1% of the bulbs to burn out before any are replaced, after how many hours should the department plan to replace all bulbs?

   (b) If two bulbs are selected at random from those that are replaced, what is the probability that at least one of them has burnt out?

   (c) What is the probability that at least 10 bulbs are burnt out in a selection of 500 bulbs that were replaced?

   (d) The city has 50,000 street lights, and each bulb costs $10. In addition, each time the bulbs are changed it costs $2,500 in labour. On average how much will the council save per year (365 days) if it changes its policy so that no more than 2% of the bulbs are burnt out before replacement?

   (e) Another way to reduce cost is to buy a better bulb. The council is considering using bulbs costing $15 each that have a mean lifetime of 4,000 hours but the same standard deviation as the current bulb. Make a recommendation to the council as to which of the three options, 1% failure rate, 2% failure rate or the new bulb, should be adopted.

Chapter 9

Sampling Distributions and the Central Limit Theorem

Introduction, Sampling Distribution of the Sample Mean

Case 1: Normal Population with σ known

1. A simple random sample of size $n = 8$ is taken from a population with mean 40 and variance 12. The population distribution is assumed to be normal. Let $\overline{X}$ be the mean of the sample. Answer the following questions.

   (a) What is the distribution of $\overline{X}$?

   (b) Find P $\left(\overline{X} > 42\right)$.

   (c) Suppose you need to choose a sample of size n such that P $\left(\overline{X} > 42\right) < 0.025$. What is the minimum n required to achieve this?

2. A simple random sample of size $n = 15$ is taken from a population with mean 20 and variance 5. The population distribution is assumed to be normal. Let $\overline{X}$ be the mean of the sample. Answer the following questions.

   (a) What is the distribution of $\overline{X}$?

   (b) What is P$(\overline{X} < 21)$?

   (c) Suppose you need to choose a sample of size n such that P$(|\overline{X} - 20| > 1.5) < 0.05$. What is the minimum n required to achieve this?

3. In a given population the mean μ_X of a random variable X can be estimated using the random variable $\overline{X}$ which denotes the sample mean, for a given sample size. If $X \sim N(2, 4^2)$, give the distribution of $\overline{X}$ in terms of the sample size n, and obtain the smallest value of n for which $P(|\overline{X} - \mu_X| < 1)$ is at least 0.8.

Case 2: Normal Population with σ unknown – the t distributions

4. Suppose X has a t-distribution with k degrees of freedom. Evaluate the following using Statistical Tables.

(a) $\mathrm{P}(X > 2.7638)$ and $\mathrm{P}(X < -2.7638)$ when $k = 10$.

(b) $\mathrm{P}(|X| < 3.4668)$ when $k = 24$.

(c) t such that $\mathrm{P}(|X| > t) = 0.05$ when $k = 14$.

5. Suppose X has a t-distribution with k degrees of freedom. Evaluate the following using Statistical Tables, and then use `R` to check your answers.

(a) $P(X > 2.1448)$ and $P(X < -2.1448)$ when $k = 14$.

(b) $P(|X| < 1.7459)$ when $k = 16$.

(c) t such that $P(|X| > t) = 0.01$ when $k = 16$.

Case 3: Population Distribution not necessarily Normal

6. The mean grocery bill at a large supermarket is \$65, with a standard deviation of \$20. If $X_1, \ldots, X_{45}$ is a random sample of 45 grocery bills, find an approximate value for $\mathrm{P}(\overline{X} < 60)$.

7. A bank manager knows that homeowners in a certain suburb have monthly mortgage payments with a standard deviation of \$84. What is the probability that the mean payment of a random sample of 49 homeowners in that suburb will not deviate from the mean payment for the entire suburb by more than \$24? What assumptions have you made in calculating this probability? Is it necessary to assume monthly mortgage payments are normally distributed?

8. A simple random sample of size $n = 80$ is taken from a population with mean 26 and variance 12. Let $\overline{X}$ be the mean of the sample. Answer the following questions.

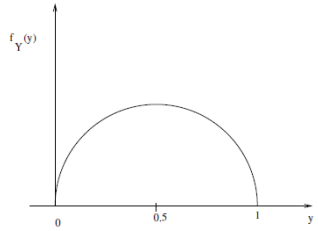
(a) What is the distribution of $\overline{X}$?

(b) Is is the distribution of $\overline{X}$ specified in (a) exact or an approximation? Give reason.

(c) What is $P(|\overline{X} - 26| < 0.5)$?

(d) Suppose you need to choose a sample of size n such that $P(|\overline{X} - 26| > 0.2) < 0.05$. What is the minimum n required to achieve this?

9. A continuous random variable Y has $E(Y) = 0.5$, $SD(Y) = 0.25$ and a probability density function $f_Y(y)$ with the following graph:



Let $Y_1, Y_2, \ldots, Y_{30}$ be a random sample from the distribution of Y.

(a) What is the approximate distribution of $\overline{Y} = \frac{1}{30}\sum_{i=1}^{30} Y_i$?

(b) Find an approximate value for $P(\overline{Y} > 0.45)$.

10. Let $\hat{p}$ be the sample proportion of heads from n tosses of a coin whose probability of heads is p.

(a) Write down an approximate distribution for $\hat{p}$.

(b) Now assume that $n = 100$ and the coin is fair. Using your answer to (a), find an approximate value for $\mathrm{P}(\hat{p} > 0.6)$.

(c) What is the minimum number of tosses of this fair coin that would ensure that $\mathrm{P}(|\hat{p} - 0.5| < 0.01)$ is at least 0.95?

11. (a) Discuss two aspects of a good sampling method.

(b) Give one example of a situation where it is impossible or impractical to make measurements on the whole population.

(c) A researcher decides to use a sample to obtain information on investor attitude toward embryonic stem cell research. Discuss in one or two sentences each of the following sampling schemes:

i. The researcher selects a sample of 200 people at random from the phone book.

ii. The researcher takes a sample of 200 people in her neighbourhood.

iii. The researcher takes a random sample of 200 publications in a medical journal and contacts the lead authors.

Chapter 10

Estimation and Confidence Intervals

Introduction, Point Estimation, Common Estimators

1. A researcher takes the salaries (in \$100,000) of a random sample of five CEOs of large corporations and computes the following statistic:

$$\tilde{X} = \frac{1}{15} \sum_{i=1}^{5} i \; X_i,$$

where $X_1, X_2, \cdots, X_5$ are the salaries. Let μ denote the mean income of all CEOs in Australia, and σ the standard deviation.

   (a) Find the mean and variance of $\tilde{X}$.

   (b) Find the mean and variance of the sample mean $\overline{X}$ of the five salaries.

   (c) How do your answers compare? Which is the better estimator of the mean salary of all CEOs in Australia, $\tilde{X}$ or $\overline{X}$?

   (d) Assume that the salaries of CEOs are normally distributed with a standard deviation of 2. Find $\mathrm{P}\left(\tilde{X} > \overline{X}\right)$.

Confidence Intervals, Required Sample Size Calculations

Population Mean: Confidence Intervals and Required Sample Sizes

2. (a) Would the 95% confidence interval for μ constructed from the same data be wider or narrower than a 90%?

   (b) Would a 95% confidence interval for μ constructed from a sample size of 400 be wider or narrower than that based on a sample size of 200?

   (c) How might 95% confidence intervals for μ constructed from two different samples of size 200 differ from each other?

(d) How large should be a sample in order to get a 90% confidence interval estimate of the population mean within 0.2 of its true value given the population variance is 0.3?

3. A random sample from a normally distributed population produced the following data:

$$6.2, 5.8, 7.1, 6.3, 6.9, 5.7, 6.5$$

(a) Enter the data in `R` and obtain the mean and standard deviation.

(b) By following the steps below, construct a 95% confidence interval for the population mean μ.

i. What is the degree of freedom of the t-distribution?

ii. Determine the required critical value (use quantiles of the t-distribution).

iii. Hence, from this information, determine a 95% confidence interval for μ.

4. A population mean is estimated from a sample of size 20. The sample standard deviation is 6, and the confidence interval is $(23.70, 31.38)$.

(a) What is the value of the sample mean?

(b) What level of confidence was used in computing the confidence interval?

(c) Compute a 95% confidence interval for the mean.

5. Discuss the three factors that affect the width of a confidence interval for a population mean, and the effect of each.

6. An ecologist obtains estimates of the number of ants (in 100,000) for last March for a random sample of 200 ant hills in the Northern Territory. She then calculates a 99% confidence interval as $(254.78, 278.46)$. State if each of the following statements are true or false, and justify your answer.

(a) The analyst should use the normal distribution critical value to calculate the confidence interval.

(b) The mean number of ants per ant hill for the 200 ant hills is 260,000.

(c) The standard error of the sample mean is 5.89.

(d) The 95% confidence interval for the mean number of ants per ant hill for a random sample of 200 ant hills for March next year will be the same.

(e) The 95% confidence interval for the mean number of ants per ant hill for the same 200 ant hills for March next year will be the same.

(f) The 95% confidence interval for the mean number of ants for the 200 ant hills for March next year will most likely overlap with this confidence interval.

(g) A 95% confidence interval for the mean number of ants for another random sample of 200 ant hills for last March will will most likely overlap with this confidence interval.

Population Proportion: Confidence Intervals and Required Sample Sizes

7. A banking survey of customers was conducted to get an estimate of the customer satisfaction for home loans. Suppose 100 home loan customers of the Big Four banks (NAB, Westpac, ANZ and Commonwealth) are surveyed and 85% of them are satisfied with their home loans.

   (a) Give a point estimate for the true proportion of customers who are satisfied with their home loans.

   (b) Give a 90% confidence interval for the true proportion of customers who are satisfied with their home loans.

   (c) Explain the model that you have used in obtaining the confidence interval in (b).

   (d) How large should be a sample in order to get a 90% confidence interval estimate the population proportion to be within 0.05 of its true value?

8. Last year, 20% of the employees in a large firm used public transport to commute to and from work. To determine if a recent company campaign encouraging the use of public transport has been effective, 50 randomly chosen employees were interviewed and it was found that 14 of them were currently using public transport.

   (a) Justify the use of a Bernoulli trials model in this problem.

   (b) Obtain a point estimate (with estimated standard error) and an approximate 95% confidence interval for p.

9. Discuss the three factors that affect the width of a confidence interval for a population proportion, and the effect of each.

10. What is the minimum sample size required to estimate a population proportion to within 0.01 with 95% confidence?

Chapter 11

Basic univariate statistical model

1. The basic univariate statistical model is

$$y_i = mean + \epsilon_i$$

or

$$y_i = \mu + \epsilon_i$$

where y_i is the response variable, and ϵ_i is a random error term, $i = 1, 2, \ldots, n$.

   (a) Explain what the term *mean* represents.

   (b) What determines how the mean is specified?

   (c) A common assumption of most models is that the error term is normally distributed. How can this assumption be verified?

Chapter 12

Hypothesis Testing for a Population Mean or Population Proportion

Introduction, Hypothesis Testing for a Population Mean

1. Suppose that we are testing $H_0 : \mu \leq \mu_0$ against $H_1 : \mu > \mu_0$ and that the test rejects H_0 in favour of the alternative hypothesis at the 2.5% level of significance.

   (a) What can we say would happen if we performed the same test using a 1% level of significance?

   (b) What can we say would happen if we performed the two tailed test $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ at the 5% level of significance?

2. Mineraltech Industries manufactures carbide drill tips used in drilling oil wells. Mineraltech claims that, under typical drilling conditions, the life of a carbide tip has mean drilling depth of at least 32 units. Some customers disagree with Mineraltech's claim and assert that Mineraltech is overstating the average drilling depth. A random sample of 25 carbide tips provides an observed mean of 29.5 units and a standard deviation of 4.0 units.

   (i) Is there enough evidence at the 2.5% level of significance in support of the customers' assertion?

   (ii) Discuss any assumptions underpinning the above analysis.

   (iii) Compute a 95% confidence interval for the mean drilling depth μ of the tips. Does this interval contain μ?

3. From past data it is known that the average stay of tourists in Hong Kong hotels has been 3.4 nights. Recent changes in the nature of tourism to Hong Kong may have changed this average. A tourism industry analyst took a random sample of 30 hotel rooms in Hong Kong, and found that the sample mean number of nights spent by tourists was $\overline{x} = 3.7$ with a sample standard deviation of $s = 0.83$. Assume that the number of nights spent by tourists in Hong Kong hotels is normally distributed.

   Let μ denote the mean number of nights tourists stay in Hong Kong hotels.

(a) Write down the model equation. State and discuss any assumptions of the model.

(b) The analyst wishes to answer the question: "Is there enough evidence to suggest that the average stay of tourists in Hong Kong hotels has increased?" Carry out the relevant test using $\alpha = 0.025$.

(c) Find a 95% confidence interval for the average stay of tourists in Hong Kong hotels.

4. A hole-punch machine is set to punch a hole 1.84 cm in diameter in a strip of metal in a manufacturing process. To test the accuracy, technicians have randomly selected 10 punched holes and measured the diameters. Measurements are given below:

$$1.81,\ 1.89,\ 1.86,\ 1.83,\ 1.82,\ 1.85,\ 1.82,\ 1.87,\ 1.85,\ 1.84$$

(a) Use `R` to test whether the average diameter of the punched holes is significantly more than 1.84 cm at the 2.5% level of significance.

(b) State and discuss the assumptions in deriving the test in (a).

5. A city health department wishes to determine the mean bacteria count per unit volume of water at a lake. The regulation says that the bacteria count per unit volume of water should be less than 200 for safety use. To test the claim that the water is safe, a researcher collected 10 water samples of unit volume and found the bacteria count to be:

$$198, 195, 215, 212, 194, 207, 210, 197, 196, 210$$

(a) Enter the data in `R` and use it to answer the following questions.

(b) Do these data indicate that there is no cause for concern? As part of your answer state and test appropriate hypotheses (use a 2.5% significance level).

(c) What are the implications of your analysis for the council?

(d) Which of Type I and Type II error is more serious in this case?

Hypothesis Testing for a Population Proportion

6. A restaurant chain regularly surveys its customers. On the basis of these surveys, the management of the chain believes that at least 75% of its customers rate the food as excellent. A consumer testing service wants to examine this by asking 460 customers to rate the food. Seventy percent rated the food as excellent. Does this give support against the management's belief? Use a significance level of 2.5%. State and discuss any assumptions of the model.

7. A company that makes computer keyboards has specifications that are designed to ensure that no more than 3% of the production is defective. The company has recently been receiving more customer complaints than usual. To investigate if the proportion of defective keyboards has increased, a quality control officer takes a random sample of 500 keyboards and finds 20 of these to be defective. Let p denote the proportion of keyboards that are defective.

   (a) State the null and alternative hypotheses that the officer is testing.

   (b) Give an expression for the test statistic and state its distribution under the null hypothesis.

   (c) Find the observed value of the test statistic.

   (d) Conduct the hypothesis test at the 2.5% level of significance, and state your conclusion.

   (e) Find a 95% confidence interval for the proportion of defective items produced by the company.

   (f) Verify any assumptions required in the above analysis.

8. Last year, 20% of the employees in a large firm used public transport to commute to and from work. To determine if a recent campaign encouraging the use of public transport has been effective, 50 randomly chosen employees were interviewed and it was found that 14 of them were currently using public transport. Is there enough evidence to suggest that the campaign was effective? Carry out the relevant test using $\alpha = 0.025$.

Chapter 13

Two-Sample Hypothesis Testing: Difference in Two Population Means

Hypothesis Testing for Difference in Two Population Means: Paired Samples

1. Fifteen pairs of monozygotic twins were studied. In each pair, one twin had schizophrenia while the other did not. Measurements were taken on both twins using magnetic resonance imaging to measure the volumes in cm^3 of regions and subregions of the brain. The question of interest is whether the twin without schizophrenia has a larger brain volume.

   (a) What is the appropriate analysis for this data?

   (b) Let d_i = brain volumes of the twins without schizophrenia − brain volumes of the twins with schizophrenia $i = 1, 2, \ldots, 15$. Also, let μ_D be the mean difference in brain volumes. From the data, $\overline{d} = 0.1987$, with standard deviation $s_d = 0.2383$.

      i. State the hypotheses of interest.

      ii. Give an expression for the test statistic, state its distribution and calculate its observed value.

      iii. Test the hypothesis at the 2.5% level of significance.

      iv. What should you conclude regarding the difference in brain volumes for the two groups?

      v. State any assumptions of the analysis.

      vi. What can you conclude about the wider population regarding brain volume and schizophrenia?

      vii. What are the consequences of a Type II error in this situation?

      viii. Compute a 95% confidence interval for the difference in mean brain volumes for non-schizophrenic people and schizophrenic people.

2. A retail company introduces a sales improvement course for its employees.

   (a) For a random sample of 6 employees who took the course the following sales results were recorded.

Employee	1	2	3	4	5	6
Sales before course	12	18	25	9	14	18
Sales after course	18	24	24	14	19	21
d =Difference (After − Before)	6	6	-1	5	5	3

It can be shown that the mean difference is $\overline{d} = 4$ and the standard deviation of the difference is $s_d = 2.68$. Is there enough evidence to suggest that the course was a success at the 2.5% significance level?

(b) If the employee names had not been recorded, and the sales results in the table (see (a) above) were incorrectly treated as two independent samples (i.e. "unpaired") from populations with assumed equal variance, what conclusion could be made, at the 2.5% significance level, about the success of the course? Comment on any difference in conclusion from that of part (a).

3. A human resources manager is investigating if worker productivity drops on Fridays. She selects 25 factory workers at random who produce the same item, and counts the number of items each produces on Wednesday and Friday for a randomly selected week. The data were analysed in `R` and some of the output is given below.

```
> with(Diff, (t.test(Wednesday, Friday, alternative='greater',
conf.level=.95, paired=TRUE)))

     Paired t-test

data:  Wednesday and Friday
t = 4.3052, df = 24, p-value = 0.0001215
alternative hypothesis: true difference in means is greater than 0

sample estimates:
mean of the differences
                   4.44
```

Define the random variable D as

$$D = \text{ the number of items produced on Wednesday} - \text{ the number of items produced on Friday},$$

and let μ_D be the mean of D.

(a) State the null and alternative hypotheses to answer the question: Is worker productivity less on Fridays compared to Wednesdays? (1 mark)

(b) Give an expression for the test statistic and state its distribution.

(c) Test the hypotheses at a 2.5% level of significance. What should the human resources manager conclude regarding productivity on Fridays?

(d) Calculate a 95% confidence interval for the mean difference in productivity between Wednesdays and Fridays.

Hypothesis Testing for Difference in Two Population Means: Independent Samples

4. A social researcher is investigating the time that houses are on the market in Sydney and Brisbane before they are sold. She takes random sample of houses that have just been put on the market in the two cities, and records the number of days each takes to sell. The data is tabulated below.

City	Number of days on market
Sydney	132, 138, 131, 127, 99, 126, 134, 126, 94, 161, 133, 119, 88
Brisbane	118, 85, 113, 81, 94, 93, 56, 69, 67, 54, 137

Assume that the variances of the number of days houses are on the market in Sydney and Brisbane are equal.

(a) The data is available in the file Sales.txt. Read the file into `R`. Examine the data to see how it is stored. Obtain relevant summary statistics for each sample, and copy down the sample means and sample variances.

(b) Without using `R`, construct a 95% confidence interval for difference between the mean number of days that houses are on the market in Sydney and Brisbane.

(c) Are houses on the market for a longer period of time in Sydney than in Brisbane? Use an appropriate hypothesis testing procedure to answer this question (without using `R`). Choose $\alpha = 0.025$.

(d) State and discuss any assumptions of the model.

(e) Now perform the hypothesis test in `R`, and obtain a 95% confidence interval for difference between the mean number of days that houses are on the market in Sydney and Brisbane. (Note that to obtain the confidence interval you will need to conduct a two-sided hypothesis test.)

5. A pathology research company thinks that its current provider of IT technical support puts them on hold too much and as a consequence is costing them money. However before they will consider switching from their current provider to a competitor, the company will have to be convinced that they will spend less time on hold. The company decides to perform a test by placing independent random samples of 12 calls to each technical support line and recording the amount of time (in minutes) that they spend on hold. The summary data are given below.

	Current Provider	The Competitor
Sample size	$n_1 = 12$	$n_2 = 12$
Sample mean	$\overline{x}_1 = 6.2$	$\overline{x}_2 = 4.7$
Sample standard deviation	$s_1 = 3.4$	$s_2 = 2.6$

Assume that $\sigma_1^2 = \sigma_2^2$, and denote the common variance by σ^2.

(a) Derive a 95% confidence interval for difference between the mean hold times for the two IT technical support groups.

(b) Is there a significance difference between the mean hold times for the two IT technical support groups? Use the confidence interval in (a) to answer this question.

(c) Should the company switch technical support provider? Use an appropriate hypothesis testing procedure to answer this question.

(d) State any further assumptions that you have made in your analysis in (a)-(c).

6. A provider of medical supplies is studying differences between two of its major outlet stores, Store A and Store B. The provider is particularly interested in the time taken, from the day of order, for customers to receive their supplies. Data concerning a sample of delivery times for a popular product are summarised below:

	Store A	Store B
Sample size	$n_1 = 21$	$n_2 = 31$
Sample mean	$\overline{x}_1 = 41.2$	$\overline{x}_2 = 43.7$
Sample standard deviation	$s_1 = 2.4$	$s_2 = 3.1$

Let μ_1 and μ_2 denote the respective mean delivery times, and σ_1^2 and σ_2^2 the respective variances of the delivery times for Store A and Store B. Assume that $\sigma_1^2 = \sigma_2^2$, and that both populations are normal.

(a) Find a 95% confidence interval for the mean delivery time for Store A.

(b) Find a 95% confidence interval for $\mu_1 - \mu_2$.

(c) Carry out an appropriate hypothesis test to answer the question: "Is there evidence of a difference in the average delivery times for the two stores?"

(d) Could you draw the same conclusion from your answer to (b)?

7. A study by researchers attempted to determine whether there is a significant difference in the purchasing strategies of the industrial buyers in two countries based on the cultural dimension labelled as "integration". Integration is being in harmony with one's self, family, and associates. For the study, 46 Taiwanese buyers and 26 mainland Chinese buyers were contacted and interviewed. Buyers were asked to respond to 35 items using a 9-point scale with possible answers ranging from 1 = no importance to 9 = extreme importance. The resulting statistics for the two groups are given in the table below.

Integration	
Taiwanese Buyers	Mainland China Buyers
$n_1 = 46$	$n_2 = 26$
$\overline{x}_1 = 5.42$	$\overline{x}_2 = 5.04$
$s_1 = 0.58$	$s_2 = 0.39$

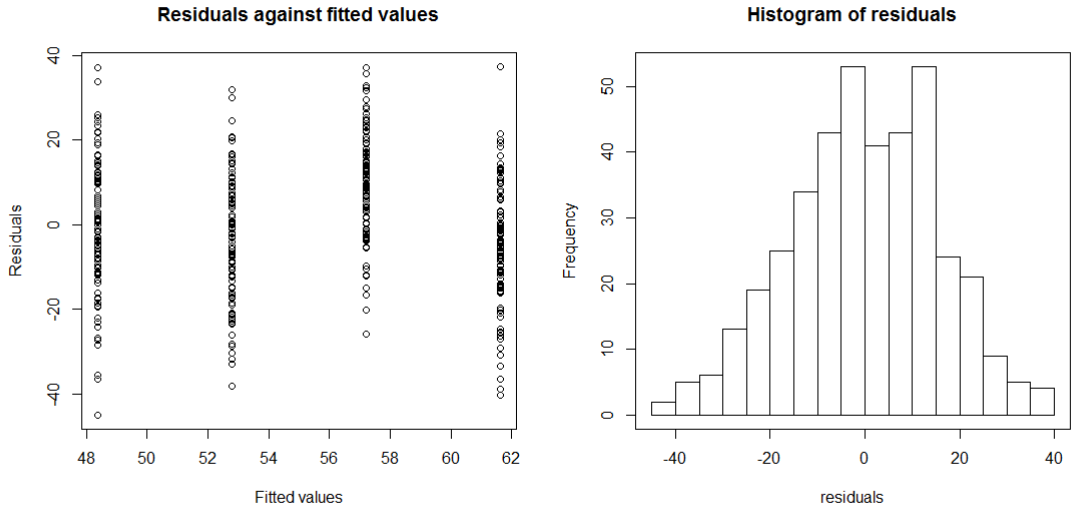
Using $\alpha = 0.01$ and assuming that the two population variances are the same, test whether there is a significant difference between the buyers of the two countries on integration. State any other assumptions of the analysis. What are the business implications of the result of this test?

Chapter 14

One-Way Analysis of Variance

1. The manager of a large company seeks to determine whether there is any difference in the mean lifetimes of the four brands of superior lightbulbs the company currently uses. The lifetimes (in thousands of hours) of 100 lightbulbs of each brand were analysed in R , and some of the output along with some diagnostics are given below.

```
> AnovaModel.1 <- aov(Lifetime ~ Brand, data=Bulbs)

> summary(AnovaModel.1)
             Df Sum Sq Mean Sq F value Pr(>F)
Brand         3  24757    8252   39.48 <2e-16 ***
Residuals   396  82765     209
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> with(Bulbs, numSummary(Lifetime, groups=Brand, statistics=c("mean", "sd")))
      mean       sd data:n
1 47.39226 15.76686    100
2 49.18473 15.05899    100
3 67.44263 12.49788    100
4 56.02555 14.29834    100

Multiple Comparisons of Means: Tukey Contrasts

Fit: aov(formula = Lifetime ~ Brand, data = Bulbs)

Linear Hypotheses:
           Estimate Std. Error t value Pr(>|t|)
2 - 1 == 0    1.792      2.045   0.877  0.81693
3 - 1 == 0   20.050      2.045   9.807  < 0.001 ***
4 - 1 == 0    8.633      2.045   4.223  < 0.001 ***
3 - 2 == 0   18.258      2.045   8.930  < 0.001 ***
4 - 2 == 0    6.841      2.045   3.346  0.00489 **
4 - 3 == 0  -11.417      2.045  -5.584  < 0.001 ***

1   2   3   4
"a" "a" "c" "b"
```



(a) State the three assumptions about the lifetimes of lightbulbs of each brand, which underlie the analysis of variance procedure.

(b) State the null and alternative hypotheses for the test.

(c) At the 1% level of significance, what do you conclude from the test about the mean lifetimes of the four brands of lightbulbs?

(d) Find a 99% confidence interval for the mean lifetime of Brand 2 lightbulbs.

(e) Find a 99% confidence interval for the difference in mean lifetimes of Brand 2 and Brand 3 lightbulbs.

(f) Use the Tukey analysis to determine which group means are different and state your conclusions (use $\alpha = 0.01$).

2. Four independent samples from four normal populations with a common variance yielded the following summary statistics:

group number (j)	1	2	3	4
sample size (n_j)	10	15	5	10
sample mean ($\overline{x}_j$)	6.6	4.6	5.9	5.3
sample stdev (s_j)	0.39	0.51	0.43	0.47

(a) Show that $\overline{\overline{x}} = 5.4375$ (the 'grand mean') and SSB $= 25.294$ (the 'between groups sum of squares', to 3dp).

(b) Given also that SSW $= 7.738$ (the 'within groups sum of squares', to 3 d.p.), produce the corresponding ANOVA table, and determine whether there are significant differences between the 4 population means at the 1% level of significance.

3. Fligh Airlines recently introduced a daily early-bird flight between Hong Kong and Singapore. The vice president of marketing for Fligh Airlines wants to determine if Fligh's average passenger load on this new flight is different from that of each of two major competitors, who run similar aircrafts with the same capacity. Ten early morning flights were selected at random from each of the three airlines and the number of empty seats on each were recorded. An ANOVA was performed on the data and the results, along with data summaries and some diagnostics, are given below.

   (a) State two assumptions of the analysis and use the output given below to verify them.

   (b) Perform a test of the relevant hypotheses.

   (c) Find a 99% confidence interval for the mean number of empty seats on the Fligh's flights.

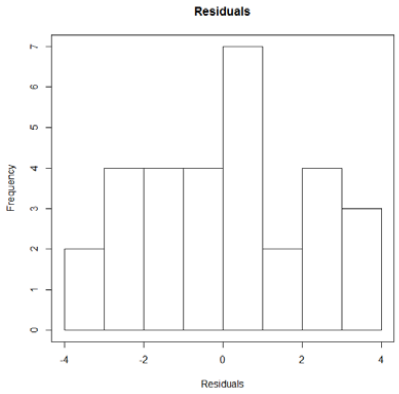
   (d) Find 99% confidence intervals for the difference in the mean number of empty seats for Fligh Airlines and each of Competitor 1 and Competitor 2. Comment on the implications of these intervals.

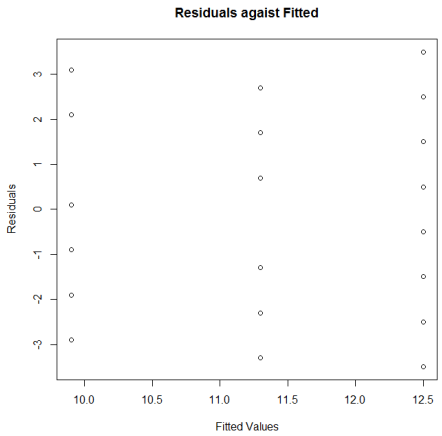
   (e) Report the conclusions of the Tukey-Kramer analysis using $\alpha = 0.05$.

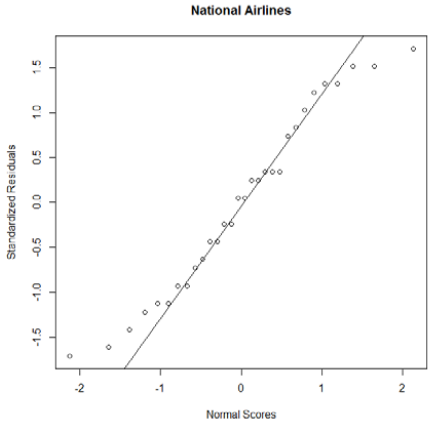
   (f) Write a short statement to the vice president informing her of the conclusions of the analysis.

```
Df Sum Sq Mean Sq F value Pr(>F)
Airline      2  33.87  16.933   3.643 0.0397 *
Residuals   27 125.50   4.648
---

mean       sd data:n
Competitor1 11.3 2.162817     10
Competitor2 12.5 2.173067     10
Fligh     9.9 2.131770     10

Multiple Comparisons of Means: Tukey Contrasts

Fit: aov(formula = Seats ~ Airline, data = Airlines)
Linear Hypotheses:
                                Estimate Std. Error t value Pr(>|t|)
Competitor2 - Competitor1 == 0   1.2000     0.9642   1.245   0.4381
Fligh - Competitor1  == 0     -1.4000     0.9642  -1.452   0.3295
Fligh - Competitor2  == 0     -2.6000     0.9642  -2.697   0.0311 *
---
Competitor1 Competitor2    Fligh
       "ab"         "b"         "a"

       Bartlett test of homogeneity of variances

data:  Seats by Airline
Bartlett's K-squared = 0.0034, df = 2, p-value = 0.9983
```

Airline	mean	variance	n
National	9.90	4.5444	10
Competitor 1	11.3	4.6778	10
Competitor 2	12.5	4.5444	10







Chapter 15

Simple Linear Regression

1. Data were collected by a bank wishing to examine the relationship (if any) between household income and home loan amount (in units of \$,000). The data are in the file `Loan.txt`.

   (a) Which of the two variables would you choose to be the response variable in a simple linear regression analysis?

   (b) Use `R` to produce a scatterplot of the data.

   (c) Fit a simple linear regression model to the data in `R`. Using your model output complete the following.

      i. Using your output complete the following.

```
lm(formula = LoanAmount ~ Income, data = Loan)

Coefficients:
            Estimate   Std. Error   t value   Pr(>|t|)

(Intercept)

Income
---

Residual standard error:_______  on ____   degrees of freedom
Multiple R-squared: _______  , ______ Adjusted R-squared:
F-statistic: ______  on _____ and ___ DF,  p-value: __________
```

      ii. Use the output to write down an estimate b_1 for the regression slope parameter β_1. Interpret the meaning of b_1 in terms of income and loan amount.

      iii. Is there evidence that the a higher income allows households to borrow more for a home loan?

      iv. From this regression, what is the predicted loan amount for a household with an income of $\$120,000$ per annum? Comment on the reliability of this prediction.

      v. Now produce some diagnostic plots. Comment on whether the residuals appear to be normally distributed.

vi. Now produce a scatterplot of residuals against fitted values. Comment on the appropriateness of the linear model. Also comment on the equal variance assumption.

2. **House Data: Regression of Price against Age**

Read the *House.txt* file in to `R`. We will perform a regression analysis of Price against the age of the houses sold. The data was collected in 2010.

(a) Produce a scatterplot of Price against AgeHouse, and describe any observations.

(b) Fit a linear regression to the data. Use your output to answer the following questions.

i. Write down the regression equation.

ii. Test appropriate hypotheses to determine if there is a significant linear relationship between Price and AgeHouse. State your conclusion.

iii. Examine the histogram of residuals and normality plot associated with the regression. Do the residuals appear normally distributed?

iv. Also investigate if a linear model is appropriate.

(c) National Realty wants you to investigate the relationship between the selling price of a house (in \$1,000) and the area of the block of land on which it is situated (in m^2). Using the `House` data, you decide to perform a simple linear regression between Price and Area.

i. First, decide which of the two variables should be chosen as the response variable. Then specify the regression model, and explain each term in the model.

ii. What are the assumptions that must be satisfied to ensure that a simple linear regression is appropriate?

iii. Fit the simple linear regression and produce an appropriate set of diagnostic plots that can be used to assess whether or not the assumptions of the regression model are justified.

iv. From your output, write down the estimated regression equation between Price and Area.

v. Give an interpretation for the estimate of the slope parameter in the estimated regression equation.

vi. Do the diagnostic plots suggest any violation of the assumptions?

(d) Is there a significant (linear) relationship between Price and Area? State the hypotheses to be tested, and read off the appropriate p–value for this test from your output.

(e) Predict the selling price for a house with area equal to (i) $900m^2$; (ii) $1900m^2$. Comment on the reliability of these predictions.

Chapter 16

Multiple Linear Regression

1. Absenteeism is a major problem for employers in most countries, reducing potential output by an estimated 10%. Economists M. Chaudhary and I. Ng (*Canadian Journal of Economics,*, August 1992) conducted a research project to better understand the causes of this problem. They randomly selected 100 organisations to participate in a year long study. For each organisation, the average number of days absent per employee was recorded, along with several other variables described below:

   - Wage : the average employee wage
   - PctPT: percentage of part time employees
   - PctU: the percentage of unionised employees
   - AvShift: availability of shift work (yes, no)
   - UMRel: union-management relationship (good, bad)

   A linear regression analysis was conducted with Absent (average number of days absent per employee) as response, and some of the output is given on the following page.

   (a) Specify the multiple linear regression model between Absent and the explanatory variables, and explain each term in the model.

   (b) Is there sufficient evidence to conclude that the availability of shift work is related to absenteeism? Justify your answer.

   (c) Can we infer that in organisations where union and management relations are poor, absenteeism is high? Justify your answer.

   (d) Write down the fitted regression model between Absent and the explanatory variables, using only the significant terms.

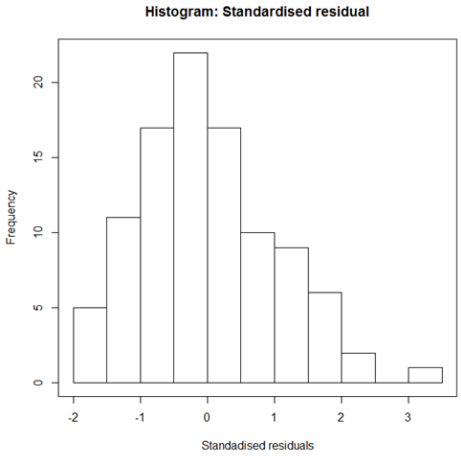
   (e) State and verify the assumptions of the linear regression model using the output.

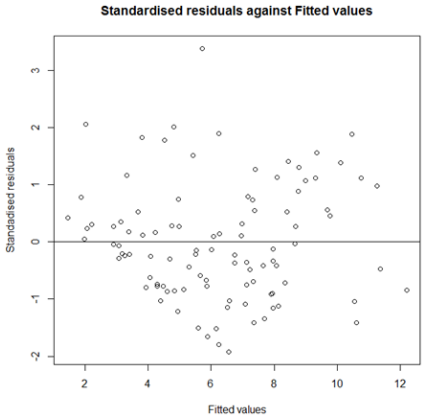
   (f) Which variable, `Av Shift` or `U/M Rel`, has the greatest affect on absenteeism in the workplace according to this data?

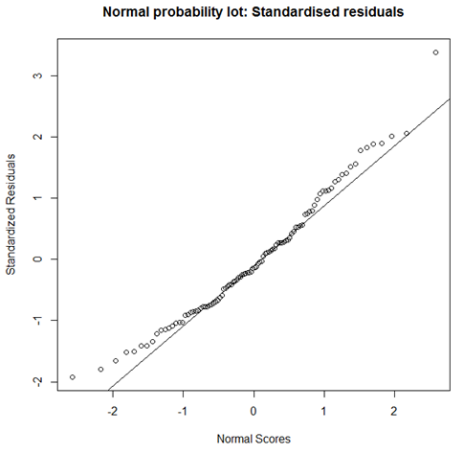
   (g) Compute a 95% confidence interval for the coefficient of the percentage of unionised employees.

   (h) How can this model be improved? Justify your answer.

```
Call:
lm(formula = Absent ~ Wage + PctPT + PctU + AvShift + UMRel,
    data = Absent)

Residuals:
    Min      1Q  Median      3Q     Max
-4.3547 -1.7679 -0.3207  1.2461  7.6769

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.026e+01  1.172e+00   8.756 8.12e-14 ***
Wage           -2.033e-04  3.573e-05  -5.691 1.43e-07 ***
PctPT          -1.069e-01  2.949e-02  -3.624 0.000471 ***
PctU            5.985e-02  1.240e-02   4.826 5.38e-06 ***
AvShift[T.yes]  1.562e+00  5.027e-01   3.107 0.002497 **
UMRel[T.good]  -2.637e+00  4.922e-01  -5.357 5.99e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.356 on 94 degrees of freedom
Multiple R-squared:  0.5323,    Adjusted R-squared:  0.5075
F-statistic:  21.4 on 5 and 94 DF,  p-value: 3.084e-14
```







2. As a further analysis, the following loglinear model was fitted to the data:

$$\ln \text{Absent} = \beta_0 + \beta_1 \text{Wage} + \beta_2 \text{Pct PT} + \beta_3 \text{Pct U} + \beta_4 \text{Av Shift} + \beta_5 \text{U/M Rel} + \epsilon$$

Some of the output from the analysis is given on the following page.

(a) Using the analysis of the previous question, justify fitting the above model to the data.

(b) Write down the fitted regression model between ln(Absent) and the explanatory variables.

(c) Is there sufficient evidence to conclude that the availability of shift work is related to absenteeism? Justify your answer.

(d) Can we infer that in organisations where union and management relations are poor, absenteeism is high? Justify your answer.

(e) State and verify the assumptions of the regression model using the output.

(f) Compare the log model to the linear model fitted in the previous question. Which is better? Justify your answer.

(g) Between `UMRel` and `AvShift`, which variable has the greatest affect on absenteeism in this model? How does this compare with the model in Question 1?

(h) Compute a 95% confidence interval for the coefficient of the percent of unionised employees, and compare your answer to that in Question 1g.

(i) Write a statement reporting the results of the analysis, referring to the factors that affect worker absenteeism.

```
Call:
lm(formula = log(Absent) ~ Wage + PctPT + PctU + AvShift +
UMRel,
    data = Absent)

Residuals:
     Min       1Q   Median       3Q      Max
-0.92443 -0.25501  0.00075  0.24517  1.08780

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.251e+00  1.976e-01  11.397  < 2e-16 ***
Wage           -3.381e-05  6.020e-06  -5.617 1.97e-07 ***
PctPT          -1.863e-02  4.970e-03  -3.749 0.000307 ***
PctU            1.110e-02  2.090e-03   5.310 7.32e-07 ***
AvShift[T.yes]  2.833e-01  8.470e-02   3.345 0.001185 **
UMRel[T.good]  -3.714e-01  8.293e-02  -4.479 2.11e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.397 on 94 degrees of freedom
Multiple R-squared:  0.526,     Adjusted R-squared:  0.5008
F-statistic: 20.86 on 5 and 94 DF,  p-value: 5.722e-14
```